Study Information

Hypotheses

H*: Participants in the contingent feedback condition will adopt an inferior one-dimension categorization strategy more often than participants in the full-information condition.

Expected pattern of results: Participants in the contingent feedback condition will have a higher average 1D score compared to participants in the full-information condition.

Design Plan

Study type

Experiment - A researcher randomly assigns treatments to study subjects, this includes field or lab experiments. This is also known as an intervention experiment and includes randomized controlled trials.

Blinding

No blinding is involved in this study.

Is there any additional blinding in this study?

Participants will be randomly assigned to one of two conditions: contingent and full-information. The participant will be aware of their condition because of the way it affects their individual experience, but unaware of the existence of the alternate condition and how it differs. Participants will also be blind to the hypothesis of the study.

The research team will build the data analysis pipeline before the start of data collection.

Because the experiment will be conducted online, researchers will have no direct interaction with the participants. Thus, no further blinding considerations need to be made.

Study design

This is a two-group design (Learning Condition: contingent v. full-information), with learning conditions manipulated between subjects.

Participants in both conditions act as a “beekeeper,” collecting honey from many beehives in a computer-based task. The instructions specify that each hive contains a single variety of bees. Most hives are “friendly” and will allow honey to be harvested, but some hives have been invaded by dangerous bees that sting upon any attempt at harvesting them.

The bees themselves are illustrations that vary according to four attributes. They may have: single or double wings;

two or six legs;
striped or spotted bodies;
antennae or no antennae.

For each participant (unbeknownst to them), two of the four attributes are chosen as relevant, and one of the four combinations of these two relevant attributes is chosen as dangerous. All other bees are friendly.

For example, the two relevant attributes chosen for Participant A may be wings and body pattern; of the four combinations of wings (single/double) and bodies (striped/spotted), the relevant dimension may be single wings and striped body. Out of the 16 bee graphics, four bees meet this set of criteria—such is the case for each pair of relevant attributes.

Participants start the experiment with a $0.40 bonus and go through a series of 64 trials. In each trial, they are shown one bee from a hive and given the option between attempting to harvest from the hive and avoiding the hive. Attempted harvests increase the participant's bonus by $0.02 if the bees are friendly, but have the potential to decrease the bonus by $0.10 if the bees are unfriendly and "sting" the participant. Avoided hives have no effect on the bonus.

All participants receive feedback when they attempt to harvest from a hive; this comes by seeing whether they increased or decreased their bonus. However, feedback after hive avoidance during the learning phase varies according to condition. Full-information participants are told after avoiding a hive whether the bees were friendly or dangerous, and what the resulting payoff would have been, had they harvested; contingent participants are not provided with this information.

During the learning phase (64 trials), every block of 16 trials contains each variety of bee exactly once, randomized with the constraint that every set of eight bees contains six friendly and two unfriendly stimuli. All participants are informed of the total number of trials in the learning phase, and the remaining number of hives is displayed throughout the trials.

Following the learning phase is a 32-trial surprise test phase, with no feedback on the outcomes of their actions (as well as no view of changes to their bonus).

After the completion of the test phase, the participant will be informed of their bonus.
The study will conclude with a questionnaire aimed at gauging participants' intuitions about the task and excluding those who relied on the use of external memory devices, like writing down information, during the experiment. This questionnaire asks a subset of questions included in

the original experiment. There were several questions that the original experiment asked that were unanalyzed, so we do not ask them here.

No files selected

Randomization

Participants will be randomly assigned to one of two conditions: contingent and full-information. This randomization will be done by randomly assigning a condition when the script loads for each participant. However, as we near the end of data collection, we may adjust the experiment script to assign participants exclusively to one condition in order to balance condition assignment. Online experiments are particularly difficult to balance using any other method, because so many participants are completing the experiment in parallel and dropout/exclusions are unpredictable.

Within each condition, two of the four binary dimensions will be chosen as relevant. This will be counterbalanced across participants, alternating between the six combinations based on the moment each participant loads the experiment script.

Given the two relevant dimensions, one combination of dimensions will be considered "dangerous" and the remaining three "friendly". This randomization will be handled by built-in jsPsych functions for random sampling.

Within the 64-trial learning phase of the experiment, stimuli will be randomized using built-in jsPsych functions under the following constraints:
Each of the 16-trial blocks1 contains each of the 16 bee varieties.
Within each 16-trial block, every eight trials will contain two dangerous and six friendly bee stimuli.

The same randomization under constraint will occur during the 32-trial surprise test phase of the experiment.

1 These trial blocks are not made apparent or distinguishable to the participant; the phases will run cohesively as 64 trials and 32 trials, respectively.

Sampling Plan

Existing Data

Registration prior to creation of data

Explanation of existing data

This project is being preregistered prior to any new data collection.

Data collection procedures

We will conduct the experiment online using participants from the labor market Prolific. (The original study recruited participants from Mechanical Turk)

We will restrict participation to Prolific users who have indicated that they are fluent in English to ensure adequate understanding of the instructions. By Prolific user requirements, all participants will be over the age of 18.

We will pay participants a base rate of $2.50, with the possibility of earning up to an additional $2.00 for performance. This is a higher pay rate than the original study ($1.25 base, $1.80 max performance bonus).

The protocol for the design is described in some detail in response to question 8 (study design) and question 9 (randomization).

No files selected

Sample size

The target analytic sample size is 474 participants.

Sample size rationale

Power calculations were done in accordance with the guidelines of the Social Sciences Replication Project (SSRP). However, instead of carrying out two rounds of data collection to achieve the stage 1 sample size target followed by the stage 2 sample size target as in Phase 1 of SCORE, in Phase 2 only one data collection effort will be carried out to achieve a single target analytic sample size.

All replications will aim to achieve the target analytic sample size, which is determined by a power analysis achieving 90% power to detect 50% of the original effect size (formally referred to as the Stage 2 sample size). However, if this sample size is not attainable for the replicating lab, the sample size will instead be calculated by achieving 90% power to detect 75% of the original effect size (formally referred to as the Stage 1 sample size).

Stopping rule

The planned sample size is 474. After achieving this sample, sampling will stop and planned analyses will be run.

In order to achieve the necessary sample size, once the planned sample size is recruited, the data will be checked for exclusions and the exact number of additional participants needed to replace excluded participants will be recruited. This process will continue until the target sample size is met.

No hypothesis testing will be conducted until the target analytic sample is met.

Variables

Manipulated variables

We are manipulating one variable: whether the participant will be provided with feedback upon avoiding a beehive during the learning phase.

In the contingent condition, no feedback is provided to the participant when they avoid a beehive; all learning will be based on instances where the participant chose to harvest, risking getting stung.

In the full-information condition, the participant will be informed of the outcomes and payoffs that would have occurred, had they chosen to harvest.

Note: During the test phase, neither group receives feedback—even when they choose to harvest.

No files selected

Measured variables

We will measure the accuracy of the participant's guesses in terms of whether or not their choice (to collect or avoid a hive) corresponds to the 'correct' choice given the features of the stimulus ('friendly' or 'unfriendly').

No files selected

Indices

Participants who have learned something about the patterns of harmful and unharmful bees may answer based on one of two heuristics: either (1) a 2D heuristic that correctly chooses to harvest or avoid based on the two different relevant dimensions, or (2) a 1D heuristic, where the participant is choosing whether to harvest based on only 1 of the two relevant dimensions, for example only based on whether or not the bee has antennae.

Thus, we will calculate two behavioral scores: a 2D score and a 1D score. The 2D score indicates the proportion of correct choices, equivalently describing the proportion of participant's choices that align with the true, two-dimensional structure of the task. The 1D score denotes the proportion of a participant's choices that align with making a decision based on only one of the relevant dimensions.

The 1D score will be calculated by finding the proportions of responses that would align with each of the two one-dimensional rules, and taking the greater of these two proportions.

A value of 1 indicates perfect adherence to a rule; a value of 0.5 would indicate random selections on the participant's behalf. If a participant receives a 1D score of 1, they would receive a 2D score of 0.75, and vice versa.

No files selected

Analysis Plan

Statistical models

To test H*, we will compare the 1D categorization scores (see Indices) for subjects in the contingent condition versus subjects in the full information condition. We will use a two-tailed independent samples t-test.

No files selected

Transformations

We will not use any transformations.

Inference criteria

[Criteria for a successful replication attempt for the SCORE project is a statistically significant effect (alpha = .05, two tailed) in the same pattern as the original study on the focal statistical evidence (H*). For this study, a successful replication will show a statistically significant difference in 1D categorization scores, with subjects in the contingent information condition having a higher mean score than subjects in the full information condition. Note, this is a departure into frequentist testing from Bayesian testing (acknowledged below).

Data exclusion

Participants will be excluded for meeting any of the following criteria:

1. The participant indicates they used an external memory device (e.g., writing down information or taking photos of the bees using a cellphone) during the task.
2. The participant requires more than two attempts to pass a comprehension quiz after reading the instructions.
3. The data set from the participant is incomplete.

Missing data

Any subjects with missing data will be excluded.

Exploratory analysis

We are not planning any exploratory analyses of the data.

Other

Other

We will Prolific, an online labor market designed specifically for online research. Prolific requires higher base pay than Amazon Mechanical Turk, which the original researchers used; we will adjust the pay accordingly.

To conform with SCORE's requirements, we will use 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. Thus, given the best available evidence we have (the original data), we expect the choice of analysis strategy to have minimal impact.

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