Martin Buchner
Julian Rose
Magnus Johannesson
Mandy Malan
Jörg Ankel-Peters

Seeking Scientific Consensus – An Expert Survey on the Replication Debate between Acemoglu et al. (2001) and Albouy (2012)

UNIVERSITÄT DUISBURG ESSEN

Open-Minded

ifso institute for socio-economics

uni-due.de/sozioekonomie/wp

ifso working paper 2026 no.63

Seeking Scientific Consensus – An Expert Survey on the Replication Debate between Acemoglu et al. (2001) and Albouy (2012)

Martin Buchner[1,2], Julian Rose[1,3], Magnus Johannesson[4], Mandy Malan[1] and Jörg Ankel-Peters[1,5*]

[1]RWI – Leibniz-Institute for Economic Research, Essen, Germany; [2]University of Duisburg-Essen, Germany; [3]LMU Munich, Germany; [4]Stockholm School of Economics, Sweden; [5]University of Passau, Germany.

June 2026

Abstract

The publication of contradictory replications often sparks persistent disputes between replicators and original authors. We investigate whether experts converge toward consensus in the prominent debate between Acemoglu, Johnson, and Robinson (AJR, 2001) and Albouy (2012). We recruited 352 experts, including many senior and highly cited economists, primarily from the pool of scholars citing one of the debate articles. We find no consensus on whether the AJR results hold after Albouy's replication, indicating no prevailing interpretation of the debate more than a decade later. Our study demonstrates a potential approach to assessing scientific consensus formation in replication debates and contested literatures.

Keywords: replication, scientific consensus, scientific credibility, expert survey, institutions and growth.

Acknowledgements

We thank Rohan Alexander, Abel Brodeur, David Card, Harry Collins, Cara Ebert, Krisztina Kis-Katos, and Colin Vance for valuable comments and suggestions. We also thank the conference participants at the German Development Economics Conference 2025, the 7th Perspectives on Scientific Error Workshop, the Leibniz Open Science Day 2024, the META-REP Conference 2024, the 2024 MAER-Net Colloquium, the Paul Meehl Graduate School PhD Day 2024, and the 14th Conference of the French Experimental Economics Association. We also thank participants at research seminars at ZEF Bonn, University of Innsbruck, Hasselt University, and University of Kassel for their helpful suggestions. The online appendix is available at https://osf.io/fx8p5/files/osfstorage/6a15b1d676c7f1578dc17c88. Prior to data collection, we uploaded a detailed pre-analysis plan (PAP) on March 27, 2024. It is available on OSF at https://osf.io/fx8p5/. Any analyses that deviate from the PAP are clearly indicated. We gratefully acknowledge funding from the German Research Foundation (Deutsche Forschungsgemeinschaft, DFG) through the DFG Priority Program META-REP (SPP 2317) and from Coefficient Giving. *All correspondence to Jörg Ankel-Peters (Joerg.peters@rwi-essen.de).

1. Introduction

For academic fields relying on the Popperian ideal, knowledge is built by testing theories. Falsification and replication are constitutive elements in this epistemic process. Whether this process should lead to consensus or whether persistent disagreement is even necessary for scientific progress remains an open question in the philosophy of science (Beatty and Moore 2010). At the same time, there is broad agreement that consensus plays an important role in fields where science strives for an authoritative role in societal debates (Hulme 2022). Economics is one such discipline (Fourcade et al. 2015, Frey et al. 1984, Kronlund 2023, Martini 2014, McCloskey 1983). Yet, replication – one of the central mechanisms of scientific scrutiny – often does not lead to convergence of views, especially when replication results are non-confirmatory (Ankel-Peters et al. 2025, Collins 1992, Humphreys 2015, Freese and Peterson 2017, Ozier 2021).

In this paper, we examine whether expert opinions have converged in one of the most prominent replication debates in contemporary economics between Acemoglu, Johnson, and Robinson (2001, 2012, henceforth AJR2001 and AJR2012) and Albouy (2012; henceforth Albouy2012), more than a decade after the exchange. The original paper, the replication, and the reply by the original authors all appeared in the *American Economic Review*. We elicited expert opinions from 352 respondents through a structured, anonymized survey conducted between April and May 2024. Our pre-analysis plan specifies the primary research question as whether experts agree more with AJR or Albouy. We find that expert opinions are widely dispersed, with only a slight tendency to side with the replicator, indicating no clear convergence of views.

Identifying who is an expert in a specific academic debate is inherently difficult. We therefore follow Collins and Evans (2002)'s foundational work on expertise and aim to capture the views of scholars with both *interactional* and *contributory expertise*. Rather than relying on open or mass-distributed survey invitations (e.g., via social media), we recruited participants in a targeted manner, primarily based on citations of the three debate papers, complemented by authors and citers of closely related work. We treat citation behavior as an approximate indicator of topic-specific expertise (Teplitskiy et al. 2022). In total, we invited 3,022 scholars, of whom 309 (10.2%) participated. An additional 43 participants were recruited via mailing

lists of two professional networks. As shown below, the final sample consists predominantly of senior economists with strong publication records and substantial citation counts, which we interpret as indicative of relevant expertise.

AJR2001's original contribution is to empirically demonstrate a causal effect of institutions – such as "more secure property rights and less distortionary policies" (p.1369) – on long-run economic development. Their empirical approach relies on an instrumental variable (IV) based on historical settler mortality. AJR2001's core claim is that variation in settler mortality shaped colonial settlement patterns and, through this channel, the development of institutions that persist to the present.

The paper is widely regarded as providing a causal empirical foundation for the "institutions" view of development, which competes with theories that put human capital, geography, or culture at their center (Alesina and Giuliano 2015, Easterly and Levine 2016, Gallup et al. 1999, Glaeser et al. 2004, McArthur and Sachs 2001, Tabellini 2010). It has had a profound and lasting influence on both academic research and policy discourse and was recognized with the Sveriges Riksbank Prize in Economic Sciences in Memory of Alfred Nobel in October 2024. Our data were collected prior to this award.

In a replication using the same underlying data, Albouy2012 raises concerns about the reliability of AJR2001's settler mortality variable. He focuses on the use of proxy data and the imputation of missing values for a substantial share of countries. Albouy2012 proposes alternative specifications that question both the first-stage strength of the IV and the validity of the results, concluding that the empirical evidence is insufficient to sustain the causal interpretation advanced by AJR2001. In their reply, AJR2012 defend their data construction and conclude that the "big picture from AJR (2001) remains intact and remarkably robust" (p. 3081). While they acknowledge the data concerns raised by Albouy, they argue that his proposed adjustments are overly restrictive and show that his findings are "largely driven by one outlier, Gambia" (p. 3078).

We refer to AJR2001, Albouy2012 and AJR2012 as the *debate papers*. The exchange leaves the interpretation of AJR2001's contribution unresolved, with both sides contesting each other's methodological decisions. Such persistent disagreement between original authors and

replicators is a common feature of replication debates. Ideally, the scientific community resolves such disputes through a process of collective evaluation, in which the prevailing interpretation emerges as "the outcome of social interactions among scientists" (Freese and Peterson, 2017, p. 149). Our paper examines whether expert opinions have converged toward such a prevailing interpretation – or even toward a consensus.

Our online questionnaire first elicited respondents' prior familiarity with the debate papers and their initial assessments. Participants were then provided with neutral summaries of each paper and asked to evaluate key arguments before stating their final position on whether AJR2001's contribution holds. The analysis follows a detailed pre-analysis plan, including the heterogeneity analysis (Malan et al. 2024). In addition, we report non-prespecified evidence on heterogeneity that is clearly identified as exploratory. All pre-specified research questions are mapped to the locations of their reported results in Table A1.

Our paper contributes to the literature on how the economics profession processes conflicting empirical claims following replication. As replication becomes more widespread (Brodeur et al. 2026), disputes following non-confirmatory replications often center on whether the original analysis or the replication is implemented correctly. In many cases, there is no independent yardstick to adjudicate between competing methodological choices. This situation of epistemic deadlock is known in the sociology of science as the "experimenters' regress" (Collins 1992). We argue that such situations also characterize many replication debates in economics (Ankel-Peters et al. 2025; Campbell et al. 2024). Our paper examines whether, in the absence of such a yardstick, the community of experts converges toward a prevailing interpretation of the evidence.

We build on previous work on unresolved replication debates (Ozier 2021, Humphreys 2015, and Roodman 2025) and on approaches that elicit expert beliefs to assess replicability, such as replication markets (Camerer et al. 2016, 2018; Dreber et al. 2015; Forsell et al. 2019). We extend this literature by eliciting and aggregating the views of a large sample of experts on both the original study and its replication, and by assessing whether these views converge toward a prevailing interpretation or consensus. More generally, we aim to encourage further research on the role of expert knowledge in synthesizing contested empirical literatures. Structured approaches to eliciting expert judgment may complement existing forms of evidence synthesis

and policy advice (Aspinall 2010; Hemming et al. 2020). Relying solely on unstructured, organic processes of consensus formation may be insufficient in fields where empirical findings inform urgent policy decisions.

2. The Debate Papers

2.1. Acemoglu, Johnson and Robinson (2001)

AJR2001 document a "precisely estimated and large" (p.1371) effect of institutions on economic performance. Because institutions are endogenous to income, cross-country regression estimates are biased. AJR2001 therefore argue that historical settler mortality influenced European settlement patterns, which in turn shaped early institutions. These early institutions, so the broader theoretical argument goes, evolved into current ones that ultimately determine today's economic performance. AJR2001 employ an instrumental variable (IV) strategy, using European settler mortality (henceforth settler mortality) as an instrument for institutional quality. The identification assumption is that settler mortality affects current income *only* through its impact on institutions. Their empirical strategy rests on the argument that the disease environment shaped European settlement patterns and the type of colonization that emerged, which in turn affected institutions. These early institutional arrangements are assumed to have persisted over time and continue to affect institutional quality today.

AJR2001 approximate settler mortality using historical records on mortality rates of European soldiers, bishops (in Latin America, to fill gaps), and sailors stationed in various colonies between the 17th and 19th centuries. The primary sources of these records are Curtin (1989, 1998) and Curtin et al. (1995). AJR2001 provide only limited details on how the data were constructed, but refer to the appendix from an earlier version of the paper, Acemoglu et al. (2000), for further details.

2.2. Comment by Albouy (2012)

Albouy2012 critiques the empirical construction of the settler mortality instrument rather than the broader institutional theory underlying AJR2001.[1] Albouy2012 questions AJR2001 on two

[1] It is worth noting that skepticism toward IVs was far less pronounced at the time of Albouy's critique of AJR than it is today (Brodeur et al. 2020, Casey and Klemp 2021, Haveresch et al. 2025, Lal et al. 2024, Mellon 2025).

main grounds. First, for 36 out of 64 countries, AJR2001 have no actual data on settler mortality and hence impute it from neighboring countries with similar disease environments. Albouy2012 casts doubt on this procedure. Since neighboring countries often have widely differing mortality rates, Albouy2012 argues that AJR2001's imputed series is highly sensitive to which comparison country is chosen. Albouy2012 further points to more direct flaws in the assignment of mortality rates. One example is AJR2001's apparent assignment of mortality rates originating in western Mali to six other countries, including countries as distant as Angola and Uganda, even though these countries' neighbors exhibit widely varying mortality rates, ranging from 78.2 to 2,004. Albouy2012 concludes that the imputation procedure is "not just unreliable but often deeply flawed, generating rates that may be far too high or too low" (p. 3064).

Second, AJR2001 frequently rely on mortality data for soldiers and African laborers to approximate the rates of settlers. Albouy2012 argues that these groups faced systematically different mortality risks, potentially inducing measurement error correlated with institutional quality or income.

To address these concerns, Albouy2012 implements three modifications: (i) dropping the 36 countries with imputed data, (ii) introducing dummy variables for mortality rates based on less comparable proxies – European soldiers on campaign and African laborers – and (iii) incorporating new settler mortality data from a later paper by Acemoglu, Johnson, and Robinson (2005). Under these adjustments, the first-stage estimates become insignificant in most specifications. Albouy2012 concludes that it is impossible to "*disentangle the effect of settler mortality from that of other variables that may explain institutions and growth, such as geography, climate, culture, and pre-existing development*." (Albouy2012, p. 3073)

2.3. Reply by Acemoglu, Johnson, and Robinson (2012)

AJR2012 defend both the validity of their settler mortality data and the robustness of their results. They reject dropping 36 out of 64 countries, arguing that "*there is a great deal of well-documented comparable information on the mortality of Europeans in those places during the relevant period*" (AJR2012, p. 3107). They provide country-by-country justification for their imputations and show that their results are robust to alternative mortality imputations.

AJR2012 further argue that Albouy2012's results for the restricted sample are primarily driven by Gambia as an extreme outlier in settler mortality rates. Applying alternative outlier treatments, such as dropping Gambia or capping mortality rates, they report estimates similar to those in AJR2001, even with the restricted sample. Moreover, they argue that Albouy2012's distinction between campaign and non-campaign episodes is overstated (p. 3079). AJR2012 conclude that their original findings remain valid: "*[t]he big picture from AJR (2001) remains intact and remarkably robust: Europeans were more likely to move to places that were relatively healthy, and when they moved in larger numbers, they imposed better institutions, which have tended to persist from the colonial period to today*." (AJR2012, p. 3081)

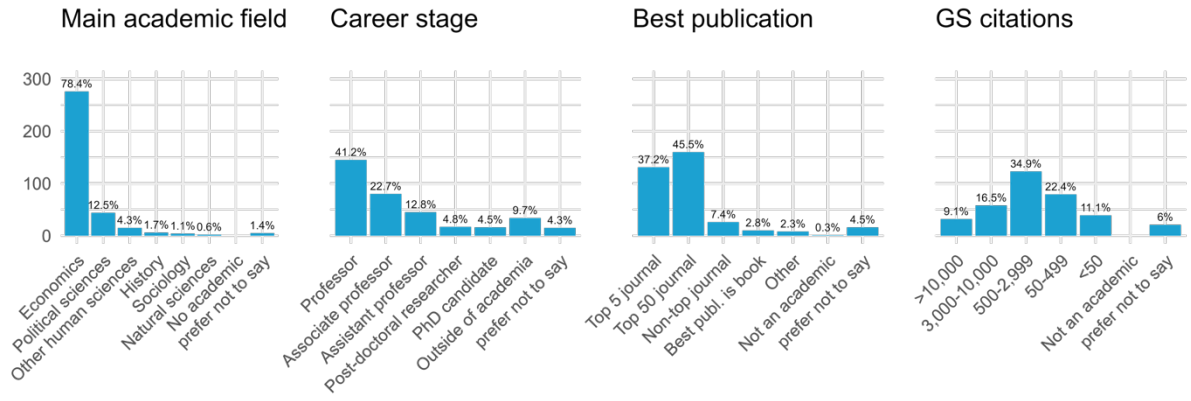
3. The Expert Survey

We construct a sampling frame of 3,022 potential experts, primarily based on citations of the three debate papers.[2] Personalized survey links allow us to track recruitment channels while preserving anonymity and to assess differences across recruitment strategies. Using Scopus (in October 2023), we identified corresponding authors of published papers citing one or more of the debate papers: 2,493 citers of AJR2001, 58 citers of Albouy2012, and 78 citers of AJR2012. Additionally, we identified a short list of papers that either use a similar identification strategy (e.g., Black et al. 2015; Markevich and Zhuravskaya 2018) or that voice criticism of this type of causal historical research (e.g., Deaton 2010; Assenova and Regele 2017). Citers of the former group of papers and authors and citers of the latter are also included in our sampling frame (85 and 491 in total, respectively). We expanded recruitment via two academic networks: the Institute for Replication (I4R) mailing lists and the Development Economics Committee of the German Economic Association.

The survey was launched on March 28, 2024 and closed on May 9, 2024, with two reminders sent in between.[3] In total, 352 respondents completed the survey, of whom 309 came from the citation-based contact list (10.2% response rate) and 43 from the mailing lists. Figure 1 shows the sample composition. Most respondents are economists, with a large share of senior scholars and strong publication records. A majority have substantial citation counts and experience publishing in leading journals, indicating substantial domain-specific expertise.

[2] Online Appendix C provides comprehensive details on the survey implementation.
[3] To incentivize participation, participants were offered a lottery with fifty respondents being randomly selected for a USD 20 Amazon voucher.

Figure 1: Descriptive statistics of respondents (n = 352)



The questionnaire first elicited respondents' prior familiarity with each debate paper, *before* providing any information. Unless otherwise stated, responses were recorded on 0–100 scales. Respondents then rated how convincing they found AJR2001's empirical analysis, Albouy2012's comment and AJR2012.[4] The questionnaire subsequently presented short, neutral summaries of each debate paper. After each summary, respondents were asked to assess the papers' key methodological choices and claims. Finally, respondents stated their final verdict on the debate, our primary research question, and indicated whether their prior beliefs had changed.

4. Results

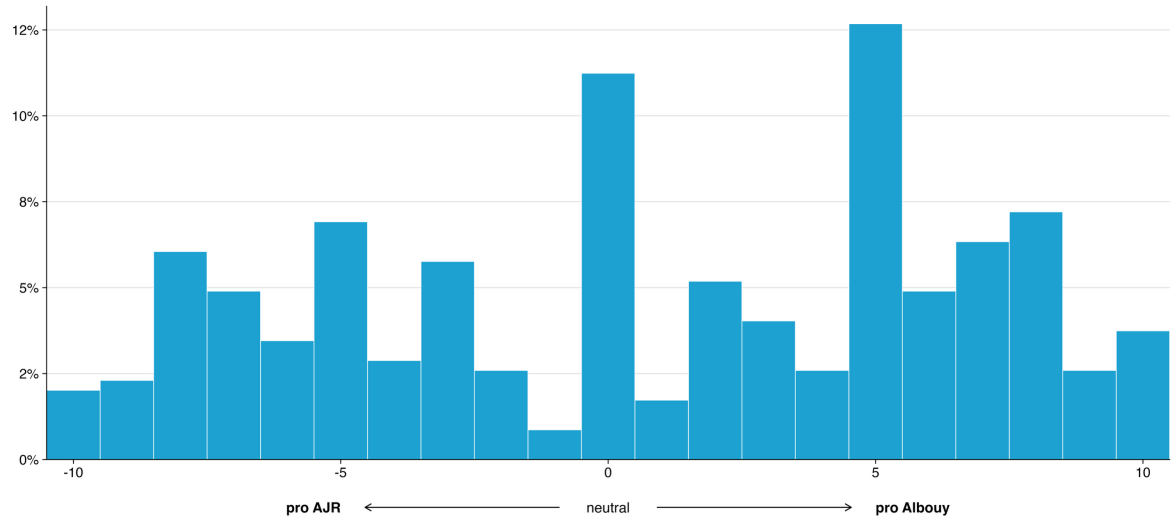
4.1. Experts' final verdict

We present the main findings from the pre-specified analyses below; Online Appendix A reports the full set of results. Figure 2 shows the distribution of responses to the pre-specified primary research question: *With whom do respondents agree more – AJR or Albouy?* After reviewing summaries of all three debate papers, respondents positioned their views on a scale from −10 ("fully agree with AJR") to +10 ("fully agree with Albouy"). The distribution is widely dispersed across the spectrum, with no clear clustering around a dominant position. Approximately 38% of respondents favor AJR (values below zero), 51% favor Albouy (values above zero), and 11% report a neutral position. While a pre-specified *t*-test suggests a slight

[4] The questionnaire structure is illustrated in Figure A1 and the full questionnaire is provided in Online Appendix C.

tendency toward Albouy (mean = 0.76, p = 0.014)[5], this difference is small relative to the wide dispersion in responses.

Figure 2: Respondents' final verdict on the debate (n=347)



Notes: The survey asked the question 'With whom do you agree more?' Scale: -10 = 'I fully agree with AJR,' +10 = 'I fully agree with Albouy.' This question is pre-specified as primary research question 1 in the PAP.

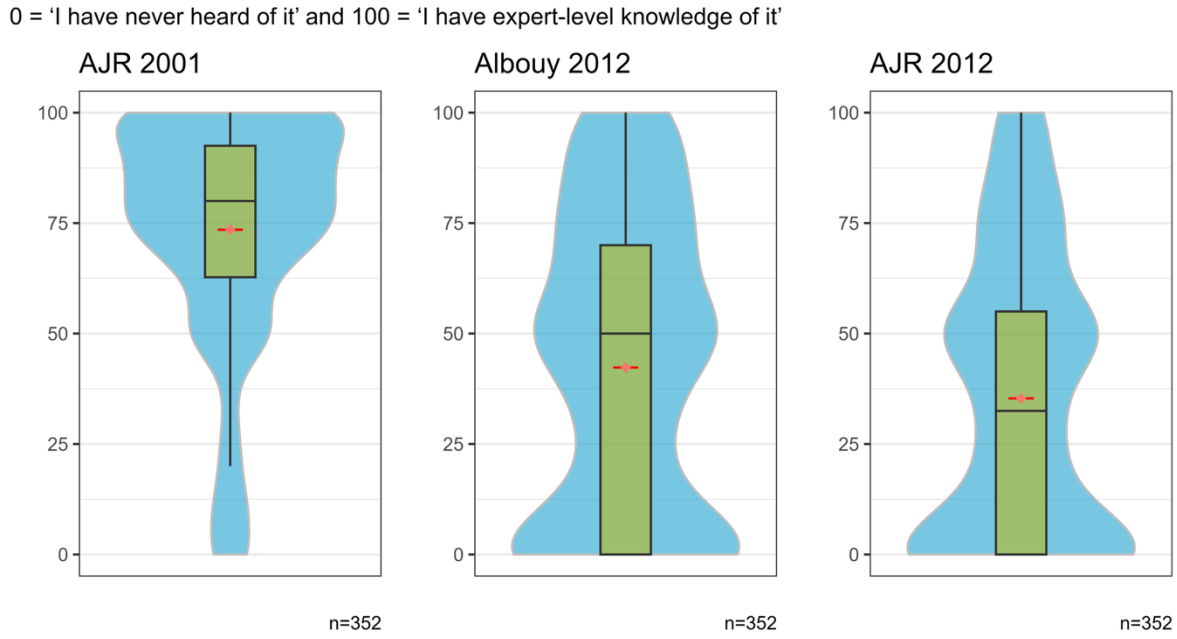
4.2. Respondents' prior familiarity and evaluation of main arguments

At the beginning of the questionnaire, we elicited respondents' familiarity with the debate and their prior assessment of the debate papers. Respondents are generally familiar with AJR2001, while familiarity with Albouy2012 and AJR2012 is substantially lower (Figure 3). Among those with prior knowledge, Albouy2012 is rated as more convincing than both AJR2001 and AJR2012 before receiving additional information.[6]

[5] Following the recommendations of Benjamin et al. (2018), we interpret two-sided *p*-values below 0.05 as "suggestive evidence" and those below 0.005 as "statistically significant evidence".

[6] AJR2001 vs Albouy2012: mean diff.= -10.59 (*p*-value<0.001); AJR2012 vs Albouy2012: mean diff. =-12.85 (*p*-value < 0.001). Scale is from 0 to 100.

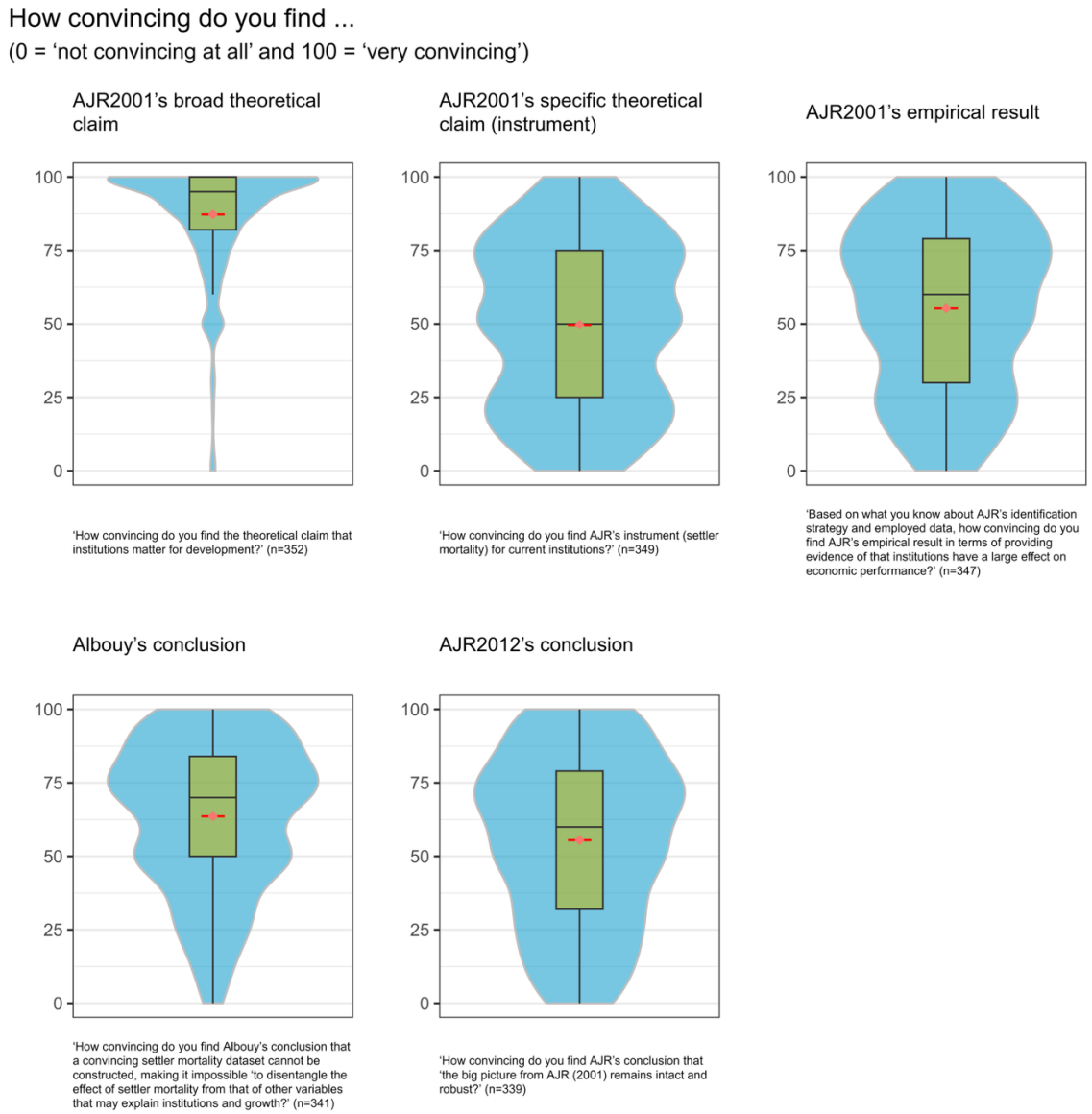
Figure 3: Respondents' familiarity with the debate papers



Notes: Box plots indicate the 25th, 50th (median), and 75th percentiles. Red dots represent mean values. This question is pre-specified as secondary research question 1 in the PAP.

After being presented with neutral summaries of the three debate papers, respondents assessed key methodological choices and claims (Figure 4). Agreement with AJR2001's broad theoretical claim that institutions matter for development is very high. In contrast, there is little agreement on the specific theoretical claim – linking settler mortality to institutions – and for the empirical results.

Figure 4: Respondents' evaluations of main arguments in AJR2001 and conclusions in Albouy2012 and AJR 2012



Notes: Box plots indicate the 25th, 50th (median), and 75th percentiles. Red bars indicate mean values. This question is pre-specified as secondary research question 3 in the PAP.

Evaluations of Albouy2012 and AJR2012 show a similar pattern. While respondents tend to agree with Albouy2012's conclusion, this agreement is far less pronounced than for the broad theoretical claim in AJR2001. For AJR2012, respondents also tend to agree with the overall conclusion, although responses are substantially more dispersed. Overall, the paper-specific

assessments are consistent with the distribution of final verdicts in Figure 2, reinforcing the conclusion that expert views do not converge.[7]

Finally, we examine whether respondents update their ex-ante beliefs after reviewing the summaries. A larger share report becoming less convinced by AJR2001 than more convinced, although the magnitude of this shift is modest. We do not observe comparable changes for Albouy2012 or AJR2012.

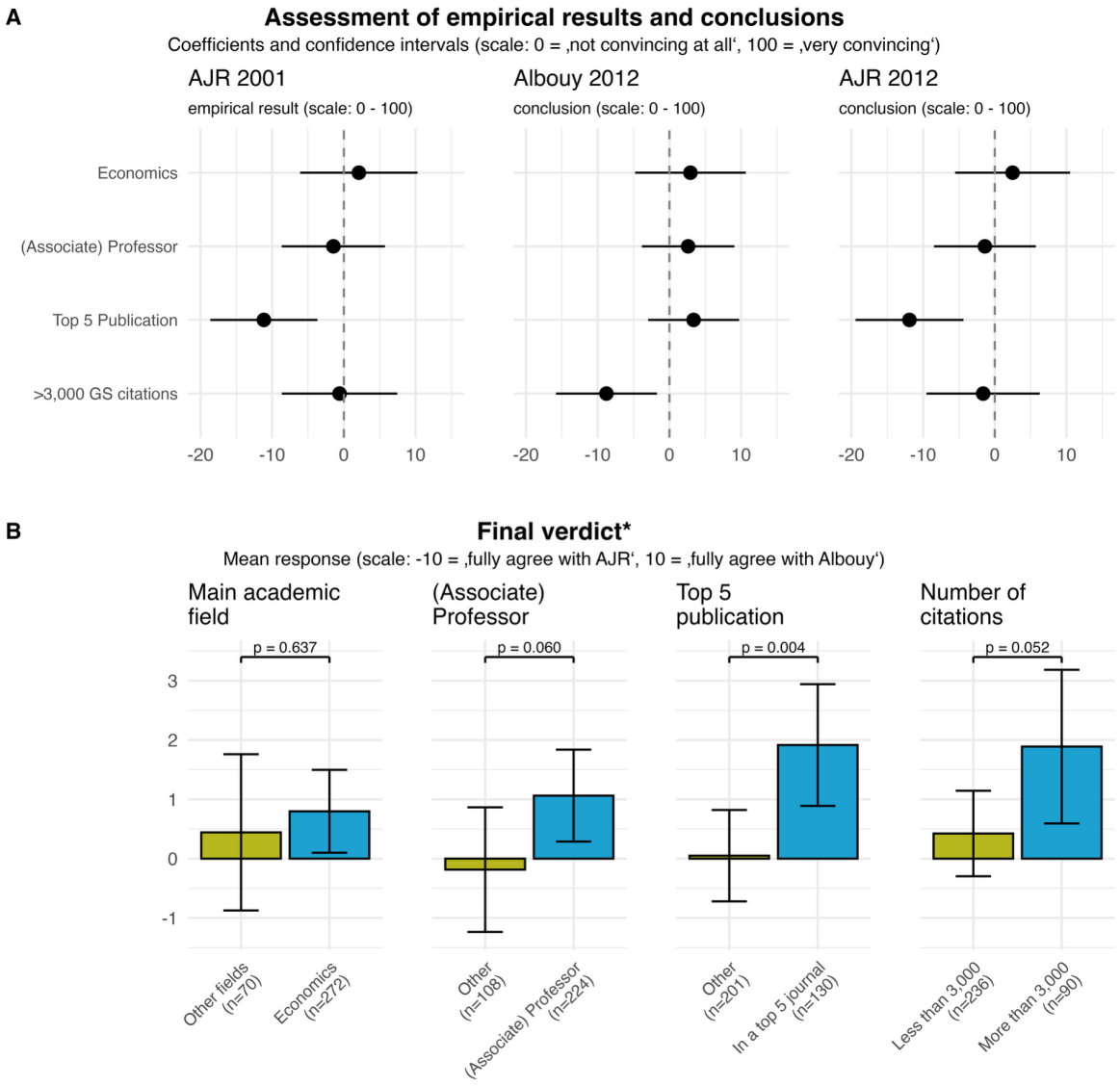
4.3. Heterogeneity analysis

We now examine heterogeneity by academic background and professional standing, first using multivariate regressions that include multiple indicators of seniority and academic standing simultaneously. The estimates are generally imprecise but suggest that respondents who published in a Top 5 journal are less convinced by both AJR2001 and AJR2012 (see Figure 5, Panel A). While this is consistent with the idea that more senior respondents are more critical of AJR, the pattern is not uniform: respondents with higher citation counts appear less convinced by Albouy2012's conclusion. We complement this with bivariate comparisons of the final verdict (Figure 5, Panel B), which avoid potential multicollinearity across respondent characteristics and are reported in full in Online Appendix B. These bivariate comparisons are not pre-specified, but they are based exclusively on the respondent characteristics included in the pre-specified multivariate analyses, and we report all pairwise comparisons for those characteristics. The patterns suggest that more senior and academically established respondents tend to lean more toward Albouy.[8]

[7] We furthermore pre-specified and elicited respondents' expectations about how other experts evaluate the empirical conclusions of the debate papers. The expected shares of participants rating the AJR2001, Albouy2012, and AJR2012 conclusions above 50 closely align with the corresponding shares in Figure 4; see Online Appendix A.2 for detailed results.

[8] We additionally pre-specified two further heterogeneity analyses. First, our ex-ante categorization into AJR-friendly and Albouy-friendly respondents (based on citation and authorship patterns) does not yield statistically significant differences. Second, respondents with greater prior familiarity with Albouy2012 tend to lean more toward Albouy in their final verdict, while respondents more familiar with AJR2012 tend to lean more toward AJR; both relationships are statistically significant. These results are presented in Online Appendix A.3.

Figure 5: Assessment of debate papers and final verdict across respondent characteristics



Notes: Panel A reports coefficient estimates from multivariate OLS regressions; horizontal error bars indicate 95% confidence intervals. The dependent variables are responses to the following questions: "How convincing do you find AJR2001's empirical result?" and "How convincing do you find Albouy2012's/AJR2012's conclusion?". Panel B reports mean final verdicts by respondent characteristics. Error bars indicate 95% confidence intervals around subgroup means; brackets report p-values from two-sided Welch t-tests of subgroup mean differences. "(Associate) Professor" refers to associate and full professors (as opposed to PhD students, post-doctoral researchers, assistant professors, and those outside academia). "Top 5 publication" indicates at least one Top 5 journal in their field (e.g., economics, political science, history). ">3,000 GS citations" refers to total Google Scholar citations. *Analyses shown in Panel B were not pre-specified.

5. What does the absence of a prevailing interpretation imply?

Empirical work on consensus in economics remains limited. A notable early contribution is Frey et al. (1984), who operationalize consensus in terms of the dispersion of expert views, with lower dispersion indicating a higher degree of consensus. The concept of scientific

consensus is not uniquely defined, and its role in scientific progress remains debated (Beatty and Moore 2010; Solomon 2007). It is commonly understood as a high degree of convergence in expert opinion, typically reflected in a dominant majority view and limited dispersion of beliefs, though not necessarily unanimity (Jorm 2025). A weaker notion is that of a prevailing interpretation, which still requires directional convergence of expert assessments. The evidence presented does not support convergence in expert views toward a prevailing interpretation, let alone a consensus. Expert opinions are widely dispersed across the spectrum, with no clear clustering around a particular position.

One explanation for this persistent disagreement is the absence of an independent yardstick to adjudicate between competing empirical claims. In the AJR-Albouy debate, disagreement centers on data construction and related methodological choices. Even for highly qualified experts, no external validation criterion is readily available. This corresponds closely to what Collins (1992) describes as the "experimenters' regress": when the correctness of an empirical result depends on the correctness of the methods used to obtain it, and these methods themselves are contested, there is no clear way to resolve disagreement.

This finding admits three interpretations. First, the absence of consensus is not only unsurprising but also without negative implication. Science is an ongoing debate, and pluralist interpretations of evidence are even desirable for open and unbiased inquiry (Gräbner and Strunk 2020, Hulme 2022, Stirling 2010). Second, the finding raises questions about how self-correction operates in economics, which is often associated with a Popperian view of science. If no prevailing interpretation of whether results are replicable and robust emerges even in a prominent debate over several years, it becomes difficult to determine which empirical claims should be considered robust.

A third and more general interpretation is that the absence of convergence to consensus is in tension with economics' claim to an influential role in societal debates (McCloskey 1983, Hulme 2022). More specifically, consensus or at least a prevailing interpretation is necessary for the effective functioning of evidence-based policy. In settings where such convergence is lacking, policy analysis may nonetheless convey a degree of certainty that is not warranted by the underlying evidence (Manski 2013).

While the AJR-Albouy debate is particularly prominent, the pattern we observe is not unique. Evidence from a systematic analysis of robustness replications published as comments in the *American Economic Review* shows that such replications rarely lead to convergence in the literature (Ankel-Peters et al. 2025). Comments are infrequently cited and do not affect the citation trajectories of the original papers, even when they raise substantive concerns. Moreover, surveys of authors and replicators reveal that there is often no agreement on whether the original contribution holds. In the absence of an independent adjudication criterion, such disagreements are unlikely to be resolved even among domain experts because competing assessments rest on contested methodological choices rather than externally verifiable standards.

6. Conclusion

When AJR received the Nobel Prize in October 2024, the committee emphasized that their work provided "*solid evidence for a causal effect of these institutions on long-run prosperity*" (Teorell 2024). The experts in our survey do not converge to such a clear conclusion. While there is broad agreement on the general importance of institutions, opinions diverge substantially on the empirical approach and on whether AJR2001's contribution holds. We find no prevailing interpretation of the debate between AJR and Albouy. We thereby document that expert disagreement can persist even in prominent replication debates, and that replication does not necessarily lead to a shared evaluation of empirical claims.

Our approach has limitations. In particular, participation in the survey is voluntary, and we cannot fully rule out selection into the sample, for example by experts with particularly critical views of AJR. Moreover, alternative approaches to eliciting consensus could deliver other outcomes. However, given the substantial dispersion in responses, we are confident that our assessment of a lacking consensus or prevailing interpretation are robust.

Our findings point to the potential value of incorporating expert assessments when evaluating contested empirical literatures. This may be particularly relevant for the growing literature on the robustness of empirical evidence, which is likely to generate further replication debates (Brodeur et al. 2024, 2026; Campbell et al. 2024). More broadly, similar challenges arise in other empirical literatures where large bodies of research do not yield a clear picture. In such

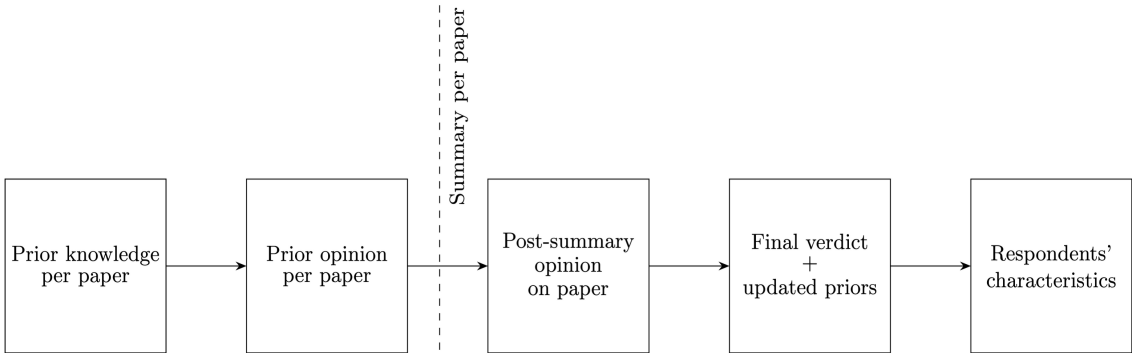
settings, structured expert surveys may provide a complementary approach to synthesizing evidence.

Appendix

Table A1 Correspondence between pre-analysis plan and presented results

Research question	Pre-specified research question	Presented in: Descriptive Results	Presented in: Hypothesis test / OLS results
Primary 1	With whom do experts agree more: AJR or Albouy?	Section 4.1 & Appendix A.1	Section 4.1 & Appendix A.1
Secondary 1	How familiar are experts with the original paper, comment, and reply?	Section 4.2 & Appendix A.2	n.a.
Secondary 2	How convincing do experts find the original paper, comment and reply based on their prior knowledge of these papers?	Section 4.2 & Appendix A.2	Section 4.2 (Footnote 6) & Appendix A.2
Secondary 3	How do experts evaluate the original paper, comment, and reply?	Section 4.2 & Appendix A.2	n.a.
Secondary 4	How do experts believe other experts evaluate the paper, comment and reply?	Section 4.2 (Footnote 7) & Appendix A.2	Appendix A.2
Secondary 5	Has the experts' priors on the paper, comment and reply changed after reading the summaries provided in the survey?	Section 4.2 & Appendix A.2	Appendix A.2
Exploratory 1	To what extent do experts likely to be AJR-friendly or Albouy-friendly agree with AJR/Albouy?	Section 4.3 (Footnote 8) & Appendix A.3	Section 4.3 (Footnote 8) & Appendix A.3
Exploratory 2	Do experts with different backgrounds have systematically different opinions?	Section 4.3 & Appendix A.3	Section 4.3 & Appendix A.3
Exploratory 3	To what extent do experts with different levels of familiarity with the papers respond differently?	Section 4.3 (Footnote 8) & Appendix A.3	Section 4.3 (Footnote 8) & Appendix A.3

Figure A1: Questionnaire Structure



References

Acemoglu, D., Johnson, S., & Robinson, J. A. (2000). *The colonial origins of comparative development: An empirical investigation.* National Bureau of Economic Research (Cambridge, MA) Working Paper No. 7771.

Acemoglu, D., Johnson, S., & Robinson, J. A. (2001). The colonial origins of comparative development: An empirical investigation. *American Economic Review*, *91(5)*, 1369–1401.

Acemoglu, D., Johnson, S., & Robinson, J. A. (2005). *A response to Albouy's 'A reexamination based on improved settler mortality data.'* Manuscript. Available online at: https://economics.mit.edu/sites/default/files/publications/A%20Response%20to%20Albouys%20A%20Reexamination%20Based%20on%20Imp.pdf (Accessed May 14, 2026).

Acemoglu, D., Johnson, S., & Robinson, J. A. (2012). The colonial origins of comparative development: An empirical investigation: Reply. *American Economic Review*, *102(6)*, 3077–3110.

Albouy, D. (2012). The colonial origins of comparative development: An empirical investigation: Comment. *American Economic Review*, *102(6)*, 3059–3076.

Alesina, A. & Giuliano, P. (2015). Culture and institutions. *Journal of Economic Literature, 53(4),* 898–944.

Ankel-Peters, J., Fiala, N., & Neubauer, F. (2025). Is economics self-correcting? Replications in the American Economic Review. *Economic Inquiry*, *63*, *463–485*.

Aspinall, W. (2010). A route to more tractable expert advice. *Nature, 463,* 294–295.

Assenova, V. A. & Regele, M. (2017). Revisiting the effect of colonial institutions on comparative economic development. *PLoS ONE*, *12(5),* e0177100.

Beatty, J., & Moore, A. (2010). Should we aim for consensus? *Episteme*, 7(3), 198–214.

Benjamin, D. J., Berger, J. O., Johannesson, M., Nosek, B. A., Wagenmakers, E.-J., Berk, R., Bollen, K. A., Brembs, B., Brown, L., Camerer, C., Cesarini, D., Chambers, C. D., Clyde, M., Cook, T. D., De Boeck, P., Dienes, Z., Dreber, A., Easwaran, K., Efferson, C., …, & Johnson, V. E. (2018). Redefine statistical significance. *Nature Human Behaviour*, *2*, 6–10.

Black, D. A., Sanders, S. G., Taylor, E. J., & Taylor, L. J. (2015). The impact of the Great Migration on mortality of African Americans: Evidence from the Deep South. *American Economic Review*, *105(2)*, 477–503.

Brodeur, A., Cook, N., & Heyes, A. (2020). Methods matter: p-Hacking and publication bias in causal analysis in economics. *American Economic Review*, *110(11)*, 3634–3660.

Brodeur, A., Esterling, K., Ankel-Peters, J., Bueno, N. S., Desposato, S., Dreber, A., Genovese, F., Green, D. P., Hepplewhite, M., Hoces de la Guardia, F., & Johannesson, M. (2024). Promoting reproducibility and replicability in political science. *Research & Politics*, *11(1)*.

Brodeur, A., Mikola, D., Cook, N., Fiala, L., Brailey, T., Briggs, R., De Gendre, A., Dupraz, Y., Gabani, J., Gauriot, R., & Haddad, J. (2026). Reproducibility and robustness of economics and political science research. *Nature*, *652*(8108), 151–156.

Camerer, C. F., Dreber, A., Forsell, E., Ho, T.-H., Huber, J., Johannesson, M., Kirchler, M., Almenberg, J., Altmejd, A., Chan, T., Heikensten, E., Holzmeister, F., Imai, T., Isaksson, S., Nave, G., Pfeiffer, T., Razen, M., & Wu, H. (2016). Evaluating replicability of laboratory experiments in economics. *Science, 351(6280),* 1433‑1436.

Camerer, C. F., Dreber, A., Holzmeister, F., Ho, T.-H., Huber, J., Johannesson, M., Kirchler, M., Nave, G., Nosek, B. A., Pfeiffer, T., Altmejd, A., Buttrick, N., Chan, T., Chen, Y., Forsell, E., Gampa, A., Heikensten, E., Hummer, L., Imai, T., …, & Wu, H. (2018). Evaluating the replicability of social science experiments in Nature and Science between 2010 and 2015. *Nature Human Behaviour*, *2,* 637–644.

Campbell, D., Brodeur, A., Dreber, A., Johannesson, M., Kopecky, J., Lusher, L., & Tsoy, N. (2024). *The robustness reproducibility of the American Economic Review*. I4R Discussion Paper Series No. 124.

Casey, G. & Klemp, M. (2021). Historical instruments and contemporary endogenous regressors. *Journal of Development Economics*, *149*, 102586.

Collins, H. (1992). *Changing order: Replication and induction in scientific practice*. University of Chicago Press.

Collins, H. M. & Evans, R. (2002). The third wave of science studies: Studies of expertise and experience. *Social Studies of Science*, *32(2)*, 235–296.

Curtin, P. D. (1989). *Death by migration: Europe's encounter with the tropical world in the 19th Century*. Cambridge University Press.

Curtin, P. D. (1998). *Disease and empire: The health of European troops in the conquest of Africa*. Cambridge University Press.

Curtin, P. D., Feierman, S., Thompson, L., & Vansina, J. (1995). *African history: From earliest times to independence* (2nd ed.). Longman.

Deaton, A. (2010). Instruments, randomization, and learning about development. *Journal of Economic Literature*, *48*, 424–455.

Dreber, A., Pfeiffer, T., Almenberg, J., Isaksson, S., Wilson, B., Chen, Y., Nosek, B. A., & Johannesson, M. (2015). Using prediction markets to estimate the reproducibility of scientific research. *Proceedings of the National Academy of Sciences*, *112(50)*, 15343–15347.

Easterly, W. & Levine, R. (2016). The European origins of economic development. *Journal of Economic Growth*, 21, 225–257.

Forsell, E., Viganola, D., Pfeiffer, T., Almenberg, J., Wilson, B., Chen, Y., Nosek, B. A., Johannesson, M., & Dreber, A. (2019). Predicting replication outcomes in the Many Labs 2 study. *Journal of Economic Psychology, 75*(Part A SI).

Fourcade, M., Ollion, E., & Algan, Y. (2015). The superiority of economists. *Journal of Economic Perspectives, 29(1)*, 89–114.

Freese, J. & Peterson, D. (2017). Replication in social science. *Annual Review of Sociology*, *43(1)*, 147–165.

Frey, B. S., Pommerehne, W. W., Schneider, F., & Gilbert, G. (1984). Consensus and dissension among economists: An empirical inquiry. *American Economic Review, 74(5)*, 986–994.

Gallup, J. L., Sachs, J. D., & Mellinger, A. D. (1999). Geography and economic development. *International Regional Science Review*, *22(2)*, 179–232.

Glaeser, E. L., La Porta, R., Lopez-de-Silanes, F., & Shleifer, A. (2004). Do institutions cause growth? *Journal of Economic Growth, 9*, 271–303.

Gräbner, C., & Strunk, B. (2020). Pluralism in economics: Its critiques and their lessons. *Journal of Economic Methodology*, *27(4)*, 311–329.

Haveresch, N., Ankel-Peters, J., & Bensch, G. (2025). A Slippery Slope: Topographic Variation as an Instrument. I4R Discussion Paper Series No. 278.

Hemming, V., Hanea, A. M., Walshe, T., & Burgman, M. A. (2020). Weighting and aggregating expert ecological judgments. *Ecological Applications*, *30(4)*, e02075.

Hulme, M. (2022). Scientific consensus-seeking. In: *A critical assessment of the intergovernmental panel on climate change*. De Pryck, K., Hulme, M., eds. Cambridge University Press, 178–186.

Humphreys, M. (2015). What has been learned from the deworming replications: A nonpartisan view [Blog post]. Available online at: http://emiguel.econ.berkeley.edu/wordpress/wp-content/uploads/2020/11/What_Has_Been_Learned_from_the_Deworming_Replications__A_Nonpartisan_View_Macartan_Humphreys_Blog.pdf (Accessed May 14, 2026).

Jorm, A. (2025). *Expert consensus in science*. Palgrave Macmillan.

Kronlund, A. (2023). From political science to politicizing science? A study of the discipline's presence in the debates of the United States Congress, 1981–2021. *Parliaments, Estates and Representation, 43(3)*, 287–305.

Lal, A., Lockhart, M., Xu, Y., & Zu, Z. (2024). How much should we trust instrumental variable estimates in political science? Practical advice based on 67 replicated studies. *Political Analysis*, *32(4)*, 521–540.

Malan, M., Ankel-Peters, J., Buchner, M., Fiala, N., Johannesson, M., & Rose, J. (2024). Pre-analysis plan: Settling settler mortality – An expert survey. OSF. https://osf.io/fx8p5/.

Manski, C. F. (2013). *Public policy in an uncertain world.* Harvard University Press.

Markevich, A., & Zhuravskaya, E. (2018). The economic effects of the abolition of serfdom: Evidence from the Russian Empire. *American Economic Review, 108(4–5)*, 1074–1117.

Martini, C. (2014). Seeking consensus in the social sciences. In: *Experts and consensus in social science*. Martini, C. and Boumans, M., eds. Springer International Publishing, 115–130.

McArthur, J. W. & Sachs, J. D. (2001). *Institutions and geography: Comment on Acemoglu, Johnson and Robinson (2000)*. NBER Working Paper No. 8114.

McCloskey, D. N. (1983). The rhetoric of economics. *Journal of Economic Literature, 21(2)*, 481–517.

Mellon, J. (2025). Rain, rain, go away: 194 potential exclusion-restriction violations for studies using weather as an instrumental variable. *American Journal of Political Science*, *69(3)*, 881–898.

Ozier, O. (2021). Replication redux: The reproducibility crisis and the case of deworming. *The World Bank Research Observer, 36(1),* 101–130.

Roodman, D. (2025). *Opinion on the replication debate over Heyes and Saberian (2019)*. I4R Discussion Paper Series No. 227.

Solomon, M. (2007). *Social empiricism.* MIT Press.

Stirling, A. (2010). Keep it complex. *Nature*, *468,* 1029–1031.

Tabellini, G. (2010). Culture and institutions: Economic development in the regions of Europe. *Journal of the European Economic Association*, *8(4)*, 677–716.

Teorell, J. (2024). Presentation speech: The Sveriges Riksbank Prize in economic sciences in memory of Alfred Nobel 2024 [Speech transcript]. Available online at: https://www.nobelprize.org/prizes/economic-sciences/2024/ceremony-speech/ (Accessed May 14, 2026).

Teplitskiy, M., Duede, E., Menietti, M., & Lakhani, K. R. (2022). How status of research papers affects the way they are read and cited. *Research Policy*, *51(4)*, 104484.

ifso working paper

ifso working papers are preliminary scholarly papers emerging from research at and around the Institute for Socio-Economics at the University of Duisburg-Essen.

All **ifso working papers** at uni-due.de/soziooekonomie/wp

ISSN 2699-7207

UNIVERSITÄT DUISBURG ESSEN

Open-Minded

ifso institute for socio-economics

Institute for Socio-Economics
University of Duisburg-Essen

Lotharstr. 65
47057 Duisburg
Germany

uni-due.de/soziooekonomie
wp.ifso@uni-due.de