Replication of Syntax-induced pattern deafness, Study 1 by Endress & Hauser (2009, *PNAS*)

Julie Fortuna
jfortuna@stanford.edu

Introduction

Many perceptual systems, including vision and audition, force biased interpretations. A similar phenomenon is observed in language by Endress & Hauser as adult humans failed to recognize a perceptual pattern (sequences starting or ending with two words from the same syntactic category) of repetition. This replication only aims to reproduce the results from experiment 1 in the original study. Experiment 1 tests the hypothesis that adult native English speakers could learn simple repetition-based rules involving the syntactic categories of nouns and verbs. Repetition-based rules are used because they are simple to learn, but rarely occur in natural language. In experiment 1, the subjects were not able to learn the repetition-based rule, and did not perform significantly better than randomly guessing.

Methods

Power Analysis

An a priori power analysis using G*Power 3 indicated that a total sample of 81 people would be needed to detect the original effect size (d=0.28) with 80% power using a one-sample t-test with alpha at 0.05. The sample size required for 90% power to detect the original effect size is 111. The sample size required for 95% power to detect the original effect size is 140.

The original study did not find an effect for experiment 1, so we chose to replicate with 80% power to give a chance of seeing whether the replication would find any difference from chance.

Planned Sample

Eighty people will participate in the experiment. All participants will be recruited on Amazon Mechanical Turk, and meet the qualifications of United States residency and a 95% HIT acceptance rate. Participants will earn $0.50 per HIT.

Materials

“Words were recorded by using a Sennheiser ME67 directional microphone connected to a PC running Audacity (available at http://audacity.sourceforge.net/), and saved in the aiff file format (44.1 kHz, 16 bit, mono). English words were recorded from different female native speakers of American English [...] Depending on the difficulty the speaker experienced producing words without list prosody, words were either recorded in isolation or embedded in short sentences, and then excised from these sentences.”

In the replication of the study, participants used the “Q” and “P” key to mark their

answers, rather than the button box used in the original experiment. The same word recordings that were used in the original study were also used in the replication, although the recordings were converted to wav format for web use. Figure 1 below displays the interface of the replication.

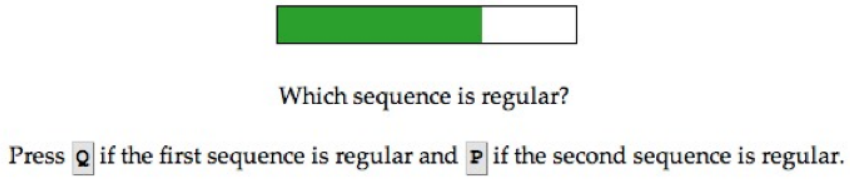

Figure 1. The interface of the replication for the test phase.

Procedure

"Familiarization Phase. Participants were informed that they would hear a number of three-word sequences, and were instructed to memorize them." Before the triplets were played, participants tested their audio by pressing a button to hear a word. They then needed to type the word they heard correctly to move on to the next task. "Participants then listened to a total of 40 triplets." "[...] participants had to learn a repetition pattern over the syntactic categories of nouns (N) and verbs (V); we only selected unambiguous examples of these categories, such that no noun could be interpreted as a verb (e.g., ''run,'' as in ''a run'' and ''to run'') and vice versa." The set of nouns and verbs used is available in the original paper's supporting information. The particular nouns and verbs were always the same within a condition. "We familiarized half of the participants with AAB triplets (i.e., either NNV sequences, such as window-napkin-annoy, or VVN sequences, such as scavenge- listen-camel); the remaining participants were familiarized with ABB triplets (i.e., VNN or NVV)." "Triplets were presented in random order with the constraint that words could not occur in consecutive triplets and that no more than three triplets of the same type could occur in a row."

"Test Phase. Before participating in the test phase, participants were informed that the triplets they had listened to conformed to some regularity. They were informed that they would listen to pairs of new triplets and that in each pair, one of the triplets conformed to the regularity of the familiarization triplets. Then they were presented with 20 test pairs." "In half of the test trials, they had to choose between a NNV triplet (AAB) and a VNN triplet (ABB); in the remaining trials, they had to choose between a VVN triplet (AAB) and a NVV triplet (ABB)." "Participants were asked to indicate which of the two triplets was like the familiarization triplets." "For each test pair type, the correct choice occurred equally often first and second in a test pair."

The replication did not have any differences in the procedure from the original study apart from the addition of the audio check at the beginning of the experiment.

Analysis Plan

Participants were told they could only complete the task once. If they completed the experiment more than once, only the data from their first trial was used in the analysis.

In addition to recording which answer the participant chose and the accuracy, the

experiment also recorded the time it took each participant to choose an answer. If a participant took more than 3 standard deviations more time than the average time to choose an answer on any question, it was determined that the worker was off task, and all the worker's data was excluded.

The percentage correct was used to determine whether a participant learned the repetition pattern - a t-test was used for this. The difference in performance for AAB as opposed to ABB was also tested, also using a t-test.

Differences from Original Study

The participants in the original study were drawn from the Harvard University Study Pool, so they were likely to be college-aged and high achievers. The replication draws from workers on Amazon Mechanical Turk, and is likely to have participants from a greater range of demographics.

The most significant difference in the replication from the original study is the lack of control over the experimental setting. While the original study was a proctored lab experiment, this replication allowed people to participate in a variety of settings. This is potentially a problem for a learning based study like this one, where participants might be saving the audio recordings or taking notes, which are activities that would be noticed in a lab-based experiment. Although we considered having the participant press the space bar to hear each sequence in the familiarization phase to ensure that the worker remained on task, we ultimately went with the original design, where the familiarization sequences were played one after another, with a one second pause between each sequence. We went with the original study design to make it more difficult for participants to take notes on the sequences and instead checked that the participant was on task by excluding those with very long response times. Additionally, with the online format, some participants may rush through the task in an attempt to get paid quickly and randomly guess. The plan to exclude data where participants answered much more quickly or slowly than the average response time attempts to account for this.

(Post Data Collection) Methods Addendum

Actual Sample

Data from eighty participants was gathered from Mechanical Turk. Of the eighty, two participants were excluded from the analysis for completing the task twice, and another three were excluded for taking more time than than three standard deviations above the mean to give a response to one of the questions. Data from another five participants was gathered for a final sample size of eighty. Forty participants were in the AAB condition, and the other forty participants were in the ABB condition. All participants were United States residents and met the 95% HIT acceptance rate qualification on Mechanical Turk.

Differences from pre-data collection methods plan

No differences from the pre-data collection methods plan.

Results

Data preparation

In accordance with the analysis plan, two participants who completed the task more than once were excluded.

Three participants responded to questions with slower response times than three standard deviations from the mean, and were also excluded. After those exclusions, the mean question response time was 1845.03 ms and the standard deviation was 2686.91 ms.

Confirmatory analysis

Participants in the replication failed to learn the repetition patterns [percentage correct: M = 49.9%, SD = 11.0%), $t(79) = -0.05$, $p = 0.52$]. There was no difference in performance for AAB as opposed to ABB [$t(71.82) = 1.80$, $p = 0.08$].

Exploratory analyses

Although not described in the original study, I performed a Shapiro-Wilk test of normality on the data to make sure a t-test was appropriate, and found a normal distribution does not well describe the data [$W = 0.94$, $p < 0.001$]. I then ran a Wilcoxon signed rank test and did not find a significant difference in the participants' responses from chance [$V = 1172.5$, $p = 0.42$]. A Wilcoxon rank sum test also did not find a significant difference in performance for AAB as opposed to ABB [$W = 964.5$, $p = 0.11$].

Discussion

Summary of Replication Attempt

The primary result of the confirmatory analysis was that participants failed to learn the repetition patterns. This finding aligned with the finding of experiment 1 in the original study. Figure 2, below, is a graph comparing the results of experiment 1 from the original study on the left with the replication on the right.

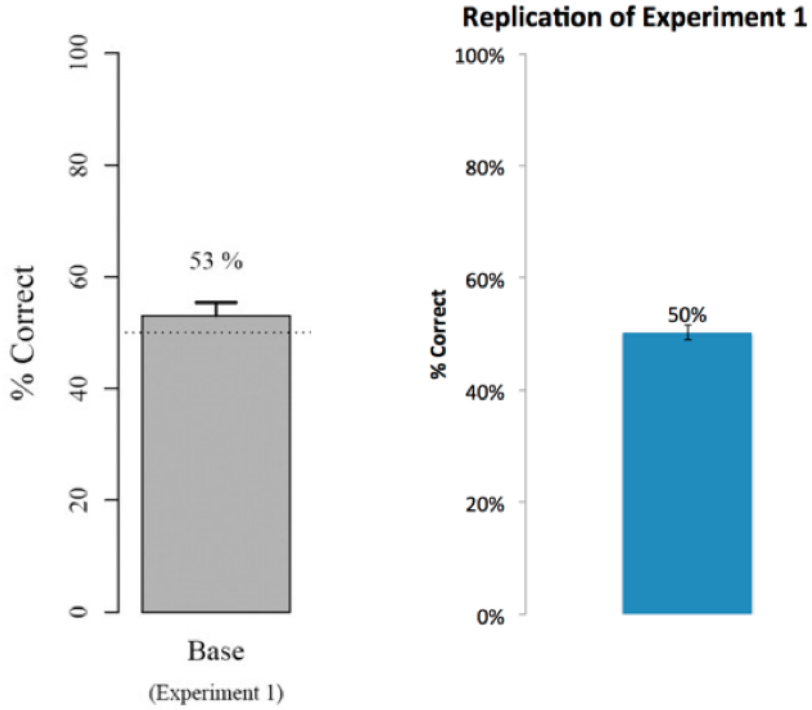

Figure 2. Comparison of results of the original study on the left and the replication on the right.

Commentary

Even though the replication confirmed the findings of the original study, the null effect of the replication could be caused by confounding variables. In the original study, participants could not walk away or mute the volume during the familiarization phase, but these are both possibilities in the replication that would not support the claim that people failed to learn the repetition patterns. If we knew a significant number of replication participants did not cooperate in the familiarization phase, we would not be confident in claiming that the replication result aligned with the results of the original study.

While we considered making the familiarization phase more interactive for the replication to ensure participants were on task, we realized this would make it much easier for the participants to take notes. If a significant number of participants took notes during the familiarization phase, this would cast doubt on whether participants were learning the patterns or just referencing their notes.

One of the original authors was very concerned about this issue of participant behavior, and he also brought up the potential issue of participants taking notes during the experiment addressed above. Although we cannot be sure the design decisions we made in the replication of experiment 1 completely eliminate the concerns of participant attentiveness, we now have a system in place for running these types of experiments on Mechanical Turk, and replications of the other experiments in the original study may give us more information about whether running this type of learning study is viable on a platform like Mechanical Turk.