

The Sources of Researcher Variation in Economics

Executive Summary

[submitted]

We use a rigorous three-stage many-analysts design to assess how different researcher decisions—specifically data cleaning, research design, and the interpretation of a policy question—affect the variation in estimated treatment effects. A total of 146 research teams each completed the same causal inference task three times each: first with few constraints, then using a shared research design, and finally with pre-cleaned data in addition to a specified design. We find that even when analyzing the same data, teams reach different conclusions. In the first stage, the interquartile range (IQR) of the reported policy effect was 3.1 percentage points, with substantial outliers. Surprisingly, the second stage, which restricted research design choices, exhibited slightly higher IQR (4.0 percentage points), largely attributable to imperfect adherence to the prescribed protocol. By contrast, the final stage, featuring standardized data cleaning, narrowed variation in estimated effects, achieving an IQR of 2.4 percentage points. Reported sample sizes also displayed significant convergence under more restrictive conditions, with the IQR dropping from 295,187 in the first stage to 29,144 in the second, and effectively zero by the third. Our findings underscore the critical importance of data cleaning in shaping applied microeconomic results and highlight avenues for future replication efforts.