ADDRESSING THE REPLICATION CRISIS:

Testing the Generalizability of Published Results

Professor Josh de Leeuw
Miles Bader '24
Rachel Ostrowski '24

THE REPLICATION CRISIS

Scientific research faces an ongoing replication crisis, in which many published results are not being independently verified—and, when they are, often do not replicate successfully. Several trusted results can be called into question, thus making it challenging to extrapolate on credible science and plan future studies accordingly. Many efforts to solve the crisis focus on taking preventative measures, such as pre-registration of research questions and analysis plans—however, these measures do little to address the not-yet-replicated body of scientific knowledge that already exists.

THE SCORE PROJECT

The Center for Open Science has organized the SCORE project to investigate whether it is possible to predict whether a given study will replicate, reproduce, and/or generalize successfully. If the reliability of a result can be predicted without requiring an independent replication, the scientific community would easily be able to retroactively reassess a massive body of research. To generate the necessary data for this meta-analysis, SCORE has created a database of individual studies for review by independent research teams. This summer, we took on two of these projects: a replication and a generalization.

REPLICATION

A replication involves re-running an experiment, while matching the original methodology as closely as possible. The goal is simply to determine whether the same pattern of results in a published experiment can be recreated.

GENERALIZATION

The goal of a generalizability study is to determine the extent to which a published claim can be replicated when changes are made to the original experimental design. A successful generalization demonstrates the relevance of the original claim to conditions beyond those of the initial experiment.

RICH AND GURECKIS

Rich & Gureckis (2018) examines the concept of a learning trap—the creation of a false belief despite trustworthy observations, often due to an inductive leap based on limited data.

Participants take on the role of a virtual beekeeper visiting a series of unique beehives, and are tasked with deciding whether or not to harvest each hive. The bees vary in pattern, wings, legs, and antennae. Most of the 16 bee varieties are friendly, and allow for harvesting—but some aren't and will "sting" the participant. A successful harvest increases the participant's bonus and a sting drops their pay.

While specific combinations of two of the features (e.g., double wings and no antennae) determine if a bee is dangerous, the study measures whether the participant has fallen into a learning trap, heuristically making judgments based on only one feature and whether feedback about rejected hives reduces the trap.

By conducting experiments online, researchers can reach a bigger audience and amass data more quickly. Both of our experiments were implemented using the jsPsych JavaScript library, optimized for rapid development of online experiment environments. Recruitment of subjects took place through the web-based labor market Prolific.

GRIFFITHS AND TENENBAUM

Griffiths & Tenenbaum (2011) explores the extent to which subjects' intuitions regarding duration and extent align with a Bayesian statistical model.

Participants are posed with a hypothetical in which they arrive in a new city and get in the first cab they see, noting the serial number of 103 on the back of the cab. Based on this information alone, the subject is tasked with guessing the total number of cabs in the fleet.

The experiment then assesses whether the subject adjusts their guess in a Baysian manner as they are given more serial numbers in the fleet.

To generalize, subjects were given numbers at several different orders of magnitude, with a random number generator case used as a priorless control. For the claim to generalize, it would need to hold at each magnitude as well as in the priorless case.

UNDERGRADUATE RESEARCH SUMMER INSTITUTE

We would like to thank Josh de Leeuw, Tessa Charles, Becky Gilbert and other contributors to jsPsych, Brian Daly, and Susie Painter for their contributions to this project.

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

Griffiths, T. L., & Tenenbaum, J. B. (2011, August 29). Predicting the Future as Bayesian Inference: People Combine Prior Knowledge With Observations When Estimating Duration and Extent. Journal of Experimental Psychology: General. Advance online publication. doi: 10.1037/a0024899

de Leeuw, J. R. (2015). jsPsych: A JavaScript library for creating behavioral experiments in a web browser. Behavior Research Methods, 47(1), 1-12. doi:10.3758/s13428-014-0458-y.