Replication of
Now You See It, Now You Don’t: Repetition Blindness for Nonwords
by Alison L. Morris and Mary L. Still
(2008, Journal of Experimental Psychology: Learning, Memory, and Cognition)

Robin Melnick
rmelnick@stanford.edu

Introduction

Morris and Still (2008) explored representations thought to mediate early stages of word recognition. The authors added to the body of earlier work that had compared priming effects for words and nonwords.by looking at short-term repetition effects found using masked priming paradigms. Previous studies addressing the question of whether masked repetition priming can be found for nonwords as well as words had produced mixed results. Morris and Still and others had in turn employed repetition blindness paradigms (RB; Kanwisher, 1987) as another method for investigating the role of lexical and sublexical representations in short-term repetition effects. When a series of words is displayed in rapid serial visual presentation (RSVP; Forster, 1970) and viewers are asked to identify all the words shown, a repeated word will often go unreported. Morris and Still reported that some prior work had also shown RB given non-identical words with similar orthography, suggesting that some RB effects might have a sublexical locus. This led, finally, to the idea that the same paradigm might be used to find RB with nonwords. Here again, however, earlier results had been mixed. In particular, Cotheart and Langdon (2003) had actually found a repetition advantage for nonwords, but Morris and Still suspected that these results might have been due to participants’ “informed guessing” strategies.

In Morris and Still 2008, the researchers first replicated Coltheart and Langdon 2003 then modified participant instructions specifically to suppress such possible “informed guessing,” ultimately yielding the predicted robust RB results for both words and nonwords. The present work will attempt to replicate Morris and Still 2008’s final trial, Experiment 6, which had participants write down their responses in addition to reporting them to the researcher. While Morris and Still found a greater amount of RB for words than for nonwords, both were significant: words, *F*1(1, 23) = 40.54, *F*2(1, 46) = 33.41, both *p*s < .001; nonwords, *F*1(1, 23) = 13.71, *F*2(1, 46) = 9.69, both *p*s < .005.

Methods

Power Analysis

The original paper does not provide detail to find the original effect size, but we can build a power analysis on either of a couple of approaches. First, a small pilot of the replication procedure was run: just five participants, but we can calculate an initial take on effect size from these results. The effect most crucial to the Morris and Still 2008 Experiment 6 finding is RB for nonwords. Using just nonword critical items and data from just the five pilot subjects, there was an extremely small (though in the predicted direction) effect, i.e., a non-significant RB effect—a decrease in “both” score, where the participant correctly recalls both orthographic near-neighbors. The pilot yielded mean “both” score of 0.081 for neighbor (“repeat”) items against 0.089 for the control items. We can take effect size here as Cohen’s d using the pooled standard deviation—the square root of the mean square error from ANOVA—here yielding *d*=0.06. At just 0.80 power, this would require 1056 subjects to yield <0.05 significance. If we assume instead at least a small but slightly more reasonable effect of *f*=0.1, we would need 393, 526, or 651 subjects, respectively, to achieve 80%, 90%, or 95% power.

Planned Sample

The pilot trial reveals that this is not a short task, taking an average of a little more than 35 minutes to complete. At a proposed participant pay rate of $4.50 per hour equivalent, this is about a $2.50 task. For purposes of the class, running at full power would be cost-prohibitive. At present, a sample of 40 subjects is planned, working within the experimenter's budget for this effort.

Materials

Materials used for the replication were verbatim from the original study, employing the original stimuli lists as provided by the authors. Morris and Still 2008 describe these materials as follows, starting with the baseline materials created for the work's earlier Experiments 1 and 2 (not replicated in the present study):

> Following Coltheart and Langdon (2003), we selected 72 monosyllabic words and 72 pronounceable nonwords; nonwords were selected from the ARC Nonword Database (Rastle, Harrington, & Coltheart, 2002). All items were 4 letters in length and were monosyllabic. Orthographic neighborhood size (orthographic N) was included as a factor in the design; half the words and nonwords had an orthographic N of 12 or greater, and the other half had an orthographic N of 5 or less. High-N and low-N words were similar in print frequency (15 per million; Francis & Kucera, 1982). (…)
>
> The high-N and low-N items were each divided into three sets of 12 items for counterbalancing purposes, and the sets were matched on N and print frequency (mean N sizes per set were 14.8, 14.7, and 14.3 for the high-N words; 13.9, 13.9, and 13.9 for the high-N nonwords; 3.8, 3.8, and 3.8 for the low-N words; and 3.1, 3.2, and 3.1 for the low-N nonwords). Sequences of 3 words or 3 nonwords were created with the first and last words or nonwords designated C1 and C2. These were then made into 6-item RSVP streams by displaying rows of symbols (e.g., %%%%, ####) as the first, second, and sixth items in each RSVP stream. The words or nonwords occupied the third, fourth, and fifth positions in each RSVP stream. Stimulus sets 1 and 2 were used to create the repeated condition, where C1 and C2 were identical. The control condition was created from the repeated condition by substituting a word from Stimulus set 3 for C1. Words or nonwords intervening between C1 and C2 were selected from a separate pool of items having N sizes between 5 and 12. Two versions of the stimulus list were created such that each participant viewed 12 three-word lists in the high-N repeated condition, 12 in the high-N control condition, 12 in the low-N repeated condition, and 12 in the low-N control condition. The same was true for the 3-nonword lists. Each participant also viewed 40 RSVP streams with only 2 words or nonwords (a row of symbols was substituted for the intervening word or nonword). Each participant saw each critical word or nonword only once. In both versions of the stimulus list, word and nonword lists were randomly intermixed.
>
> (Morris & Still, 2008:150)

Next, the original authors modified these materials for their Experiment 3 (again, not replicated in the present work):

> Each item (…) appeared in three conditions: repeated, control, and neighbor. To create the neighbor condition, words and nonwords from Experiments 1 and 2 were paired with

> an orthographic neighbor C1 (e.g., *cast* was paired with *cart*). The control condition was created from the repeated condition by substituting an orthographically nonsimilar word or nonword for C1. Words or nonwords intervening between C1 and C2 were selected from a separate pool of items having N sizes between 4 and 12; these words were orthographically dissimilar to C1 and C2. Four versions of the stimulus list were created such that each participant viewed 6 three-word lists in the high-N repeated condition, 6 in the high-N neighbor condition, 12 in the high-N control condition, 6 in the low-N repeated condition, 6 in the low-N neighbor condition, and 12 in the low-N control condition. The same was true for the 3-nonword lists. Thus, the proportion of repeated trials was smaller than in Experiments 1 and 2. Each item appeared in repeated, neighbor, and control conditions, counterbalanced across participants. In addition to these trials, participants also viewed 40 RSVP streams with only 2 words or nonwords (a row of symbols was substituted for the intervening word or nonword). Participants viewed each critical word or nonword only once. In all versions of the stimulus list, word and nonword lists were randomly intermixed.
>
> (Morris & Still, 2008:156)

Finally, the materials for the original Experiment 6—the target for replication in the present work—were created as follows:

> Materials were the same as those used in Experiment 3 (…) except that the repeated condition was omitted. Half of the three-item trials were neighbor trials, and the other half were control trials.
>
> (Morris & Still, 2008:162)

Procedure

The primary difference in procedure between the present replication and the original study's Experiment 6 was that the replication was conducted over the web, rather than at a standalone computer in a lab setting. This means that in the replication, participants were unsupervised as they entered their responses, whereas in the original, participants first verbally reported their responses to the experimenter then also wrote them down themselves. In the replication, the next item was displayed as soon as the participant hit <Enter> after recording a response. In all other details, the replication followed the original. Morris and Still describe the procedure as follows, first regarding the baseline procedure in the original work's Experiment 1:

> (…) (I)n the present experiment, participants named all words or nonwords displayed (…). Each trial began with a "+" displayed in the middle of the computer screen for 560 ms, followed by a blank interval of 560 ms. The RSVP stream was then displayed with each item appearing for 126 ms. Following the RSVP stream, a "?" was displayed to indicate that the participant should report the items. (…) All stimuli were displayed in a white font (36 pt. Chicago) on a black background. The experimental trials were preceded by 10 practice trials.
>
> (Morris & Still, 2008:150)

Where the original experiment delivered its RSVP streams using the PsyScope software package, in the present study online participants were redirected from an Amazon Mechanical Turk (AMT) website page to the experimenter's own website[1] based on custom Javascript and PHP written for the purpose. In a further nod to the practicalities of an unsupervised procedure,

[1] http://www.stanford.edu/~rmelnick/cgi-bin/254/index.php

a progress bar was also added to the response screen to give participants an indication of how far along they were through the experiment.

Finally, the original authors' procedure description for both Experiments 5 (not replicated) and 6 (the target of the current replication) additionally note:

> The procedure was the same as in Experiment 1, except that participants were not told to expect repeated words or nonwords. Instead, they were warned that "some of the words or nonwords will look similar to each other" and were told to read them carefully.
> (Morris & Still, 2008:161)

For the replication this was reflected in the pre-experiment participant instructions.

Analysis Plan

The differences in analysis strategy for the replication are limited to the pre-analysis preparation of data from an unsupervised online task, i.e., cleaning and data exclusion rules. Using best practices for validating crowdsourced data (cf. Munro et al., 2010), no participants are turned away pre-test for being outside the target demographic profile. This encourages truthful responses to related pre-test questions. All participants are paid regardless of their responses. Post-test, but pre-analysis, however, we exclude based on age and language. First, non-adults (*age* < 18) are removed. Next, while the AMT platform allows pre-screening by location so only web participants connecting from within the U.S. ever reach the site, beyond this, anyone whose first language is not English is also excluded pre-analysis. The question used to tease out this information simply asks participants to list all languages they speak at an intermediate level or better, with a seemingly offhand aside, "(native language first)." No indication is given that the researchers are only seeking native English speakers. This again encourages truthful responses. Finally, any participants not providing a response to all items, including practice items, is excluded pre-analysis.

The a priori planned analysis of resulting data mirrors as closely as possible the reported analysis from the original study. Repetition Blindness is measured by comparing the percentage of trials in which both C1 and C2 were reported for the "repeated" condition—meaning orthogonal near-neighbors since no full repeats were part of the Experiment 6 materials used in the replication—compared with the control condition. The critical statistic is repeated-measures analyses of variance (ANOVA), including interaction terms, with three factors: (1) condition (repeated/neighbor or control); (2) lexicality (word or nonword); and (3) neighborhood size (high-N or low-N).

Differences from Original Study

The original study employed 24 participants. As noted in the preliminary power analysis above, the planned sample size here is a larger 40, though this remains far short of the size suggested for sufficient power with a small effect.

Recapping from the procedure section, while the actual cognitive task is aimed at being as close as possible to the original, the online platform creates obvious differences in the mechanics, setting, and population. Given that the details of the task and instructions themselves are respected in the replication, there is nothing in the difference inherently at odds with the claims of the original study, but a few potential challenges suggest themselves immediately. First, the pilot work suggests that the RSVP task is one that requires extraordinary focus and concentration. It would be a difficult task in the lab, and crowdsource participants can generally be expected to be less focused, not more. Second, as previously noted, the original study had participants state their responses to the experimenter verbally before writing them down. While only the written responses were used in subsequent analysis, interaction with an observer could affect the results in a variety of ways: It almost certainly encourages focus on the task. The immediate verbal response could reinforce recall more quickly than writing alone. Though not explicitly mentioned in the original report, the experimenter may have acted—consciously or unconsciously—to encourage participants to at least make attempts at all items observed. (Pilot results indicate that in the current online task subjects in many cases only attempt one word in response to the two or three words displayed.)

Finally, the crowdsource population is expected to be substantially more diverse than that employed in the original study, though beyond that the original subjects were undergraduates at Iowa State University and native speakers of English, no further demographic detail was supplied in the original study.

(Post Data Collection) Methods Addendum

Actual Sample

Forty participants—18 male, 22 female, 18 to 60 years old (mean=32.1)—were recruited via Amazon Mechanical Turk. (This does not include the five pilot participants, and their results are not included in the data analysis.) All participants were paid $2.51 for their work. Average task completion time, including the brief demographic survey, was 26.7 minutes, about 25% faster than the initial pilot subjects averaged.

A total of 4971 individual stimulus responses were collected. One participant was excluded post-test as a self-reported non-native English speaker. Four other participants failed to complete the entire survey and were also excluded, leaving 35 subjects for analysis.

Differences from pre-data collection methods plan

The data was analyzed as planned.

Results

Data preparation

As in the original study, the “both” score was calculated as ‘1’ if both critical items—orthographic near-neighbors—are recalled correctly, otherwise ‘0’. Neither the intervening item (word, nonword, or symbols) nor order was considered, i.e., a response scored a ‘1’ even if the order of recall was reversed and regardless of whether the intervening item, if present in the stimulus, was correctly recalled.

Confirmatory analysis

Critical pair recall. Figure 1 shows the mean percentage of correct recall of both critical items for the various conditions in the replication of Morris and Still (2008) Experiment 6. Repeated-measures ANOVAs with the factors of condition (neighbor or control), lexicality (word or nonword), and neighborhood size (high-N or low-N) demonstrate that joint report of words was higher than for nonwords (31.4% vs. 12.6%), $F_1(1, 34) = 24.61$, $F_2(1, 136) = 13.06$, both *p*s < 0.001. Contrary to prediction, control pairs were not reported at a higher rate than “repeated” (i.e., orthographic near-neighbor) pairs; in fact, neighbors were slightly higher on average, though non-significantly so (22.2% vs. 21.9%), $F_1(1, 34) = 0.14$, $F_2(1, 136) = 0.21$, both *p*s > 0.5. In another contrast from the original study, the main effect of neighborhood size was significant by subjects, though not by items (26.0% vs. 18.2%), $F_1(1, 34) = 20.18$, $p < 0.0001$, $F_2(1, 136) = 2.61$, $p > 0.1$. (Neighborhood size was not significant by either analysis in the original study.) Both the Neighborhood Size x Condition and Neighborhood x Lexicality interactions were similarly significant by subjects but not by items, $F_1(1, 34) = 4.60$, $p = 0.04$, $F_2(1, 136) < 0.01$, $p > 0.5$, and $F_1(1, 34) = 22.84$, $p < 0.0001$, $F_2(1, 136) = 0.98$, $p > 0.1$, respectively. These interactions reflect a small reversal of repetition effect for different neighborhood sizes. As Figure 1 indicates, in each high-N / low-N pairing there is a small RB effect with the lower neighborhood-size items and a small repetition advantage for the higher neighborhood-size items. The Neighborhood Size x Lexicality interaction indicates that this reversal is more pronounced for words than for nonwords. Neither of these interactions were significant in the original study.

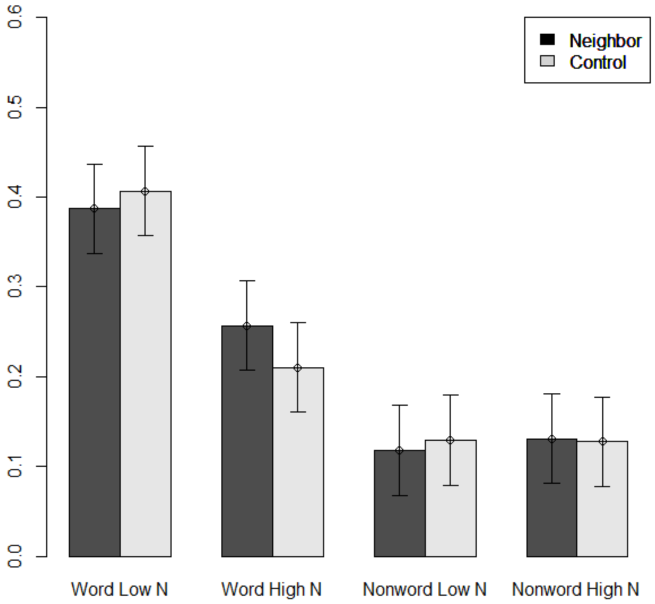

Figure 1. Mean percentages of joint reports of critical words and nonwords as a function of the neighborhood size and condition (neighbor/repeat vs. control) in the replication. Error bars represent 95% confidence intervals. *N* = orthographic neighborhood size.

Exploratory analyses

An additional analysis suggested in the original study's discussion sought a better estimate of the availability of partial orthographic information by calculating the percentage of trials on which at least just two letters of a word or nonword were reported correctly, rather than the entire word. The present replication engaged this alternative calculation, as well. The primary difference in result is that the overall "both" scores rise across the board—to a grand mean of 28.0% using the partial recall scoring vs. 22.1% for whole words—but the general character of the results do not otherwise change substantially. Specifically, there remains no significant RB effect, neither overall, nor within either word or nonword groups.

Discussion

Summary of Replication Attempt

As in Morris and Still (2008) Experiment 6, the present replication found that words were better recalled following RSVP display than nonwords. The primary conclusion of the original study, however, was that RB effects appeared for both words and nonwords, suggested as evidence for a sublexical representation mediating the early stages of word recognition. In contrast, the replication undertaken here did not find any significant RB effects, neither as a main effect nor within either word or nonword groups.

It is worth emphasizing that the main thrust of the original work was specifically finding RB effects with nonwords, since multiple prior studies had already found a repetition blindness effect for actual words. While the present work fails to replicate the nonword RB finding put forth by the original authors as a counter to an earlier work by Cotheart and Langdon (2003) that had actually suggested a repetition advantage for nonwords, the present results certainly do not support the Cotheart and Langdon repetition advantage finding either. No significant difference between repeated (neighbor) and control items was found here. It is in fact a substantial concern for the present replication that the RB effect for words, found in both of the aforementioned works, was not replicated here.

Commentary

That the RB effect for words—a finding that formed a baseline for both of the competing earlier investigations with regard to nonword RB—was not replicated here (let alone the subtler nonword RB effect that was the differentiating element in the original study specifically explored here) would seem to suggest that the method employed for this replication is inadequate in one or more respects for the present purpose. As discussed in the section above on differences from the original study, any of several sources could be the culprit. It is a long and challenging task, requiring substantial concentration by the participant. Here, a different population; with different motivations and levels of patience and focus; the lack of present supervision throughout the task; the change from a combination of verbal and written response to written only—any of these could contribute to the failure to replicate.