Replication of “Is It Better to Select or to Receive? Learning via Active and Passive Hypothesis Testing”

Markant & Gureckis, 2014
Journal of Experimental Psychology: General

Kyle MacDonald
kylem4@stanford.edu
Psychology 254 - Winter 2015
Stanford University

Introduction

Markant and Gureckis (2014) investigated the effects of active vs. receptive hypothesis testing on the rate of participants' learning of categories. They tested two different types of category structures: a rule-based category, which varied along 1 dimension, and an information-integration category, which varied along 2 dimensions. In the active learning condition, the learner chose specific observations from the category to test his or her beliefs, whereas in receptive learning condition, the data were passively encountered. They found that participants in the active condition learned the category structure faster and achieved a higher overall accuracy rate compared to participants in the receptive learning condition, but only for the rule-based category.

The target finding for this replication is the advantage in overall accuracy for the active learning condition compared to the passive learning condition for the rule-based category (varying along one dimension).

Methods

Power Analysis

We calculated Cohen's d for the t-test that compared the overall accuracy rate of the active and receptive learning conditions (d = 0.47). Next we conducted a post hoc power analysis using G*Power and found that the original study achieved a power of 0.43. Finally, we conducted an a priori power analysis for the proposed replication in order to achieve 80%, 90%, and 95% power to detect that effect size. The results were:

- 80%, n = 84 (42 in each group)
- 90%, n = 158 (79 in each group)
- 95%, n = 238 (119 in each group)

Planned Sample

Our planned sample size was 48 participants, 24 in each condition. This sample size allowed us to achieve a 47% power to detect the an effect the size of the original finding. This sample size was chosen to maximize power within our funding constraints. Participants were recruited from Amazon Mechanical Turk and restricted to workers with U.S. IP addresses and with a >85% approval rate.

Materials

We used the same stimuli as the original study (described below).

"Stimuli were defined by a two- dimensional continuous-valued feature space, where one dimension corresponded to the size (radius) of a circle and the second dimension corresponded to the angle of a central diameter (see example in Figure 5B). Stimuli of this type have been used in many studies of perceptual classification (e.g., Garner & Felfoldy, 1970; Nosofsky, 1989; Shepard, 1964), and previous work has established that these two

dimensions are, for most participants, separable and independent (Nosofsky, 1989). Stimuli could be assigned a value on each dimension within the range [1,600]. These values were converted for display such that there was a limited range of possible orientations and sizes. The orientation of the stimulus could vary over 150°, ensuring that a full rotation of the stimulus was not possible. The minimum radius and orientation were randomized so that the optimal decision boundary corre- sponded to a unique location in perceptual space for each participant.

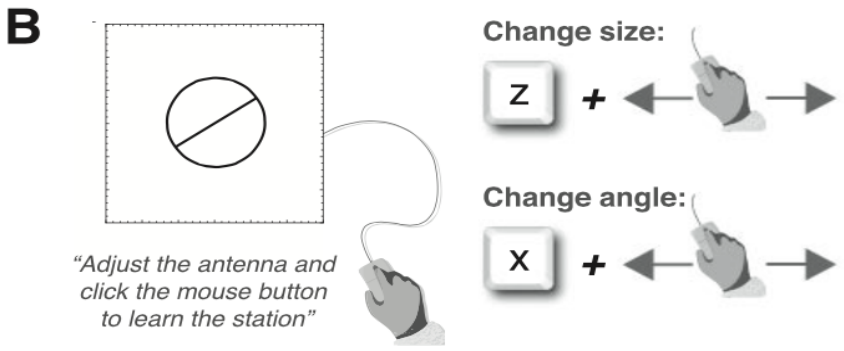

Figure 1: A depiction of a stimulus (left) and the interface used in the self-directed learning condition.

One hundred twenty-eight training stimuli were created for the R training condition using samples from two bivariate Gaussian distributions (see Figure 5A), with mean and covariance parameters slightly modified from Ashby et al. (2002). For classification trials, a uniform grid of 256 unique test items was generated over the feature space for use in all conditions. For each test block, eight stimuli were randomly chosen (without replacement) from each quadrant of the stimulus space to avoid random biases in the test distribution, for a total of 32 items in each block. The order of individual test items within each block and the order of the eight test blocks were both randomized for each participant."

Condition	μ_x	μ_y	σ_x^2	σ_y^2	cov_{xy}
Rule based (RB)					
1. Category A	220	300	2,000	9,000	0
Category B	380	300	2,000	9,000	0
2. Category A	300	220	9,000	2,000	0
Category B	300	380	9,000	2,000	0
Information integration (II)					
1. Category A	250	350	4,538	4,538	4,463
Category B	350	250	4,538	4,538	4,463
2. Category A	250	250	4,538	4,538	−4,463
Category B	350	350	4,538	4,538	−4,463

Table 1: Category distribution parameters used in the receptive learning condition

Procedure

We followed the exact procedures used in the original study (described below).

"Participants were told that the stimuli in the experiment were "loop antennas" for televisions and that each antenna received one of two channels (CH1 or CH2). They were told that the channel received by any antenna depended in some way on the two dimensions described above, and the participant's goal was to learn the difference between the two types of items. Participants were instructed that the antennas were sometimes "noisy" and could pick up the wrong channel "on occasion" and that it would be beneficial to integrate over a number of trials during learning (i.e., that they should learn what channel was "most often" received by a particular type of antenna). In this experiment, however, the feedback associated with each item was deterministic. The experiment consisted of eight blocks, with each block divided into a set of 16 training trials followed by 32 test trials.

Training phase: All conditions. The overall design of the training phase roughly matched the "observational learning" procedure used by Ashby et al. (2002). In that study, participants viewed a stimulus for a short, fixed duration followed by the corresponding category label of the stimulus for a fixed duration (the "no response/after" condition). Critically, participants were not asked to make an explicit prediction, and corrective feedback was never provided. The observational learning procedure is ideal for studying self-directed learning because we wanted to limit the conflicting demands of sampling informative items and sampling items that would result in "correct" feedback under a supervised procedure.

Training: S condition. On each training trial, the participant "designed" a TV antenna and learned about its category membership. Each trial began with the presentation of a randomly generated stimulus in the center of the screen. The participant could then alter its size and orientation by moving the mouse from left to right while holding down either the *Z* or *X* key (see Figure 5B). The direction of motion and mapping of keys to features were randomized across participants. Only one dimension could be changed at a time, but participants could make any number of changes and use as much time as needed. When the stimulus was the desired size and orientation, participants pressed the mouse button to reveal the category label, which appeared above the stimulus and was visible for 1,500 ms. Querying the category label was not permitted until the participant had made a change to the initial stimulus. Trial duration was recorded starting with the initial presentation of the stimulus until the end of the trial.

Training: R condition. In the R condition, participants were unable to interact with the stimuli in any manner. Instead, in each trial they were presented with a stimulus generated from the category distributions described in Table 1. On each trial, a fixation cross was presented, followed by the stimulus (for 250 ms), followed by the category label and stimulus together. Out of concern that in this passive, observational condition participants might not pay attention during the learning phase (relative to the S participants who interacted with the display), the participant was required to press a key corresponding to the displayed category in order to end the trial. The stimulus and label remained visible on the screen until the verification response was registered.

Test: All conditions. Each set of 16 training trials was followed by 32 test trials. On

each test trial, a single item was presented in the center of the display, and participants were asked to classify the item according to the channel the item was most likely to receive. No feedback was provided after their judgment. Following their response, participants were then asked to rate how confident they were about their response using a scale ranging from 1 (*"not at all" confident*) to 5 (*"extremely" confident*). Participants made classification and confidence responses at their own pace. At the end of each block, participants were told their cumulative accuracy during the block they just completed, as well as their accuracy during the preceding test block."

Analysis Plan

We will follow the same exclusionary criteria and analysis plan as the original study (described below). The key analysis is a comparison of classification accuracy on test trials between the active learning and receptive learning conditions.

Classification accuracy on test trials:

"Responses during test blocks were scored according to whether the participant identified the correct category of each test item (as determined by the true category boundary). Three participants (one each in the RB/S, RB/Y2, and II/S conditions) were excluded from further analysis because their overall performance (averaged across blocks) was more than three standard deviations below the mean of their group."

"In the RB task, overall accuracy was marginally higher in the S condition than in the R condition, $t(57)$ 1.82, p .07, and significantly higher than both yoked conditions: Y1, $t(57)$ 2.33, p .02; Y2, $t(56)$ 2.34, p .02. There was no difference between accuracy in the R condition and either yoked condition: Y1, $t(58)$ 1; Y2, $t(57)$ 1, and no difference between the two yoked conditions, $t(57)$ 1."

Differences from Original Study

Our web-based replication was different from the original study in the following ways:

Sample: We recruited participants from Amazon Mechanical Turk as opposed to undergraduates. And we plan to recruit 48 participants as opposed to 60. This sample size was chosen to maximize our power within our funding constraints. We expected our sample to conform to the typical demographics of Mturk, with a median age in the 30s. Thus our sample will be older than that of the original study.

Web-based experiment: The replication experiment will be presented to participants through their web browsers. Thus we will be unable to control for differences in physical characteristics of participants' screens such as size and brightness. We will also be unable to control how participants interact with the stimuli in the active learning condition (e.g., with a mouse vs. a trackpad).

Length of experiment: We will run six training/test blocks as opposed to the eight run in the

original study. We chose to reduce the number of blocks because of length and funding constraints. Also, participants' accuracy reached an asymptote by block six in the original study.

Implications of design changes: None of the previously mentioned differences from the original study are hypothesized to impact the general phenomenon under investigation.

(Post Data Collection) Methods Addendum

Actual Sample

48 adults participated via Amazon Mechanical Turk. Participants were randomly assigned to either the active or receptive learning conditions. The dimension of the category boundary (radius vs. orientation) was counterbalanced across participants within each condition. All participants had gained approval for ≥85% of previous work; had US IP addresses; and indicated that they were ≥18 years old. Participants were paid $2.50 for about 20 minutes of their time. One participant was excluded from analysis because their mean classification accuracy was greater than three standard deviations below their group's mean.

Sample demographics were as follows. More men (n=28) than women (n=20) participated. Participants ranged in age from 22-67 years (median age: 33 years). The majority of participants (n=46) reported English as their native language; 1 participant reported Russian and 1 participant reported Spanish. The majority of participants reported their highest educational attainment as receiving a Bachelor's Degree (n=22); 15 participants reported having completed some college (no degree), 9 participants reported having received an Associate's Degree, and 2 participants reported having received a graduate or professional degree.

Differences from pre-data collection methods plan

None.

Results

Data preparation

Data preparation was minimal[1]. The raw data was stored as separate JSON strings for each participant. These strings were merged to create a large array, which was then converted into a data frame for analysis.

[1] All data, data processing, and analysis scripts can be found at: https://github.com/kemacdonald/Act-Learn

Confirmatory analysis

We directly followed the proposed analysis plan. Following Markant and Gureckis (2014), we compared mean classification accuracy between the active and receptive learning conditions. Overall accuracy was significantly higher in the active condition than in the receptive condition, $t(45) = 2.36$, $p = 0.02$.

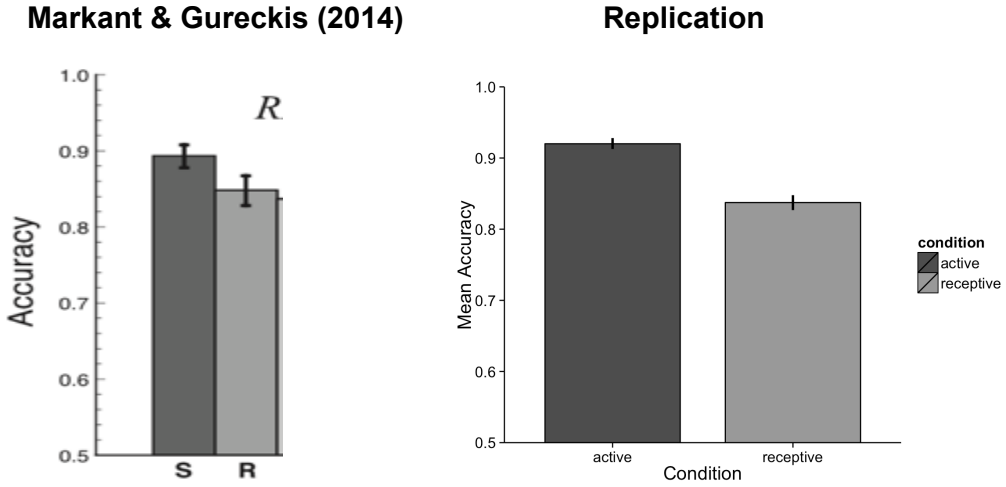

Figure 1: Overall classification accuracy for each condition averaged across all test blocks for the original study (left) and the replication (right).

We also analyzed mean accuracy for each condition (active vs. receptive) as a function of test block. Unlike the original study, we found an immediate advantage for active learners starting in the first test block.

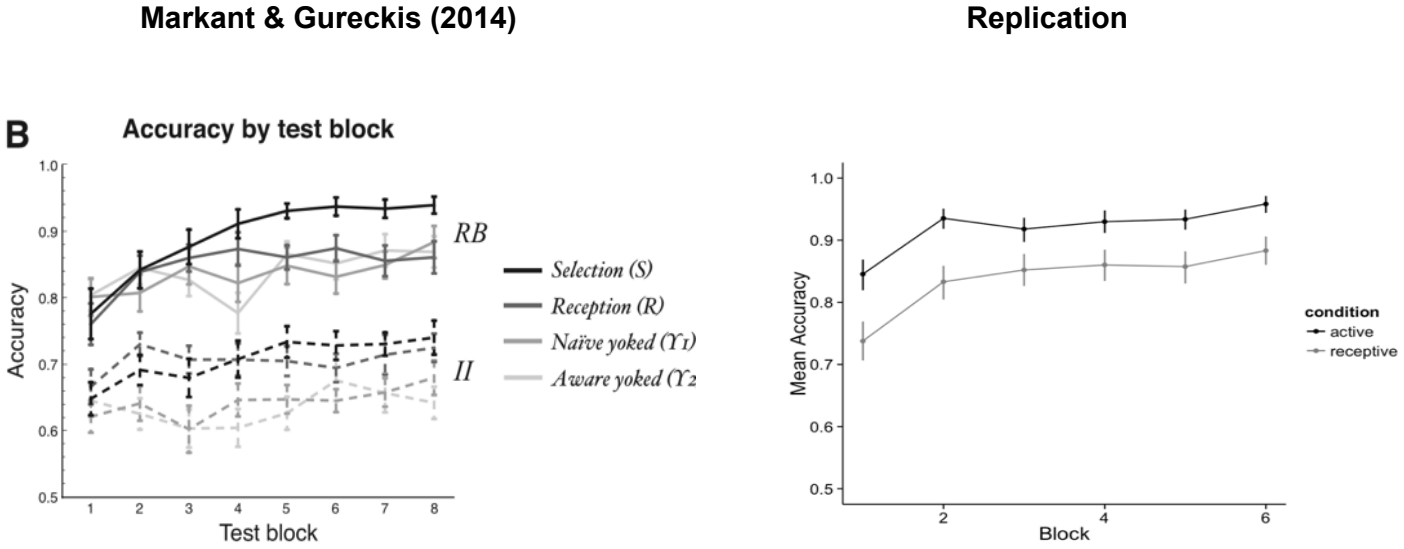

Figure 2: Mean classification accuracy for each condition as a function of test block for the original study (left) and the replication (right).

Discussion

Summary of Replication Attempt

Our data provide strong evidence for a successful replication of the original results reported in Markant and Gureckis (2014). We found a comparable advantage in overall classification accuracy for active learners over receptive learners in a web-based experiment with two fewer training/test trial blocks. We also found an immediate advantage for active learners that was not present in the original study.

References

Markant, D. B., & Gureckis, T. M. (2014). Is it better to select or to receive? Learning via active and passive hypothesis testing. *Journal of Experimental Psychology: General*, *143*(1), 94. [link]