Michael Ganslmeier Tim Vlandas August 14th, 2025

Even honest research results can flip – a new approach to assessing robustness in the social sciences

Estimated reading time: 6 minutes

When academic studies get things wrong, it is often blamed on misconduct and fraud. Yet as **Michael Ganslmeier** *and* **Tim Vlandas** *argue, even good-faith research, conducted using standard methods and transparent data, can produce contradictory conclusions.*

Recent controversies around research transparency have reignited longstanding concerns about the fragility of empirical evidence in the social sciences. While some discussions have centred on misconduct and fraud, an equally important challenge lies in the sensitivity of results to defensible modelling choices: what if the more widespread issue runs

deeper, not in individual misconduct, but in how we conduct empirical research?

In a new study, we set out to measure the fragility of findings in political science by asking how much do empirical results change when researchers vary reasonable and equally defensible modelling choices?

To answer this question, we estimated over 3.6 billion regression coefficients across four widely studied topics in political science: welfare generosity, democratisation, public goods provision and institutional trust – although we only report results for the latter three in this blog post. Each topic is characterised by well-established theories, strong priors and extensive empirical literatures.

Our results reveal a striking pattern: the same independent variable often yields not just significant and insignificant coefficients but also a very large number of both statistically *significant positive* and statistically *significant negative* effects, depending on how the model is set up. Thus, even good-faith research, conducted using standard methods and transparent data, can produce contradictory conclusions.

A new approach to sensitivity analysis

Recent advances – such as pre-registration, replication files and registered reports – have significantly improved research transparency. However, they typically begin from a pre-specified model, and even when researchers follow best practices, they still face a series of equally plausible decisions: which years or countries to include, how to define concepts like “welfare generosity”, whether and which fixed effects to use, whether and how to adjust standard errors and so on.

Each of these choices may seem minor on its own, and many researchers already use a wide range of robustness checks to explore their impact.

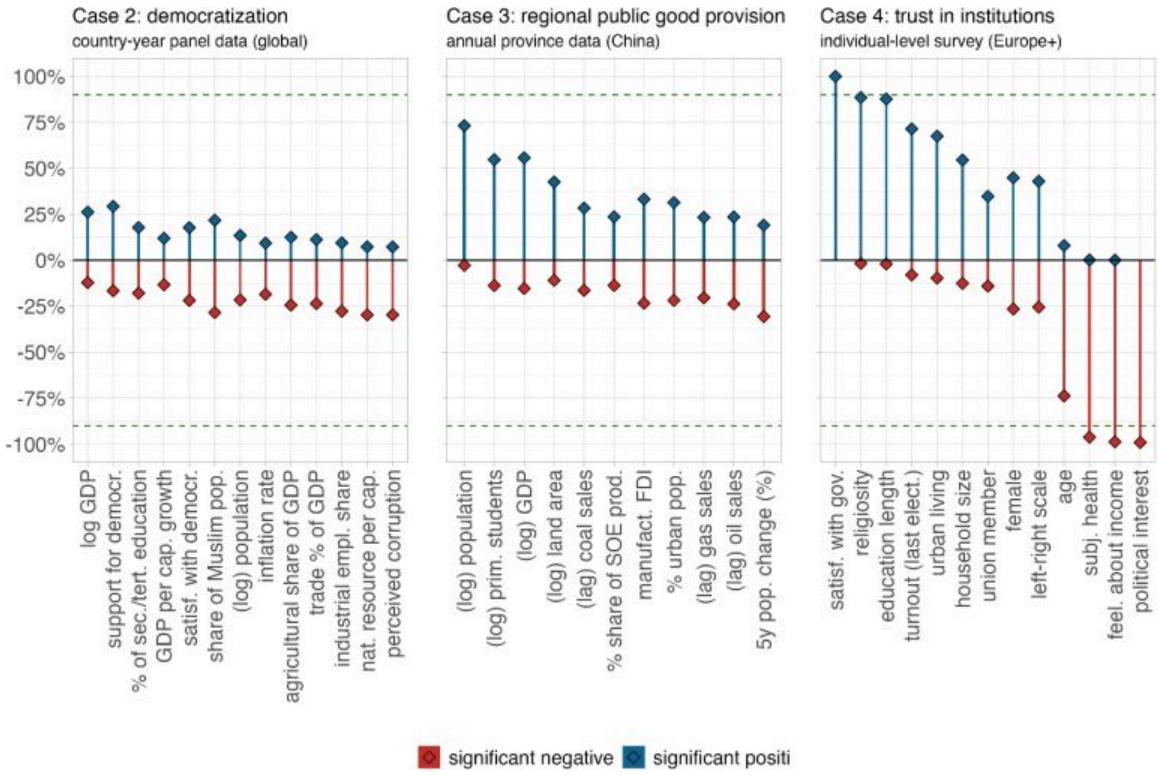
But collectively, these decisions define an entire modelling universe and navigating that space can profoundly affect results. Standard robustness checks often examine one decision at a time, which may miss the joint influence of many reasonable modelling paths taken together.

To map that model space systematically, we combined insights from extreme bounds analysis and the multiverse approach. We then varied five core dimensions of empirical modelling: covariates, sample, outcome definitions, fixed effects and standard error estimation. The goal was not to test a single hypothesis, nor indeed to replicate prior studies, but instead to observe how much the sign and significance of key coefficients change across plausible model specifications.

The fragility of empirical research

For many variables commonly used to support empirical claims, we found many model specifications where the estimated effect was positive and statistically significant as well as others where it was strongly negative and statistically significant (Figure 1).

Figure 1: Share of significant coefficients in the model space for three topics



Note: The panels present the share of (positive and negative) significant coefficients (blue and red, respectively) of all independent variables in the unrestricted model universe for the three test cases: democratisation, regional provision and institutional trust. The dashed line indicates 90%. The figure is adapted from the authors' accompanying article in the Proceedings of the National Academy of Sciences (PNAS).

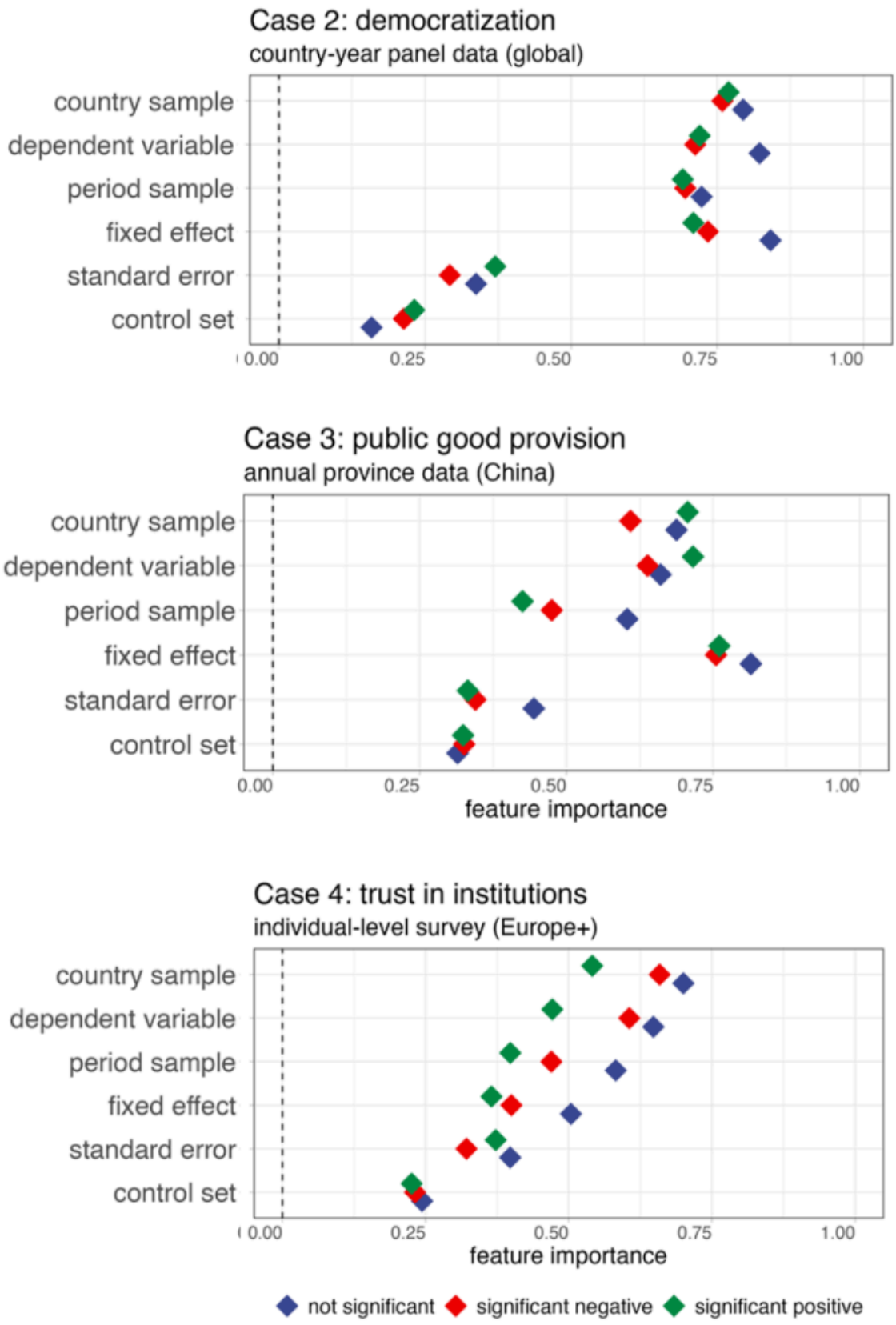
One clear implication is that conventional robustness checks, while valuable, may still be too limited in scope. Researchers frequently vary control variables, estimation techniques or subsamples to assess the stability of their findings. But by examining modelling decisions in isolation, these checks are typically applied sequentially and independently. Our results suggest that this approach can miss the larger picture: it is not just which decisions are made but how their combination determines the stability of empirical results.

By systematically exploring a wide modelling space, while automating thousands of reasonable combinations of covariates, samples, estimators and operationalisations, our approach can assess the joint influence of modelling choices. This allows us to identify patterns of fragility that are invisible to conventional checks.

The sources of model uncertainty

In our study, we estimated the feature importance scores for these different model specification choices. To do so, we first extracted a random set of 250,000 regression coefficients from the unrestricted model universe for each topic. Then, we fitted a neural network to predict whether an estimate is “negative significant”, “positive significant” or “not significant”.

Figure 2: Feature importance scores of model specification decisions



Note: The panels show the feature importance scores (SHAP values) for different model specification choices. The figure is adapted from the authors' accompanying article in the Proceedings of the National Academy of Sciences (PNAS).

Figure 2 shows that the greatest source of variation is not driven by the control variables per se, but rather by decisions on sample construction –

which countries or time periods are included – and how key outcomes are defined. These upstream decisions, often made early and treated as background, exert the strongest influence on whether results are statistically significant and in which direction.

Lessons for empirical research

To be clear, the implication of our findings is not that quantitative social science is futile. On the contrary, our work underscores the value of systematically understanding where results are strong and where (and why) they might be less stable.

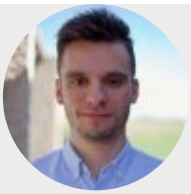
With this new approach, we hope to provide an additional tool that researchers can use to carry out systematic robustness checks and to increase transparency. To that end, we provide our code which future research can use to analyse and visualise the model space around a result.

For more information, see the authors' accompanying paper in the Proceedings of the National Academy of Sciences (PNAS).

Note: This article gives the views of the authors, not the position of EUROPP – European Politics and Policy or the London School of Economics. Featured image credit: Lightspring / Shutterstock.com

---

About the author



Michael Ganslmeier

Dr Michael Ganslmeier is an Assistant Professor in Computational Social Science at the University of Exeter and a Visiting Fellow in the Department of Methodology at the London School of Economics and Political Science. His research investigates political and electoral behaviour in advanced economies. He also studies the economics and politics of climate change in collaboration with the International Monetary Fund and the World Bank.

Tim Vlandas

Dr Tim Vlandas is an Associate Professor of Comparative Social Policy at the University of Oxford, a Fellow of St Antony's College and an associate member of Nuffield College. His research has appeared in over 50 publications in leading international academic journals and has been cited by prominent organisations such as the OECD, ILO, UN, World Bank, European Parliament, ECB and the UK House of Commons.

Posted In: Latest Research | LSE Comment | Politics

Related Posts