Best Practices for Reproducibility, Research Assessment Reforms, and Implications for Experimental Economists

Zacharias Maniadis[1]

University of Cyprus and University of Southampton

Abstract: Scientists are under pressure to adhere to best practices for enhancing reproducibility, such as preregistration and data sharing. This tendency will certainly increase with the unfolding reforms in researcher assessment, and it brings new challenges. Heterogeneity in the amenability of different domains to reproducibility-enhancing practices raises an issue of possible inequity: will different scientific domains bear disparate adjustment costs? Is this justified and efficient? To illustrate the problem, we consider recent concerns expressed by experimental economists, namely that they are unfairly burdened relative to other economics domains. Our analysis indicates that such fairness concerns may have merit, only insofar as research assessment does not fully internalize the costs of adjusting to new practices.

1. The Credibility Crisis and Developments in Researcher Assessment

The last two decades have seen great interest in studies and methods for measuring and enhancing research reproducibility.[2] The 'credibility crisis' in biomedical, social and behavioral sciences (Munafo et al., 2017) has raised concerns about researchers' practices. Most of the discussion has taken place in the context of experimental science, especially using the benchmark of clinical trials (Ioannidis, 2005; 2022).[3] During the scientific debate over the alleged crisis, novel norms for improvement and best practices have been developed (Ioannidis, 2014; Simmons, Nelson, and Simonsohn, 2011; Fanelli, 2013). Such best practices include preregistration and registered reports, sharing of data, code, script, and experimental material, and adoption of transparency checklists (Altman et al., 2008; Moher at al., 2001), among others.

The adoption of such best practices is receiving increasing institutional support in science, incorporated within movements for reforming research evaluation (Adler, Ewing, and Taylor, 2009). DORA (The American Society for Cell Biology, 2012) provided several recommendations for revisiting research assessment practices at the research institution, funder, and publisher level. The 'Hong Kong Principles for Assessing Researchers' (Moher et al., 2020) came to complement DORA by focusing on practices pertaining to open science, integrity and reproducibility. Developed by the World Conference on Research Integrity, the Hong Kong

[1] This project has received funding from the European Union's Horizon 2020 Research and innovation Programme under Grant Agreement number: 857636 — SInnoPSis — H2020-WIDESPREAD-2018-2020/H2020-WIDESPREAD-2018-04. This project has also received funding from the European Union's Horizon Europe research and innovation programme under grant agreement No [101079196] (Twinning for Excellence in Management and Economics of Research and Innovation [TWIN4MERIT]). We would like to thank Fabio Tufano, John PA Ioannidis, Andreas Ortmann and two anonymous referees for valuable comments.

[2] We adopt the glossary formulated by the consortium iRISE: https://www.irise-project.eu/). *Replicability* of a study is the extent to which design, implementation, analysis, and reporting of a study enable a third party to repeat the study and assess its findings. The Replicability of a study is independent from whether the study was indeed replicated, or results of the study were reproduced. On the other hand, *Reproducibility* is the extent to which the results of a study agree with those of replication studies**.** Finally, *Computational Reproducibility* is the ability of reported results to be reproduced by repeating the same computational steps on the original data (LeBel et al. 2018).

[3] As we shall discuss later, the reason for this focus this is not that experimental science faces a more acute credibility problem, but because it serves as the paradigmatic case, where replicability can be scientifically assessed.

principles focused on rewarding research integrity and good practices, such as transparency, replication and research synthesis. The European commission has been providing formal support not only to the Hong Kong principles (see scoping report of 2021), but also (and especially) to the Coalition for Advancing Research Assessment (CoARA).[4]

Clearly, as this discussion shows, there are strong pressures for researchers to adopt such best practices, which are likely to be codified further in explicit statements, policies, and assessment criteria of various institutions. For instance, the coalition CoARA has over 700 signatories, including research institutions, funders, societies and associations. In this opinion piece, we shall discuss a policy problem that stems from these developments. How can scientific best practices be imposed upon diverse research domains (e.g. within a university), and how can the principles of efficiency and fairness be protected in this process? To our knowledge, this problem has not been addressed before, but it has important implications for science, as reflected in relevant concerns publicly expressed by experimental economists. To shed light on the problem, we introduce a novel methodological taxonomy and conduct a careful review of the meta-research literature.

2. The Policy Problem

The imposition of 'best practices' can be analyzed as a policy problem, using the lenses of economic theory (Gall, Ioannidis and Maniadis, 2017). However, in practice, most of these policy suggestions are implemented without ex-ante rigorous modelling and evaluation.[5] This is mainly driven by a lack of policy analysis expertise on the part of proponents for change (belonging to diverse disciplines), as well as the inherent complexity of social phenomena (Gall, Ioannidis and Maniadis, 2017). Yet, it is important to take a step back and theorize about the general policy problem: as the CoARA principles (which incorporate reproducibility practices) are implemented at the university level for signatory institutions, which are the concrete implications for hard sciences, biomedicine, engineering, social sciences and the humanities? Is it possible to impose changes uniformly? And how is each subdomain likely to adjust to the new reality?

A New Taxonomy

We argue that to analyze this complex problem the focus should be on the underlying methodology, rather than scientific discipline. To advance our understanding, we employ a basic taxonomy of research methodologies, illustrated in Table 1. To explain the contents of the table,

[4] CoARA is a "*collective of organisations committed to reforming the methods and processes by which research, researchers, and research organisations are evaluated*." CoARA's main declaration argues in favor of valuing "*… diverse outputs (FAIR data sets, replication studies, registered reports, pre-prints*) […]*"* and of diversifying *"… indicators (Open science badges; Publons, ORCID, open peer review; CRediT …"*

CoARA – Coalition for Advancing Research Assessment.

[5] Retrospective evidence is accumulating, and so far, messages are mixed. Brodeur, Cook and Neisser (2024) find no evidence for improvement, stemming from data and material-sharing policies. Brodeur et al. (2024) show that preregistrations, as practiced by economists, lead to no noticeable reduction in p-hacking or publication bias. However, the authors find that, if accompanied by a Pre-Analysis Plan, preregistrations can lead to both reduced p-hacking and reduced publication bias. Moreover, Brodeur, Mikola, and Cook (2024) find that increasing adherence to best practices seems to make a difference for computational reproducibility.

note that experimental work typically involves several components: controlled generation of new data, data analysis, analytical coding, and the use of a strict experimental protocol. For each of these components, there are corresponding best practices targeting enhanced reproducibility. For instance, preregistration aims to discipline new data generation, pre-analysis plans to discipline data analysis, code-sharing to help reproduce analytical coding, etc. On the other hand, non-experimental empirical work includes the above components except the controlled generation of new data. For non-empirical methods, reproducibility is more difficult to define, and such best practices are a relatively new issue (this will be discussed further below). In general, as we move down the rows of Table 1, the 'components' tend to decrease, with an accompanying decrease in established best practices.

Table 1: A Taxonomy of Methods and Corresponding Best Practices

Methodology		**Corresponding Best Practices**
Empirical	1. Experimental Analysis	Preregistration Pre-Analysis Plans Protocol Sharing Data Sharing Code Sharing
	2. Observational Analysis	Pre-Analysis Plans Data Sharing Code Sharing
Non-Empirical	3. Theoretical and/or Computational Analysis	Code/Analysis Sharing
	4. Qualitative Analysis	Protocol Sharing

The highly simplifying scheme of Table 1 is useful, because it helps us visualize the policy problem: a research institution wishing to enhance reproducibility will have to adopt and monitor different sets of practices for different domains, because these domains use different methodologies. Because of the skewness of the relevant literature towards empirical (and especially experimental) methods, disciplines employing them have (comparatively speaking) readily available paths towards established best practices.[6] However, outside these methods (lower two rows in Table 1), reproducibility becomes more difficult in terms of how to define, let alone how to improve it. Now we shall illustrate these points using the example of economics.

The Case of Experimental Economics

A case study will now achieve a dual objective: it will illustrate the significance of the policy problem for science, as well as operationalize the taxonomy provided in Table 1. The latter task will be achieved by providing instances of the heterogeneity of the content of reproducibility — and corresponding best practices — across different domains. Experimental economics research is increasingly expected to adhere to the practices presented at row 1 of Table 1. As this

[6] See the numerous early proposals by Bakker, van Dijk, and Wicherts (2012), Nosek, Spies, and Motyl (2012), Ioannidis (2014), Simmons, Nelson, and Simonsohn (2011), and Fanelli (2013), all of which are grounded on the experimental methodology.

expectation has been increasingly enforced (formally and informally) by top journals,[7] it has raised concerns among the experimental economics community. Is this adjustment a burden that lies upon experimental sciences alone? What about non-experimental empirical research, and even economic theory? If, as the argument goes, the onus of adjustment (within economics) falls disproportionately upon experimental economics, this could potentially be seen as an unfair development. To illustrate the possible disparity in adjustment costs between different domains, it is important to review the current reality of adoption of reproducibility practices across different fields, both within economics, and beyond.

One of the key areas that drive the concerns of experimental economists is comparisons with non-experimental empirical economics. How is applied microeconomics expected to adhere to similar best practices? First, in this domain, material and data-sharing policies by journals are like in experimental economics.[8] However, preregistration seems difficult to apply to non-experimental work,[9] as also discussed in other empirical disciplines. Ioannidis (2022) argues that *"Pre-registration probably adds less trust to the research work, when existing information is accessible in ways that one cannot guarantee that multiple iterations and explorations have not already happened before one decides to register a protocol—which may reflect the seemingly best-performing models."*

When it comes to comparisons with theory domains (row 3 of Table 1), data-sharing is not an issue, but important issues remain: mathematical proofs should be thoroughly available for reproducibly purposes. For complex computational work, such as for large macroeconomics and environmental models, the issue of improved code-sharing is also crucial. There is documented difficulty to reproduce results in several domains of macroeconomics, and a general need to improve reproducibility-related practices in these domains.[10] Interestingly, Ioannidis (2022) argues that preregistration is possible even for formal mathematical models, if they pertain to forecasting, and, in general, if their predictions can be confronted with data. However, the reproducibility discussion in these domains is still at an early stage.

Outside economics, experimental psychology has played a major role in the debate about reproducibility (see Pashler and Wagenmakers, 2012) and has developed best practices almost identical to experimental economics, with increasing acceptance: "... *growing acceptance of stricter Open Science practices among psychology researchers ... psychology journal editors, by adopting moderately stringent or even stricter policies, could align with this cultural shift without risking a decline in submissions. On the contrary, journals ignoring Open Science may face challenges in attracting papers in an evolving academic landscape*." (Rudenko, 2024). It seems

[7] Many economics and social science journals have imposed requirements for the sharing of data, code, script, and experimental material. Brodeur, Cook and Neisser (2024) explain how " *'Top Five' journals in economics explicitly require data and code to be provided (with possible exemptions)."*

[8] For instance, in a descriptive analysis conducted in 2019 and deposited on Zenodo, Trisovic and Mattern document that, in the top 37 economics journals, about 65% require data sharing, and an additional 27% recommend it.

[9] Of course, for non-experimental empirical analysis, only Pre-Analysis Plans are possible, not preregistration. Theoretical arguments for preregistration are strong (Chambers-Tzavella, 2022), and the empirical jury is still out regarding their effectiveness. Our focus is not on this matter, but on the issue of proportionality and fairness in the imposition of novel practices.

[10] For instance, Dawid et al. (2019) note that for large agent-based models, publications present results *"without truly providing all the details that would be necessary to fully understand all aspects of the emerging dynamics, or to actually facilitate a full re-implementation of the model from scratch. Even if the source code of the model were available, it remains costly to run the simulation model properly and to reproduce the exact same setup as the analyses presented in the paper. [...] a perceived lack of transparency and replicability is a relevant issue and might contribute to the perception of agent-based models as "black-boxes".*

that experimental psychologists are open to improving what they see as a problematic situation (e.g. Open Science Collaboration, 2015).

Biomedicine has long been at the forefront of the meta-research movement for reproducibility, being the first to develop practices such as clinical trials registration (Zarin et al., 2017), data and material sharing, and transparent reporting (Altman et al., 2008, Moher at al., 2001). In many countries, regulators, editors, professional associations and national health institutes have supported such changes. Even so, adoption has been far from perfect, even for registration of clinical trials. Sharing of raw data has increased from <1% to about 20%, and other practices may be even less common, despite some clear progress. This shows that resistance to the adoption of these practices can be substantial across diverse fields.

In the humanities, the meaning of 'reproducibility' is still developing, especially in language studies. In history, Burrows (2023) argues that, traditionally, "*... the goal was to ensure that the sources on which an argument was based could be identified and consulted, as the primary means of distinguishing fact from interpretation, and of assessing whether the conclusions drawn from the evidence were justifiable*." In recent years, with quantitative and digital history, reproducibility is associated with public availability of digitized resources. Burrows (2023) further encourages the sharing of code and metadata for statistical analysis in digital humanities.

3. Assessing Fairness Concerns for Experimental Economics and Beyond

Now that the issue of heterogeneity has been raised and documented, and its methodological roots have been specified, we can investigate its potential implications for fairness. Although the discussion is focused on the field of economics, such analysis will likely generate general insights for science. Since different branches of economics use methodologies that fall into the first three rows of Table 1, heterogeneity in the type of presumed 'best practices' is inevitable across sub-domains of economics. The question, then, is on fairness issues that this potentially raises. A list of potential concerns will help advance the discussion:

A. No rigorous scientific or policy analysis guides us on how the overall institutional adjustment should take place.[11] Hence, institutions are likely to resort to ad hoc and casual box-ticking exercises that cannot be trusted and implemented by researchers.
B. It adapting to the open science era, institutions are likely to readily promote presumed best practices. Experimental domains are 'easier targets' for showing progress, with relatively established research on best practices and their relationship with reproducibility (e.g. Bakker, van Dijk, and Wicherts, 2012; Nosek, Spies, and Motyl, 2012).
C. On the contrary, in certain observational and computational domains, it may be difficult not only to advance, but even to define reproducibility. For such domains (e.g. macroeconomics, environmental), the jury is still out, on key concepts, such as what constitutes a 'replication' with new data. This difficulty may naturally result in weaker demands for adjustment in terms of 'best practices', not because there is no true need, but because it is difficult to substantiate what a replication is, let alone a 'best practice' for promoting reproducibility.[12]

[11] In fact, the current article constitutes a simple first step in this direction.
[12] The long quote by Dawid et al. (2019), presented in Footnote 10, illustrates the great challenges these fields face even in pursuing mere computational reproducibility.

D. If experiments are rendered more costly to produce because of the need to implement best practices, this may result in an ostensible decline in experimenters' productivity, even though they are doing 'the right thing' for science. This may hamper their career prospects relative to similar researchers in other subfields of economics.

In assessing point A, it is clearly the case that institutional reforms rarely happen with full organization and rigor in the real world, so this can be thought of as a rather moot argument for the current discussion. That said (from the pragmatic point of view), we do encourage greater focus on this important policy problem in the medium and long term, and especially for policy-oriented economists.

Point B deserves serious consideration, in the sense that experimental domains could truly be viewed as low-hanging fruits for institutions willing to exhibit positive changes. However, this should be a source of pride for experimentalists, for being able to define the gold standards of reproducible research, rather than a point of protest. Once more, although we are somewhat dismissive of this argument from the practical short-run policy perspective, we think that in the longer term it deserves some serious attention. After all, it is important for society to heed the difference between 'absence of evidence' and 'evidence of absence' of a potential reproducibility problem in non-experimental domains. In a similar vein, Point C does not seem to matter much in terms of the debate about short-term policy. However, there is a need for deeper conceptual, theoretical and empirical examination of reproducibility in these non-experimental domains in the longer horizon.

We argue that Point D may be the one which is likely to hold most water in terms of the current discussion. Rigorous research assessment is crucial for modern science, and Point D adds to the reasons for this. As sub-domains adjust at disparate paces to the adoption of novel (and costly) practices, these practices must be appropriately weighted in research assessment. The main challenge is to adequately capture the higher cost – but also the higher quality – of studies incorporating reproducibility-enhancing practices. In Ioannidis and Maniadis (2024) we argued that skillful normalization of bibliometrics across subdomains of science is crucial, for these metrics to deal with the challenges and gaming risks of modern science. Importantly, such normalization is also feasible given the existing scientometric resources, and we should move swiftly in this direction.

4. Conclusions

Research practices that promote reproducibility have been increasingly encouraged, rewarded and even imposed to some degree in science. This is an important trend that is only likely to accelerate, given the demands and support - by institutions such as the European Union - for reforming research assessment and practice. As an offshoot of this development, we have argued that scholars in some areas of science have expressed dissatisfaction, feeling that they are facing disproportionate adjustment demands relative to other subdomains.

We have introduced a taxonomy to illustrate the existence of natural heterogeneity in the amenability of different domains of science to such adoption of best practices. Our analysis reveals that some domains will inevitably face costlier adjustment than others, at least in the short run. We have addressed potential fairness arguments against such disparity, and assessed them to be generally weak, apart from the potential implications of the disparity on researcher

assessment. In sub-domains with disproportionate requirements to conform with, adjustments should be accompanied by changes in assessment standards.

It is important to note that the introduced taxonomy merely reflects the current state of our knowledge: so far, the debate about reproducibility has been taking place disproportionally within a limited set of methodologies. The current analysis raises the issue of advancing the reproducibility debate in more diverse methodological domains, difficult as this may appear.

References

Adler R, Ewing J and Taylor P (2009). Citation statistics: a report from the International Mathematical Union (IMU) in cooperation with the International Council of Industrial and Applied Mathematics (ICIAM) and the Institute of Mathematical Statistics (IMS). *Statistical Science*, 24.1 1-14.

Altman D et al. (2008). EQUATOR: reporting guidelines for health research. *The Lancet*, 371(9619), 1149-1150.

American Society for Cell Biology (2012). San Francisco declaration on research assessment (DORA).

Bakker M, van Dijk A, & Wicherts JM (2012). The rules of the game called psychological science. *Perspect Psychol Sci*. 7(6):54354.

Brodeur A, et al. (2024). Do Preregistration and Pre-analysis Plans Reduce p-Hacking and Publication Bias? Evidence from 15,992 Test Statistics and Suggestions for Improvement. *Journal of Political Economy Microeconomics* 2.3, 527-561.

Brodeur, A, Cook N, and Neisser C. (2024). P-hacking, data type and data-sharing policy. *The Economic Journal*, 134.659, 985-1018.

Brodeur A, Mikola D, & Cook N (2024). Mass Reproducibility and Replicability: A New Hope. IZA Discussion Paper, No. 16912.

Burrows, T (2023). Reproducibility, verifiability, and computational historical research. International Journal of Digital Humanities 5.2, 283-298.

Camerer CF, Dreber A, et al. (2016). Evaluating replicability of laboratory experiments in economics. *Science*, 351, no. 6280, 1433-1436.

Chambers CD, and Tzavella L (2022). The past, present and future of Registered Reports. *Nature human behaviour* 6, no. 1: 29-42.

Dawid H, Harting P, Van der Hoog S, & Neugart M (2019). Macroeconomics with heterogeneous agent models: fostering transparency, reproducibility and replication. *Journal of Evolutionary Economics*, *29*, 467-538.

European Commission, Directorate-General for Research and Innovation (2021). Towards a reform of the research assessment system—Scoping report. Publications Office. Accessed 29 Apr 2024.

Fanelli D (2013). Redefine misconduct as distorted reporting. *Nature*, 494(7436):149.

Gall T, Ioannidis JPA, and Maniadis Z (2017). The credibility crisis in research: Can economics tools help? *PLoS biology*, 15.4, e2001846.

Hicks D., Wouters P, Waltman L, De Rijcke S, & Rafols I (2015). Bibliometrics: the Leiden Manifesto for research metrics. *Nature*, *520* (7548), 429-431.

Ioannidis, JPA (2005). Why most published research findings are false. *PLoS medicine* 2.8, e124.

Ioannidis JPA (2014). How to make more published research true. *PLoS Med.*, 11(10):e1001747.

Ioannidis JPA (2022). Pre-registration of mathematical models. *Mathematical Biosciences*, 345, 108782.

Ioannidis, JPA, and Maniadis Z (2024). Quantitative research assessment: using metrics against gamed metrics."*Internal and Emergency Medicine* 19.1, 39-47.

Moher D, et al. (2020) The Hong Kong Principles for assessing researchers: Fostering research integrity. *PLoS biology*,18.7, e3000737.

Munafò MR, et al. (2017). A manifesto for reproducible science. *Nature human behaviour,* 1.1 1-9.

Nosek BA, Spies JR & Motyl M (2012). Scientific utopia: II. Restructuring incentives and practices to promote truth over publishability. *Perspectives on psychological science,* 7(6) 61531.

Open Science Collaboration (2015). Estimating the reproducibility of psychological science. *Science*, *349* (6251), aac4716.

Pashler H and Wagenmakers EJ (2012). Editors' introduction to the special section on replicability in psychological science: A crisis of confidence? *Perspectives on psychological science*, 7.6 528-530.

Rudenko N (2024). Are Psychologists Ready for Stricter Journal Policies on Open Science? Insights into Pre-registration, Data Sharing, and Open Peer Review Across Tenure and Methodology, Diss. Harvard College Cambridge, Massachusetts.

Simmons JP, Nelson LD, and Simonsohn U (2011). False-positive psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychol Sci.*, 22(11):135966.

Zarin, DA, Tse, T, Williams RJ, & Rajakannan, T (2017). The status of trial registration eleven years after the ICMJE policy. *The New England journal of medicine, 376*(4), 383.