

A comment on “Malaria suitability, urbanization and persistence: Evidence from China over more than 2000 years (European Economic Review)”

PLEASE ADD NAMES

November 13, 2025

Abstract

Summarize in a few sentences the original study, focusing on the main results in the original abstract in terms of the main claims which you attempt to reproduce or replicate. Provide information, if relevant, on the magnitude and statistical significance of the main results. Then report all your reproduction and replication results.

If there are too many robustness tests per claim to report individually, then report a summary measure such as the fraction of tests that replicate (i.e., statistically significant in the same direction as the original result) for each claim and the average relative size of the tests for each claim.

Definitions:

- **Computational Reproducibility:** Ability to duplicate the results of a prior study using the same data and procedures as were used by the original investigator.
- **Recreate Reproducibility:** Tests the extent to which results can be reproduced using only the information provided in the original study.
- **Robustness Reproducibility:** Tests whether results are robust to alternative plausible analytical decisions using the same data.
- **Direct Replicability:** Ability to duplicate the results using new data but the same procedures.
- **Conceptual Replicability:** Ability to duplicate results using new data and different procedures.

Example Abstract:

Speedy Analyst et al. (2022) examine the effect of a policy implemented in the fictional country Labas. In their preferred analytical specification, the authors find that the policy (PROSCOL) increased educational attainment of the treated group by 33 percentage points and decreased fertility by 9.8%. Their point estimates are statistically significant at the 5% and 10% levels, respectively.

First, we successfully computationally reproduce the main claims of the paper, but uncover two minor coding errors which have no effect on the study’s main results. Second, we directly replicate the results by adding more years to the sample. We find that adding more years decreases the size of the point estimate by one-third for education and by one-fourth for fertility. The fertility estimate becomes statistically insignificant at the 10% level, while the education estimate remains significant at the 5% level. Third, we test the sensitivity of the results to changing how standard errors are clustered. Clustering at the region level makes both point estimates statistically insignificant at the 10% level.

1 Introduction

Briefly describe the main data sources, method, policy or treatment, time period, and population for which the estimates apply. Then describe the main scientific claims (descriptive or causal) and robustness checks relevant for your replication. Quote the original part of the study that contains the main scientific claim(s), including page number(s).

Structure your summary as:

“The paper tested the effect of X on Y for population P, using method M. The main results show an effect of magnitude E (with units and standard errors).”

Explain how you obtained the data and codes and whether the original authors responded to your requests. Indicate the repository where your programs and data are located.

Describe any coding errors uncovered and their implications. For robustness reproduction, clearly describe your robustness checks. For replication, describe the new data. Summarize your results precisely, e.g.:

“Implementing this robustness increases the main point estimate for Y by X% and it remains statistically significant at the 5% level.”

Maintain a professional and constructive tone throughout the report.

Example:

Ms. Speedy Analyst, Mr. Speedy Ansensialyst, and John Doe (2022), henceforth SSD, investigate the impact of a program called PROSCOL in the country of Labas. In 2000, its government introduced an antipoverty program in northwest Labas (Ravallion, 2001). The program aims to provide cash transfers to poor families with school-age children. To be eligible to receive the transfer, households must have certain observable characteristics that suggest they are poor.

SSD test the effect of the policy (PROSCOL) on school enrollment and fertility for low-income families, using a difference-in-differences approach comparing the treated region (Labas) to untreated regions before and after the policy. The data come from the Labas Social Survey from 1998 to 2002. SSD’s main results are:

“We show that the policy (PROSCOL) increased school enrollment rates for the treated group by 30 percentage points and decreased the number of children born by 0.10 per family (mean of the dependent variable is 3.4). Our point estimates are statistically significant at the 5% level.”

In this report for the Institute for Replication (Brodeur et al., 2024), we assess the computational reproducibility of SSD’s results and test their robustness and replicability by:

1. Adding more years to the sample.
2. Changing how standard errors are clustered.

In the original study, SSD use data from 1998–2002 and cluster standard errors at the region/year level. We extend the sample to 1998–2004 and cluster at the region level. The authors provided the raw data upon request.

We successfully reproduce the main tables using the provided code, although small discrepancies arise from coding errors. Two minor issues were found: (a) the control variable **Age** was coded using the household head’s age instead of the mother’s, and (b) the gender dummy variable was treated as continuous in one regression.

Extending the sample decreases the point estimates by roughly one-third (education) and one-fourth (fertility). Fertility becomes insignificant at the 5% level, while education remains significant. Changing clustering renders both effects insignificant.

2 Computational Reproducibility

Describe the completeness of the replication package. Note whether there were:

- Cleaning code: none
- Analysis code: partial
- Raw or analysis data: partial (no raw data)

State whether the study is:

[label=()]Computationally reproducible from analysis data.

We used the replication package provided by XXX. Cleaning codes were not included, but only analysis data were provided.

We did not identify coding errors, but we added some code to reproduce some of the figures (Figure 1 and 2). We could produce code to replicate figures in the Appendix.

We could not replicate part of Figure 2, because the raw data were not available.

A minor point is that in some table (e.g. Table C2) the note mentions the inclusion of the control Yellow river, but this is dropped due to collinearity.

The raw data are not provided, which prevent the reproduction of Figure A1 and B3.

We could not use the code for the Conley’s SE, however when using alternative coding in R, we could produce reasonably similar standard errors.

Example:

Two minor coding errors were uncovered:

1. **Age** variable used the wrong definition.
2. The gender dummy was treated as continuous.

Correcting these errors had no significant effect on sign, magnitude, or significance.

3 Robustness Reproduction and Replication with New Data

Clearly state the types of reproduction/replication you conducted. For robustness reproduction, describe your robustness checks and how they impact the main estimates. For replication, describe the new data. Report all checks without selective reporting.

Example: We conduct a replication by extending the time period (1998–2004) and a robustness reproduction by changing how standard errors are clustered (region instead of region/year). We chose these because the original study was underpowered and may have within-region correlation across years.

3.1 Regression Model

We rely on the same difference-in-differences specification as the original study, comparing treated and untreated regions before and after policy implementation. The analysis is at the region/year level for education and at the family level for fertility.

3.2 Results: Educational Attainment

Extending the time period to 2004 reduces the effect size from 30 to 21 percentage points; it remains significant at the 5% level. Changing clustering increases the standard errors, making the estimates statistically insignificant.

3.3 Results: Fertility

Extending the time period reduces the main effect from -0.098 to -0.075 , which becomes statistically insignificant. Changing clustering further increases standard errors, also leading to insignificance.

3.4 Combined Checks

Applying both extensions simultaneously yields the same pattern: smaller effects and loss of statistical significance.

4 Conclusion

We successfully reproduce and replicate the key findings of SSD (2022) with only minor discrepancies. The main results are broadly robust but sensitive to specification changes. The education result remains stable, while the fertility result weakens. Future work could extend these analyses using alternative data or methods.

References

- Analyst, S., Analyst, S., and Doe, J. (2022). *The impact of PROSCOL on educational attainment and fertility*. The Journal, 1(1), 1–10.
- Brown, A. N. and Wood, B. D. K. (2018). Which tests not witch hunts: a diagnostic approach for conducting replication research. *Economics*, 12(1), 20180053.
- Brodeur, A. et al. (2024). *Mass Reproducibility and Replicability: A New Hope*. I4R Discussion Paper Series.
- Ravallion, M. (2001). *The mystery of the vanishing benefits: An introduction to impact evaluation*. The World Bank Economic Review, 15(1), 115–140.