RUHR
ECONOMIC PAPERS

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

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

technische universität dortmund
rwi
#1176

Imprint

Ruhr Economic Papers #1176

Responsible Editor: Manuel Frondel
RWI – Leibniz-Institut für Wirtschaftsforschung e.V.
Hohenzollernstraße 1–3 | 45128 Essen, Germany
Fon: +49 201 8149-0 | email: rwi@rwi-essen.de
www.rwi-essen.de

The Institute has the legal form of a registered association;
Vereinsregister, Amtsgericht Essen VR 1784

Published by

RWI – Leibniz-Institut für Wirtschaftsforschung e.V.
Hohenzollernstr. 1-3, 45128 Essen, Germany

Ruhr-Universität Bochum (RUB), Department of Economics
Universitätsstr. 150, 44801 Bochum, Germany

Technische Universität Dortmund, Department of Economic and Social Sciences
Vogelpothsweg 87, 44227 Dortmund, Germany

Universität Duisburg-Essen, Department of Economics
Universitätsstr. 12, 45117 Essen, Germany

Bergische Universität Wuppertal, Schumpeter School of Business and Economics
Gaußstraße 20, 42119 Wuppertal

RWI is funded by the Federal Government and the federal state of North Rhine-Westphalia.

ISSN 1864-4872 (online)
ISBN 978-3-96973-361-5
DOI https://dx.doi.org/10.4419/96973361

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.

September 2025

Abstract

Consensus is crucial to authoritative science, as is replicability. Yet, in economics and the social sciences, the publication of contradictory replications often sparks fierce debates between replicators and original authors. This paper investigates whether experts can reach a consensus on a famous yet unsettled debate about the robustness of the seminal paper by Acemoglu, Johnson, and Robinson (AJR, 2001) following a replication by Albouy (2012). We recruited 352 experts mainly from the pool of scholars citing one of the involved or similar articles. Through a structured online questionnaire, we assess the extent to which these experts align with AJR or Albouy. Our findings indicate no consensus on whether the original results hold after Albouy's replication, although there is a slight tendency among experts to side with the replicator. Exploratory heterogeneity analysis suggests that experts with greater academic credentials are more likely to align with Albouy. Our study demonstrates a potential way to scope scientific consensus formation and navigate replication debates and contested literatures.

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

Acknowledgements

We thank Abel Brodeur, David Card, 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/68dccff174a6e133fffd05d5. 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 Open Philanthropy. *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 must lead to a consensus or whether disagreement is even needed for scientific deliberation is a matter of an ongoing debate in the philosophy of science (Beatty and Moore 2010). However, there is consensus that consensus is needed in fields where science strives for an authoritative role in societal debates (Hulme 2022). Economics is one of those disciplines, and so are parts of other social sciences (Fourcade et al. 2015, Frey et al. 1984, Kronlund 2023, Martini 2014, McCloskey 1983). At the same time, replication itself in many cases does not lead to consensus, especially when replication results are non-confirmatory. This commonly results in disputes between replicators and replicated scholars regarding who has implemented the analysis correctly and therefore regarding the replicability of the original result. Absent an independent standard to judge correctness, a circular problem arises in the knowledge generation process, known as the "experimenters' regress" in the sociology of science (Collins 1992).

In economics and the social sciences, there is mounting evidence that the experimenters' regress prevails when replications challenge published studies (Ankel-Peters et al. 2025, Auspurg and Brüderl 2024, Freese and Peterson 2017, Humphreys 2015, Ozier 2021). In this paper, we examine whether a consensus among experts has emerged in a seminal replication debate in economics between Acemoglu, Johnson, and Robinson (2001, 2012, henceforth AJR2001 and AJR2012) and Albouy (2012; henceforth Albouy2012), more than 10 years after publication. The original paper, the replication and the reply by the original authors appeared in the *American Economic Review*. The journal editors did not issue an editorial note or statement on the matter, and thus did not take sides.

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 economic growth. Their empirical approach relies on the instrumental variable (IV) method, which is frequently used in economics (Angrist and Kruger 2001). In brief, IVs exploit naturally occurring variation that must be exogenous to the socio-economic relationship under evaluation – and hence akin to an experiment as it is conducted in medical trials. More

specifically, AJR2001 use historical settler mortality as an IV, assuming that it – exogenously – affected early institutions which in turn determine today's institution. AJR2001's abstract prominently summarizes the approach and the findings:

> *"Exploiting differences in European mortality rates as an instrument for current institutions, we estimate large effects of institutions on income per capita. Once the effect of institutions is controlled for, countries in Africa or those closer to the equator do not have lower incomes."* (AJR2001, p. 1369)

AJR2001 reinforced the institution-focused narrative in economics thinking, which competes with theories that put human-capital (Bolt and Bezemer 2009, Easterly and Levine 2016, Glaeser et al. 2004), geography (Gallup et al. 1999, McArthur and Sachs 2001) or culture (Alesina and Giuliano 2015, Tabellini 2010) at their center. The paper has had a profound impact on the economics discipline, with around 18,000 citations on Google Scholar until 2024, culminating in the Sveriges Riksbank Prize in Economic Sciences in Memory of Alfred Nobel awarded to the authors in October 2024.

In a replication using the same data as AJR2001, Albouy2012 questions the reliability of AJR2001's settler mortality data and highlights two key concerns. First, Albouy identifies questionable decisions regarding data imputation and approximation that, if handled differently, render the AJR2001 results statistically insignificant and uninterpretable. The imputation affects AJR2001's central variable, the approximation of historical settler mortality. Albouy2012 reveals that AJR2001 could retrieve settler mortality data only for 28 countries out of 64 included in the analysis; the remaining 36 are imputed from other countries with disease environments that AJR2001 consider sufficiently similar. Second, Albouy2012 discloses that most AJR2001's settler mortality rates are derived from (non-combatant) mortality data of soldiers *on campaign*. Albouy2012 argues that on campaign, also excluding combats, soldiers systematically faced higher mortality than soldiers in barracks due to poor shelter and hygiene. Albouy2012 proposes remedies such as discarding the 36 countries with imputed settler mortality data and adding a dummy variable indicating where campaign data were used. Applying these remedies turns AJR's results insignificant.

In sum, Albouy2012 fundamentally questions the feasibility of AJR2001's empirical strategy:

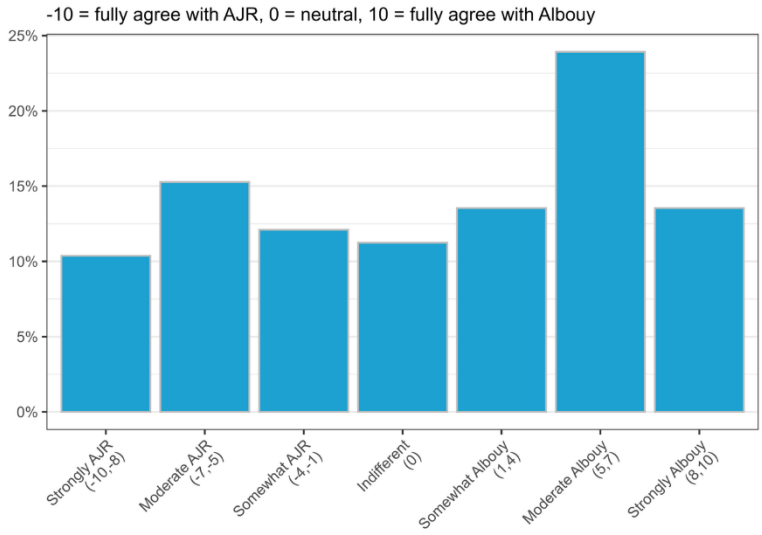
> *"Given the limited data sources currently available, it seems unlikely that a convincing set of settler mortality rates can be constructed. As such, cross-country growth regressions cannot disentangle the effect of settler mortality from that of other variables that may explain institutions and growth, such as geography, climate, culture, and preexisting development, leaving the AJR theoretical hypotheses without a strong empirical foundation."* (Albouy2012, p. 3073)

In their reply, AJR refute Albouy2012's critique (AJR2012) and state that the "big picture from AJR (2001) remains intact and remarkably robust" (p. 3081). AJR2001 do not dispute the missing mortality data for the 36 countries, but argue that discarding all those 36 countries does not do justice to what their historical sources know about these countries. They also show that Albouy2012's finding is "largely driven by one outlier, Gambia" (p. 3078). Moreover, AJR2012 argue that Albouy2012's distinction between campaign and non-campaign episodes is overstated, since there are "no systematic differences in mortality rates" (p. 3079) between the two. They also point out inconsistencies in Albouy2012's classification of campaigns. To the best of our knowledge, the exchange between AJR and Albouy did not extend beyond these publications.

We henceforth refer to these three papers – AJR2001, Albouy2012 and AJR2012 – as the *debate papers*. There is no clarity between the debating authors about whether AJR2001's contribution holds. Both sides accuse each other of having made incorrect methodological decisions – hence a classical experimenters' regress situation. Ideally, the scientific community solves such a situation by organic self-correction, that is, whether AJR2001's contribution holds or not should be "the outcome of social interactions among scientists" (Freese and Peterson, 2017, p. 149). Our paper scrutinizes whether the prevailing interpretation in the scientific community has iterated towards such a consensus. We use expert opinions for this assessment, which we elicited by means of a survey conducted between April and May 2024 and thus prior to the Nobel award. Our findings suggest that no consensus has emerged. Figure 1 shows the results of our pre-specified primary research question: *With whom do respondents agree more – AJR or Albouy?* After reviewing summaries of all three debate papers, we asked experts to position themselves on a scale from -10 ("I fully agree with AJR") to 10 ("I fully agree with Albouy"). While a pre-specified *t*-test suggests a subtle tendency towards Albouy (mean = 0.76, t(346) =

2.47, p = 0.014)[1], the distribution of responses remains widely dispersed across the spectrum. About 38% of experts are in what one might refer to as the pro-AJR camp (i.e. below zero). About 51% are pro-Albouy (i.e. above zero); 11% are indifferent (i.e. they report a zero).

Figure 1: 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.'

Our analysis is based on an anonymized structured survey among 352 respondents. Since it is difficult to delineate who is an expert in a particular academic discussion, we follow Collins and Evans (2002)'s foundational work on expertise and aim to capture the opinion of scholars with both *interactional* and *contributory expertise* on the subject matter. We therefore did not mass-email the survey or post it on social media. Instead, we recruited participants primarily based on citations of the three debate papers, complemented by authors and citers of a short list of similar articles. We assume that citing behavior is an approximative indicator for expertise on the cited topic (Teplitskiy et al. 2022) but we acknowledge that citations may also reflect strategic considerations (Rubin and Rubin 2021).

In total, we invited 3,022 scholars, of whom 309 (10.2%) participated in the survey. An additional 43 participants were recruited via mailing lists of two professional networks. As we demonstrate, most of our respondents are economics professors with Top 50 journal

[1] 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".

publications and well-cited Google Scholar profiles. We therefore consider them to be *experts* on the subject. Our online questionnaire first elicited prior knowledge and beliefs about the three debate papers. We then provided respondents with a summary of each paper before they were asked to assess AJR2001's approach and the core of Albouy2012's criticism as well as AJR2012's rebuttal. Last, respondents were asked to give their final verdict on whether the contribution of AJR2001 holds (depicted in Figure 1), and whether their views had changed during the survey. Our analysis follows a detailed pre-analysis plan, if not stated otherwise (Malan et al. 2024). Table 1 in Section 4.1 lists the pre-specified research questions and where the corresponding results are presented.

Our paper builds on previous endeavors to shed light on this and other unsettled replication debates, for example by Ozier (2021), Humphreys (2015) and Roodman (2025). It is also related to replication markets that use expert predictions to assess replicability of published findings (Camerer et al. 2016; Camerer et al. 2018, Dreber et al. 2015; Forsell et al. 2019). We push further by adding the perspective of a large number of experts on both the replicated paper and the replication and by attempting to seek a potential consensus. More generally, the aim of our paper is to encourage further research into the use of expert knowledge when resolving replication debates and interpreting academic literatures, particularly contested ones. Different approaches, for example on how to select experts and how to tap into their knowledge, are possible and should be piloted, too (see Aspinall 2010, Fraser et al. 2023, Hemming et al. 2020, Martinez I Coma and van Ham 2015). We contend that surveying experts is an important avenue to catalyze the synthesis of evidence. Such traceable approaches can also complement policy advisory panels. Leaving this evidence synthesis to organic consensus seeking processes does not do justice to the urgency of many policy issues studied in the empirical social sciences.

2. Results

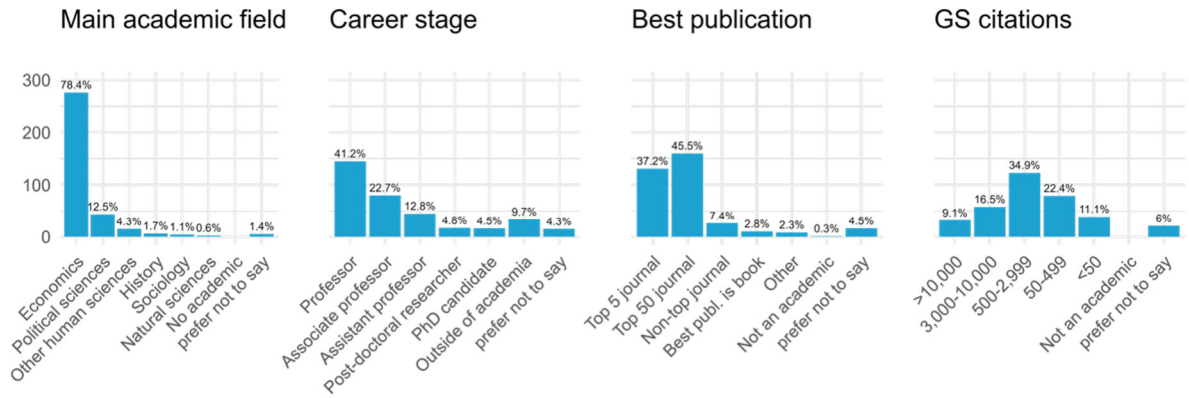
2.1 Recruitment of Experts

Using the Scopus database, we identified scholars who cited AJR2001 (2,493 citers), Albouy2012 (2012; 58 citers), and AJR2012 (78 citers). Additionally, we identified authors of studies that use similar identification strategies (85) and both authors and citers of papers that are critical of this type of causal historical research (491), including those who authored other

replications and critiques of AJR. Responding participants from this population of in total 3,022 potential experts form our main sample. We expanded recruitment through two academic networks, specifically the Institute for Replication (I4R) mailing lists and the Development Economics Committee of the German Economic Association. We used personalized links so we can track and separate recruitment channels in our analysis, while still fully preserving respondents' anonymity.

Of the 352 respondents who completed the survey, 309 came from our contact list of citers and authors—a decent 10.2% response rate. The other 43 responded to the survey shared through the mailing lists. As can be seen in Figure 2, the majority are economists (78%). Figure also demonstrates that we successfully recruited *experts*. Two thirds of our sample are professors (41% are full professors, 23% are associate professors), and 83% have published in at least one Top 50 journal in economics, with 37% having published in a Top 5 journal. A total of 213 respondents (61%) report 500 citations or more on Google Scholar. About 9% even have more than 10,000 Google Scholar citations.

Figure 2: Descriptive statistics of respondents (n = 352)

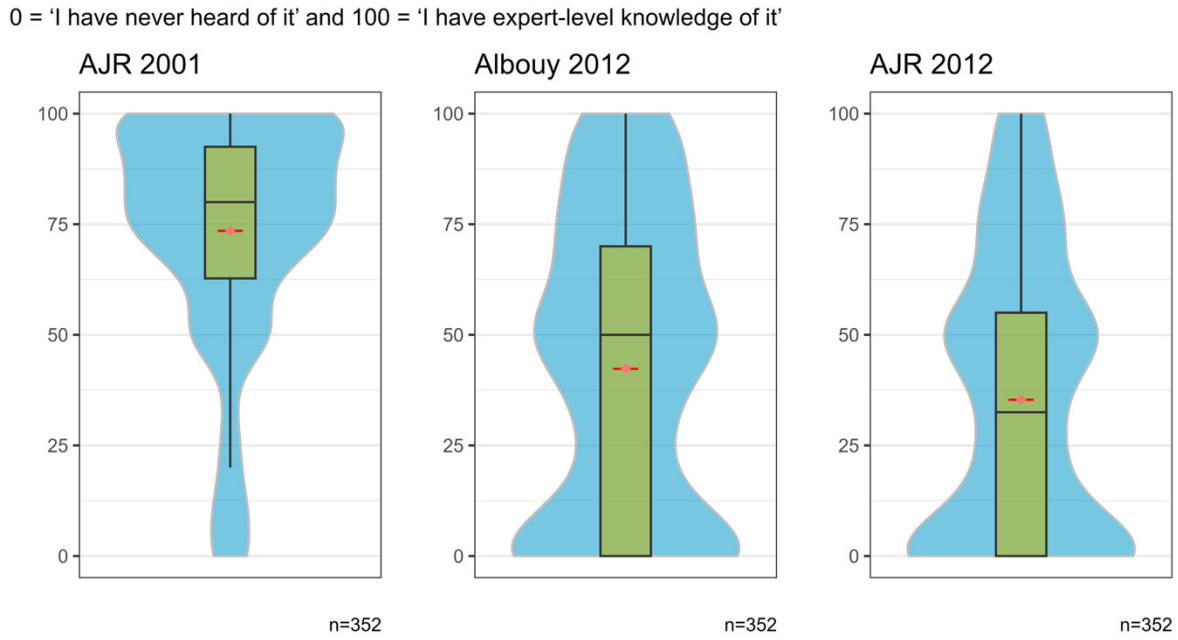


2.2 Main Results

Our pre-analysis plan specifies one Primary Research Question (PRQ), five Secondary Research Questions (SRQ), and three Exploratory Research Questions (ERQ). We first asked respondents to rate their familiarity with each paper before we provided any further information (SRQ1 PAP). Respondents are largely familiar with AJR2001 on a 0–100 scale from "never heard of it" to "expert-level knowledge" (mean score: 73.51), while Albouy's comment (42.32) and AJR's reply (35.32) are less well known, as shown in Figure 3.

We then asked respondents with prior knowledge (i.e., an above zero familiarity rating) to evaluate how convincing they found the respective debate paper on a scale from 0 ("not convincing at all") to 100 ("very convincing") – still before the survey provided any background information (SRQ2 PAP). We pre-specified pairwise *t*-tests, which show that Albouy's comment is found to be more convincing than AJR's original study and reply (AJR2001 vs. Albouy2012: mean difference = -10.59, t(210) = - 3.59, p < 0.001; AJR2012 and Albouy2012: mean difference = -12.85, t(174) = -4.20, p < 0.001; see Appendix 3.3 for details).

Figure 3: Respondents' familiarity with the debate papers

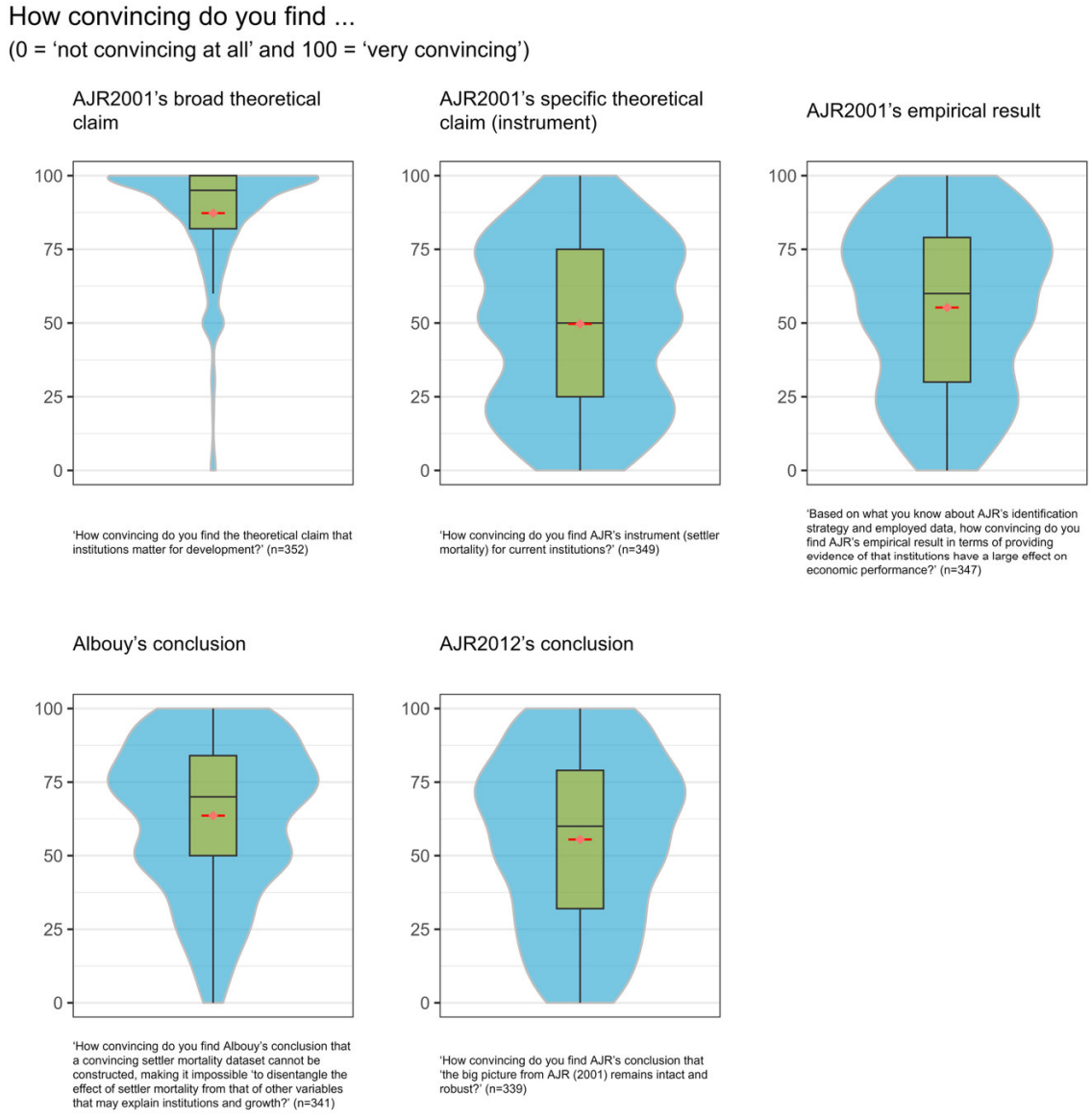


Notes: Box plots indicate the 25th, 50th (median), and 75th percentiles. Red dots represent mean values.

Next, respondents were presented with a summary of each of the three debate papers and were then asked to evaluate the paper's analytical decisions and main arguments (SRQ3 PAP). The upper panels of Figure 4 present the detailed analysis of AJR2001. Note that respondents were asked these questions before the survey presented the summary of Albouy2012's critique. It is striking that respondents overwhelmingly agree with AJR2001's broad theoretical claim that institutions matter for development – a level of agreement that can probably be called a consensus. This overwhelming agreement vanishes for AJR's specific theoretical claim, the rationale of the instrument, that historical settler mortality shaped European settlement patterns, which in turn determined early institutions and, ultimately, current institutions. Responses are almost evenly distributed across the spectrum. The agreement pattern for the

overall empirical results shown in the upper right panel of Figure 4 appears very similar, though a tad more in agreement with AJR2001 than for the specific theoretical claim. In sum, the results reveal that the power of AJR2001 is the appeal of its broad theoretical claim, not so much the empirical analysis and neither the specific theoretical claim.

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.

The lower panels of Figure 4 display the respondents' assessment of the respective conclusions of Albouy2012 and AJR2012. Respondents tend to agree with Albouy2012's overall conclusion. Taking the almost-consensual agreement with AJR2001's theoretical claim as a benchmark, though, shows that the agreement with Albouy2012 is a lot more hesitant. For AJR's reply,

AJR2012, many respondents again tend to agree with their conclusion, however, there is visibly more opposition. In sum, the paper-specific patterns in Figure 4 mirror the overall verdict in Figure 1, enhancing confidence in our interpretation of a lacking consensus. The full analysis of respondents' agreement with Albouy2012 and AJR2012 is provided in Appendix 3.4.

We furthermore elicted respondents' expectations about how other experts would evaluate the empirical result of AJR2001, Albouy2012 and AJR2012. Their expectations are indeed in line with what we observe across the distribution of experts (SRQ4 PAP; see Appendix 3.5 for detailed results and *t*-tests), underpinning that we tap into the knowledge of a scientific community that forms judgements about debates based on "social interactions", as Freese and Peterson (2017, p. 149) suggested.

Next, the survey asked for the respondents' final verdict, depicted in Figure 1 (PRQ1 PAP), showing that there is no consensus among the respondents on whether AJR's claim holds or not in the light of Albouy's replication. We then asked respondents whether reading the summaries had led them to update their priors about the papers (SRQ5 PAP). Indeed, it had a noteworthy impact on respondents' assessment of AJR2001, although not a massive one. About 20.4% reported being "less" or "much less" convinced by AJR2001 than they were before the survey, whereas 8.3% say they are "more" or "much more" convinced by AJR2001. On average, the effect is negative and statistically significant, with a mean difference of -0.13. This effect is modest in magnitude given the scale from -2 ("much less convinced") to +2 ("much more convinced"; $t(313) = -3.92$, $p < 0.001$; see Appendix 3.6). We do not observe a similar significant effects on whether participants were more or less convinced by Albouy2012 or AJR2012 (Albouy: mean = -0.05, $t(211) = -1.10$, $p = 0.27$; AJR2012: mean = 0.02, $t(181) = 0.40$, $p = 0.687$).

Next, we examine whether respondents' academic backgrounds and professional standing are associated with how they assess the debate papers, as shown in Figure 5. We find some indication that academically more influential respondents are more critical of AJR. Note that the regression results on the debate papers' conclusions are pre-specified (ERQ 2 PAP), while

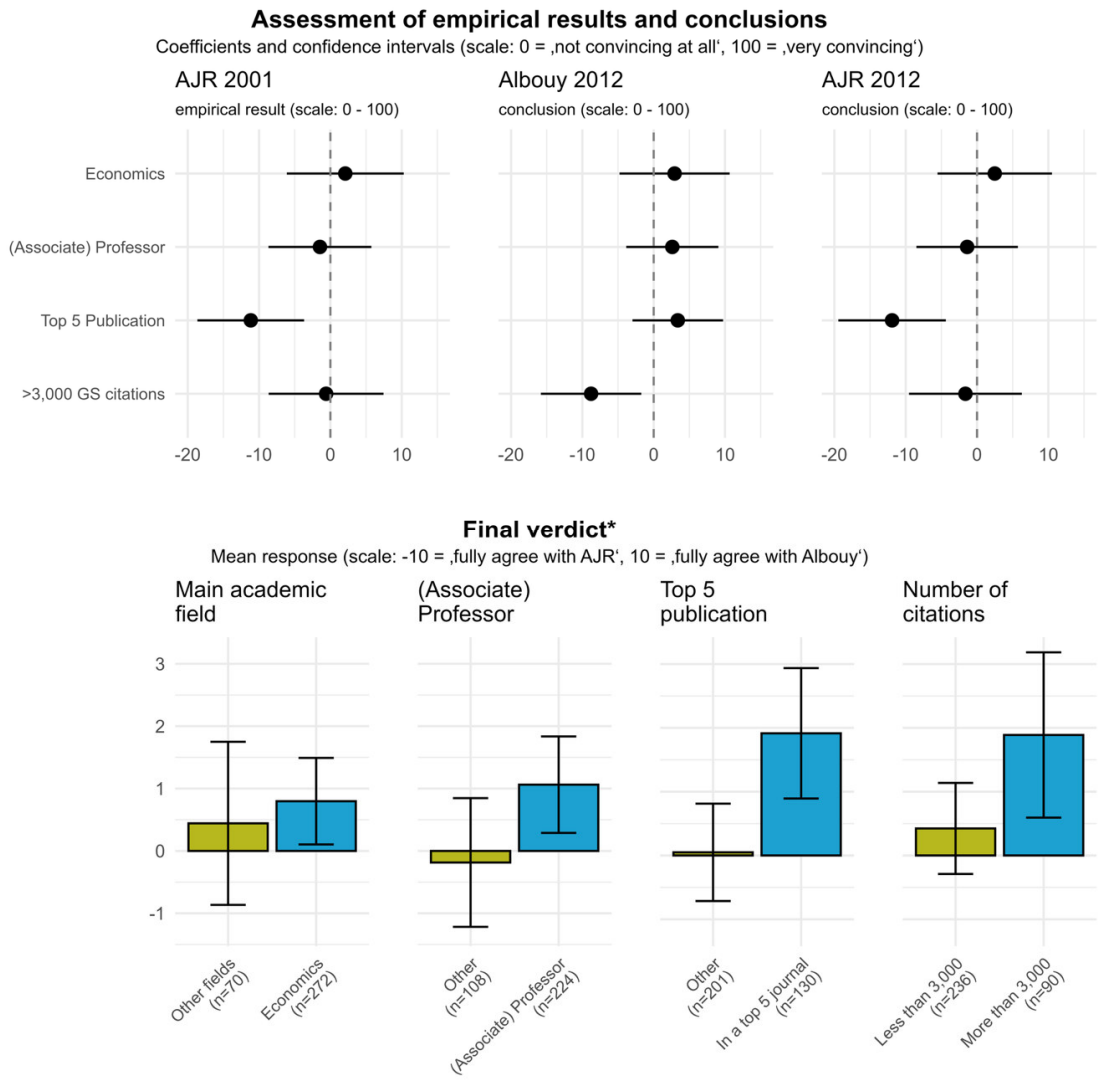
the bivariate analyses of the final verdict are not.[2] Overall, the regression results are imprecise and inconclusive, suggesting, on the one hand, that respondents with a Top 5 journal publication are 11.2 points (on a 0-100 scale) less convinced by the empirical result of AJR2001 ($p = 0.004$) and 11.9 points less convinced by the conclusion of AJR2012's reply ($p = 0.002$). On the other hand, those with more than 3,000 citations are 8.8 points less convinced by Albouy2012's conclusion ($p = 0.015$).

The bivariate examination of heterogeneity underlying the final verdict, depicted in the lower panels of Figure 5, is clearer. Respondents at associate or full professor level lean towards Albouy, as do respondents with at least one Top 5 publication and those with more than 3,000 citations. Yet, only the difference in means between respondents with and without a Top 5 publication is supported by a *t*-test ($t(262.61) = -2.87$, $p = 0.004$; see Appendix 4.2 for full results).

We also examine further sources of heterogeneity in respondents' evaluations. In our PAP, we categorized respondents into AJR-friendly and Albouy-friendly groups based on which paper the respondent cited or authored (ERQ1 PAP). We do not find significant differences in respondents' final verdicts between the two groups (see *t*-test results in Appendix 3.7). We further examine whether differences in respondents' familiarity with the papers, according to our first survey questions, affect their evaluations (ERQ3 PAP). In a regression of the final verdict on familiarity with AJR2001, Albouy2012, and AJR2012, we find that respondents more familiar with Albouy2012 are more pro-Albouy ($\beta = 0.044$, $p = 0.001$), while those more familiar with AJR2012 are more pro-AJR ($\beta = -0.036$, $p = 0.009$; see Appendix 3.9 for details and additional analyses).

[2] We provide more detailed analyses in Appendix 4.2, underpinning that the non-prespecified parts of Figure 5 are not driven by how we aggregated the data.

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



*Notes: Coefficient estimates in upper panel derived from multivariate OLS regressions with responses to the the following questions as dependent variables: "How convincing do you find AJR2001's empirical result?", "How convincing do you find Albouy's conclusion?", and "How convincing do you find AJR2012's conclusion?". "(Associate) Professor" refers to respondents who are associate or full professors (as opposed to PhD students, post-doctoral researchers, assistant professors, and those outside academia). "Top 5 publication" indicates whether the respondent has at least one Top 5 journal in their academic field (e.g., economics, political science, history). ">3,000 GS citations" refers to respondent's total citations on Google Scholar. Bars in lower panel represent the mean final verdict across expert characteristics. Error bars indicate 95% confidence intervals. *Analyses shown in the lower panel were not pre-specified.*

3. Discussion and Conclusion

When AJR received the Nobel prize in October 2024, Jan Teorell, member of the committee, stated in his award ceremony speech:

> *"By creatively using historical data on how vastly different the societal institutions were that emanated from this colonial experiment, they could provide solid evidence for a causal effect of these institutions on long-run prosperity."* (Teorell 2024)

The experts recruited for our survey do not seem to agree with this summary. We show that while there is a quasi-consensus on the broad theoretical claim that AJR are often associated with, there is much less enthusiasm about the empirical approach, and perhaps more surprisingly, also no agreement about the specific theoretical claim in AJR. Moreover, we conclude from the results presented in the previous section that no consensus on the replication debate between AJR and Albouy exists. Different perspectives on this are possible. The optimistic perspective is that the absence of a 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 und Strunk 2020, Hulme 2022, Stirling 2010). Pessimistically, though, our findings raise fundamental concerns about replicability in economics. If no consensus is emerging in this prominent debate and after several years, how then should any replication debate iterate towards a consensus? Accepting that such an unresolved controversy is the usual outcome of a non-confirmatory replication logically implies that replicability is unattainable in the first place. This would be at odds with the self-understanding of economics as an authoritative science that contributes hard evidence with quasi-factual status to the public discourse.

We acknowledge that the debate between AJR and Albouy – due to AJR2001's prominence – is not representative of any other replication debate in economics or the social sciences. We nevertheless contend that our observation is qualitatively transferable to most other replication debates about the robustness of previously published findings. Transferability may be less clear for direct replications, which repeat the same analysis on new data. Here, the study design quality, for example statistical power, can be agreed upon more objectively than for analytical choices underlying replication debates about robustness. The underlying reason for the lacking consensus among impartial experts for AJR vs. Albouy2012, we argue, is the vehement disagreement between the involved scholars. Similar vehement disagreements can be observed in virtually any non-confirmatory robustness replication (see Ankel-Peters et al. 2025, Humphreys 2015, Ozier 2021, and Roodman 2025). In fact, the AJR-Albouy debate is a perfect example of Collins' experimenters' regress: "since experimentation is a matter of skil[l]ful practice, it can never be clear whether a second experiment has been done sufficiently well to count as a check on the results of a first." (Collins 1992, p. 2). We have shown that not

only replicator and original experimenter disagree about the *skillful practice* – experts do as well.

Our approach is subject to legitimate criticism. We believe we have compiled a heterogeneous and somewhat representative population of potential respondents, by addressing citers and authors of a pool of diverse articles. The response rate of 10.2% is also decent, but nevertheless it is possible that invited experts with a particularly critical view of AJR. were more likely to participate in our survey. We cannot observe the underlying self-selection mechanisms within the pool of contacted scholars and must accept this caveat. Another criticism is that alternative approaches to seeking consensus could deliver other outcomes – for example, the Delphi method, a structured, multi-stage survey process in which experts are consulted over several rounds, gradually working toward a consensus or forecast (Dalkey and Helmer 1963). But we are confident that the qualitative interpretation of no (clear) consensus would hold across other approaches.

More importantly, it is our ambition to inspire further research on using expert knowledge to assess ongoing debates in the social sciences. This direction of research should indeed not (only) replicate our specific approach, but it should test different ways of exploring consensus or dissent. This will be useful for the emerging literature on the robustness of empirical evidence in the social sciences, which will likely yield many new replication debates (Brodeur et al. 2024a, 2024b; Campbell et al. 2024). But also beyond replication, several important social sciences literatures reveal no clear picture on where the evidence leads, despite a large body of rigorous empirical research. Here, expert surveys might help to accelerate the synthesis of the emerging evidence and, perhaps, even the consensus-finding process.

4. Materials and Methods

4.1 Adherence to pre-analysis plan

Table 1 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 2.2 & Appendix 3.1	Section 2.2 & Appendix 3.1
Secondary 1	How familiar are experts with the original paper, comment, and reply?	Section 2.2 & Appendix 3.2	n.a.
Secondary 2	How convincing do experts find the original paper, comment and reply based on their prior knowledge of these papers?	Section 2.2 & Appendix 3.3	Section 2.2 & Appendix 3.3
Secondary 3	How do experts evaluate the original paper, comment, and reply?	Section 2.2 & Appendix 3.4	n.a.
Secondary 4	How do experts believe other experts evaluate the paper, comment and reply?	Section 2.2 & Appendix 3.5	Appendix 3.5
Secondary 5	Has the experts' priors on the paper, comment and reply changed after reading the summaries provided in the survey?	Section 2.2 & Appendix 3.6	Section 2.2 & Appendix 3.6
Exploratory 1	To what extent do experts likely to be AJR-friendly or Albouy- friendly agree with AJR/Albouy?	Section 2.2 & Appendix 3.7	Appendix 3.7
Exploratory 2	Do experts with different backgrounds have systematically different opinions?	Section 2.2 & Appendix 3.8	Section 2.2 & Appendix 3.8
Exploratory 3	To what extent do experts with different levels of familiarity with the papers respond differently?	Section 2.2 & Appendix 3.9	Section 2.2 & Appendix 3.9

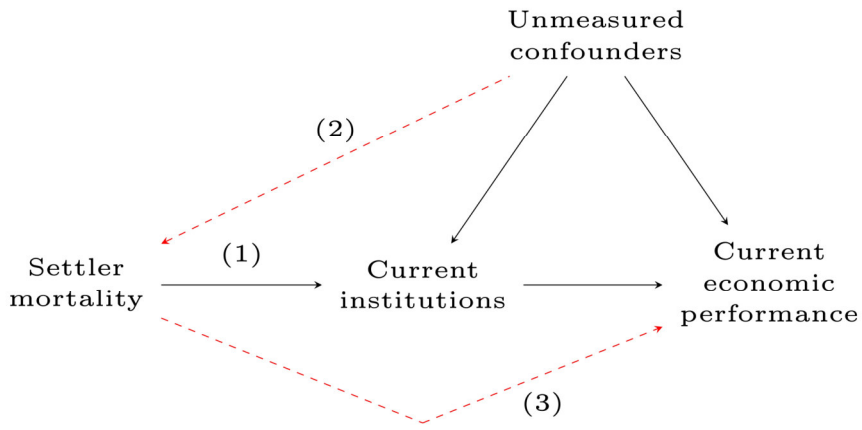
4.2. The Debate: AJR vs. Albouy

4.2.1 Original Paper by D. Acemoglu, S. Johnson and J. A. Robinson (2001)

AJR2001 investigate the causes of large differences in economic performance across countries. They broadly hypothesize that better institutions – such as well-defined property rights and less distortionary policies – foster greater investment in physical and human capital, leading to higher income levels. However, they argue that the effect of institutions on economic performance cannot be reliably estimated using simple cross-country OLS regressions as institutions are endogenous to income levels. Thus, the specific theory AJR2001 put forward is that historical settler mortality influenced European settlement patterns, which in turn shaped early institutions. These early institutions, so the theoretical argument goes, evolved into current ones that ultimately determine today's economic performance.

AJR2001 employ an instrumental variable estimation, using European settler mortality (henceforth settler mortality) rates as an instrument for institutional quality, measured as the average "risk of expropriation" index from 1985 to 1995. They find a large and precisely estimated effect of institutions on economic performance. The instrumental variable (IV) approach – one of the most widely used methods for identifying causal relationships in empirical economics – exploits variation that is assumed to be exogenous to the relationship under evaluation (Angrist and Krueger 2001). AJR2001 draw on variation in historical settler mortality, claiming it influences current institutional quality without having a direct effect on current economic performance. In other words, for the IV to be valid, it must only affect current economic performance through the instrumented variable, here, institutions.

***Figure* 6: Schematic depiction of AJR's empirical strategy**



Notes: Settler mortality (measured as deaths per annum per 1000 between 17th and 19th centuries) is the instrumental variable used to estimate the effect of current institutions (independent variable; measured as average protection against expropriation risk 1985-1995) on current economic performance (dependent variable; measured as GDP per capita PPP in 1995). The underlying assumptions are: (1) the instrument is correlated with the independent variable, (2) that there are no confounding variables that affect both the outcome and instrument, and (3) that there is no direct relationship between the instrument and the outcome (exclusion restriction).

Figure 6 illustrates the identification strategy, which – beyond the standard IV assumptions discussed in the figure notes – relies specifically on two conditions. First, the disease environment, particularly the prevalence of malaria and yellow fever, influenced European settlement patterns. AJR2001 argue that in countries with low mortality rates, Europeans were more likely to settle in small numbers and to establish extractive institutions designed to transfer resources to the colonizers. In contrast, where Europeans settled in greater numbers, they implemented European-style institutions that emphasized private property and

protection against government power. Second, AJR2001 assume that these early institutional structures set up by the colonizer have had a lasting impact on the quality of institutions observed today. While central to AJR's analysis, the specifics of their IV identification strategy are only tangentially relevant to Albouy2012 critique. It is worth noting that scepticism 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, Lal et al. 2024, Mellon 2025).

AJR2001 approximate settler mortality using historical records on mortality rates of European soldiers (the primary source for most regions including most of Africa and Asia), bishops (in Latin America, to fill gaps), and sailors stationed in various colonies between the 17th and 19th centuries. The primary source of these records is the work of Philip D. Curtin (1989, 1998; Curtin et al. 1995). AJR2001 state that they approximate settler mortality using data on soldiers, bishops, and sailors. However, they provide only few details on how these figures were constructed, and those are provided in a footnote, and no details on the imputation methods later critized by Albouy2012. Instead, the authors refer to the data appendix from an earlier version of the paper, Acemoglu et al. (2000), for further details.

4.2.2 Comment by David Albouy (2012)

Albouy2012 questions AJR2001 mainly based on two key concerns. First, for 36 out of 64 countries, AJR2001 lack settler mortality data from the country itself and instead impute values from other countries, arguing that disease environments are similar. Albouy2012 points out that the same dataset shows neighboring countries often have very different disease environments. For instance, he highlights that among the six countries assigned mortality rates based on Mali – including countries as distant as Angola and Uganda – their respective neighboring countries have–actually observed mortality rates ranging from 87.2 to 2,004. Presenting several additional examples of mortality rate assignments, he concludes that AJR's imputation method is "not just unreliable but often deeply flawed, generating rates that may be far too high or too low" (p. 3064).

Second, AJR2001's settler mortality rates are largely derived from data on soldiers and African laborers, rather than actual settlers. Albouy2012 argues that these mortality rates are not comparable to those of settlers, as African laborers and soldiers typically faced higher mortality rates than settlers. In the case of soldiers, the dataset also mixes mortality from

soldiers on campaign with those stationed in barracks during peacetime. Even without actual fighting, soldiers on campaign experience higher mortalities due to lack of shelter and hygiene conditions. Albouy2012 concludes that these proxies do not capture the disease environment of settlers at the time. Furthermore, he suggests that mortality rates were assigned endogenously, with higher rates disproportionately assigned to countries with weaker institutions and lower income per capita. AJR2001 do not mention this specific characteristic of the mortality rate data in their original paper.

Albouy2012 addresses these concerns by three key modifications. First, he removes the 36 imputed countries. Second, he introduces dummy variables for countries where mortality rates are based on campaign data or African laborer data. Third, he incorporates new settler mortality data from a later paper by Acemoglu, Johnson, and Robinson (2005).

Albouy2012's results show that the AJR2001 findings do not hold when these modifications are made. First, the first-stage estimates become insignificant and thus weak in most models. Second, to address this weak IV issue, he estimates Anderson-Rubin–confidence intervals which become unbounded for almost all specifications. Based on these results, 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)

4.2.3 Reply by D. Acemoglu, S. Johnson, and James A. Robinson (2012)

AJR2012 defend both the validity of their settler mortality data and the robustness of their findings. They argue that there is no basis for discarding 36 out of 64 countries from their sample, emphasizing 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 detailed country-by-country justification to substantiate their imputation. AJR2012 further demonstrate that their results remain robust to alternative mortality rates imputations (AJR 2000).

In addition, AJR2012 argue that Albouy2012's results for the smaller sample are primarily driven by a single extreme outlier in settler mortality rates, Gambia. They apply various outlier

management techniques, such as dropping Gambia or capping mortality rates, and show that their original results remain robust even within Albouy2012's restricted sample.

AJR2012 also disagree with Albouy2012's distinction between campaign and non-campaign mortality rates, arguing that the differences were not as large as Albouy2012 suggests. Military campaigns, so their argument goes, often did not involve actual combat, making the classification less meaningful. They underpin this by comparing mortality rates between episodes that can be reliably classified as campaigns and other periods where they find no systematic differences. Ultimately, AJR2012 assert that their original findings remain valid and robust, concluding that "*[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)

4.3 Expert recruitment, Survey Design, and Outcomes

We recruited experts primarily based on citations of the three debate papers. Using Scopus, we identified corresponding authors with valid email addresses available in the database who cited these works. In total, we identified 2,493 unique scholars who cited AJR2001, 58 scholars who cited Albouy2012, and 78 who cited AJR2012, based on citation records as of October 2023. Scopus includes only citations appearing in published journal articles and book chapters; citing working papers and grey literature are therefore excluded.

In addition, we identified 85 authors of 59 studies employing comparable identification strategies, for example Markevich and Zhuravskaya (2018) and Black et al. (2015). Furthermore, we identified 12 authors of critical papers, including those who authored other replications and critiques of AJR.[3] Because of this low number of directly identified authors of critical papers, we also included the corresponding authors of articles that cite at least one of those critical papers. Additionally, we included all 63 authors who published in the African Economic History Network (AEHN) working paper series between 2012 and 2021. These

[3] Among these critical papers, some specifically address AJR2001 and raise doubts about its findings and underlying assumptions (e.g., Assenova and Regele 2017, Olsson 2004), while others offer more general critiques of causal empirical designs using historical data (e.g., Deaton 2010, Conley and Kelly 2025 [working paper version used for recruitment]). See Appendix 5.1 for details.

efforts resulted in a total of 491 scholars linked to critical literature on AJR. See Appendix 5.1 for the comprehensive list of similar studies and critical papers considered for this recruitment strategy.

To avoid multiple contacts per individual, we deduplicated the pool of identified scholars, ensuring that each expert received only one invitation. This procedure resulted in a final pool of 3,022 potential participants. For 384 of these, the provided email addresses were later found to be invalid; in 276 cases, we retrieved updated contact details through manual web searches and contacted them again.

We supplemented this pool by distributing invitations via two academic mailing lists: (i) the Institute for Replication (I4R) and (ii) the Development Economics Committee of the German Economic Association. Personalized survey links were used in all invitations to allow differentiation between recruitment channels in subsequent analysis, while fully preserving respondent anonymity.

The survey was launched on March 28, 2024, for the group of identified citers and authors, with follow-up reminders sent two and four weeks after the initial invitation. The two mailing list invitations were sent on April 16, 2024. To incentivize participation, we implemented a lottery in which fifty randomly selected respondents were awarded a USD 20 Amazon voucher. Participation in the lottery was strictly optional, and any identifying information (i.e., email address and country of residence) was stored separately from the survey responses to maintain full anonymity. The survey closed on May 9, 2024. Response counts by recruitment strategy are reported in Appendix 1.1. The survey was conducted on *onlineumfragen.com* and analyzed using R version 4.5.0 (R Core Team 2024).

In total, 475 recipients answered at least the initial data protection and conent question, and 352 of them completed the entire survey. Two respondents declined consent and did not proceed further. Of the remaining 121 who drop out, the majority (91) did so after the consent question and before reaching the paper summaries. See Appendix 4.1 for more details. Our final sample only consists of respondents with complete responses.

Out of the 352 resondents, 309 are citers and authors of the above mentioned papers. Most are in the group of debate paper citers (250), followed by authors (or citers) of critical papers (38),

authors of similar papers (11), and 10 who both cited a debate paper and authored a paper categorized as similar or critical. Response rates per group are reported in Appendix Table A1. In our PAP, we categorized the different recruitment channels into AJR-friendly and Albouy-friendly, depending on whether the cited paper is rather supportive or critical of AJR or the method used by AJR. Despite yielding substantially more respondents (253 vs. 56), the AJR-friendly group recorded a virtual identical response rate as the Albouy-friendly group, indicating no meaningful selection bias by this classification. Additionally, 43 respondents were recruited via the two mailing lists.

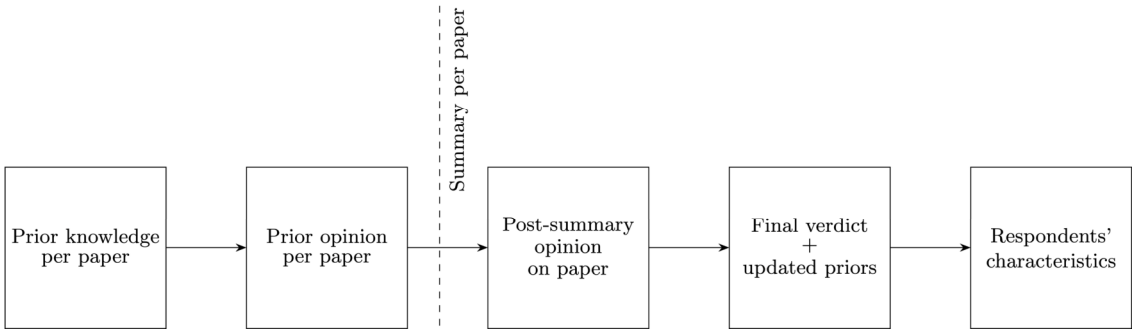
The questionnaire's structure is illustrated in Figure 7 and the full questionnaire is provided in Appendix 2. The survey first elicited the respondents state-of-knowledge about the debate papers, *before* we provided any information, on a scale from 0 to 100 (0 = 'I have never heard of it' and 100 = I have expert-level knowledge of it'). Respondents with some prior knowledge then evaluate how convincing they find AJR's (2001) empirical analysis, Albouy2012's comment and AJR's reply (0 = 'not convincing at all', 100 = 'very convincing').

Each paper is summarized by a short and neutral description of its key arguments, accompanied by a link to the full text.[4] After each summary, respondents assessed the key analytical steps (0 = 'not convincing at all', 100 = 'very convincing'). For AJR2001, the questions focus on the theoretical claims (broad and specific) and the empirical results. For Albouy2012, the questions focus on the decision to drop 36 countries with assigned mortality values, the inclusion of dummies for campaign and African laborer data, and the overall conclusion. For AJR2012, the questions address their counterarguments against Albouy2012's analytical choices and the overall conclusion of their reply.

In the final verdict section,respondents position themselves on a scale ranging from fully agreeing with AJR (-10) to fully agreeing with Albouy (10). Respondents with knowledge about the respective paper before the survey are also asked whether their priors have changed, using a 5-point Likert scale ranging from -2 ("yes, much less convinced")' to 2 ("yes, much more convinced").

[4] In order to ensure access to ungated versions, the AJR2001 link pointed to the AER website; the Albouy2012 link went to ResearchGate; and the AJR2012 link directed to MIT's DSpace.

Figure 7: 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.'* Unpublished.

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*.

Angrist, J. D. & Krueger, A. B. (2001). Instrumental variables and the search for identification: From supply and demand to natural experiments. *Journal of Economic Perspectives*, *15(4),* 69–85.

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.

Auspurg, K. & Brüderl, J. (2024). Toward a more credible assessment of the credibility of science by many-analyst studies. *Proceedings of the National Academy of Sciences*, *121(38),* p.e2404035121.

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.

Bolt, J. & Bezemer, D. (2009). Understanding long-run African growth: Colonial institutions or colonial education? *The Journal of Development Studies*, *45(1)*, 24–54.

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. (2024a). Promoting reproducibility and replicability in political science. *Research & Politics*, *11(1)*.

Brodeur, A., Mikola, D., Cook, N., Brailey, T., Briggs, R., de Gendre, A., Dupraz, Y., Fiala, L., Gabani, J., Gauriot, R., Haddad, J., Lima, G., Ankel-Peters, J., Dreber, A., Campbell, D., Kattan, L. Fages, D. M., Mierisch, F., Sun, P., …, & Zhong, Y. (2024b). *Mass reproducibility and replicability: A New Hope*. I4R Discussion Paper Series No. 107.

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

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

Conley, T. G. & Kelly, M. (2025). The standard errors of persistence. *Journal of International Economics, 153*, 104027.

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

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

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

Dalkey, N. C. & Helmer, O. (1963). An experimental application of the Delphi method to the use of experts. *Management Science*, *9(3)*, 458–467.

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.

Fraser, H., Bush, M., Wintle, B. C., Mody, F., Smith, E. T., Hanea, A. M., Gould, E., Hemming, V., Hamilton, D. G., Rumpff, L., Wilkinson, D. P. Pearson, R., Singleton Thorn, F., Ashton, R., Willcox, A., Gray, C. T., Head, A., Ross, M., Groenewegen, R., …, & Fidler, F. (2023). Predicting reliability through structured expert elicitation with the repliCATS (Collaborative Assessments for Trustworthy Science) process. *PLoS ONE*, *18(1)*, e0274429.

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.

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 February 14, 2025).

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/.

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.

Martínez i Coma, F. & van Ham, C. (2015). Can experts judge elections? *European Journal of Political Research*, 54, 305–325.

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

McArthur, J. W. & Sachs, J. D. (2001). *Insitutions 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.

Olsson, O. (2004). *Unbundling Ex-Colonies: A Comment on Acemoglu, Johnson, and Robinson, 2001*. Working Papers in Economics 146, University of Gothenburg, Department of Economics.

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

R Core Team. (2024). R: A language and environment for statistical computing (Version 4.5.0) [Computer software]. R Foundation for Statistical Computing. https://www.R-project.org/.

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

Rubin, A. & Rubin, E. (2021). Systematic bias in the progress of research. *Journal of Political Economy*, *129(9)*, 2666–2719.

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 June 12, 2025).

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.