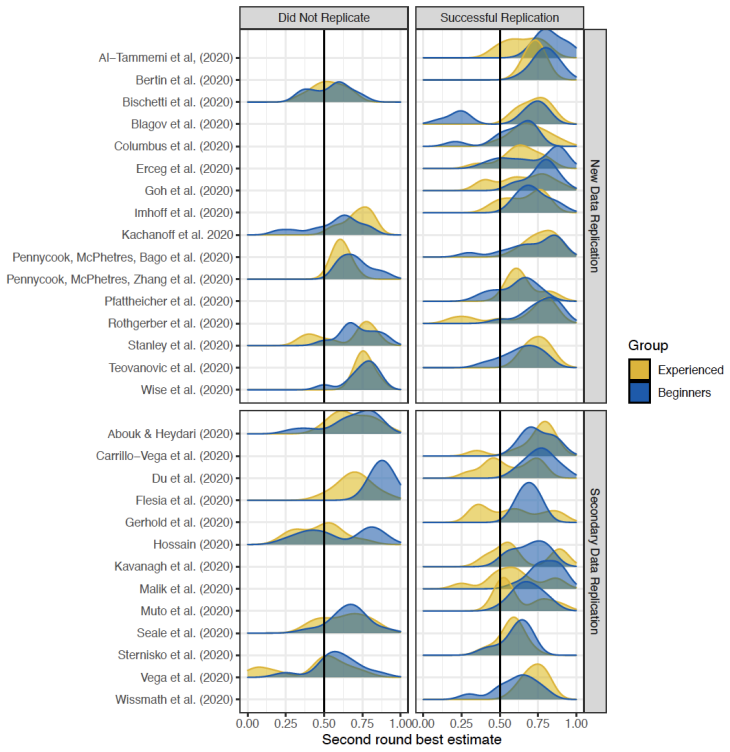
Predicting the Replicability of Social and Behavioral Science Claims in COVID-19 Preprints

PI Yang Liu participated in a multi-year, multi-institutional project funded by DARPA's high-profile SCORE program [1], which seeks to measure the replicability of scientific findings. The project tackles the replicability crisis, a growing issue in scientific research where many published results, even from prestigious journals like *Science* and *Nature*, fail to be reproduced. To address this, the team developed automated tools that generate "confidence scores" for research claims, helping to assess their likelihood of being replicated. One major outcome of this project, titled “Predicting the Replicability of Social and Behavioral Science Claims in COVID-19 Preprints,” [2] was accepted for publication in *Nature Human Behaviour*.

While replication is vital for validating the reliability of published research, replicating every study is both costly and impractical. To address this, the project explored faster and more cost-effective alternatives, such as structured prediction elicitation from crowds of ordinary participants, to quickly and efficiently evaluate scientific claims. This approach is especially critical during crises, like the COVID-19 pandemic, when timely and reliable data are essential for policy decisions. The team collected judgments from participants on 100 claims from COVID-19-related preprints using an interactive elicitation method and conducted 29 high-powered replications to assess the accuracy of these predictions.



The findings showed that after peer interaction, participants with less task expertise (‘beginners’) made significantly larger adjustments to their estimates and confidence levels compared to those with more experience (‘experienced’). Despite the uncertainty of the context, both groups were able to predict the replicability of "fast science" claims with better-than-chance accuracy—69% for beginners and 61% for experienced participants (Figure 1). As a key contribution, PI Liu introduced an innovative method, surrogate scoring rules [3], for analyzing judgment error rates without needing ground-truth outcomes, enabling timely insights into the reliability of scientific claims even before replication studies are completed.

Figure 1. Smoothed distribution of participants’ best estimates for each of the 29 known-outcome research claims with ≥0.8 power with an α = 0.05, organised by type of replication (new or secondary data) and success (did or did not replicate). Experienced participants are shown in yellow, and beginners in blue.

References

[1] DARPA SCORE program:
https://www.darpa.mil/program/systematizing-confidence-in-open-research-and-evidence
[2] Alexandru Marcoci et al. Predicting the replicability of social and behavioural science claims in COVID-19 preprints. Nature Human Behaviour, 2024.
[3] Liu et al. Surrogate Scoring Rules. ACM Conference on Economics and Computation (EC), 2020.