Ankel-Peters, Jörg et al.

Working Paper

A Protocol for Structured Robustness Reproductions and Replicability Assessments

IZA Discussion Papers, No. 17691

Provided in Cooperation with:
IZA – Institute of Labor Economics

Suggested Citation: Ankel-Peters, Jörg et al. (2025) : A Protocol for Structured Robustness Reproductions and Replicability Assessments, IZA Discussion Papers, No. 17691, Institute of Labor Economics (IZA), Bonn

This Version is available at:
https://hdl.handle.net/10419/314588

I Z A Institute
of Labor Economics
Initiated by Deutsche Post Foundation

DISCUSSION PAPER SERIES

IZA DP No. 17691

A Protocol for Structured Robustness Reproductions and Replicability Assessments

Jörg Ankel-Peters
Abel Brodeur
Anna Dreber
Magnus Johannesson
Florian Neubauer
Julian Rose

FEBRUARY 2025

I Z A Institute
of Labor Economics

Initiated by Deutsche Post Foundation

DISCUSSION PAPER SERIES

IZA DP No. 17691

A Protocol for Structured Robustness Reproductions and Replicability Assessments

Jörg Ankel-Peters
RWI and University of Passau

Abel Brodeur
University of Ottawa and IZA

Anna Dreber
Stockholm School of Economics

Magnus Johannesson
Stockholm School of Economics

Florian Neubauer
RWI

Julian Rose
RWI, University of Passau and LMU

FEBRUARY 2025

ISSN: 2365-9793

IZA – Institute of Labor Economics

Schaumburg-Lippe-Straße 5–9
53113 Bonn, Germany

Phone: +49-228-3894-0
Email: publications@iza.org

www.iza.org

ABSTRACT

A Protocol for Structured Robustness Reproductions and Replicability Assessments*

Robustness reproductions and replicability discussions are on the rise in response to concerns about a potential credibility crisis in economics. This paper proposes a protocol to structure reproducibility and replicability assessments, with a focus on robustness. Starting with a computational reproduction upon data availability, the protocol encourages replicators to prespecify robustness tests, prior to implementing them. The protocol contains three different reporting tools to streamline the presentation of results. Beyond reproductions, our protocol assesses adherence to the pre-analysis plans in the replicated papers as well as external and construct validity. Our ambition is to put often controversial debates between replicators and replicated authors on a solid basis and contribute to an improved replication culture in economics.

JEL Classification: A11, C18

Keywords: replication, reproducibility, robustness, research transparency, meta-science

Corresponding author:
Jörg Ankel-Peters
RWI – Leibniz Institute for Economic Research
Hohenzollernstraße 1-3
45128 Essen
Germany
E-mail: peters@rwi-essen.de

* This work was supported by the German Research Foundation (Deutsche Forschungsgemeinschaft, DFG), Grant No. 3473/1-1 within the DFG Priority Program META-REP (SPP 2317). We also thank Open Philanthropy for financial support of this work. Conflict of interest: The corresponding author Jörg Ankel-Peters is co-editor of Q Open.

1. Introduction

The replication crisis has become a significant concern in various scientific fields, including economics. Several studies highlight the gravity of the issue. For economics, Askarov et al. (2023), Brodeur et al. (2020, 2023), Ferraro & Shukla (2020, 2022), and Ioannidis et al. (2017) have provided compelling evidence for the prevalence of biases in publication processes, underpowered study designs and, more generally, questionable research practices. Low rates of replicability in direct replications (Camerer et al. 2018) resonate with concerns about limited generalizability of empirical findings (Holzmeister et al. 2024; Masselus et al. 2024; Peters et al. 2018; Vivalt 2020). Replications as a potential remedy remain rare in economics (Ankel-Peters et al. 2023; Finger et al. 2023).

Since replications do not seem to occur naturally, systematic and coordinated replication and reproduction programs have emerged (Brodeur, Esterling, et al. 2024). Following several large-scale replicability and computational reproducibility projects conducted over the past decade (Camerer et al. 2016, 2018; Chang & Li 2015; Open Science Collaboration 2015), the boundary has recently been pushed to meta-robustness reproductions (Brodeur, Mikola, et al. 2024; Campbell et al. 2024). Such projects have a higher explanatory power and visibility than stand-alone replications, and reduce adverse effects on the careers of replicators (Brodeur, Esterling, et al. 2024).

Terminologically, we follow Dreber & Johannesson (2024) and distinguish between reproductions that redo the original analysis using the same data (computational reproduction) and those that conduct additional analyses (robustness reproduction) as well as replications that collect new data to implement the same study design (direct replication) and a similar study design (conceptual replication). This paper proposes a protocol to structure the post-publication assessment in empirical economics, focusing on computational and robustness reproductions.[1]

A significant challenge in the reproducibility discourse is determining the failure or success of robustness reproductions where results are often subject to intense debates between replicated authors and replicators (Ankel-Peters et al. 2024; Ozier 2021). While our protocol cannot resolve these debates – for example because perspectives on different analytical choices and

[1] An application example can be found in Rose et al. (2024).

robustness results are often subjective (Simonsohn et al. 2020) – it can help structure the exchange of arguments. Most notably, we draw on three tools to present the results of the robustness checks: Reproducibility indicators following Dreber & Johannesson (2024), a visual dashboard that assesses the alignment of original and robustness results in terms of statistical significance, sign of the effect, and effect size (Bensch 2024), and specification curves (Simonsohn et al. 2020).

The protocol encourages replicators to pre-register all robustness checks before implementation, thereby countering concerns about so-called null hacking by selective reporting in reproductions (Bryan et al. 2019). Beyond reproduction, the protocol also considers external validity, construct validity, and pre-specification diligence, which are also critical for assessing generalizability and replicability and are often insufficiently discussed in the economics literature. All these dimensions require subjective judgements in which we recommend replicators to be conservative in their assessment and follow an '*in dubio pro reo*' principle.

If the protocol is applied in a meta-reproduction project, we advise including a review process with referees from within the project team to ensure high-quality of and consistency across reproductions. The remainder of the paper is written from the perspective of replicators who adopt the protocol, and it follows the sequential workflow of our protocol with six key elements: Computational reproduction, code inspection, robustness reproduction, testing the paper's supportive analyses, pre-specification, as well as construct and external validity. While we strongly recommend starting with the first three elements and conducting them in order, the latter three can be implemented in any order.

2. The Replicability Assessment Protocol

2.1. Data Availability and Computational Reproduction

Computational reproducibility refers to the ability to reproduce the results of a study using the original data and code, following the exact analytical procedures. Therefore, replicators begin by assessing whether the necessary material to reproduce the original paper is available and the ease of use. The precondition to conduct reproductions is the availability of data. The original code is also helpful and may in some very complex cases even be indispensable. Some

journals have data editors and a priori provide the data (which still is often hampered by restrictions of proprietary data, see Vilhuber (2023)), but for many journals obtaining the data implies contacting the authors (Askarov et al. 2022; Brodeur, Cook, & Neisser 2024; Christensen & Miguel 2018; Pérignon et al. 2019). Most leading journals have data policies that oblige the authors to share the replication package, but still replicators are dependent on the authors' willingness to cooperate. In their assessment of the replication package, replicators distinguish between the availability of raw data and analysis data.

Any issues with the replication package encountered during the computational reproduction, such as missing data or code, or the need to manually install specific packages, should be documented. When reporting the results, replicators categorize findings based on whether they are fully, partially, or non-reproducible, from both raw data and from analysis data. The distinction between raw and analysis data is crucial because, unlike analysis data, the availability of raw data allows for a comprehensive vetting of the analysis, including decisions that otherwise remain hidden. The decisive criterion for a study to be computationally reproducible is whether the results obtained via re-running the original code match those presented in the paper. If they do not, we consider the paper to be not computationally reproducible. Replicators should explicitly state whether deviations are considered minor or consequential. Similarly, we consider a paper not computationally reproducible if code is missing, making computational reproducibility not feasible using the provided package.

2.2. Code Inspection

When doing the computational reproduction, the replicators carefully inspect the code to identify *objective* (i.e. unambiguous) coding errors. Such errors may include mis-specifying a variable, accidentally coding a variable differently than claimed in the paper, handling missing data erroneously, or making mistakes in sensitive commands such as merging, appending, reshaping, or dropping data, and others. These objective errors contrast with subjective analytical decisions affecting the robustness of results that are informed by methodological or economic appraisals of the respective researcher – and that may be controversial (see Section 2.3). In case errors are found, the replicator corrects them, and the corrected specification is then used as the basis for the robustness reproduction.

2.3. Robustness Reproduction

A crucial step at the beginning of the robustness reproduction is to identify the main outcomes of the original paper and tracking the corresponding coefficients within the paper. For this, the logic of the paper's argument needs to be recapitulated, guided by the title and abstract, sometimes also the introduction and the conclusion. A useful question to consider is: Which results are essential for the paper's main argument to hold? For example, if a paper finds an unsurprising effect for the full sample analysis (e.g. a null finding) and the pitch of the paper entirely circles around a significant heterogeneity, this heterogeneity should also be the center of the robustness reproduction. Conversely, if a paper's claim is based on the full sample analysis, a heterogeneity analysis with no prominent role in the paper should not be included in the robustness reproduction.

The robustness reproduction then involves testing the stability of these main outcomes to different assumptions, model specifications, estimation techniques, samples, and alterations in the data cleaning and processing. The goal is to assess the sensitivity of the main outcomes to various analytical choices.

To avoid null hacking and selective reporting, the replicators may register robustness checks after inspecting the data and code but before implementing them. That is, robustness checks should be specified after coding errors and other flaws have been removed. Robustness checks can also address crucial coding decisions to process the raw data towards the analysis data that are not discussed in the article.

The protocol suggests several dimensions for robustness checks: Definition of the sample, outcome and treatment variables, inclusion and definition of control variables, outlier management, handling of missing observations, and model selection. Replicators are free to choose additional robustness checks (e.g., adding or using other data), but the protocol emphasizes that each robustness check and its purpose should be clearly described, including a justification in terms of their economic and methodological legitimacy vis-à-vis the specification choices in the replicated paper.

Replicators are encouraged to conduct all theoretically and statistically reasonable robustness checks. This decision about reasonable robustness checks is often pivotal and, in case of a

negative robustness reproduction, regularly leads to controversies between replicators and original authors (Ankel-Peters et al. 2024; Simonsohn et al. 2020; Steegen et al. 2016).

Based on these robustness specifications, replicators assess the reproducibility of a study by combining three approaches: First, reproducibility indicators proposed by Dreber & Johannesson (2024). The reproducibility indicators check the degree of reproducibility for single outcomes in a study, or a group of studies, and they are directly comparable across studies. They measure the share of robustness tests supporting the original finding, changes in the average estimated effect, and the variation across robustness tests. The second approach is a reproducibility dashboard that visualizes for each outcome variable whether original and robustness results align in terms of statistical significance, sign of the effect, and effect size. In addition, it provides information on relative effect sizes and variation indicators. The reproducibility indicators and the reproducibility dashboard can be created using the Stata command *repframe* (Bensch 2024). Third, specification curves show the results of all implemented robustness checks in a multiverse. While the indicators and the dashboard summarize the overall reproduction results, specification curves are more detailed and also reveal the consequences of individual robustness checks on effect size and significance (Simonsohn et al. 2020).

2.4. Testing the Paper's Supportive Analyses

The previous section of the protocol focused on the robustness of the main outcomes of the replicated paper and how they are estimated. Most papers offer additional statistical analyses to support the findings or address caveats. Such supportive analyses can be tested in this section of the protocol. Prevalent examples are identification strategy checks, weak first stage tests, statistical power analysis as well as attrition or multiple hypotheses testing corrections.

Our protocol recommends replicators to identify those statistical analyses in the paper that are important for the established argument – and test them. For example, they should critically reflect on the identification strategies by applying placebo tests or checking the parallel trends assumption for difference-in-differences studies, and other relevant checks.

2.5. Pre-Analysis Plan

Although pre-registration and pre-analysis plans (PAPs) have increased significantly in recent years, they are most prevalent in experimental studies and hardly existent for work based on secondary data (Ferguson et al. 2023; Ofosu & Posner 2021). Even among experimental studies, many are registered, but not pre-specified, that is, they lack a distinct pre-analysis plan (Brodeur, Cook, Hartley, et al. 2024). A PAP outlines the analytical steps that researchers plan to take before examining the data and, thereby, helps reducing the risk of data mining and p-hacking. By adhering to a pre-specified analysis plan, researchers can ensure that their findings are not driven by post-hoc adjustments or selective reporting. Other critical elements are how vigorously the PAP determines the analysis and whether the paper adheres to it (Ofosu & Posner 2021).

Our protocol prescribes to check whether the original study includes a PAP and if so, whether the authors' analyses follow it. This step is critical for understanding whether authors have used the PAP to tie their hands and reduce researcher degrees of freedom on the results.

2.6. External and Construct Validity

External validity refers to the extent to which the results of a study can be generalized to other settings, populations, or time periods (Peters et al. 2018). Construct validity is concerned with whether results of a study are generalizable to policy-relevant scaled versions of the treatment or theoretically relevant replicated versions of it (Esterling et al. 2023; Masselus et al. 2024). Our protocol checks whether sufficient information is provided about the treatment under evaluation for the reader to assess a paper's construct and external validity. Specific information about the operationalized treatment is also relevant if other researchers want to implement a sufficiently similar treatment design in a direct or conceptual replication. Both external and construct validity are crucial for assessing the broader applicability and relevance of a study's findings (Esterling et al. 2023). Our protocol includes a structured checklist-like assessment of external and construct validity, guiding replicators through a series of questions to evaluate these aspects. This step examines whether the assessed paper is not only internally consistent but whether it also provides the necessary information for the reader to evaluate to which broader contexts it is generalizable.

3. Conclusion

The reproducibility assessment protocol presented in this paper provides a structured and standardized approach to reproducing and evaluating the robustness of published economics studies and offers a comprehensive framework for assessing the reproducibility of research findings. This methodological focus can be complemented on a case-by-case basis, for example if replicators sense that issues so far underrepresented in our protocol are critical to a replicated paper. Our protocol can be extended into these directions, so that a modular approach can be applied: replicators pick the protocolled dimensions that are most suitable for a replicated paper.

References

Ankel-Peters, J., Fiala, N., & Neubauer, F. (2023). 'Do Economists Replicate?', *Journal of Economic Behavior & Organization*, 212: 219–32. DOI: 10.1016/j.jebo.2023.05.009

Ankel-Peters, J., Fiala, N., & Neubauer, F. (2024). 'Is economics self-correcting? Replications in the American Economic Review', *Economic Inquiry*, ecin.13222. DOI: 10.1111/ecin.13222

Askarov, Z., Doucouliagos, A., Doucouliagos, H., & Stanley, T. D. (2022). 'The Significance of Data-Sharing Policy', *Journal of the European Economic Association*, jvac053. DOI: 10.1093/jeea/jvac053

Askarov, Z., Doucouliagos, A., Doucouliagos, H., & Stanley, T. D. (2023). 'Selective and (mis)leading economics journals: Meta-research evidence', *Journal of Economic Surveys*, joes.12598. DOI: 10.1111/joes.12598

Bensch, G. (2024). 'Repframe. A Stata package to calculate, tabulate and visualize Reproducibility and Replicability Indicators based on multiverse analyses.'

Brodeur, A., Carrell, S., Figlio, D., & Lusher, L. (2023). 'Unpacking p-Hacking and Publication Bias', *American Economic Review*, 113/11. DOI: 10.1257/aer.20210795

Brodeur, A., Cook, N., & Heyes, A. (2020). 'Methods Matter: p-Hacking and Publication Bias in Causal Analysis in Economics', *American Economic Review*, 110/11: 3634–60. DOI: 10.1257/aer.20190687

Brodeur, A., Cook, N. M., Hartley, J. S., & Heyes, A. (2024). 'Do Preregistration and Preanalysis Plans Reduce p-Hacking and Publication Bias? Evidence from 15,992 Test Statistics and Suggestions for Improvement', *Journal of Political Economy Microeconomics*, 2/3: 527–61. DOI: 10.1086/730455

Brodeur, A., Cook, N., & Neisser, C. (2024). 'p-Hacking, Data type and Data-Sharing Policy', *The Economic Journal*, 134/659: 985–1018. DOI: 10.1093/ej/uead104

Brodeur, A., Esterling, K., Ankel-Peters, J., Bueno, N. S., Desposato, S., Dreber, A., Genovese, F., et al. (2024). 'Promoting Reproducibility and Replicability in Political Science', *Research & Politics*, 11/1: 20531680241233439. DOI: 10.1177/20531680241233439

Brodeur, A., Mikola, D., & Cook, N. (2024). 'Mass Reproducibility and Replicability: A New Hope', *I4R Discussion Paper Series, No. 107*. DOI: 10.2139/ssrn.4790780

Bryan, C. J., Yeager, D. S., & O'Brien, J. M. (2019). 'Replicator Degrees of Freedom Allow Publication of Misleading Failures to Replicate', *Proceedings of the National Academy of Sciences*, 116/51: 25535–45. DOI: 10.1073/pnas.1910951116

Camerer, C. F., Dreber, A., Forsell, E., Ho, T.-H., Huber, J., Johannesson, M., Kirchler, M., et al. (2016). 'Evaluating Replicability of Laboratory Experiments in Economics', *Science*, 351/6280: 1433–6. DOI: 10.1126/science.aaf0918

Camerer, C. F., Dreber, A., Holzmeister, F., Ho, T.-H., Huber, J., Johannesson, M., Kirchler, M., et al. (2018). 'Evaluating the replicability of social science experiments in Nature and Science between 2010 and 2015', *Nature Human Behaviour*, 2/9: 637–44. DOI: 10.1038/s41562-018-0399-z

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*.

Chang, A. C., & Li, P. (2015). 'Is Economics Research Replicable? Sixty Published Papers from Thirteen Journals Say "Usually Not"', *Finance and Economics Discussion Series*, 2015/83: 1–26. DOI: 10.17016/FEDS.2015.083

Christensen, G., & Miguel, E. (2018). 'Transparency, Reproducibility, and the Credibility of Economics Research', *Journal of Economic Literature*, 56/3: 920–80. DOI: 10.1257/jel.20171350

Dreber, A., & Johannesson, M. (2024). 'A framework for evaluating reproducibility and replicability in economics', *Economic Inquiry*, ecin.13244. DOI: 10.1111/ecin.13244

Esterling, K. M., Brady, D., & Schwitzgebel, E. (2023). 'The necessity of construct and external validity for generalized causal claims', *I4R Discussion Paper Series No. 18*, 18.

Ferguson, J., Littman, R., Christensen, G., Paluck, E. L., Swanson, N., Wang, Z., Miguel, E., et al. (2023). 'Survey of open science practices and attitudes in the social sciences', *Nature Communications*, 14/1: 5401. DOI: 10.1038/s41467-023-41111-1

Ferraro, P. J., & Shukla, P. (2020). 'Feature—Is a Replicability Crisis on the Horizon for Environmental and Resource Economics?', *Review of Environmental Economics and Policy*, 14/2: 339–51. DOI: 10.1093/reep/reaa011

——. (2022). 'Credibility Crisis in Agricultural Economics', *Applied Economic Perspectives and Policy*, 1–17. DOI: 10.1002/aepp.13323

Finger, R., Grebitus, C., & Henningsen, A. (2023). 'Replications in agricultural economics', *Applied Economic Perspectives and Policy*, 45/3: 1258–74. DOI: 10.1002/aepp.13386

Holzmeister, F., Johannesson, M., Böhm, R., Dreber, A., Huber, J., & Kirchler, M. (2024). 'Heterogeneity in effect size estimates', *Proceedings of the National Academy of Sciences*, 121/32: e2403490121. DOI: 10.1073/pnas.2403490121

Ioannidis, J. P. A., Stanley, T. D., & Doucouliagos, H. (2017). 'The Power of Bias in Economics Research', *The Economic Journal*, 127/605: F236–65. DOI: 10.1111/ecoj.12461

Masselus, L., Petrik, C., & Ankel-Peters, J. (2024). 'Lost in the design space? Construct validity in the microfinance literature',. DOI: https://doi.org/10.31219/osf.io/nwp8k

Ofosu, G. K., & Posner, D. N. (2021). 'Pre-Analysis Plans: An Early Stocktaking', *Perspectives on Politics*, 1–17. DOI: 10.1017/S1537592721000931

Open Science Collaboration. (2015). 'Estimating the reproducibility of psychological science', *Science*, 349/6251: aac4716. DOI: 10.1126/science.aac4716

Ozier, O. (2021). 'Replication Redux: The Reproducibility Crisis and the Case of Deworming', *The World Bank Research Observer*, 36/1: 101–30. DOI: 10.1093/wbro/lkaa005

Pérignon, C., Gadouche, K., Hurlin, C., Silberman, R., & Debonnel, E. (2019). 'Certify reproducibility with confidential data', *Science*, 365/6449: 127–8. DOI: 10.1126/science.aaw2825

Peters, J., Langbein, J., & Roberts, G. (2018). 'Generalization in the tropics – Development policy, randomized controlled trials, and external validity', *The World Bank Research Observer*, 33/1: 34–64. DOI: 10.1093/wbro/lkx005

Rose, J., Neubauer, F., & Ankel-Peters, J. (2024). 'Long-term effects of the Targeting the Ultra Poor program—A reproducibility and replicability assessment of Banerjee et al. (2021)', *Q Open*, qoae031. DOI: 10.1093/qopen/qoae031

Simonsohn, U., Simmons, J. P., & Nelson, L. D. (2020). 'Specification curve analysis', *Nature Human Behaviour*, 4/11: 1208–14. DOI: 10.1038/s41562-020-0912-z

Steegen, S., Tuerlinckx, F., Gelman, A., & Vanpaemel, W. (2016). 'Increasing transparency through a multiverse analysis', *Perspectives on Psychological Science*, 11/5: 702–12. DOI: 10.1177/1745691616658637

Vilhuber, L. (2023). 'Reproducibility and Transparency versus Privacy and Confidentiality: Reflections from a Data Editor', *Journal of Econometrics*, 235/2: 2285–94. DOI: 10.1016/j.jeconom.2023.05.001

Vivalt, E. (2020). 'How Much Can We Generalize From Impact Evaluations?', *Journal of the European Economic Association*, 18/6: 3045–89. DOI: 10.1093/jeea/jvaa019

Appendix

A Protocol for Structured Robustness Reproductions and Replicability Assessments

Jörg Ankel-Peters, Abel Brodeur, Anna Dreber, Magnus Johannesson, Florian Neubauer, and Julian Rose

This protocol serves as a comprehensive guide to conduct structured reproducibility assessments of empirical papers in economics. The focus of the protocol is on computational and robustness reproducibility, complemented by a checklist on pre-specification as well as construct and external validity.

1. Data Availability and Computational Reproduction

The replicator should assess whether all necessary material is available to reproduce the results of the original paper. To find reproduction packages, check the journal's website, the authors' websites and as a last resort reach out to the original authors.

Use the provided data and scripts of the original paper and run the code without any corrections/changes. This is to check whether the results match the ones in the paper. Besides the change of working directory, only install missing packages and record them but do not do more to try and run the code. For all other errors due to missing data or incorrect code, document these problems if you were able to identify them, but do not yet attempt to solve them. Note: In case you are not able to computationally reproduce the results of the original paper, but you are able to locate and solve the errors, then you should do so in Section 2 "Code Inspection". Similarly, any coding errors that do not prevent computational reproduction, but you think need to be corrected should be part of the code inspection (see Section 2).

If essential code or data is missing, or if you encounter challenges in computationally reproducing the paper for other reasons, consider reaching out to the original authors for clarification.

Use the exemplifying Table 1 to present the results of the computational reproduction:

Table 1: Results of the computational reproduction

	Fully	Partially	No
Raw data provided		x	
Cleaning code provided	x		
Analysis data provided	x		
Analysis code provided	x		
Reproducible from raw data		x	
Reproducible from analysis data	x		

Reproducible from Analysis Data:

- Mark as "Fully" if the analysis data and analysis code are provided, and the reproduced results are identical to the paper's results.
- Mark as "Partially" if the analysis data and analysis code are provided, but there is slight deviation in the reproduced results from the paper.

Reproducible from Raw Data:

- Mark as "Fully" if both raw data and cleaning code are provided, and utilizing these inputs leads to identical results as in the paper.
- Mark as "Partially" if raw data and cleaning code are provided, but utilizing these inputs leads to slight deviations in the results.
- Mark as "No" if no raw data is provided, if raw data is provided without cleaning code, or the code fails to run.

2. Code Inspection

In this section, carefully check the code to identify any objective coding errors. Objective errors refer to errors such as mis-specifying a variable or including missing observations in the generation of variables. This contrasts with analytical decisions, which can be challenged in Section 4. Correct any coding errors and check whether the key results are affected. Document the changes. Considering the volume of code involved, the process can be efficiently guided by focusing on critical parts, such as:

1. **Main Outcome(s):** Carefully review all code sections that are directly related to generating the main outcomes of the original study. This includes any calculations, transformations, or statistical procedures used to produce the primary results. Additionally, check whether the description in the paper matches the code.
2. **Treatment Variable(s):** Carefully review all code that handles the treatment variable, including any data manipulations, assignments, or computations related to the treatment group or conditions. Additionally, check whether the description in the paper matches the code.
3. **Control Variables:** Carefully review all code relevant to the control variables used in the pertinent estimations. This involves scrutinizing how these variables are handled and integrated into the analysis. Additionally, check whether the description in the paper matches the code.
4. **Estimation Process:** Carefully review all code sections that implement the statistical models or methods in the estimation process itself. This encompasses code used to derive the main outcomes. Additionally, check whether the description of the estimation in the paper matches the code.

Throughout the process of checking these four items, be vigilant in searching for coding errors such as:

- variable misspecification
- coding switcheroo (e.g., variable accidentally differently coded as claimed)
- outlier management
- missing data handling
- sensitive commands such as merging, appending, reshaping, and dropping data.

In case errors are detected, report the corrected results for all main outcomes. If the original analysis path (combination of analytical decisions) is considered a reasonable analysis path, this analysis path with the corrected results will also be included as one analysis path in the multiverse robustness tests as well as in the dashboard and specification curve (see Section 5).

3. Scoping

Replicators should carefully read the original paper. The focus should be on identifying the main research question, one or several main outcomes, their presentation (Tables, Figures), and the methodology. The main outcomes are those that are mentioned most prominently in the abstract (note: both statistically significant and null results qualify)[2] and will be subjected to robustness tests to assess the robustness reproducibility of these results. For selection of the main outcomes, the replicator may dissect the abstract into claims and subsequently track the corresponding coefficient(s) in the paper. In case several coefficients correspond to one claim, the replicator may choose the one most prominently discussed in the paper (e.g., effect size discussed in the introduction). If no distinction can be made, the replicator can pick one of the qualifying coefficients randomly in a reproducible way.

To provide an overview of the main results, Table 2 can be used. It should contain essential information about the main outcomes, such as where the outcomes can be found in the original paper, the coefficients with accompanying standard errors, t/z-values, and p-values, information on outcome pre-specification, as well as information on the estimation strategy like the level of analysis, number of observations, estimations method, use of fixed effects, control variables, etc. Replicators can also compile a brief summary of the original paper and include it in the introduction of the reproduction.

Table 2: Example for recording details on main outcomes

	(1) Consumption	(2) Food Security
Name of display item	Original Paper, Table 1	Original Paper, Table 1
Column	Column 2 - Panel D	Column 3 - Panel D
Estimate	0.579	0.431
Standard Error	0.175	0.062
t/z-value	3.309	6.952
p-value	0.001	0.04
Level of analysis	Household	Household
Type of variable	Standardized index	Standardized index
Units	Deviation from baseline value	Deviation from baseline value
Number of observations	880	885
Sample	Full sample	Full sample
Estimation method	OLS	OLS
Fixed Effects	Hamlet fixed effect	Hamlet fixed effect
Standard Error type	Clustered at household level	Clustered at household level

[2] Note that for a null result the calculation of the reproducibility indicators is slightly different, see Section 5 for details.

Control variables	Baseline value of outcome	Baseline value of outcome
Outcome pre-specified	No	No
Outcome construction pre-specified	No	No
Minimum detectable effect size (MDE)	0.49	0.17

Source: https://bitss.github.io/ACRE/scoping.html#read-sum

4. Robustness reproduction

In the following, the protocol provides a concise guide outlining **possible** robustness checks along several dimensions (see Subsection 4.3). The specific implementation of these checks may vary depending on the paper. Replicators should carefully assess each dimension and incorporate all relevant ones in their robustness checks. Robustness checks are not limited to those listed here. The robustness tests should be carried out in a multiverse fashion (see Simonsohn et al. 2020 and Steegen et al. 2016), so that all reasonable combinations of the analytical decisions that replicators decide to vary in the robustness tests are estimated. Each such reasonable combination of analytical decisions in the multiverse is henceforth referred to as an "analysis path". Before proceeding with the implementation of these checks, the protocol recommends to first compile a comprehensive list of the specific checks the replicators intend to conduct for each of the selected main outcomes and formally register them, as described in step 4.1.

4.1. Registering robustness checks

At this stage, replicators have engaged intensively with the original paper, data, and code. They should compile a list of all robustness checks that they came up with during the first steps of the reproduction protocol alongside a short explanation for each robustness check. Replicators should, in a pre-analysis plan, preregister all analytical decisions they want to vary (e.g., outcomes, covariates, outlier management, etc.) alongside all choices within each analytical decision (e.g., which specific covariates to include in which specification). The original choice within each analytical decision should also be included if that is defined as a reasonable choice (note that if the replicators think that the original choice within one analytical decision is not a reasonable choice and they only defined one reasonable choice for this decision, they should include this choice as one analytical decision with only one option in the multiverse robustness

checks). An option is to upload the pre-specification document to OSF folder before implementing any robustness checks.

When implementing the pre-registered robustness checks, replicators should include the full multiverse of all reasonable combinations of analytical decisions and choices they have made. Reasonable refers to all choices and combinations that are sensible and can be backed by logical argumentation (and that are non-redundant). Note that one of the reasonable combinations of analytical decisions (one analysis path) will be the original analysis if replicators interpret this as one of the reasonable analysis paths.[3]

Should replicators realize during the implementation of the pre-registered robustness checks that they would like to apply more robustness tests than preregistered, they have the flexibility to introduce additional robustness checks if they find them to be pertinent during the reproduction process. However, it is crucial that any such additional checks are clearly labelled as "non-pre-registered" for transparency. Also, these additional robustness checks should be conducted in a multiverse fashion so that replicators add analytical decisions or added choices within each analytical decision and thereafter implement the multiverse analysis of all the reasonable combinations of analytical decisions (all the analysis paths).

4.2. Role of control variables

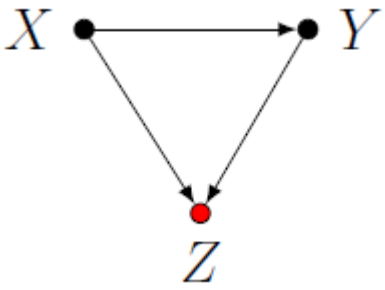
The protocol recommends discussing the inclusion of each control variable in the context of omitted variables and bad controls. Replicators compile a list of all control variables and assess for each variable whether they are a collider (opening a path to a collider) or a necessary control to avoid omitted variable bias. This exercise is repeated to justify the omission/inclusion of control variables in all robustness checks. Some stylized "Directed Acyclic Graphs" (DAGs) are shown below.

A. Bad controls

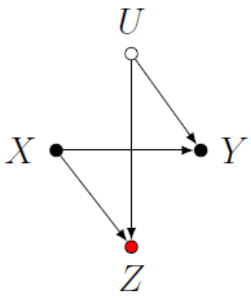
The following examples are from Cinelli et al. (2022):

[3] In case the result for this analysis path was not computationally reproduced in step 2, it should still be included as an anlysis path if it is interpreted as a reasonable analysis path but it would give different results from the original analysis. Similarly, if the replicators discover coding errors in step 4, the original anlysis path after correcting for these coding errors should still be included as an analysis path if it is interpreted as a reasonable analysis path.

Model 1:



Model 2:

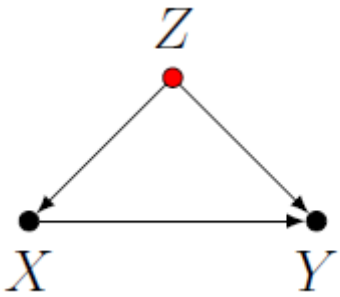


Model 1: Collider bias, controlling for Z opens the path X → Z ← Y, but also the colliding path due to the latent parents of Y. (Example: Assume one wants to estimate the effect of attending lectures (X) on good grades (Y). One might then want to control for a poll about attending lectures and good grades (Z), but those who attend and have good grades are more likely to fill out the poll. This therefore introduces a collider bias.

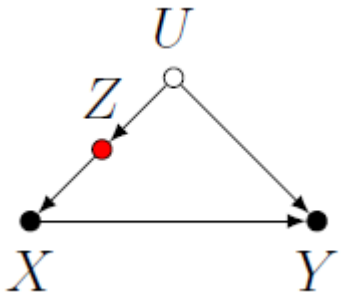
Model 2: Controlling for Z induces "selection bias" since it opens the colliding path X → Z ← U → Y. (Example: Assume that job satisfaction has no relationship with income, but education has. At the same time, there is an unobserved variable "work-life balance" that influences income and job satisfaction but has no relationship with education. Controlling in this scenario for income opens a pathway between education and job satisfaction.)

B. Good controls (leading to omitted variable bias if not controlled for)

Model 3:



Model 4:



Model 1: Z is a common cause of both X and Y (confounder) → controlling closes the back-door path from X to Y.

Model 2: Z is not a common cause of both X and Y, but controlling for Z blocks the back-door path from X to Y due to unobserved confounder U.

4.3. Dimensions and Robustness Checks

This section provides a list of dimensions that are relevant to the data processing and analytic decision of researchers (think of each dimension as one analytic decision that can be potentially varied in the robustness tests; but it is also fine to define and include more than one analytic decision within a dimension that is varied in the robustness tests). Researchers take hundreds of decisions when preparing data for analysis. While many of these decisions are taken subconsciously, they can have a large impact on the results. In this part of the reproduction, the protocol aims to show to what extent the results of the original paper depend on these decisions. Along several dimensions, replicators assess the reasonable analytical decisions and choices in each decision for each of the main outcomes. In doing so, they consider the following dimensions (and any other dimensions and analytical decisions that might be relevant for the specific paper):

- Definition of the analysis sample
- Definition of the outcome variable(s)
- Definition of treatment variable(s)
- Inclusion and definition of control variables
- Outlier management
- Missing observations
- Model selection

Table 3 provides an example to present the robustness checks by showing for each main outcome the change to the original data manipulation alongside an explanation. It includes each analytical decision that is varied or deviates from the original analysis as one row in the table. Note that if the original analysis is considered as one of the reasonable choices for an analytical decision that is varied in the robustness checks, then this original analysis choice should be included among the alternative options in Table 3. Note further that if replicators do not think that the original choice is a reasonable choice and only define one reasonable choice for this analytical decision, they should include this analytical decision as one row in the Table with only one choice in the column “Choice(s) in robustness checks”. When

presenting their analyses, replicators should sufficiently describe and motivate their implemented robustness checks.

Table 3: Example of robustness checks

Nr.	Main outcome in Table 2	Analytical decision	Rationale & Challenge	Choice(s) in Robustness Checks (include also the original analysis choice if that is considered one of the reasonable analytical choices)
1.	(1)	Outcome Variable	The authors use indices for all outcomes. This reduces the problem of MHT and allows to include several dimensions into one measure. Yet, indices can be created in many different ways.	1. Original choice 2. Alter definition of outcome index by including a more comprehensive set of input variables
2.	(1)	Control Variables	Timing of EL4. The endline took place over 5 months. In particular for income, food security, and consumption, the timing of the survey could matter.	1. Original choice 2. Add controls for the month in which the interview took place
3.	(1)	Control Variables	The authors do not use any baseline controls. At baseline, slight imbalance for income.	1. Original choice 2. Add baseline controls for all outcome variables.

Note: Column "Main outcome in Table 2" should include a reference to the main claims (main outcomes) from Table 2 in Section 3 that the numbered item refers to. Column "Analytical decision" captures the part of the analysis the robustness check refers to. Column "Rational & Challenge" captures the reason for the robustness check and which problems of the original paper it tries to address. Column "Choice(s) in Robustness Checks" includes the analytical choices included in the multiverse robustness checks (note that this can be one choice if the original analysis choice is not considered reasonable and only one alternative reasonable choice is included in the robustness checks).

As mentioned above, the robustness tests of each main outcome are estimated as a multiverse analysis for all the reasonable combinations of the analytical decisions (the combinations of the rows in Table 3).

5. Reporting of reproducibility indicator results

Replicators should report a table including all the results for each of the main outcomes for the reproducibility indicators described in this section alongside a short interpretation of the results. Different indicators need to be calculated for those original main outcomes reported as statistically significant results and those reported as null results.

The indicators are calculated based on the data extracted from the original study and all robustness checks. In extracting that data adhere to the following general principles:

A. The correct sign of the effect size and the t/z-value of each analysis path in the multiverse robustness tests need to be recorded, as this is important for the calculation of the indicators.
B. Some "main outcomes" may involve more than one regression coefficient, such as, for example, the coefficient of a variable and the coefficient of the squared variable for studies testing non-linear relationships. In such cases where the test of a hypothesis of a "main result" depends on the statistical significance of two coefficients both coefficients should be recorded.
C. For studies using instrumental variables only the 2nd stage results should be included to calculate the reproducibility indicators (i.e., the p-value, effect size and t/z-value of the 2nd stage regression coefficient is used for estimating the reproducibility indicators).
D. In case the original paper already includes one or more (plausible) robustness tests, replicators should include these analysis paths (robustness tests) in the calculation of the reproducibility indicators in the same way as the analyses paths were added.

To calculate the indicators, the Stata program "repframe" can be used, which is explained in more detail and can be downloaded here: https://github.com/guntherbensch/repframe (Bensch 2024). See also Dreber & Johannesson (2024) on the calculation of these reproducibility indicators.

Important: <u>If the command "repframe" is used, note that the following procedures (except the specification curve) are all performed by the command so that no further data manipulations are required.</u> <u>Only note that</u> in the two cases where (i) different effect size units are used (e.g. log and non-log) or (ii) the original analysis is included as one analysis path in the multiverse robustness, the help file of the command with instructions on the options "sameunits()" and "orig_in_multiverse()" needs to be consulted.

Reproducibility Indicators <u>for original results reported as statistically significant:</u>

Important: Replicators should estimate the following five indicators only if an original main result is reported as statistically significant in the original paper. An original main result reported as statistically significant in the original paper is defined as: a two-sided p-value<0.05 or an original result with a two-sided p-value>0.05 reported as significant by, for instance, using a star for p<0.10 in reporting results or referring to this result as significant at the 10% level, marginally significant or some similar term. If a main result is not defined as an "original main result reported as statistically significant" according to the above definition, it is defined as a "null result"; for main outcomes reported as null results, the three indicators further down are calculated.

1. **Statistical significance indicator**: Share of robustness tests analysis paths in the same direction as the original effect and a two-sided p-value <0.05.

   Note:

   - The original analysis path **should not be included** among the analysis paths in the estimation of the statistical significance indicator.
   - For "main outcomes" involving two coefficients, both these coefficients need to have an effect in the same direction as the original result and have a two-sided p-value<0.05 to count as a statistically significant robustness test analysis path for the statistical significance indicator.

2. **Relative effect size indicator**: Mean effect size of all the robustness tests analysis paths divided by the original effect size.

   Notes:

   - The original analysis path **should not be included** among the analysis paths in the estimation of the mean effect size of all the robustness tests analysis paths.
   - Only the robustness tests analysis paths that use the same effect size units as the original result should be included. This excludes, for example, robustness tests analysis paths with effect sizes in logs, while the original results are in levels.
   - The "relative effect size indicator" should not be estimated for "main outcomes" involving two coefficients.

3. **Relative t/z-value indicator**: Mean t/z-value of all the robustness tests analysis paths divided by the t/z-value of the original effect.[4]

   Notes:

   - The original analysis path **should not be included** among the analysis paths in the estimation of the mean t/z-value of all the robustness tests analysis paths.
   - For "main outcomes" involving two coefficients, the "relative t/z-value indicator" should be based on the average t/z-value of these two original coefficients and the t/z-value of a robustness test analysis path should be based on the average t/z-value of these two coefficients in the robustness test analysis path.[5]

4. **Variation indicator: effect sizes**: The standard deviation in effect sizes of all the robustness tests analysis paths divided by the standard error of the original effect size.

   Notes:

   - This indicator should not be estimated for "main outcomes" involving two coefficients.
   - If the original analysis is included as one analysis path in the multiverse robustness test that analysis path **should be included** in the estimation of the standard deviation.
   - Only the robustness tests analysis paths that use the same effect size units as the original result **should be included** in the estimation of the standard deviation. This excludes, for example, robustness tests analysis paths with effect sizes in logs, while the original results are in levels.

5. **Variation indicator: t/z-values**: Standard deviation of t/z-value of all the robustness tests analysis paths.[6]

   Notes:

[4] If replicators encounter a different test statistic, please convert the p-value of the test to the equivalent z-value. This can apply both to the original t/z-value if the original test is based on some other test statistic, and/or the robustness checks if some or all robustness checks use some other test statistic.

[5] In case the command "repframe" is not used: if the main result involves two coefficients and one has a positive and one a negative coefficient, the two original t/z-values must both be assigned positive signs in estimating the average t/z-value of the original result; and for the average t/z-value of a robustness test analysis path the t/z-value of a coefficient must be assigned a positive sign if the coefficient is in the same direction as the original coefficient and a negative sign if the coefficient is in the opposite direction of the original coefficient.

[6] If a different test statistic is encountered, the p-value of the test needs to be converted to the equivalent z-value. This can apply both to the original t/z-value if the original test is based on some other test statistic, and/or the robustness checks if some or all robustness checks use some other test statistic.

- If the original analysis is included as one analysis path in the multiverse robustness test, that analysis path **should be included** in the estimation of the standard deviation.
- For "main outcomes" involving two coefficients, the "variation indicator: t/z-values" should be based on the average t/z-value of these two coefficients in the robustness test analysis path.[7]

Reproducibility Indicators <u>for original results reported as null results</u>:

Important: The following three indicators are only estimated if an original main result is reported as a null result. For main outcomes reported as statistically significant in the original paper, the five indicators above apply.

1. **Statistical significance indicator**: Share of robustness tests analysis paths with a two-sided p-value >0.05

   Notes:

   - The original analysis path **should not be included** among the analysis paths in the estimation of the statistical significance indicator.
   - For "main outcomes" involving two coefficients, such a "main result" will count as a null result (P>0.05) in the statistical significance indicator if at least one of these two coefficients have a p-value>0.05.
2. **Variation indicator: effect sizes**: The standard deviation in effect sizes of all the robustness tests analysis paths divided by the standard error of the original effect size.

   Notes:

   - This indicator should not be estimated for "main outcomes" involving two coefficients.
   - If the original analysis is included as one analysis path in the multiverse robustness test, that analysis path **should be included** in the estimation of the standard deviation.

[7] In case replicators are not using the command "repframe" if the main result involves two coefficients and one has a positive and one a negative coefficient, the two original t/z-values must both be assigned positive signs in estimating the average t/z-value of the original result; and for the average t/z-value of a robustness test analysis path the t/z-value of a coefficient must be assigned a positive sign if the coefficient is in the same direction as the original coefficient and a negative sign if the coefficient is in the opposite direction of the original coefficient.

- Only the robustness tests analysis paths that use the same effect size units as the original result should be included in the estimation of the standard deviation. This excludes, for example, robustness tests analysis paths with effect sizes in logs, while the original results are in levels.

3. **Variation indicator: t/z-values**: Standard deviation of t/z-value of all the robustness tests analysis paths.[8]

   Notes:

   - If the original analysis is included as one analysis path in the multiverse robustness test, that analysis path **should be included** in the estimation of the standard deviation.
   - For "main outcomes" involving two coefficients, the "variation indicator: t/z-values" should be based on the average t/z-value of these two coefficients in the robustness test analysis path.[9]

Aggregated reproducibility indicators on the paper level:

For papers with more than one main outcome included in the robustness tests, the Stata command "repframe" will automatically calculate the robustness indicators on the paper level as the average of the indicators across the main outcomes (e.g., if there are three main outcomes reported as statistically significant in the original paper, the paper level of the "statistical significance indicator" is estimated as the average of the "statistical significance indicator" for these three main outcomes). This estimation is done separately for main outcomes reported as statistically significant in the original paper and main outcomes reported as null results in the original paper (i.e, if an original paper includes at least one main outcome reported as statistically significant and at least one main outcome reported as a null result, there will be two aggregated values of the "statistical significance indicator" for this paper).

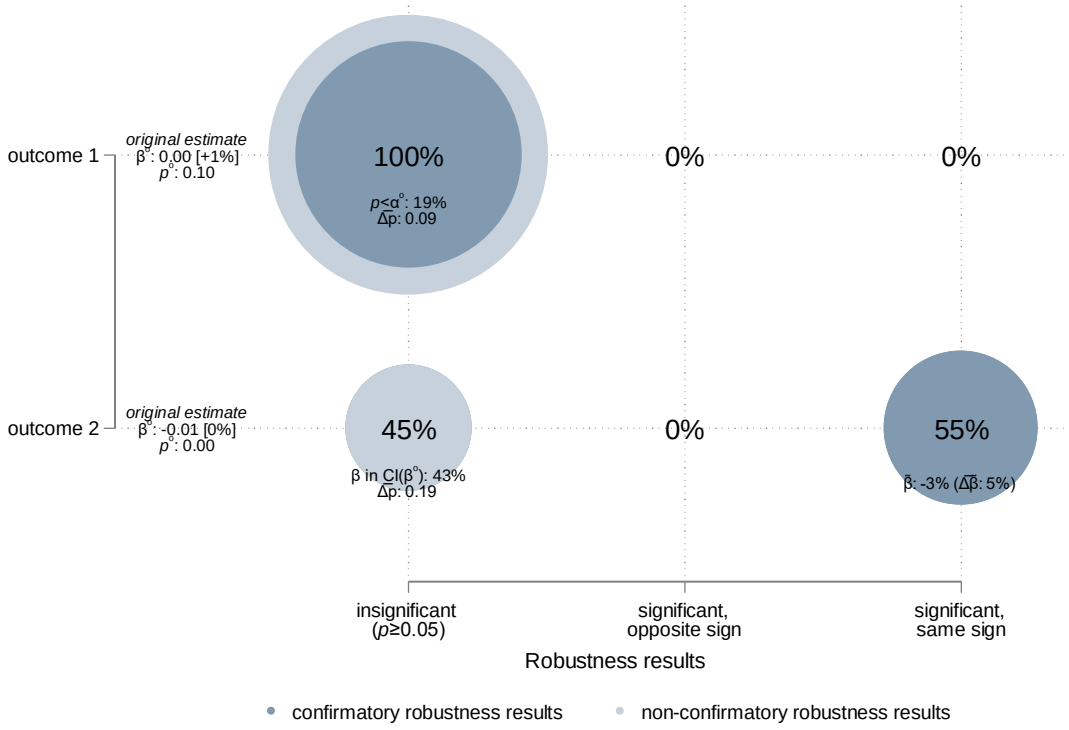
Reproducibility Dashboard

[8] If repliactors encounter a different test statistic, they need to convert the p-value of the test to the equivalent z-value. This can apply both to the original t/z-value if the original test is based on some other test statistic, and/or the robustness checks if some or all robustness checks use some other test statistic.

[9] In case replicators are not using the command "repframe": if the main result involves two coefficients and one has a positive and one a negative coefficient, the two original t/z-values must both be assigned positive signs in estimating the average t/z-value of the original result; and for the average t/z-value of a robustness test analysis path the t/z-value of a coefficient must be assigned a positive sign if the coefficient is in the same direction as the original coefficient and a negative sign if the coefficient is in the opposite direction of the original coefficient.

The Stata command "repframe" also visualizes a so-called *Reproduciblity Dashboard* (see Figure 1). The dashboard highlights confirmatory and non-confirmatory robustness check results and provides indicators tailored to subsets of the analysis paths for which the indicators are most informative. As can be taken from the figure, they are tailored to whether original and robustness results are statistically significant or not. The Reproducibility Dashboard thereby provides a complementary set of statistical significance indicators, relative effective size indicators and variation indicators. Details on the indicators can be found at https://github.com/guntherbensch/repframe.

Figure 1: Sample Reproducibility Dashboard



Notes: α = significance level; β = estimate; CI = confidence interval; Δ = mean absolute deviation; ° = original estimate; bars over certain values refer to mean values, tildes to median values. A version of the dashboard is applied in Ankel-Peters et al. (2023).

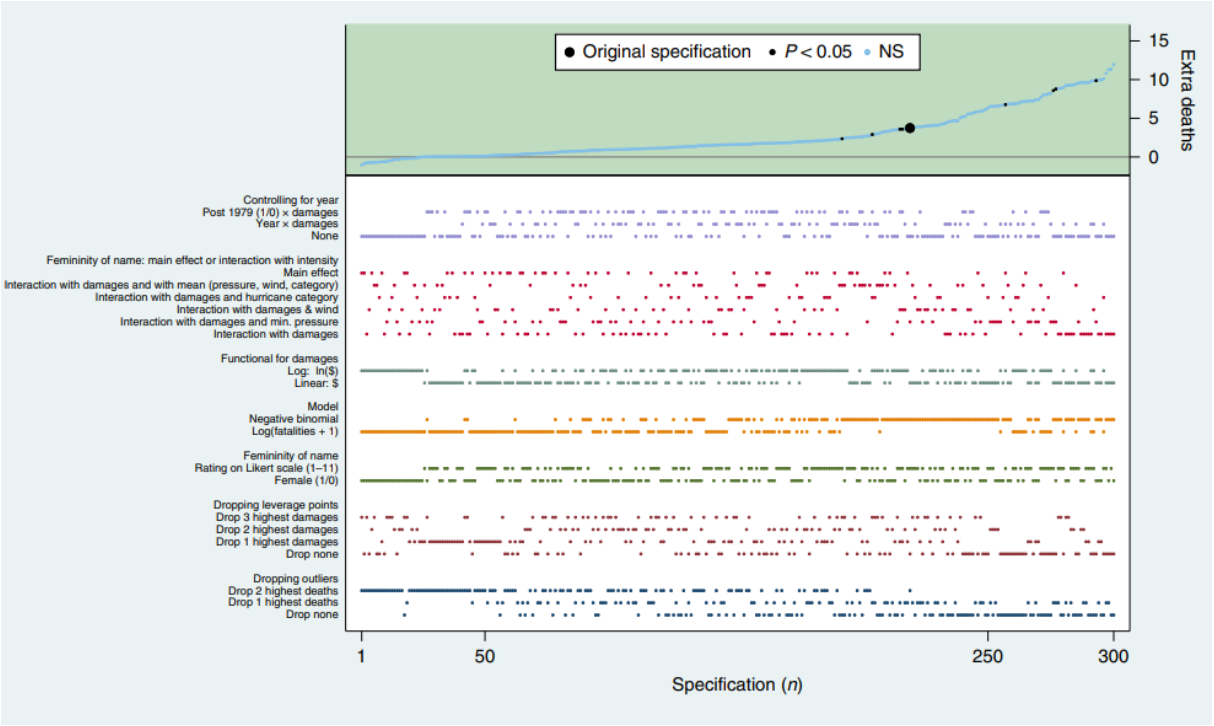
Specification Curve

To complement the previous two reporting tools, specification curve for each main outcome can be used. This curve is based on the multiverse robustness tests that were already implemented above (i.e. the same analytical decisions and analysis paths as above). In case the original paper already includes one or more (plausible) robustness tests, they should be included among the analysis paths in the multiverse robustness tests if they are considered

plausible analysis paths. In contrast to the previous reporting tools, the specification curves allow to assess the influence of individual robustness checks on the sensitivity of the results.

A specification curve as suggested by Simonsohn et al. (2020) and presented in Figure 2 should be generated. In case the replicators use Stata (version 16 or newer), the command “speccurve” produces a specification curve. In R, the packages "r2especcurve"[10], which is geared to the data structure necessary for the repframe command, and “specr” can be used. In case an analysis only includes few analysis paths, a specification curve may not be useful, but the results can be shown in a table instead.

Figure 2: Specification Curve



Source: Simonsohn, U., Simmons, J. P., & Nelson, L. D. (2020). Specification curve analysis. Nature Human Behaviour, 4(11), 1208-1214.

6. Testing the Paper’s Supportive Analyses

In this section, replicators are encouraged to assess additional statistical analyses provided by the original paper to support its findings. Prevalent examples are identification strategy checks, weak first stage tests, statistical power analysis as well as attrition or multiple hypotheses testing corrections. For example, replicators should critically reflect on the identification

[10] See Huntington-Klein (2024).

strategies by applying placebo tests or checking the parallel trends assumption for difference-in-differences studies, and other relevant checks.

Note that these checks are not included in the reproducibility indicator or dashboard.

7. Pre-Analysis Plan

Replicators should check whether the authors registered the study and published a pre-analysis plan (PAP). If the PAP is not publicly available, consider reaching out to the original authors. Please note, which of the following cases applies:

a. No preregistration
b. Preregistration without a PAP
c. Preregistration with a PAP

If the PAP is available, it should be compared to the original paper and deviations highlighted.

- Does the PAP specify the exact design of the study, the exact data collection and exactly how all analyses and tests will be conducted?
- Does the paper follow the pre-registered design and data collection?
- Does the paper carry out all analyses and tests reported in the paper exactly as specified in the PAP?
- Does the paper report all analyses and tests as specified in the PAP?
- Does the paper report additional analyses and tests not specified in the PAP?

8. External and construct validity

Replicators should assess the extent to which the paper provides information regarding external validity and construct validity. Following Esterling et al. (2023), Masselus et al. (2024), and Peters et al. (2018) this checklist can be used:

Construct validity:

- Does the paper provide comprehensive information about the details of the intervention, that would allow other researchers/implementers to implement an intervention that is sufficiently similar to the intervention under evaluation?
- Does the paper describe whether the intervention is an existing intervention implemented by a real-world policy agency (as opposed to an intervention designed by researchers for academic purposes)? If not, does the paper describe how the intervention deviates from what is implemented in the real world and whether these deviations are costly?

Researcher involvement and special care:

- Does the paper describe to what extent the authors (or other researchers) were involved in designing the intervention (as opposed to, for example, an NGO or a governmental agency designing the intervention)?
- Does the paper describe who implemented the intervention, and to what extent the researchers were involved in the implementation?

Hawthorne & John Henry effects:

- Does the paper describe whether participants were aware of the randomization/being part of a study?

General Equilibrium Effects:

- Does the paper discuss general equilibrium effects (GEE)?

Generalizability:

- Does the paper determine the scope of generalization regarding the intervention under study?
- Does the paper determine the scope of generalization regarding the population and sample?

9. Documentation of reproduction

To ensure reproducibility of your reproduction, provide all code, data, and a read-me file in a reproduction package. Upload the reproduction package to our server before publishing the reproduction report and/or a working paper.

10. For guidelines on ensuring reproducibility, see:

- https://blogs.worldbank.org/impactevaluations/how-make-sure-your-research-paper-reproducible-evidence-55-papers-guest-blog-post
- Reproducible Research - Dimewiki (worldbank.org)

11. References

Ankel-Peters, J., Bensch, G., & Vance, C. (2023). 'Spotlight on researcher decisions–Infrastructure evaluation, instrumental variables, and first-stage specification screening.', *Ruhr Economic Papers, No. 991.*

Bensch, G. (2024). 'Repframe. A Stata package to calculate, tabulate and visualize Reproducibility and Replicability Indicators based on multiverse analyses.'

Cinelli, C., Forney, A., & Pearl, J. (2022). 'A crash course in good and bad controls', *Sociological Methods & Research*, 004912412210995. DOI: 10.1177/00491241221099552

Dreber, A., & Johannesson, M. (2024). 'A framework for evaluating reproducibility and replicability in economics', *Economic Inquiry*, ecin.13244. DOI: 10.1111/ecin.13244

Esterling, K. M., Brady, D., & Schwitzgebel, E. (2023). 'The necessity of construct and external validity for generalized causal claims', *I4R Discussion Paper Series No. 18*, 18.

Huntington-Klein, N. (2024). 'r2especcurve: Makes Specification Curves out of R2E-Formatted Data'.

Masselus, L., Petrik, C., & Ankel-Peters, J. (2024). 'Lost in the design space? Construct validity in the microfinance literature',. DOI: https://doi.org/10.31219/osf.io/nwp8k

Peters, J., Langbein, J., & Roberts, G. (2018). 'Generalization in the tropics – Development policy, randomized controlled trials, and external validity', *The World Bank Research Observer*, 33/1: 34–64. DOI: 10.1093/wbro/lkx005

Simonsohn, U., Simmons, J. P., & Nelson, L. D. (2020). 'Specification curve analysis', *Nature Human Behaviour*, 4/11: 1208–14. DOI: 10.1038/s41562-020-0912-z

Steegen, S., Tuerlinckx, F., Gelman, A., & Vanpaemel, W. (2016). 'Increasing transparency through a multiverse analysis', *Perspectives on Psychological Science*, 11/5: 702–12. DOI: 10.1177/1745691616658637