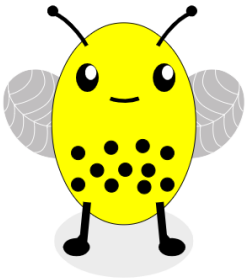
Replication of a Research Claim from Rich & Gureckis (2018), from the *Journal of Experimental Psychology: General*

Joshua R. de Leeuw
Miles Bader
Rachel Ostrowski

Project ID: Rich_JournExPsychGen_2018_LbEB - deLeeuw - 6zzo6
OSF project: https://osf.io/ty5be/
Preregistration: https://osf.io/xba78

Claim Summary: The claim selected for replication from Rich & Gureckis (2018) is the Experiment 1 claim that participants tasked with categorizing multi-dimensional stimuli were more likely to adopt an inferior one-dimension categorization strategy in a contingent feedback condition than they were in a full-information condition. This reflects the following from the paper's abstract: "in a series of experiments we present evidence that people robustly fall into this [one-dimension categorization] trap, even in the presence of various interventions predicted to meliorate it."

For the focal test selected for the SCORE project, participants were tasked with acting as a "beekeeper," collecting honey from many beehives in a computer-based task. Most varieties of bees allowed honey to be harvested, but some varieties were dangerous and stung upon attempted harvesting; in order to complete the task successfully, participants had to determine whether a given bee belonged to a safe or dangerous variety. Each bee variety was a unique illustration (example, right), differing from the other varieties according to four physical attributes. Unbeknownst to participants, two of these four attributes were randomly chosen as relevant—and, of the possible combinations of these two relevant dimensions, one combination was randomly chosen to designate the dangerous bee varieties (for every participant, 4 of the 16 varieties were dangerous).



Participants were divided into two conditions: the first received feedback only when they chose to attempt to harvest a hive, but not when they chose to avoid the hive (contingent condition), while the second received feedback regardless of whether they chose to harvest or avoid a given hive (full-information condition). After excluding participants who reported using an external memory aid, like paper and pen, responses were used to calculate two behavioral scores. The two scores were a 2D and a 1D score, which correspond, respectively, to the proportion of correct choices based on the correct two-dimensional structure of the task, and the proportion of choices that align with an incorrect one-dimensional structure based on only one of the attributes.

Replication Criteria: For this study, a successful replication will show a statistically significant difference in 1D categorization scores, with participants in the contingent information condition having a higher mean score than participants in the full-information condition.

Replication Result: A total of 574 people completed the experiment, 474 of whom met the inclusion criteria. 100 participants were excluded for reporting having used an external memory aid such as pen and paper to help remember features of the bees. Of the 474 participants included in the analytic data set, 223 were in the full-information condition and 251 were in the contingent condition. Our pre-registered analytic sample size was 474, which we met.

We tested the focal hypothesis with an independent samples t-test. 1D accuracy scores in the contingent information condition (M=0.7708, SD=0.1847) were significantly higher than 1D accuracy scores in the full information condition (M=0.6974, SD=0.1317), $t(472) = 4.92$, p = 0.00000196723. The effect size was d = 0.4571, in the same direction as the original study. We conclude that the result replicated according to SCORE criteria.

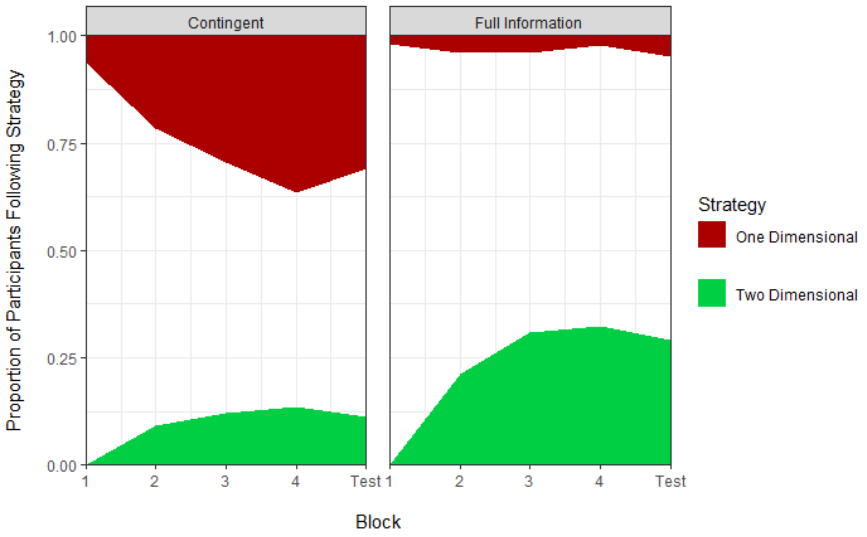


Figure 1. Proportion of participants using 1D and 2D strategies throughout the experiment. The critical comparison for **H*** is based on the proportion of 1D strategies (top red sections) at test. The figure shows that the 1D strategy was more prevalent in the contingent information condition.

Deviations From the Original Study: We used Prolific, an online labor market designed specifically for online research, rather than Amazon Mechanical Turk (MTurk). Pay rates are higher on Prolific, and higher due to inflation, so we adjusted the pay accordingly. We offered $2.50 for completing the task, and up to $2.00 in bonus pay.

The only other deviation from the original study is the choice of analysis model. To conform with SCORE's requirements, we used frequentist inference instead of the Bayesian inference used by the original study. While there are a variety of theoretical reasons to prefer Bayesian inference even in relatively simple statistical models like this one (Kruschke, 2013), in practice frequentist inference provides a similar bottom-line conclusion in most cases. To demonstrate this, we reanalyzed the data from *the original study* using our SCORE-compliant analysis plan. We found that the 1D scores for learners in the contingent information group were significantly higher than the 1D scores of learners in the full information group, $t(93) = 2.91$, p = 0.0045, 95% CI = 0.027 to 0.145. We reach the same conclusion, and the 95% frequentist confidence interval matches closely the 95% Bayesian credible interval.

Deviations from Pre-registration: We calculated the proportion of excluded participants midway through the experiment and discovered that it was much higher than anticipated. The original study excluded 1 person for using external memory aids, but after collecting data from the first 222 people we found that 61 (27.4%) reported using something to help remember the objects. This threatened to exhaust our budget for the study before reaching our target sample size. We decided to modify the description of the experiment on Prolific to explicitly ask people to not use something to help them remember. In the second round of data collection, only 39 of the 352 people reported using an external memory aid (11%). Since condition assignment was cyclical, this change affected all conditions equally.

Description of Materials Provided:

Data Component on OSF (https://osf.io/9jegx/)
The full set of raw data in JSON format is found in this component. A data dictionary for all variable values written in the experiment is available in tab-separated value format (data-dictionary.TSV). We also included a CSV of the aggregated data after pre-processing to produce tidy data.

Analysis Component on OSF (https://osf.io/vu3d4/)
The main analysis script (analysis.Rmd) and resulting R Notebook (analysis.nb.html) are uploaded here. In order to rerun the analysis, you must first download the JSON files included in the data section as described above and adjust the file path to match the download. Once you have ensured that the paths are correct and that you have downloaded the relevant R packages, this script will run with no further input required.

Methods and Materials Component on OSF (https://osf.io/87vzy/)
The experiment script, experiment.html, is found here, as well as all other scripts used to complete the experiment code. The experiment was created using jsPsych, and the supporting library scripts are also included here.

All materials of this project will be available on the OSF website.

Citation:

Rich, A. S., & Gureckis, T. M. (2018). The limits of learning: Exploration, generalization, and the development of learning traps. *Journal of Experimental Psychology: General, 147*(11), 1553–1570. https://doi.org/10.1037/xge0000466

References:

Kruschke, J. K. (2013). Bayesian estimation supersedes the t test. *Journal of Experimental Psychology: General, 142*(2), 573.