CRRESS Book

Conference on Reproducibility and Replicability in Economics and the Social Sciences

Table of contents

1 Home

The Conference on Reproducibility and Replicability in Economics and the Social Sciences is a series of virtual and in-person panels on the topics of reproducibility, replicability, and transparency in the social sciences. The purpose of scientific publishing is the dissemination of robust research findings, exposing them to the scrutiny of peers and other interested parties. Scientific articles should accurately and completely provide information on the origin and provenance of data and on the analytical and computational methods used. Yet in recent years, doubts about the adequacy of the information provided in scientific articles and their addenda have been voiced. The conferences will address the following topics: the initiation of research, the conduct of research, the preparation of research for publication, and the scrutiny after publication. Undergraduates, graduate students, and career researchers will be able to learn about best practices for transparent, reproducible, and scientifically sound research in the social sciences.

2 Contact Us

For more information, or if you are a presenter and have questions, please contact us.

CRRESS is managed by co-PIs **Lars Vilhuber and Aleksandr Michuda** (Cornell University).

The organizing committee is composed of **Vilhuber, Michuda, Ian Schmutte (UGA), and Marie Connolly (UQAM)**.

Support is provided by Sara Brooks (Cornell University) as well as the staff at the Cornell University ILR School.

Part I

Session 1 - Institutional support: Should journals verify reproducibility?

Different journals have different approaches towards enforcement of their data availability policies, ranging from a thorough and complete verification including running code and checking the output, to a cursory review of the files provided to make sure they appear satisfactory, to simply receiving the data and code package and archiving it on a website or a repository. What drives the choice of approach? What are the reasons behind such choices?

In this webinar, held on September 27th, 2022, and moderated by Lars Vilhuber, we had three panelists who are experts on this topic:

1. Guido Imbens, Professor of Economics at the School of Humanities and Sciences; Senior Fellow at the Stanford Institute for Economic Policy Research; Coulter Family Faculty Fellow at Stanford University, and editor of Econometrica,
2. Tim Salmon, Professor of Economics at Southern Methodist University and the editor of Economic Inquiry, and
3. Toni Whited, Dale L. Dykema Professor of Business Administration at the Ross School of Business at the University of Michigan and editor-in-chief at the Journal of Financial Economics.

3 Replication Packages for Journals: For and Against

It is vital to the integrity of our field to push for greater transparency in the research that we produce. In empirical work there is a substantial opportunity to hide, even unintentionally, very subtle but important details of a project in a long series of decisions regarding how to clean and merge data sets, how to calculate variables, how to calculate summary measures and how tests are actually performed. If all it is possible to see of a paper is a set of final tables, it is often quite difficult to understand exactly how authors achieved those results and to verify that they were produced accurately. By having authors make all of those decisions transparent and available for review it allows for the possibility of others to validate the work that has been performed and it allows for future researchers to be better able to build off of that work to advance our understanding of the issues. Journals can and should facilitate that process by requiring authors to provide full details about all empirical work performed. This is not an uncontroversial viewpoint and so in what follows I will talk about some of the key arguments for and against that position.

3.1 Background

I have been on the editorial boards of many different journals for over 10 years. That experience, and my experience trying to publish in journals for much longer, has made me frequently question the editorial process, how to improve it and how journals can maintain high standards for work which they publish. In July of 2021, I took over as Editor of *Economic Inquiry* and was then in position to begin putting in place some policies which I thought would be beneficial in this regard. One of the first policies that I began working on was a policy requiring authors to share data and code related to papers published in the journal. I, of course, borrowed liberally from other journals which had already adopted such policies as there were many good models out there to borrow from. When the policy was finalized, we had chosen to fund a repository on OPENICPSR for both journals operated by the Western Economic Association International (*Contemporary Economic Policy* being the other journal) and establish a policy that requires all papers published by EI which include data analysis to publish a replication archive on that or a suitable alternative site. I had many discussions along the way to arrive

at that policy and here I will explain some of the considerations which helped me to make the final choice.

3.2 Main Thoughts

3.2.1 Why Should Journal Require Replication Packages?

We can first examine the case in favor of journals operating data archive sites like ours or in general of requiring authors to post replication packages which will allow others to reproduce their work. The main point behind this push for reproducible science is that such measures are necessary not just to maintain the credibility of individual research papers but to maintain the credibility of all academic research. There have been many examples of fraudulent work being published in academic journals over the years including many cases of researchers faking data. Two of the more famous incidences of this type of fraud were by Michael LaCour and Diedrik Stapel. In the case of LaCour, he was able to publish a paper in Science, supposedly the top journal across all disciplines for academic research, in 2016 which claimed to show that contact with a homosexual individual improved one's support for gay marriage proposals.[1] This was a blockbuster finding picked up by many news outlets. It was quite humiliating to many involved when it was later discovered that the data were faked. Diedrik Stapel is a repeat offender on this issue as he was able to publish many different studies in high quality journals on the basis of faked data.[2] There are also other types of poor quality research that are fraudulent despite using real data which show up in journals as well. Among the more notorious offenders here would be Brian Wansink, the former head of a large research center at Cornell, who was also forced to retract many articles once the methods behind those articles were revealed.[3] In his case, the data existed but he engaged in methods to achieve his results which involved, to quote Cornell's Provost at the time Michael Kotlikoff, "misreporting of research data, problematic statistical techniques, failure to properly document and preserve research results, and inappropriate authorship."[4] Many of the results from these papers had also been picked up in the popular press and so the findings of research misconduct here were quite public and embarrassing to all of the research community that allowed this work to publish. Many more examples of these problems can be found on https://retractionwatch.com/ and indeed the fact that such a website exists is a testament to the fact that far too much problematic research somehow makes its way to the pages of scientific journals.

[1] The views expressed herein are those of the author and do not necessarily represent the views of the Federal Reserve Bank of Kansas City or the Federal Reserve System.

[2] Butler, C. R. & Currier, B. D. (2017). You can't replicate what you can't find: Data preservation policies in economic journals. Presentation to the International Association for Social Science Information Services & Technology (IASSIST) Conference, Lawrence, KS. Available at http://doi.org/10.17605/OSF.IO/HF3DS

[3] https://retractionwatch.com/2022/05/31/cornell-food-marketing-researcher-who-retired-after-misconduct-finding-is-publishing-again/

[4] https://statements.cornell.edu/2018/20180920-statement-provost-michael-kotlikoff.cfm

We clearly need to do better and requiring more transparency in empirical work at journals is a good start. Facing requirements to provide the all of the underlying data, explicit details on methods for data collection and code for conducting the regressions would undoubtedly deter most of the cases discussed above and many other besides. This is because being required to produce the data and make it visible to others would often unmask the underlying fraud quickly and easily. There would also be a clear public record one could check to determine legitimacy of the work. Knowing it will be harder to pass through, one hopes fewer would try and when those few still try, it should be easier to uncover the problems and deal with them as necessary. Further, not only should these requirements reduce these egregious cases of fraud, which thankfully are not that wide spread, but they will force all authors to think very carefully through their empirical processes knowing that they will be publicly viewable. This increased scrutiny should hopefully improve the quality of all research published in our journals. Preventing cases of fraud while making the details of high quality research transparently available should be a substantial boost to the legitimacy of all of our work.

It is also important that journals have policies about data availability because the ability for future researchers to reproduce existing work is necessary for the advancement of science. In many cases, one research group may wish to build upon the work already published in a journal. A first step in that process is often reproducing the initial work so that the researchers can start from there and build up. Unfortunately, if these data availability policies are not in place it is often quite difficult for a set of researchers to back out exactly what others did from a published paper alone. In one case at my own journal, a paper was submitted which was attempting to do exactly this of building off of a previous paper published at the journal. The new paper's goal was to improve on the estimation process of the previous one. The problem is that the new researchers could not reproduce the original results and so their "replication" estimation generated a result not just quantitatively different from the original authors but qualitatively different in a very meaningful way as well. This makes it then difficult to evaluate whether their improvement to the original estimation approach yielded an improvement as it is unclear that they replicated the original one correctly. That is a problem for the researchers who previously published their work as it is harder for others to build on it and it is certainly frustrating for the later researchers who cannot replicate the prior work. Having replication packages accompanying published papers can resolve this problem quickly as researchers who wish to build off of the work of others can see exactly what they did to get those results without guessing and potentially failing to identify exactly what they did.

A great example of the reason that replicating the work of others is often difficult is contained in Huntington-Klein et al (2021). This study examines the problem of replication at a deeper level than what journals usually engage in. The authors of this paper asked several teams of researchers to take the same raw data as two published papers and try to provide an answer to the same research question posed in those papers. This meant that the new researchers had to take the initial data, make all of the choices empirical researchers have to make about processing that data and specify a final regression to examine the issue. The results were that the original results often did not replicate. In some cases, the replication studies found a different sign on the key effect in question while in others, the magnitude and standard error

of the effect were quite different. Importantly, in all cases, the final number of data points considered differed despite all studies starting with the same raw data with the same number of observations. The discrepancy in the final results may have been due to the fact that different research teams often made very different choices along the way to the final specification. And thus, to really know how a team of researchers arrived at a set of results, one really needs to know more than just what was the nature of the regression conducted but you need to know all the small steps along the way to get there. Without this detailed level of information, it can be impossible to really understand how two different studies arrived at different outcomes.

It is important at this point to distinguish between two very different, though related, goals of the data availability policies of journals and how data archives may be vetted by journals. The most commonly discussed check that journals may wish to perform about a replication archive is whether one can use the archive to reproduce the results in the paper. Such a certification verifies that indeed when code is run that the results of that code reproduce what is in the paper. This verification is valuable, but a certification that the authors can re-produce their own results is not really all that useful on its own. What the paper just discussed points out is that we also need the replication sets to provide all of the details regarding how the empirical analysis was performed so that future researchers can know exactly what the authors did. With this knowledge, future researchers can begin from more robust baselines regarding published work. Without this information, we run the risk of having many parallel research programs generating seemingly conflicting results with no way to clearly determine if the conflict is due to regression specifications, different choices in data processing, errors in data processing or something else along the research chain. When designing data availability policies, we need to keep both of these goals in mind and when certifying archives as being of high quality, we need to ensure that both of these goals can be achieved.

3.2.2 Why Shouldn't Journals Require Replication Packages?

While I find the arguments above convincing regarding why journals should require replication packages, when I was contemplating putting one in place for EI, I did talk to many people who were of the opinion that journals should not be putting these requirements in place. It is worth examining their arguments against these policies to determine how convincing they are.

The first concern many would suggest about these archives is that if authors are required to post their data and their code for conducting their analysis, then others would be able to copy their work. Their concern is that the authors may have spent a great deal of time figuring out how to find the data involved, merge multiple data sets and clean them so that they work together. It may have also taken a great deal of time to implement the empirical methodology for the model in the paper. Many researchers may wish to keep that work for themselves so that they may continue to exploit that in future publications and do not want to allow others to make use of their efforts. At face value, this argument may seem somewhat convincing. While I had my own response to this, I have to say that the most convincing counter-argument

against this line of thinking came from Guido Imbens in our panel discussion on this topic. He pointed out that allowing empirical researchers to hide their methods like this is similar to allowing theorists to publish theorems while keeping the proofs hidden. A theorist could mount the same argument that the proof may have taken a long time to work out, perhaps requiring the development of special techniques in the process and they may wish to be the only ones exploiting their methods in future work. We do not allow theorists to avoid providing proofs because we need to see verification that the theorems are valid. We do not simply trust them blindly. Yet empiricists who wish to hide their methods are asking journals to blindly trust them. That should not be how publishing works. Also, while yes, making your methods and data transparent may allow others to "copy" your work, the proper way to see that is that it allows others to build off of your work. Your work can now form the foundation of the work of others and have greater impact. I would argue that the possibility that it allows others to learn more from your work is in fact one of the main reasons why journals should be requiring these packages. It is not a downside.

Another common concern about journals requiring replication packages is the suggestion that these requirements place an undue burden on authors. This can be of particular concern to certain journals as putting such requirements in place could potentially decrease the number of submissions to the journal as authors decide to submit to peer journals without such requirements. Journals likely do need to weigh this concern when considering how stringent to make their data availability policy. It is worth noting that as more and more journals adopt these policies, authors will have fewer places to submit where they can avoid these requirements and so over time concerns over this issue should diminish. It is also worth considering as a journal editor whether you want to be among the last journals not enforcing these requirements. If you are, this will mean that all those people who do not want to engage in transparent research practices will submit to your journal. As an editor, do you want to be the recipient of those submissions? Perhaps not though that decision may depend on the peer journal group for a specific journal. For journals whose peers are not yet putting these policies in place, then even high quality authors might wish to avoid the burden if they have good alternatives. For journals whose peers mostly have these requirements, then being one of the few that do not poses significant risk to the journal of receiving work for which there is a reason the authors wish to avoid transparency. Different journal editors may examine this issue and come to different conclusions on the right policy for their journal at a specific point in time. For EI, we have had the policy in place for a little less than one year and based on the current data our total submissions are slightly lower than the previous few years at this point in the cycle.[5] There are a few other possible explanations so it is not clearly attributable to this policy but the decrease is not at a problematic level even if the data policy is responsible for the entire decline.

My other view on the issue of a replication package being a burden on authors is that this is only true if authors wait until the end of the publication process to think about the reproducibility of

[5] I note that in the discussion I think I said our submissions were more or less unchanged. I re-checked the data after and with the most up to date numbers we have had a small but noticeable drop.

their work. If authors have engaged in their work in a haphazard way prior to acceptance, then it can indeed be a substantial burden to go back and document all of the data manipulation that was done and script all of the regressions performed. If, however, authors begin thinking about these issues when they begin their research, there is no real burden and in fact I would argue that engaging in your research in a way from the beginning which will make the work replicable will actually save the authors time and allow them to do higher quality work. In my own work, I admit that early in my career I did much of my data work by hand. Then when I got a referee comment suggesting a different way to conduct a regression I would have to engage in some forensic econometrics to first back out what I actually did to get the prior result. This was wasted time and not the best way to do research. Now that I have all the analysis scripted, making changes like this is much faster and I do not have to wonder exactly how I created a variable or exactly which observations may have been dropped or why. All of that is in the scripting files from the beginning. As authors begin to expect to face these requirements and learn how to take this into account from the beginning of their analysis, the burden of providing a replication package upon acceptance of the paper diminishes substantially. I expect that these practices should be becoming more common in the profession and so the concern over this element should diminish with time. We can further diminish them by making sure that replicable research is brought into Ph.D. training programs.

A final notion that some suggested to me is that there is no need for journals to require replication packages. Individuals who want to provide their data can do so on their own sites and if there are professional incentives to do so perhaps in the form of these packages being seen as signals of high quality, everyone will do this anyway. Perhaps this could be true but most do not currently publicly archive replication files absent journal requirements. Were that to start, then it could be seen as a high-quality signal when someone does it which would mean that as journal editors we should be taking it into account in our decisions whether someone provides the data archives. If we do that, it is just a backdoor way to require replication archives but with a serious downside. If we make an accept decision based on an author saying that they will post an archive, after the paper is published authors could quickly pull that archive. Essentially, this approach is not an effective way of accomplishing the goals of research transparency. In order to ensure that the data remains available, it is best that journals maintain the archives for integrity of the process so that authors cannot manipulate the archive after the paper is published.

3.3 Conclusion

I believe quite strongly in the need for transparency in research. In order to preserve and maintain the integrity of all of academic research, we need to push for ever greater transparency in how research is done. That way when there are questions about the legitimacy of a claim, those questions can be quickly and easily addressed. This level of legitimacy is a benefit to us all. The main "cost" (if one sees it that way) would be that greater transparency limits the ability of people to publish ill-founded results. It is true that greater transparency does place

greater requirements on researchers to engage in more careful and rigorous work which can survive the greater scrutiny possible with the increase to transparency. I see this as a clear benefit rather than as a cost.

Of course, the path to this greater transparency norm will not be direct and not all journals will adopt the same standards at the same time. There are some journals leading in this direction, some following and some lagging behind. There are good reasons for different journals to be in each of those stages. As journals collectively move along this path it is important that we do so in ways that are not unduly burdensome on authors. This means that while requiring replicable archives is a valuable thing, it does not make sense for different journals to impose very different and idiosyncratic requirements about file structures and things like that such that when authors prepare a replication archive they must do a great deal of work to change it from a format suitable to one journal versus another. As a journal editor I appreciate the work done by others to establish clear guidelines on these issues which other journals can adopt as well to try to harmonize these requirements where we can.

3.4 References

Huntington-Klein, N., Arenas, A., Beam, E., Bertoni, M., Bloem, J.R., Burli, P. et al. (2021) The influence of hidden researcher decisions in applied microeconomics. *Economic Inquiry*, 59: 944– 960. https://doi.org/10.1111/ecin.12992

4 Comments on Reproducibility in Finance and Economics

4.1 Introduction

Reproducibility is defined as obtaining consistent results using the same data and code as the original study. Most of the discussion of reproducibility has centered around the many obvious benefits. Reproducible research advances knowledge for several reasons. It reduces the risk of errors. It also makes the processes that generate results more transparent. This second advantage has an important educational component, as it helps disseminate not just results but processes. However, reproducibility is not without costs. Good research procedures consume resources both in terms of a researcher's own efforts and in terms of the involvement of arms-length parties in actually reproducing the research. This second cost is not just a time cost; it is pecuniary as well.

Thus, reproducibility is a good that is costly to produce and that has many positive externalities. Researchers internalize many of the benefits of reproducibility, especially in terms of research extendability and personal reputation. However, they do not internalize any of the benefits to the research community at large. Because reproducibility is costly, it is unlikely to be produced at a socially optimal rate by any individual researchers. Thus, the questions are the extent to which reproducibility should be subsidized and who should subsidize it. Should all research be reproduced by arms-length parties, and what are the least costly policies that facilitate reproducible research? The rest of this note is organized around policies regarding actual reproduction and proprietary data.

4.2 Code, Data, and Arms-Length Reproduction

One low-cost and easily implementable set of policies that enhances the reproducibility of research is journals' data and code disclosure policies. In the age of inexpensive data storage and an abundance of public repositories, the costs of these policies are small, and the policies should be implemented. They impose some costs on researchers in terms of organizing data and code, but well-organized data and code are already an essential part of the research process, so these costs should be small.

While simple to implement, this low-cost policy is not without non-pecuniary drawbacks for journals. The code and data can be incomplete, poorly documented, or unusable. Moreover, journal editors have to retract articles that, after publication, cannot be reproduced. In economics, these concerns have prompted journals to start arms-length reproduction of results. The benefit of this policy is primarily that authors and journals can be confident that the code submitted with an article actually works to reproduce the results.

However, the pecuniary costs of this policy can be substantial. It is expensive for journals to hire data editors and well-trained research assistants, and many academic journals run on tight budgets. It is often time-consuming for authors to comply with reproducibility requirements. This last issue is particularly burdensome for authors who cannot afford research assistance.

While the above issues involve costs, the following are more fundamental. Reproducibility policies give researchers incentives to do research that is easier to reproduce, thus restraining research innovation that requires either large data or intense computing. Most importantly, code that can run on data and reproduce results can still contain errors.

These arguments imply that while individual researchers are likely to underproduce reproducibility, it is also unlikely optimal for the progress of science that all research be reproduced before publication. Some papers, even those in the very best journals, rarely get read or cited, and the benefits of reproducing these papers are small.

However, ex-ante, it is hard to know which papers will attract attention and which will not. One solution that lies between data and code disclosure and arms-length reproduction is verification. It is much less expensive to verify the contents of a replication package than to do an actual reproduction. Verification might consist of checking for the existence of replication instructions, an execution script, or either data or pseudo-data. This type of service could be provided by journals or other third parties, much as copy editors fix syntax and grammar errors before articles are submitted. At that point, reproducibility would be left up to the academic community, with the more important pieces of research being subject to greater scrutiny.

A final issue with reproducibility is education. In economics and finance, students are not taught how to create reproducible research. An improvement that would go a long way toward improving the culture surrounding reproducibility would be to teach PhD students how to organize research projects and to write code in such a way that others can reproduce results easily. This type of education would lower the costs to individual researchers of making their own research reproducible.

4.3 Proprietary Data

A possibly larger challenge for reproducibility than verification or arms-length execution of code is proprietary data. A clarification is necessary because not all types of data with restricted access are completely secret, that is, available only to the data provider and a researcher. For

example, commercial data sets are not secret, just costly to obtain. Similarly, administrative datasets are not secret. They just require special permission. In contrast, proprietary data cannot be offered to the research community at large for the purposes of reproducing the results. So the question is whether journals should discourage the use of this type of data or require that verifiers have access to the data. Given the large number of studies using proprietary data, this issue is possibly more important than the issue of running code.

4.4 Conclusion

In conclusion, the reproducibility of research is essential for the advancement of science. However, it is not without costs, so blanket statements that all research should be reproducible are not feasible. Instead, feasible policies include those that lower the costs for others to replicate research. Data and code disclosure is a low-cost policy that should be implemented widely. Verification of code and data packages is a slightly more costly option. Arms-length reproduction is a much more costly alternative. Finally, perhaps the most important issue that impedes reproducibility in finance and economics is the use of proprietary data.

Part II

Session 2 - Replication and IRB

One of the most crucial dimensions that Institutional Review Boards are interested in are the protocols that researchers have in place to protect their subjects' privacy. This often leads to researchers writing in their IRB protocols that they will destroy their data once their project is complete. Understandably, however, destruction of data makes it impossible to verify and replicate work, which is increasingly becoming a vital part of modern science. How should data privacy be handled in the wake of the replication crisis? What protocols and standards should be put in place to minimize the risk of data leakage? Or should data be destroyed after some time span?

5 Sex, Lies, and Data: New Models of Informed Consent

This paper identifies gaps within the current research ethics regime that can be complemented by intersectional feminism, queer theory, sexual education, BDSM, and critical theory. When these disciplines are applied, informed consent models can better integrate revocability, ongoing consent, and contextual consent. Future work on informed consent and research ethics should incorporate these areas to preserve the dignity of research subjects and equitably benefit all members of the research process.

5.1 Introduction

In the summer of 2015, a company called Ashley Madison was hacked and over seven million users' data was leaked (Lord, 2017). Data breaches like this are unfortunately common, but as Ashley Madison was specifically created to facilitate infidelity, the social fallout of the breach was more harmful than most. After the breach and subsequent fall-out, which included many individuals being publicly named and shamed after their information was found in the customer records, many users of the site ended up getting divorced, some lost their jobs, and there were multiple confirmed reports of self-harm (Baraniuk, 2015).

Several researchers saw the data breach as an opportunity to advance their research. They used the leaked data to explore different questions about user demographics, geography, and risk-taking behavior and published the results in multiple peer-reviewed articles (Billau, 2017; Vedantam, 2016). However, in order to be published, these studies first had to be approved by their university's Institutional Review Board (IRB).

IRBs are intended to reduce the potential risks and harms in research proposals that involve human subjects (or data thereof) and are often perceived as a marker of credibility for ethical research. All IRBs use something called the Common Rule, which is a set of federal guidelines that dictate how to conduct ethical research, and there is a stipulation in it that says if data is already public, then there is little potential for risk or harm for it to be used in research. It makes no stipulations about how that data became public. In fact, IRBs have no remit to review secondary uses of data, with bio-specimen data as the sole exception. To an IRB, public

data is considered low risk regardless of how that data was obtained, and regardless of how sensitive the data is, and evaluating its use is not within their scope of review. Researchers got their studies approved by their IRBs, and academics continue to use and cite Ashley Madison data in publications. As a result, many more people see the leaked data and names of the implicated humans than would have had those papers never been published.

5.2 Nonconsensual Porn and Nonconsensual Data

There are parallels between the use of leaked data sets like the Ashley Madison hack, and the viewing of nonconsensual pornography (NCP). For the unfamiliar, NCP is "the distribution of sexually graphic images or videos of an individual without their consent in the context of an intimate relationship," (Carter, 2021). Perpetrators of NCP sometimes upload photos or videos of their victims onto pornography websites, which allow many more people to view them. Like most platforms, these websites often use engagement metrics, like view counts, which help determine if a video is "popular", with more popular media receiving more promotion. From a victim's perspective, the more people who see their image or video, the worse the mental and/or emotional harm. While the original perpetrator inflicted a significant harm, that harm gets compounded every time it gets viewed or shared by other people, and by viewing it, it raises the likelihood that someone else will view it.

When researchers use data that was collected nonconsensually, unethically, or illegally, it has the potential to amplify the harm of the original data collector by pointing other people to it. Using unethical data for research is not the moral equivalent of the original hack, but there are many circumstances where it may still be considered morally wrong. The fact that the Common Rule and IRBs don't have the mechanisms necessary to address this issue means that we need more tools to fully judge the needs and bounds of ethical research (Jordan, 2022). Fortunately, there are many disciplines and scholars that have developed theoretical tools and practices we can use to improve the ethical conduct of research. I offer several areas that have been personally productive, though there are many others that can be used.

5.3 Intersectional Feminism, Queer Theory, Sex Education, and Critical Theory

The first area that has been helpful for developing a fuller understanding of consent is intersectional feminism. Feminism is an interdisciplinary approach to addressing oppression related to gender identity and expression. Intersectionality is a theory that when someone has more than one marginalized identity (e.g., being a Black woman), the combination of those identities produces a greater risk than the sum of their parts (Crenshaw, 2006). Intersectional feminism interrogates oppression on multiple fronts including gender, race, class, disability, and others, especially how these oppressions interact with each other. These theories have much to say

about what informed consent should look like, especially when it comes to bodily autonomy and agency. An intersectional feminist approach to informed consent will ensure that participants have control over their participation in a research study, individual control over their personal data, and communal control over communal data (Sterling, 2011; Fiesler, McCann, Frye & Brubaker, 2018). Informed consent can sometimes be framed as a liability waiver for institutions. Intersectional feminism would reframe informed consent as an expression of care for the wellbeing of the person which may supersede the research goals of the researchers or the legal liability of the university.

A second helpful area is queer theory. Queer theorists critique dominant social expectations of sexual orientation and gender identity (Cohen, 1997). Informed consent can sometimes mirror hetero-patriarchal models of power by framing consent as something only certain people are qualified to give (cis, heterosexual men), while others (LGBTQ people) are disqualified, for example, the FDA's ban on gay and bisexual men donating blood (Human Rights Campaign, 2020). Integrating queer theory and listening to the LGBTQ community can aid in analyzing how certain processes can unintentionally reify discriminatory ways of thinking, and they can help us move toward more inclusive ways of achieving informed consent (Edenfield, 2019; de Heer, Brown & Cheney, 2021).

A third area is sexual education and the BDSM community. Inclusive and evidence-based sex educators have developed models of affirmative and enthusiastic consent that are sensitive to context (Center for Sex Education, 2016). Bondage/Discipline, Dominance/Submission, Sadism, and Masochism (BDSM) are consensual sexual acts that involve a power dynamic between partners. This kind of sex has a higher potential for risk because there are elements of pain or power involved and the BDSM community has developed many practices around consent and communication to mitigate those risks such as safe words (Dunkely & Brotto, 2020). These practices include experiences ranging from in-person to remote relationships, sometimes involving people's bodies while other times involving domination over personal documents or computers (Vogt & Goldman, 2018). Sex education and BDSM have richer, more nuanced theories and consent practices than most researchers at universities.

A fourth area is critical theory, in particular its analysis of power and push for social change. Critical theory takes many forms and is often integrated into existing disciplines as a way to challenge structural assumptions and redistribute resources and opportunities within them, such as critical legal studies, critical pedagogy, and others. A critical approach to consent would be attuned to who has power within an informed consent transaction, who doesn't, and attempt to redress the asymmetrical dynamic. Applying a critical theory approach to informed consent can produce some unanticipated results. For example, the Panama Papers were a set of leaked documents exposing offshore financial transactions including criminal tax evasion, money laundering, and other financial crimes around the world (Fitzgibbon & Hudson, 2021). This was likely a situation where the material was obtained unlawfully and without the consent of the people whose information was publicized. Hundreds of papers and many books have been written that detail the Panama Papers and the individuals named in them. A traditional approach to informed consent would say that the Panama Papers leak was a

wrongful violation of privacy. A critical consent approach would disregard claims of privacy by the victims because of the nature of the information leaked, namely a highly organized global financial system used by the rich and powerful to hide assets at the expense of the poor and powerless. Informed consent with a critical theory lens inverts traditional power dynamics, meaning that some rules about research ethics should be broken when following them would cause inequitable outcomes.

5.4 New Models of Consent

When you look across the areas of intersectional feminism, queer theory, the LGBTQ community, sex educators, the BDSM community, and critical theory, there are commonalities that can be pulled out to help develop a better framework of consent in research. Using these as reference points, informed consent must be at least three things: revocable, ongoing, and contextual.

Revocability means that after you give your consent to something, you can change your mind at any time and for any reason. Current research practice does this somewhat. Most consent forms state that research participants can stop their participation at any time, but there are many informal and implicit forces that discourage this. Participants may feel that by changing their minds they would inconvenience or disappoint the researchers, who are in a position of authority. Some participants may be persuaded by the sunk cost fallacy that if they've done something long enough, they might as well finish it even if they would prefer not to. In general, research participant consent is revocable only during the data collection phase. Once that data is used to publish something, there is almost nothing that a participant can do to remove their data from the study.

Ongoing consent means that the 'one and done' model of most research is inadequate. In practice, it means frequent points of intentional communication gauging the interest of the participants in continuing in the research study. This does not have to devolve into performative check-ins that likely produce consent fatigue (Ranisch, 2021). The modes of ongoing consent should adapt to the environment and be naturally integrated into the requirements of the participant experience, so that providing or revoking consent is both easy and perceived as casual. In practice, this can look like introducing several points in a study where participants have to opt-in to continuing, with the expectation that if they don't, they automatically discontinue participation. Sometimes called contextual consent or just-in-time consent, this makes leaving a study mid-way through seem less interruptive or socially uncomfortable (IF, 2022).

Contextual consent means that there isn't a template or formula for getting consent that can be applied across all research studies. When people consent to having sex, it is for specific people, places, and times. Giving consent to have sex once does not mean that you consent to sex with that person at any time and place in the future. Contextual consent adapts to the specific conditions it is being offered in and adjusts how it is communicated. In practice, this could

look like developing unique consent methods for different populations that are attentive to the cultural and rhetorical differences, even if the research protocol is otherwise the same. For some populations, the standard written forms of consent using academic language may be adequate. For other populations, offering techniques using visual, auditory, narrative, behavioral, or game-based methods might be better suited to get meaningful consent

Integrating these informed consent practices into human subject research will significantly improve the experience of participants as well as reduce the potential for unethical data collection and sharing. Academia has many justifications to ignore informed consent. Most times consent is neglected, the subsequent risks are distributed unequally among participants. The people who are put at the most risk are usually the people who are already the most vulnerable: women, children, people who are LGBTQ, non-binary, disabled, poor, people of color, and many others. Researchers have a responsibility to conduct research ethically and with respect for individual privacy needs and expectations. Using these consent practices by integrating intersectional feminism, queer theory, the LGBTQ community, sex education, BDSM, and critical theory can increase the quality of research and respect the dignity of the participants.

An admittedly difficult area for applying these frameworks is when the circumstances of data collection and informed consent are more quotidian. The examples used above are dramatic when compared with most of the common research practices such as collecting public social media data from places like Facebook, Reddit, or Twitter. When the consent obtained is ambiguous, when the data being collected isn't particularly sensitive, or when the potential for significant social benefit is high, applying these theoretical tools isn't straightforward. Reasonable people can, and often do, disagree as to what ethical values are at stake and how to adjudicate them. The unfortunate ethical gray area this produces is still better than having clear but inequitable guidelines.

5.5 Final Thoughts

It is highly unlikely that any rules developed to address the growing complexities of research ethics, especially in technology spaces, will be able to address every possible ethical dilemma that arises. The four areas offered above are tools that can help with some of the gaps currently found in research ethics, but they need to be constantly reexamined and supplemented as new issues present themselves. Ethical research is a horizon, not a place; we never arrive. By prioritizing the wellbeing of the most vulnerable and marginalized people in our communities and continually inventing better models of informed consent, we can take more confident steps towards that horizon.

5.6 References

Baraniuk, C. (2015). *Ashley Madison: 'Suicides' over website hack*. BBC. August 24.

https://www.bbc.com/news/technology-34044506

Billau, C. (2017) *Academic research uses hacked Ashley Madison data to map areas with most cheating husbands*. UToday, Alumni, Arts and Letters. January 10.

https://news.utoledo.edu/index.php/01_10_2017/academic-research-uses-hacked-ashley-madison-data-to-map-areas-with-most-cheating-husbands

Carter, C. (2021). *An Update on the Legal Landscape of Revenge Porn*. National Association of Attorneys General. November 16.

https://www.naag.org/attorney-general-journal/an-update-on-the-legal-landscape-of-revenge-porn/

Center for Sex Education, (2016). Why Comprehensive Sex Ed and Consent Education Go Hand in Hand, May 13.

https://www.sexedcenter.org/why-comprehensive-sex-ed-and-consent-education-go-hand-in-hand/

Cohen, C. J. (1997). Punks, bulldaggers, and welfare queens: The radical potential of queer politics? Glq, 3(4), 437-465.

https://doi.org/10.1215/10642684-3-4-437

Crenshaw, K. W. (2006). *Intersectionality, identity politics and violence against women of color*. Kvinder, Køn & Forskning, (2-3)

https://doi.org/10.7146/kkf.v0i2-3.28090

de Heer, B., Brown, M., & Cheney, J. (2021). Sexual consent and communication among the sexual minoritized: The role of heteronormative sex education, trauma, and dual identities. Feminist Criminology, 16(5), 701-721.

https://doi.org/10.1177/15570851211034560

Dunkley, C. R., & Brotto, L. A. (2020). The role of consent in the context of BDSM. Sexual Abuse, 32(6), 657-678.

https://doi.org/10.1177/1079063219842847

Edenfield, A. (2019). Queering consent: Design and sexual consent messaging. Communication Design Quarterly Review, 7(2), 50-63.

https://doi.org/10.1145/3358931.3358938

Fiesler, C., McCann, J., Frye, K., & Brubaker, J. R. (2018, June). Reddit rules! characterizing an ecosystem of governance. In *Twelfth International AAAI Conference on Web and Social Media*.

https://www.aaai.org/ocs/index.php/ICWSM/ICWSM18/paper/viewPaper/17898

Fitzgibbon, W., Hudson, M. (2021) Five years later, Panama Papers still having a big impact. International Consortium of Investigative Journalists. April 3, 2021

https://www.icij.org/investigations/panama-papers/five-years-later-panama-papers-still-having-a-big-impact/

IF. (2022) "Giving and Removing Consent: Just-in-time Consent". Data Patterns Catalog. IF

https://catalogue.projectsbyif.com/patterns/just-in-time-consent/

Human Rights Campaign. (2020) Blood Donations.

https://www.hrc.org/resources/blood-donations

Jordan, S. (2022) The Playbook: Data Sharing for Research. The Future of Privacy Forum.

https://fpf.org/wp-content/uploads/2022/12/FPF-Playbook-singles.pdf

Lord, N. (2017). *A Timeline of the Ashley Madison Hack*. Digital Guardian. July 27, 2017.

https://digitalguardian.com/blog/timeline-ashley-madison-hack

Ranisch, R. (2021). Consultation with doctor twitter: Consent fatigue, and the role of developers in digital medical ethics. American Journal of Bioethics, 21(7), 24-25.

https://doi.org/10.1080/15265161.2021.1926595

Sterling, R. L. (2011). Genetic research among the Havasupai: A cautionary tale. *AMA Journal of Ethics*, *13*(2), 113-117.

https://journalofethics.ama-assn.org/article/genetic-research-among-havasupai-cautionary-tale/2011-02

Vedantam, S. (2016). *Ashley Madison Hack Inspires Social Scientists to Look Behind the Names*. National Public Radio. April 28.

https://www.npr.org/2016/04/28/476060486/ashley-madison-hack-inspires-social-scientists-to-look-behind-the-names

Vogt, P. J., Goldman, A. (2018) Episode 116 Trust The Process. /reply-all/. February 28, 2018.

https://open.spotify.com/episode/0NGOihC8u0GBuxdV3fEBLb

Part III

Session 3 - Should teaching reproducibility be a part of undergraduate education or curriculum?

This session was held at the Southern Economics Association meeting on November 20th, 2020. Panelists discussed teaching reproducibility (TIER Protocol), the involvement of undergraduates for replications based on restricted-access data, and other topics.

6 Data Citations and Reproducibility

in the Undergraduate Curriculum

Data citations are the foundation of reproducibility. To develop reproducibility skills among undergraduate students we must start with basic data literacy skills such as citing data consistently.

Disclaimer: The views expressed in this article are those of the authors and don't necessarily reflect the position of the Federal Reserve Bank of St. Louis or the Federal Reserve System.

6.1 Introduction

The scholarship of teaching and learning in economics documents multiple efforts to bring the quantitative dimension of our professional work closer to the undergraduate college curriculum.

Economics educators describing data-focused assignments and projects (Wolfe, 2020; Halliday, 2019; Wuthisatian and Thanetsunthorn, 2019; Marshall and Underwood, 2019; Mendez-Carbajo, 2015 & 2019) highlight the data-finding step of these projects. Even when the datasets are directly provided to the students, (e.g., Easton, 2020) the instructors emphasize the broader literacy dimensions of the assignments. However, there is neither professional consensus about how to build data-literacy skills (Wuthisatian and Thanetsunthorn, 2019) or much actual research on their mastery among economics students (Halliday, 2019).

In this chapter, we document baseline proficiency levels among undergraduate college students related to identifying data series and their sources. We also put forward an accessible pedagogical strategy to develop basic reproducibility skills.

We argue reproducibility should be part of the undergraduate curriculum in economics because it is a valuable professional skill to be developed throughout the curriculum by consistently citing the data sources used in economic arguments. We must instill the practice leading by example and enrolling the help of librarians

6.2 Expected Proficiencies

There is a natural overlap regarding the development of data-literacy skills between economics and library science: both disciplines value it and contribute to its development.

The two seminal descriptions of data literacy expected proficiencies among undergraduate students are provided by Hansen (2012) and Pothier and Condon (2019). The first of the seven broad competencies of economics majors named by Hansen directly address data provenance. It states: "Access existing knowledge: […] Track down economic data and data sources. Find information about the generation, construction, and meaning of economic data."

The library science perspective provided by Pothier and Condon is articulated through seven expected data competencies of economics and business majors. The last one states: "Data ethics: The principles of data ethics are built on data ownership, intellectual property rights, appropriate attribution and citation, and confidentiality and privacy issues involving human subjects."

The utilitarian and ethical aspects of data reproducibility outlined above are bridged by the American Economic Association's (AEA) (2020) Data and Code Availability Policy, which clearly states "All source data used in the paper shall be cited, following the AEA Sample References." However, the scholarship documenting the collaboration in this area between instructional economics faculty and librarians is limited. Neither the calls by economics instructors (McGrath and Tiemann, 1985; Li and Simonson, 2016; Mendez-Carbajo, 2016) nor the experiences documented by librarians (Wheatley, 2020; Wilhelm, 2021; Waggoner and Yates Habich, 2020) appear to have broad impact.

6.3 Evidence of Broad Data Literacy Skills

Mendez-Carbajo (2020) documents baseline levels of data literacy competency in several areas key to the accurate and ethical use of data for communication and decision-making among high school and college students.

In the online economic education module produced by the Federal Reserve Bank of St. Louis "FRED Interactive: Information Literacy", two separate groups of high school students (N= 450) and college students (N= 912) answer seven pre-test questions. The questions are mapped to the data literacy competencies described by both Pothier and Condon (2019) and Hansen (2012).

The analysis finds effectively identical levels of average baseline data literacy competency between high school and college students. However, it also documents much higher levels of perceived self-efficacy among college students than among high school students. In other words, college students are no more knowledgeable or skilled than high school students but

are significantly more confident in their work. This finding highlights a major challenge for instructors working to develop the expected proficiencies identified in the literature: the average college student is unduly comfortable in their limited understanding of the primary sources of economic data.

6.4 Evidence of Narrow Reproducibility Skills

During the fall semester of 2020, we distributed a short online assignment to all 854 students enrolled in two different upper-division economics courses offered by a large public university in the United States.

On average, the students are slightly above 20 years of age, 49% identify themselves as female, 21% identify as non-White racial or ethnic minorities, and 92% report English is their native language. Academically, 87% of students are business, economics, or finance majors and hold a grade point average of 3.41. Also, 68% of students are currently enrolled in a statistics course required by their program and, on average, have previously completed more than one and a half economics courses.

The assignment had three sections:

- First, the students were directed to read a brief, 900-word, essay on how to create data citations with FRED®. This essay provided background on the value of good data citations for practitioners of economics and could be used as reference material for the next two sections of the assignment.
- Second, the students were directed to read two short --under 600 words, economic essays. See them here and here. Each included a line graph of economic data. In the text, the authors referenced the data series and their sources while interpreting the quantitative information presented in the graph.
- Third, the students were asked to complete three tasks: identify the data series discussed in the essay; identify the sources of the data series discussed in the essay; and identify the missing elements of a data citation provided in the essay.

The assignment was completed in its entirety by 501 students. Table 1 reports our findings.

Table 1. Data Literacy Skills

Scores, Misconceptions and Errors	**Essay A**	**Essay B**
Identifies Series Correctly	0.57	0.47
Identifies Sources Correctly	0.21	0.03
Identifies Incomplete Citation	0.18	-0.04
Can't Identify Sources	0.05	0.12
Confuses Source with Distributor	0.72	0.73

Scores, Misconceptions and Errors	Essay A	Essay B
Considers Citation to be Complete	0.25	0.40

Note: Data Literacy Scores: = (# − #) / (#)

We document very weak student data literacy competencies associated with narrow reproducibility skills. Data literacy scores related to correctly identifying the sources of the data or recognizing an incomplete data citation are very low. Moreover, we document a frequent misconception of confusing the data source with the distributor.

These findings have practical implications for instructors, whether they are librarians or economic educators. Our work suggests there is a substantial instructional opportunity to help students develop the ability to recognize data series and their sources. In that regard, disambiguating the roles of data distributors and data sources can potentially yield large benefits to students, who would be able to acquire a more sophisticated understanding of how data are created and made available.

6.5 Proposed Instructional Intervention

We propose a broad instructional intervention for economics instructors reflecting the fact that correctly citing the data is a foundational literacy skill.

- Lead students by example and consistently name the sources of all data referenced or used in your teaching.
- Embed this practice in all your teaching, regardless of the type or subject of the course.
- Enroll the help of librarians by leveraging their ongoing instructional outreach on information literacy to include data citations.

Proficiency in identifying data sources is foundational to the development of reproducibility skills. The earlier and the more frequently students are exposed to best practices in data citations, the more effortlessly they will be able to adopt sophisticated professional replicability practices.

6.6 Conclusion

Reproducibility should be part of the undergraduate curriculum in economics:

- It is a valuable professional skill that shows the background work that goes into doing economic research. Citing the sources of the data makes research work more thorough.

- This skill should be developed throughout the curriculum. This skill is not particular or exclusive to econometrics or statistics courses.
- The first step is to consistently cite the data sources used in economic arguments. This includes data tables, plots, and in-text references.
- We must instill the practice by leading by example. Economics educators should enroll the help of librarians in developing this skill among students.

6.7 References

American Economic Association. (2020). Data and Code Availability Policy. https://www.aeaweb.org/journals/
code-policy.

Easton, T. (2020). Teaching econometrics with data on coworker salaries and job satisfaction. *International Review of Economics Education*, 34, 100178. DOI 10.1016/j.iree.2020.100178.

Halliday, S. D. (2019). Data literacy in economic development. *The Journal of Economic Education*, 50 (3), 284-298, DOI: 10.1080/00220485.2019.1618762

Hansen, W. L. (2012). An expected proficiencies approach to the economics major. In *International handbook of teaching and learning economics*, ed. G. Hoyt and K. McGoldrick, 188–94. Cheltenham, UK and Northampton, MA: Edward Elgar.

Li, I., and Simonson, R. D. (2016) The value of a redesigned program and capstone course in economics. *International Review of Economics Education*, 22, 48-58, DOI: 10.1016/j.iree.2016.05.001.

Marshall, E. C., and Underwood, A. (2019). Writing in the discipline and reproducible methods: A process-oriented approach to teaching empirical undergraduate economics research. *The Journal of Economic Education*, 50 (1), 17-32. DOI: 10.1080/00220485.2018.1551100

McGrath, E. L., and Tiemann, T. K. (1985). Introducing empirical exercises into principles of economics. *The Journal of Economic Education*, 16 (2), 121-127. DOI: 10.1080/00220485.1985.10845107

Mendez-Carbajo, D. (2015). Visualizing data and the online FRED database. *The Journal of Economic Education*, 46 (4), 420-429. https://doi.org/10.1080/00220485.2015.1071222

Mendez-Carbajo, D. (2016). Quantitative reasoning and information literacy in economics. In *Information Literacy: Research and Collaboration across Disciplines* (pp. 305-322), Barbara D'Angelo, Sandra Jamieson, Barry Maid, and J anice R. Walker (editors). Perspectives on Writing. Fort Collins, Colorado: WAC Clearinghouse and University of Colorado Press. https://wac.colostate.edu/books/infolit/chapter15.pdf

Mendez-Carbajo, D. (2019). Experiential learning in macroeconomics through FREDcast. *International Review of Economic Education*, 30 (1). DOI: 10.1016/j.iree.2018.05.004.

Mendez-Carbajo, D. (2020). Baseline competency and student self-efficacy in data literacy: Evidence from an online module. *Journal of Business & Finance Librarianship*, 25:3-4, 230-243. DOI: 10.1080/08963568.2020.1847551

Pothier, W., and Condon, P. (2019). Towards data literacy competencies: Business students, workforce needs, and the role of the librarian. *Journal of Business and Finance Librarianship* 25:3-4, 123-146. DOI: 10.1080/08963568.2019.1680189

Waggoner, D., and Yates Habich, B. (2020). Collaboration is the key: faculty, librarian and Career Center professional unite for marketing class success. *Journal of Business & Finance Librarianship*, 25:1-2, 82-91. DOI: 10.1080/08963568.2020.1784658

Wilhelm, J. (2021). Joint venture: An exploratory case study of academic libraries' collaborations with career centers. *Journal of Business & Finance Librarianship*, 26:1-2, 16-31. DOI: 10.1080/08963568.2021.1893962

Wheatley, A., Chandler, M., and McKinnon, D. (2020). Collaborating with faculty on data awareness: A case study. *Journal of Business & Finance Librarianship*, 25:3-4, 281-290. DOI: 10.1080/08963568.2020.1847553

Wolfe, M. (2020). Integrating data analysis into an introductory macroeconomics course. *International Review of Economics Education*, 33, DOI: 10.1016/j.iree.2020100176

Wuthisatian, R., and Thanetsunthorn, N. (2019). Teaching macroeconomics with data: Materials for enhancing students' quantitative skills. *International Review of Economics Education*, 30, 100151. DOI 10.1016/j.iree.2018.11.001.

7 "Yes We Can!": A Practical Approach to Teaching Reproducibility to Undergraduates

This paper presents a sequence of four exercises illustrating how students can be introduced to reproducible methods of data processing and analysis in a series of small steps. Each step is modest and feasible even in introductory classes, but cumulatively they allow students to achieve state of the art standards of reproducibility. By demonstrating the feasibility of teaching reproducible methods to beginning students, these exercises support the assertion that reproducibility can and should be integrated into quantitative methods training at all levels of the curriculum.

I thank Lars Vilhuber and Aleksandr Michuda for the opportunity to participate in the Conference. I acknowledge my debt to Barak Obama and Robert T. Builder for the inspirational slogan I have borrowed from them to use in the title of this paper. The ideas about teaching reproducibility in this paper have been developed over more than a decade of collaboration with Norm Medeiros.

7.1 Background

Is it feasible to include reproducible research methods in undergraduate training in quantitative data analysis? There are reasons to believe the answer to that question is "no"—that reproducibility is an advanced topic best left to graduate school or early career training. Professional standards such as the AEA Data Editor's guidelines and the World Bank Development Impact Evaluation (DIME) manual may appear too technical and complex to introduce to undergraduates. Even the TIER Protocol, which was designed to be accessible to students at all levels, is elaborated with a degree of specificity and detail that could give the impression that incorporating reproducibility into undergraduate classes and research supervision would be a costly and disruptive undertaking.

This essay argues that, on the contrary, integrating reproducibility into the undergraduate curriculum is eminently feasible. To support this claim, we develop a simple exercise of the kind that might be assigned in an introductory quantitative methods class, and then present

four versions of the exercise: a baseline in which the issue of reproducibility is entirely neglected, and three subsequent versions that incrementally introduce essential elements of reproducibility. The additional skills students must acquire for each version of the exercise are modest, but cumulatively they prepare students in computational methods that achieve state of the art standards of reproducibility. These exercises demonstrate the feasibility of teaching reproducibility to undergraduates, and provide instructors with concrete examples of small, practical steps they can take to achieve that goal.

7.2 Main Thoughts

7.2.1 The Exercise

In all versions of the exercise, students are given an extract of data from the 2018 American Community Survey (citation), and use it to compare average incomes of prime working-age workers by race and sex. The computational tasks are (i) to construct a table showing the means of total income for groups defined by race and sex, and (ii) illustrate those group means in a bar graph. Students then write a report in which they present the table and bar graph, and comment on the patterns they observe.

The report students submit for all four versions are identical. The versions differ in the extent to which students adopt practices that enhance the reproducibility of their results, and in the documentation that is submitted with the report.

Version 1: Interactive and non-reproducible. In this baseline version, the issue of reproducibility is entirely ignored. Students open the data file by double-clicking, and then generate the table and bar graph using a menu-driven GUI or by typing commands interactively. They use a word processor to write the report, and insert the table and graph by copying-and-pasting the output displayed on their monitor. The only work students turn in is a single document—the report.

Version 2: Writing scripts, the project folder, and the working directory. Scripts are fundamental to reproducible research: it is by executing scripts written and preserved by the author of a study that interested readers are able to reproduce the results.

Instructors and students accustomed to an interactive workflow are often reluctant to adopt reproducible methods because they perceive learning to write code and work with scripts as a hurdle. But version 2 of the exercise shows that the hurdle is not as high as it might appear. Students need not master a programming language to get started: learning the syntax of a few basic commands is sufficient to begin working with scripts and take meaningful steps toward reproducibility.

Version 2 is identical to the non-reproducible version 1, except that instead of interactively typing commands or using menus, students write a script that includes all the commands needed to open the data file and generate the table and bar graph. As in version 1, students

write the report with a word processor and copy and paste the results from their monitor into the report.

Because the data file is opened by a command in the script (rather than by double-clicking), it is necessary to be explicit about where the data file is stored and which folder is designated as the working directory. The instructions for version 2 advise students to follow a very simple convention to ensure the software can find the data file:

- All the files for the exercise—the data file, the script, and the > report—should be stored in a single folder, which is referred to > as the project folder.
- Before executing the script, the user should designate the project > folder as the working directory for their software.

The instructions to version 2 also provide guidance on several best practices for writing scripts:

- **Headers.** Every script should begin with a header. Instructors may use their discretion to decide what information they ask students to include in the header for any particular script, but typically headers provide information such as the date, the name of the person writing the script, and a description of the purpose of the script. It is also useful to include a note in the header indicating to the user which folder should be designated as the working directory when the script is executed.
- **Setup.** It is usually convenient to start a script with commands that (i) declare the version of the software being used, (ii) install any other software or add-ons that will be necessary, (iii) clear memory, and (iv) specify any relevant settings for the software.
- **Open the data.** The data file should be opened by a command in the script (not by double-clicking). The command that reads the data must come before any commands that manipulate or analyze the data.
- **Comments.** Throughout the script, it is essential to write detailed and informative comments explaining the purpose of each command. These comments will be helpful to any interested reader who chooses to explore the documentation for a project. Moreover, they are valuable to the students themselves: unless they include good comments in their scripts, they may have trouble deciphering code they wrote only a few days ago.

As in version 1, students write the report with a word processor, and copy and paste the results from their monitor into the report. In version 2, however, the work they submit consists not just the report, but their entire project folder, containing the data file, their script, and the report.

The instructor should then be able to reproduce the table and bar graph simply by launching the software, setting the working directory to the project folder, and executing the script.

Version 3: Saving output. In versions 1 and 2, students copy and paste output from their monitor into the report, but their results are not preserved in any other way. In version 3, students write additional code in the script that saves the results in two output files: a text file containing the table, and a graphics file containing the bar graph. As in version 2, students store the data file, their script, and their report in a single project folder, which is again designated as the working directory. Because the project folder is designated as the working directory, that is where the two output files are saved when they are generated.

Saving the output files makes it possible to automate the process of inserting the results into the report. Instead of using a word processor and copying and pasting, students can write the report in a markup language (like Markdown or LaTeX), embedding links to the output files at appropriate points in the text.

The work students submit for version 3 again consists of the entire project folder, but in this case the project folder contains not only the data file, script, and report, but the two output files as well.

Version 4: The reproducibility trifecta: Folder hierarchy and relative directory paths. Version 3 involves a number of files of several different types, all of which are stored together in a single project folder. In version 4, students add some structure by creating several subfolders inside the project folder and distributing the various files among them. The organizational scheme in version 4 is very simple:

- The report is stored in the top level of the project folder.
- Three new folders are created in the top level of the project > folder: **Data**, **Scripts**, and **Output**.
  - The data file is saved in **Data**.
  - The script is saved in **Scripts**.
  - The output files are saved in **Output**.

For more complex projects, it is usually convenient build a more developed folder hierarchy, often including several levels of subfolders within the project folder. But the simple scheme used in version 4 is sufficient to introduce the key practices for achieving reproducibility given any folder structure adopted in a particular application.

When the files for a project are distributed in a hierarchy of folders within the project folder, the key to reproducibility lies in three practices that we refer to as the ***reproducibility trifecta***.

1. **Establish a well-defined folder hierarchy**.
   - All of the documentation for a project should be stored in a > single project folder.
   - The project folder should contain a hierarchy of subfolders in > which the various files are organized in some convenient and > sensible way.

- This structure should be established, and the hierarchy of > folders (all initially empty) should be built, before work > with the data begins.
- The folders should then be populated with the data, scripts, > and other files generated as work on the project > progresses.

1. **Be explicit about the working directory.**
   - For every script you write, choose one of your folders (either > the project folder or one of the folders inside the project > folder) to be designated as the working directory when the > script is executed.
   - We recommend the following convention: Always designate the > project folder as the working directory. When you, or an > independent investigator interested in your project, launch > the software to begin working with your scripts, they will > need to check whether the project folder is designated as the > working directory; if not, they will need to manually set the > working directory to the project folder. After that, there > should be no need to change the working directory again.[1]
   - In the header for each script, include a note that (i) indicates > which folder you have chosen as the working directory, > and (ii) reminds the user to be sure that the chosen folder is > in fact designated as the working directory before executing > the script.
2. **Use relative directory paths**.
   - A relative directory path is a path through the folders on the > computer you are using that begins in whichever folder has > been designated as the working directory and leads to a target > folder (from which, for example, you wish to open an existing > file, or in which you wish to save a newly created file).
   - In your scripts, whenever you write a command in which you need > to specify the location of a particular folder, you should do > so using a relative directory path. You should not specify a > directory path that begins in a particular folder on a > particular computer (such as the C: drive on your computer).

The three elements of the reproducibility trifecta are interrelated: when you write a relative directory path, you must know what folder is designated as the working directory (that is where the relative directory path starts), and you must know the structure of the folder hierarchy (since the relative directory path must specify how to navigate through that hierarchy to the target folder). Beginning students need guidance about how to properly synchronize their folder and file structure, the choice of the working directory, and the relative directory paths they write in their scripts. But by introducing these concepts in simple setting, version 4 of the exercise makes it easy for them to grasp how the pieces fit together.

[1]The views expressed herein are those of the author and do not necessarily represent the views of the Federal Reserve Bank of Kansas City or the Federal Reserve System.

7.3 Conclusion

7.3.1 Standards of reproducibility

The reproducibility trifecta makes it possible to achieve two important standards of reproducibility, which we refer to as (i) ***(almost) automated reproducibility*** and (ii) ***portable reproducibility.***

Automated reproducibility means that, once a user has copied the project folder onto their own computer, the computations that generate and save the results can be reproduced just by running the scripts, with no need to do anything by hand (such as editing directory paths in scripts or moving files from one folder to another).

Synchronizing the folder hierarchy, working directory, and relative directory paths according the principles of the reproducibility trifecta ensures that automated reproduction is possible—almost. Before the scripts can be executed, there is one task the user needs to complete by hand, namely setting the working directory to whatever folder has been designated by the author. Hence the qualifier "almost" before the term "automated reproducibility".

The standard of portable reproducibility is that any user should be able to perform an (almost) automated reproduction of someone else's project on their own computer. Provided they have the necessary software installed, they should be able to copy the project folder and all its contents onto their computer, and then (after setting the working directory as necessary) run the scripts that reproduce the results.

The key to achieving portable reproducibility is that all directory paths specified in the scripts must begin and end in folders on the user's computer. Because the reproducibility trifecta specifies that the working directory should be set to the project folder (or one of its subfolders), and that folder locations should be given by relative directory paths beginning in the working directory, this condition is satisfied the moment a user copies the project folder onto their own computer.

(Almost) automated reproducibility and portability are state of the art standards for professional social science research; they are among the properties that leading conventions such as the AEA Data Editor's guidelines and the DIME Manual are intended to achieve. The four versions of the exercise we have presented show that these professional standards can be introduced to students in introductory level classes via a sequence of modest, feasible innovations.

7.3.2 Bells and whistles

To make the fundamental principles and practices as transparent as possible, we have presented a simple exercise that excludes a number of important elements of documentation. But once students have a foundation in the fundamentals, it is easy to introduce additional elements

such as a read-me file, more complex directory structures, data citations, a master script, log files, and a data appendix, to name a few.

Instructors looking for a more substantial project that introduces many of these peripherals, but is still accessible to students in introductory courses, might consider the Project TIER exercise titled "Animal House in Alcohol-Free Dorms?". When students move beyond structured exercises and begin research projects of their own, they may benefit from the TIER Protocol, which gives detailed guidance about the components of a comprehensive reproduction package. Examples of all the components of the documentation described in the TIER Protocol can be found in an accompanying demo project.

TABLE 1: PROPERTIES OF THE FOUR VERSIONS OF THE EXERCISE

Version	**Elements of reproducibility introduced**	**Work submitted by students**	**How the report is written**
Version 1: Interactive	None	A report (a *.pdf* document)	Text composed with a word processor; table and figure inserted by copying and pasting
Version 2: Scripted computations	• Writing all commands in an executable script • Keeping all files in a project folder • Designating the project folder as the working directory	A project folder, containing: • a report (a *.pdf* document) • the data > file • a script	Text composed with a word processor; table and figure inserted by copying and pasting
Version 3: Saving output	• All elements of version 2 • Writing additional commands in the script that save output files to the working directory	A project folder, containing: • a report (a *.pdf* document) • the data > file • a script • two output > files	Text composed with a word processor; table and figure inserted by copying and pasting or Text composed in a markup language; table and figure imported from output files

Version	Elements of reproducibility introduced	Work submitted by students	How the report is written
Version 4: A ssembling an (almost) a utomated, portable rep roduction package	• The reproducibility trifecta: – Establishing > a well-defined `> folder > hierarchy > within the > project > folder` – Designating > the > project > folder as > the > working > directory – In scripts, > use > relative > directory > paths to > specify > locations > of > specific > folders	A project folder, containing: • a report (a > *.pdf* > document) • a **Data** > subfolder, > containing > the data > file • a **Scripts** `> subfolder, > containing > a script` • an > **Output** > subfolder, > containing > two output > files	Text composed with a word processor; table and figure inserted by copying and pasting or Text composed in a markup language; table and figure imported from output files

Part IV

Session 4: Reproducibility and confidential or proprietary data: can it be done?

What happens to reproducibility when data are confidential or proprietary? Many journals can only ask that detailed access procedures be provided in a ReadMe file, but what mechanisms could be used to conduct computational reproducibility checks on such data? Should authors temporarily share their data with the journal for the purposes of reproducibility verification, even if they are not part of the public data replication package? Is it feasible to use a network of "insiders" to run code provided as part of a data replication package to assess reproducibility? Could a "certified run" be used?

8 Reproducibility with confidential data: The experience of BPLIM

Access to confidential data is often seen as a major obstacle to reproducibility. The Banco de Portugal microdata laboratory (BPLIM) has been providing access to confidential data for research purposes since 2016. Based on our experience, we argue that our data access model is an opportunity to demand that researchers integrate good reproducibility practices into their research process. Our goal is to automate the production of a reproducibility package as part of the researcher's process of analyzing confidential data. We hope that this approach can overcome the limitations of working with confidential data and help certify reproducibility with confidential data. The approach we propose can also be adopted by researchers using non-confidential datasets.

8.1 Background

The Banco de Portugal Microdata Laboratory (BPLIM) was established in 2016 with the primary goal of promoting external research on the Portuguese economy by making available data sets collected and maintained by Banco de Portugal (BdP). By making this information available to researchers from around the world, BdP aims to support the development of evidence-based policies and insights that can benefit the Portuguese economy and society. However, given that some of these data sets contain highly sensitive information BPLIM had to implement a data access solution that preserved the confidentiality of the data.

The common approach by other Research Data Centers that make confidential data available for research involves the provision of on-site access to accredited external researchers in a secure computing environment. However, in the case of BPLIM this approach was deemed undesirable for two reasons. First, because it would limit access to a handful of researchers who were able to come to the Bank's premises. Second, because there were still concerns that a breach of confidentiality might occur if individuals from outside the bank could gain access to original data sets that contained confidential information.

After an internal debate at the Bank it was decided that the solution to be adopted by BPLIM had to be based on the following principles:

- access free of charge and only for scientific purposes;

- all data should be analyzed on the servers of the Bank;

- external researchers were granted remote access to the server;

- confidential datasets placed on the server had to always be perturbed/masked;

- researchers could always ask BPLIM staff to run their scripts on the original data.

The general workflow defined for data access at BPLIM was the following. After a research project is approved and the external researchers are accredited an account is opened on the BPLIM external server. External researchers gain access to a computing environment that does not allow users to transfer files to and from the server. They have access to a restricted area where standard software such as *Stata*, *R*, *Julia* and *Python* are available. Since there is no connection to the internet, installation of specific packages has to be requested from the staff. The datasets for the project are placed in a read-only folder. For the confidential datasets what is placed in the account of the researcher are perturbed versions of the data (noise is added to the original data). The researcher implements all scripts based on the data he/she has available and produces the (non-valid) outputs required for the project. Once researchers complete this task, they can ask BPLIM staff to rerun their scripts, this time using the original confidential data. For this process to be successful BPLIM staff must first run the scripts using the same data as the researcher to verify that the scripts written by the researchers reproduce exactly the outputs (typically graphs and/or tables). This process is done in a different server (BPLIM internal server). Only upon completion of this first step can BPLIM staff modify the scripts, this time to read the original data and regenerate the intended outputs . These outputs are then subject to standard output control checks for confidential data and delivered to the researcher.

8.2 Main Thoughts

Over time we have come to realize that this somewhat cumbersome process of running the code thrice, first by the researcher on the perturbed data, second by BPLIM staff again on the perturbed data, and finally, on the original data was in fact an exercise on reproducibility. Even though the reproducibility check was on the perturbed data it was already a very good assurance that a reproducibility check would hold on the original data.

We realized that a great deal of our work involved reproducing the results obtained by the researchers with the perturbed data and that led us to look at ways to improve our workflow. It became obvious that the process could be streamlined and would be more efficient if researchers adhered to the best practices on reproducibility. Hence, as part of our strategy, we have decided to raise awareness of our researchers to the need of implementing good practices on reproducible research. We have been doing this by several means. For example, we have held practical workshops which are designed to enhance the skills of our researchers. For these workshops

we invite leading experts to present best practices and recent developments on data analysis. We also provide direct advice to the researchers, prepare templates and documentation, and make available tools that facilitate the analysis of our datasets (particularly for more complex tasks such as building a panel or calculation of specific variables).

On the other end, it was also obvious that there was margin for improvement on our work sequence. One possible improvement was the assurance that the computing environment used by the researcher on BPLIM's external server was identical to that used by BPLIM staff when reproducing the code. Thus, for the case of researchers that work with open-source software, we have been incentivizing researchers to work with Singularity containers. This facilitates our work because we are sure that our reproducibility check is implemented in the same self-contained environment that was used by the researcher. Researchers that use Stata can resort to containers but in that case, it is easier to control the environment because we install all packages on a folder that is specific to each project and have developed tools that facilitate comparison of the Stata ado files across environments. (all tools are publicly available and can be found at https://github.com/BPLIM/Tools/tree/master/ados/General.)

More recently, we have worked on shifting the burden of the reproducibility check to the researcher itself. We are developing an application that we are presently testing with a select number of researchers. The application is targeted mainly at researchers that use BPLIM's (perturbed) confidential datasets but we hope to eventually convince all other users to take advantage of it. To illustrate, we provide a screenshot of the application:

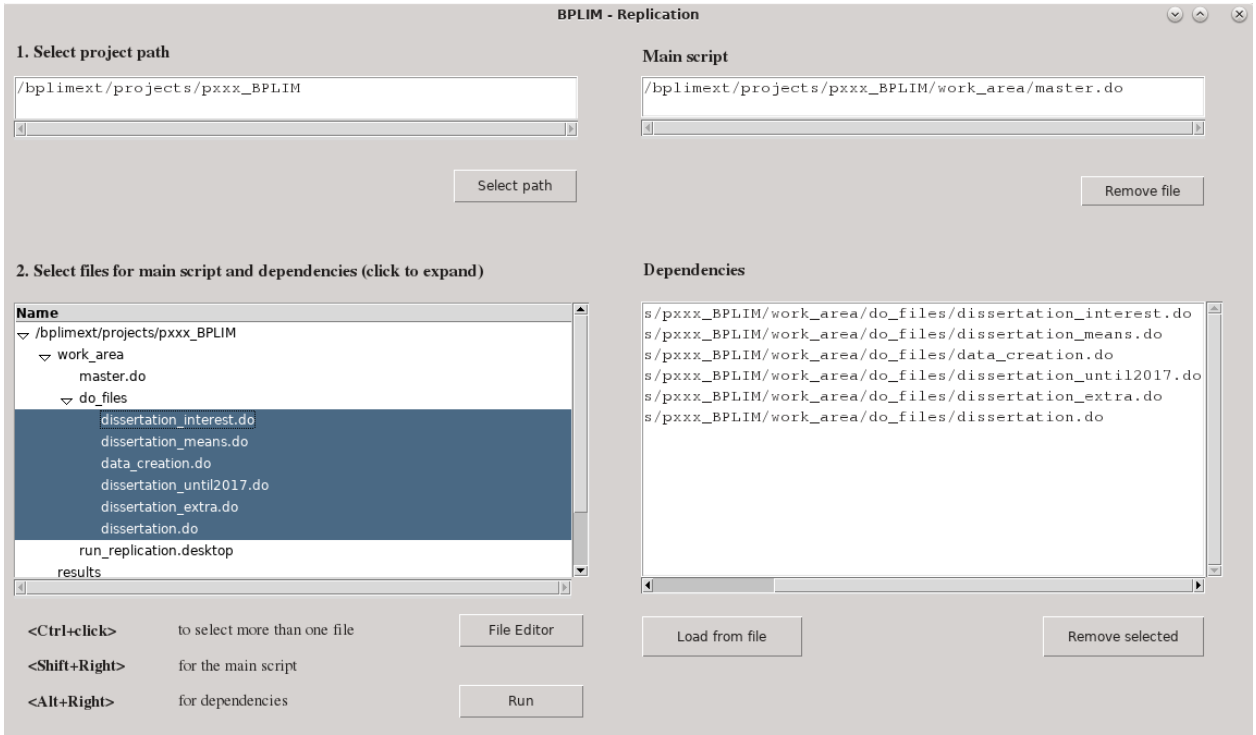

Figure - BPLIM Replication tool – selecting input files

Before requesting a replication from BPLIM using the original data the researcher must first validate his/her code by successfully submitting the scripts through BPLIM's Replication application. The process involves selecting the main script as well as all the required dependencies created by the researcher. The folder structure used by the researcher is replicated and the BPLIM datasets have to be read from the (read-only) data folder. All intermediary output files must be created during the replication (it is however possible to start from an intermediary output file. In that case, the intermediary file must have been validated in a prior run. BPLIM will then copy the file to the (read-only) data folder.).

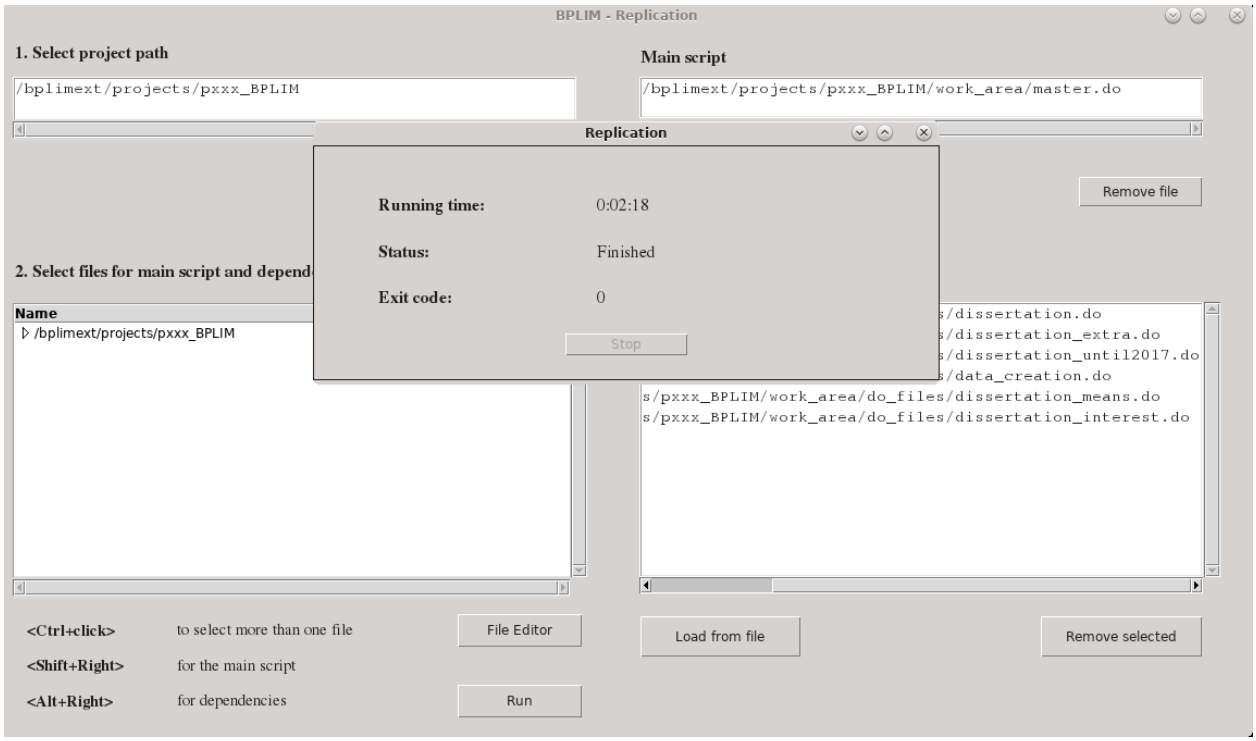

Figure - The BPLIM Replication tool - finished task

The researcher then uses the application to run the code (see Figure 2 for a completed run). The code must run from top-to-bottom and produce no errors. If the run is successful, then implementation on the original data requires only that BPLIM staff changes the relative paths to the data folder and rerun the code. A side advantage of this process is that it automatically produces a replication package for the researcher. Stored in the folder are all replication scripts, the output files, as well as two additional files, one fully characterizing the software environment, and another *json* file containing a listing of all scripts used in the replication. If we add the definition file used to produce the container (or a listing of all packages and respective versions) then we have a full replication package (except for the data).

8.3 Conclusion

Our goal at BPLIM is to make sure that all researchers create their replication packages as an integrated part of their research process. The fact that we are the ones running the code on the original data should be seen as an opportunity to request that researchers make reproducible code while implementing their research.

In the ideal situation that we envision, researchers download a template for the definition file of the Singularity container, customize that template by adding and testing the packages they need, and share with us the definition file. Based on that definition file we build the container for the project and make it available on our external server. The researcher then uses the container to implement the analysis and when he/she is ready to obtain results based on the original data he must first validate the scripts using our application. The researcher can go through this process multiple times and each time a replication package will be created. Apart from the data, all files can be publicly shared, and the replication package created at BPLIM should be easily customized for submission to any data editor.

We are already implementing this solution for all new projects that use confidential data. However, we hope that over time we can convince all researchers at BPLIM to work with Singularity containers and go through the same validation steps that are needed for projects that deal with confidential data.

Projects that are implemented in BPLIM have additional advantages when it comes to reproducibility. First, because all BPLIM datasets are versioned and registered with a Digital Object Identifier (DOI) we are sure that the original data is exactly identified. Second, the computing environment is stable, and the software packages used by researchers are specific to each project. Finally, if external researchers have used BPLIM confidential data then there is an assurance that their code was reproduced at some point.

Ultimately, whether the scientific work is reproducible depends on the researchers. But we hope that integrating reproducibility into the research process with confidential data, provides a way to alleviate the inconvenience of third parties that cannot access the original data and want to verify reproducibility of the results.

Part V

Session 5: Disciplinary support: why is reproducibility not uniformly required across disciplines?

Why do learned societies decide (or not) to implement data (and code) availability policies? What influences the level of enforcement, and the choice of "enforcer" (data editor, administrative staff, referees)? What are reasons NOT to require data sharing or code sharing?

9 Reproducibility in Economics: Status and Update

This article provides a status update on reproducibility in the Economics profession. This includes a particular emphasis on quantitative research. And a chronology of the role the American Economic Association has taken in this progress.

9.1 Background

I am an economist and study poverty and inequality and the role of the social safety net in family and child wellbeing. I want to start by positioning myself and my background a little bit, as it relates to the conversation today. Over the past decade or more, I have engaged in a variety of professional service activities that are linked to reproducibility and open science. First, I am currently a member of the Committee on National Statistics at the National Academies of Science, Engineering and Medicine, where we discuss a range of issues around data, replication and reproducibility. I was a member of the Commission on Evidence Based Policymaking, and there we made recommendations around administrative data and linkages balancing access and privacy. I mention these experiences because in Economics there has been tremendous growth in the use of confidential, proprietary or administrative data and there are significant challenges in open science goals around these data. Second, I spent a decade as a Co-Editor of journals under the American Economic Association, first at AEJ: Economic Policy and subsequently at the American Economic Review, the flagship journal of the association. While a Co-Editor at the American Economic Review, and also a member of the AEA Executive Committee, I Co-Chaired the search committee for an AEA Data Editor. This is the position that Lars Vilbuber holds.

9.2 Main Thoughts

The theme of this session is how differences in support for reproducibility exist across the social sciences. Hence I organize my comments to focus on the status of reproducibility in Economics and what might make it different from other social sciences.

9.2.1 I. The (possibly unique) role of the American Economic Association

The American Economic Association (AEA) has a lot of "market power" because it controls a large share of the high quality journal landscape. The AEA journals include the American Economic Review, the American Economic Review: Insights, and American Economic Journals (there are four of these), [as well as the Journal of Economic Literature and the Journal of Economic Perspectives]. The AEA started with the American Economic Review (AER) and is one of the "top 5" general interest publications in Economics. The AER became so clogged with quality submissions and accepted papers (with 2 year waits between acceptance and publication) that the AEA expanded the number of issues that they produced per year. This led the AER to publish a disproportionate share of the articles in Economics top 5 journals.

Following the success of the AEA journals and the strong budget situation at the association, the AEA decided to expand and add additional journals to its portfolio. We added AER: Insights, featuring short format articles with quick turnaround, in an attempt to compete with publications in Science and PNAS. We also added four top field journals to try to occupy the space just below the top 5, perhaps wrestling some of the market from for profit journals such as those published by Elsevier. This became the four American Economic Journals (AEJ) - AEJ: Applied Economics, AEJ: Economic Policy, AEJ: Microeconomics, and AEJ: Macroeconomics.

So if you put this all together, there's a lot of the journal space (and highly ranked journal space) that the American Economic Association runs.

With this "market power" and over period of time the association made a series of moves in the area of open science. This approach came about through both top down and bottom up mechanisms. There was definitely interest among the AEA Executive Committee including the lead editors of the American Economic Review, first Penny Goldberg followed by Esther Duflo. But there was also a perspective that percolated up from the membership. This may have come in part in response to the replication crises in other disciplines. Additionally, many of us engaged in training graduate students desired data for students to get practice and skills.

Overall, I think the AEA views itself as having the potential to take actions that set standards that are adopted more broadly in the profession. (I hope the association takes the same approach to dealing with sexual harassment in Economics.)

9.2.2 II. AEA Actions

So what did the AEA do?

They set up the AEA RCT Registry. As stated on the website: "Randomized Controlled Trials (RCTs) are widely used in various fields of economics and other social sciences. As they become more numerous, a central registry on which trials are on-going or complete (or withdrawn) becomes important for various reasons: as a source of results for meta-analysis; as

a one-stop resource to find out about available survey instruments and data. Because existing registries are not well suited to the need for social sciences, in April 2012, the AEA executive committee decided to establish such a registry for economics and other social sciences."

Second, and more pertinent to this conversation, the American Economic Review adopted a requirement to post data and code for all accepted papers. It initially was very much on the honor system but it was felt that this was a good first step. There were some staff who worked on confirming that files were uploaded, but not much beyond that.

After some time, there was an evaluation of how this was going. The answer was, not so well. Often times, the data and code would be incomplete. Documentation would be insufficient. It would be difficult for reuse, and it wasn't replicable. There was a recognition that you need skilled leadership and staff for this to work.

This led to the next stage of open science at the AEA. We recognized that we had the budget to do more. The AEA compensates all journal editors a pretty fair wage for the work. And the view was perhaps we should take the next step to hire someone who would work with the journal editors as well as with the AEA Executive Committee to create a more robust system for reproducibility. There was a discussion about whether or not we wanted this individual to be a practicing academic who has skills in this area, or whether we just needed to build up more staffing. We reached out to folks in other disciplines to find out what they were doing. We liked the model at the American Statistical Association who had appointed an academic researcher who also had skills in open science.

We had a search for this new position which we called the AEA Data Editor. From the job ad, the data editor would "Collaborate with journal editors and executive committee; design and oversee implementation of strategy for archiving & curating data, promoting reproducible research." In short, we needed an architect of the new system as well as a manager and implementer. We were lucky to attract Lars Vilhuber to the position. He is a practicing academic researcher who also has the skills in open science, and could work at the frontier.

9.3 Conclusion

I will conclude by providing some thoughts about where the gaps are and where we need to continue to make progress.

First, we need to continue to revisit the staffing of the Data Editor's operation. Are the resources sufficient for the needed work?

Second, we need to devise approaches to deal with confidential data. It is increasingly common for AEA journal papers to use proprietary or confidential data. Examples include government administrative data, data from firms, and so on. Researchers typically do not have the ability to post this data. Therefore, replication and reproducibility has to get over a very large hurdle with these data.

Third, we need to build on the success of pre-registration of RCTs. For example, should the journal require pre-registration? Should this be expanded to pre-registration of quasi experimental papers? And relatedly, should the AEA create an RCT Editor to check compliance with pre-registration plan?

Finally, we need do more analysis and landscaping to investigate to where the gaps are in the Economics discipline more broadly, and what role the AEA can take to continue to move us towards further progress in open science.

9.4

10 Crisis? What Crisis?

Sociology has been slow to adopt the standards and practices of scientific transparency. Existing efforts stem from the voluntary efforts of individual researchers or journals rather than from the discipline's professional association. The absence of organized, top-down initiatives to nudge the discipline toward open science reflects not only the usual problems of organizational inertia and resource constraints, but also sociology's unusually high intradisciplinary fragmentation. This fragmentation takes many forms, not the least of which is disagreement among sociologists over both the desirability and the feasibility of data and code sharing, preregistration, and reproducibility standards. Although prospects for large-scale, open science initiatives in the field seem slim, more modest forms of transparency will continue to diffuse among networks of scholars, especially those using quantitative methods.

10.1 Background

Sociology has lagged behind economics, political science, and psychology in its recognition of a replication crisis and efforts to move toward scientific transparency. Although individual researchers have voluntarily adopted some of the "best practices" of transparency, they are still in the minority within the field. Similarly, a handful of sociology journals have tried to nudge the field toward transparency, most often by requiring authors to provide replication packages with code and data for quantitative articles or by explicitly welcoming replication studies. Very few have mandated preregistration of studies, and to my knowledge no journals have a process to verify replication packages.

Notably, these efforts to introduce transparency into sociological research have all been "bottom up." The American Sociological Association, the discipline's primary professional and scholarly association and the natural locus of top-down initiatives to improve disciplinary practice, has been largely silent on scientific transparency. The flagship journal in the field, which is published by the ASA, recommends but does not mandate replication packages, and the ASA's statement of professional ethics does not mention scientific transparency.

10.2 Main Thoughts

Why has adoption of transparency standards in sociology been so slow and piecemeal? Is the absence of "top-down" leadership from the ASA merely the standard story of the challenge of coordination, organizational inertia and resource constraints? Or is there something about sociology as a field that has slowed its adoption of scientific transparency? I argue that it's both: fragmentation within the discipline contributes to organizational inertia, and the combination makes top-down leadership on scientific transparency, or for that matter any issue related to the day-to-day practice of sociology, rather unlikely.

Some elected members of the ASA have attempted to leverage the authority and power of the organization over its journals to implement scientific transparency standards. As Philip Cohen documents in his blog, he brought a proposal to the Publications Committee (on which he served as an elected member) to adopt the Open Science Badge system. This proposal failed in a vote. Sixteen months and two ad hoc committees later, an alternative proposal passed the committee: authors would declare in a footnote whether their data and code were available online, and if not why not. This proposal was referred to Council, where it was rejected and, after another 4 months, sent back down to the Publication Committee with comments. By this time, the members of the Publication Committee who had pushed hardest for transparency had either given up or rotated off of the committee.

Without insider information, an autopsy of the two proposals' death by committee is bound to be inconclusive. Perhaps committee and Council members thought the benefits of implementing such modest, unenforceable, and honor-system policies were not worth the effort. Perhaps they were concerned that any open science requirements, even weak ones, would burden already over-burdened editors and reviewers or slow an already slow editorial and publication process. Perhaps members of the permanent staff were concerned that each step toward open science would irrevocably alter the organization's contract with Sage press, from which the ASA receives a large share of its annual revenues and operating budget. Perhaps committee members were worried that footnotes or badges would stigmatize papers whose authors declared they were exempt from sharing data and code. Or, perhaps some of the members of the committees objected in principle to scientific transparency.

The latter two speculations point to the second relevant feature of sociology: it is a highly fragmented discipline, with a weaker core and more internal heterogeneity than economics, political science, or psychology. (This statement is borne out by bibliometric data, which shows sociology is more likely than other social sciences to cite outside the field.)

Fragmentation within the field tends to be fractal. At the most basic level, sociologists disagree over the goals and epistemology of sociology. Should it be a science that strives, however imperfectly, to identify objective truth or generalized social processes? Is it a normative project, in which the main goal is to reveal not what is but what should be? Or is it an interpretivist project more akin to the humanities than to economics, political science, or psychology?

Even among scholars who fall into the "sociology is a science" camp, fragmentation can be observed in the wide range of methods used within the field. Although the majority of research applies quantitative methods to secondary (survey, digital trace) data, research using qualitative, mixed, ethnographic, experimental, computational, and historical-archival methods is also common. Practitioners of these disparate methods have different ideas about what scientific transparency means, whether particular practices (e.g., preregistration, the dissemination of replication packages) are feasible for their type of work, and even whether transparency and reproducibility are appropriate goals. At the risk of oversimplifying, scholars who mainly use qualitative methods have been more reluctant to embrace scientific transparency than those who mainly use quantitative methods.

Among quantitative scholars, objections to open science are typically over practicalities rather than principle: for example, how to resolve the tension between scientific transparency and the legal, ethical, or normative constraints of making restricted access or proprietary data publicly available. In most cases, these concerns are not insurmountable. For example, journals could simply exempt research that uses proprietary or restricted access data from replication package requirements, which is precisely the approach taken by the handful of journals with replication package policies. Alternatively, journals or researchers could work with third parties to verify results, although this solution quickly runs up against resource constraints: very few editors or reviewers in sociology are paid for their labor even "in kind," relatively few sociologists enjoy institutional support for verification of their replication packages, and even fewer journals have the resources to pay for staff or students for this task. The greater challenge for open science among quantitative scholars is the general devaluation of replication studies, which are rare even at the journals (e.g., *Sociological Science)* that have explicitly embraced them as a valid and valued form of research.

Among qualitative scholars, concerns over data-sharing and other scientific transparency standards take a slightly different form than among quantitative researchers. Some of these concerns focus on the consequences of data sharing for qualitative research practice. Will data sharing undermine the trust that develops between researchers and subjects – trust on which qualitative research depends? Will subjects respond to data sharing by becoming more reluctant to participate in research, particularly on the sensitive topics that sociologists often study? Will research subjects change their behavior if they know that others in their community will have access to field notes or interview transcripts?

Other concerns about transparency in qualitative research focus on intellectual property and incentives. A tacit agreement in sociology is that researchers who collect their own data will have sole access to them as long as they wish to keep publishing off of the data, typically a matter of years rather than months. This norm seems to fly in the face of scientific practice, but in a world of extremely limited and declining federal or state funding for data collection, monopoly ownership is one way to offset the greater risks and potential career costs to individual researchers of collecting new data. Put bluntly, if qualitative researchers are required to release their field notes, transcripts or videos, or other raw data products, it's likely that fewer researchers will conduct qualitative studies, to the detriment of the field.

A final concern is that reliability, replicability, and reproducibility are not appropriate yardsticks against which to measure qualitative research. The process of producing and analyzing qualitative data is often iterative rather than linear, it is inherently intersubjective, and it relies on non-verbal cues from subjects and the embedded experiences of researchers that cannot be captured in transcripts, field notes, or other data products (Tsai et al 2016). Moreover, qualitative research is often not designed to generalize or to be replicable, but rather to generate new insights into social processes that can guide quantitative data collection efforts. In this context, transparency around the production of data or around the data themselves (e.g., through data sharing) may make little sense (Tsai et al 2016).

These concerns could, of course, be circumvented by exempting qualitative research from scientific transparency standards, or at least modifying standards to be sensitive to qualitative research (see Tsai et al for suggestions). Notably, however, some qualitative researchers whose work would likely be exempt from scientific standards still object to such standards. In some cases, this objection reflects disagreement over whether sociology is or should strive to be a science. In other cases, it seems to be rooted in the fear that adopting scientific standards, even with broad exemptions for qualitative research, would marginalize and stigmatize qualitative research within the field.

In this context, it is not surprising that the ASA, which represents the interests of all sociologists, would be reluctant to engage in top-down initiatives to mandate scientific transparency. More cynically, in the wake of rapidly declining membership, the ASA *needs* to be a big tent under which all sociologists can find shelter. The organizational inertia generated by governance structures that allow the death of transparency initiatives by committee is this reinforced by intra-disciplinary dynamics, and the desire to please (or at least not alienate) all segments of the field. The combined forces of inertia and disciplinary fragmentation forestall top-down initiatives for scientific transparency.

10.3 Conclusion

The conditions that slow sociology's adoption of transparency standards are not likely to dissipate any time soon. The most likely impetus for change is the steady diffusion of scientific transparency in quantitative work, as more individual scholars and journals adopt transparency practices voluntarily, younger scholars are trained in these practices, and cross-disciplinary collaborations create ties over which "best practices" in other fields can infiltrate sociological research networks.

A top-down approach led by the discipline's professional association seems less likely. Sociology is characterized by a live and let live ethos that allows scholars with very different perspectives on the discipline and on ways of doing sociology to coexist. The ASA embraces this ethos, making it difficult for elected representatives to nudge the discipline toward scientific transparency, or indeed any change in the day-to-day practice of research, about which various segments of the discipline hold very different views.

10.4 References

Cohen, Philip. 2021 (March 8). "The American Sociological Association is Collapsing and Its Organization a Perpetual Stagnation Machine." https://familyinequality.wordpress.com/2021/03/28/the-american-sociological-association-is-collapsing-and-its-organization-is-a-perpetual-stagnation-machine/

Tsai, Alexander C., Brandon A. Kohrt, Lynn T. Matthews, Theresa S. Betancourt, Jooyoung K. Lee, Andrew V. Papachristos, Sheri D. Weiser, and Shari L. Dworkin. 2016. "Promises and Pitfalls of Data Sharing in Qualitative Research." *Social Science & Medicine* 169(2016): 191-198.

Part VI

Session 6: Institutional support: How do journal reproducibility verification services work?

When journals conduct active verification of replication packages, including accessing data and running code, how does that work? Can journals with limited resources still assess reproducibility? What depth of verification is optimal? Do journals provide a clear indication of whether an article was successfully reproduced?

11 The role of third-party verification in research reproducibility

Research reproducibility is defined as obtaining similar results using the same data and code as the original study. In practice, to check research reproducibility, third-party verification constitutes a useful complement to the work done by journals' internal teams. Third-party verification services can also be used by individual researchers seeking a presubmission reproducibility certification to signal the reproducible nature of their research. Using the example of the cascad certification agency, which I co-founded in 2019 with Christophe Hurlin, I discuss the functioning, utility, comparative advantages, and challenges of third-party verification services. I thank Olivier Akmansoy, Jean-Edouard Colliard, Christophe Hurlin, Jacques Olivier, and Lars Vilhuber for their comments and support.

11.1 Background

The quest for a reproducible science requires *three preconditions* to be met, and I believe all three are met today in the field of economics.

The first precondition is to have a good understanding of *what research reproducibility is*. Collectively, the survey of Christensen and Miguel (2018), the report of the National Academies of Sciences, Engineering, and Medicine (2019), and the work of the American Economic Association (Vilhuber, 2021) brought some much-needed clarity to the different concepts used to describe reanalyses in economics. Currently, the consensus is increasingly favoring the notion that an empirical result is deemed reproducible if it can be recreated by running the original code of the authors on the original data. This type of tests contrasts from other forms of reanalyses such as replications, robustness analyses or extensions (Vilhuber, 2020).

The second precondition is *to recognize that the current level of reproducibility is low*. Indeed, there is significant evidence that the success rate of reproducibility studies in economics and finance remains surprisingly low, mainly due to missing code/data/information and numerous bugs (Chang and Li, 2017; Gertler et al., 2018; Herbert et al., 2021; Pérignon et al., 2023). Depending on the studies, the success rate ranges between 14 and 52%.

The third precondition is *to acknowledge that this lack of reproducibility is problematic and that we need to act to improve the situation*. Following early decisions by the American Economic Association (Duflo and Hoynes, 2018) and the Royal Economic Society, most of the other leading scientific associations and academic journals are now considering strengthening their code and data availability policies.

Now that these three preconditions have been met, the most important challenge is *implementation*. As of today, reproducibility verification is conducted by dedicated verification teams working for some academic journals or associations or by third parties (e.g., *cascad* and the *Odum Institute*). Either internal or external to a journal, the verifier checks whether the submitted material complies with a set of guidelines and attempts to regenerate all the results from the code and data provided by the authors. While verifications at journals are typically conducted once the manuscript has been conditionally accepted, third-party verifications can be made at any time in the life of a paper. After successful reproductions, third-party verification services typically award reproducibility badges or certificates, which can be added to the manuscript when submitting to an academic journal.

11.2 The advantages of an early third-party reproducibility verification

From the viewpoint of the researcher, I see three main reasons to conduct an early third-party reproducibility verification.

The first one is to *detect as early as possible mistakes or inconsistencies* in the analysis. Indeed, when preparing the materials required to request a verification, the authors regularly identify typos and mistakes and they have the opportunity to correct them at no cost. Differently, when such mistakes are discovered later in the process, and especially after publication, the research community, and in particular, journal editors, will have to decide whether this is an honest mistake or plain misconduct. In the latter case, the stigma in terms of reputation can be very large.

The second reason concerns the *cost* of conducting verification, mainly in terms of time for the researchers themselves. Today, when they target top journals in economics, researchers know that (1) the cost will be faced with probability one as most journals have systematic verification in place and that (2) the cost increases significantly with time, as more datasets, code versions, and forking paths are added to the analysis. While optimal timing will also reflect time preferences, waiting until the paper acceptance to start thinking about reproducibility is unlikely to be an optimal strategy.

The third reason to conduct an early reproducibility verification is to *build trust*, and in particular among coauthors. Indeed, most academic papers have multiple authors, and the latter tend to specialize in exploiting their comparative advantages. Furthermore, some specialized

coauthors may not have the time, nor the skills, to monitor and review tasks that are not under their operational control. In this case, a third-party verification provides some reassurance for all the parties involved. However, it is important to acknowledge that a pre-publication reproducibility verification is not an "all-risks insurance". Indeed, there are many aspects of the research that a third-party verifier does not check, such as the correspondence between the claims and equations in the paper and the content of the code, the presence of typos in the code, or whether the authors engaged in data manipulation or fabrication. These problems can subsequently be identified by other researchers by reviewing the original code and datasets (see the www.datacolada.org website for such forensic investigations).

11.3 The cascad certification agency

Christophe Hurlin and I founded *cascad* (www.cascad.tech) in 2019 with a double objective: (i) to help individual researchers signal the reproducible nature of their research by granting reproducibility certificates and (ii) to help other scientific actors (e.g., academic journals, universities, funding agencies, scientific consortia, data providers) verify the reproducibility of the research they publish, fund, or contribute to the production of.

In terms of organization, *cascad* is a nonprofit research laboratory funded by the French National Center for Scientific Research (CNRS) along with several universities and research institutions. While it operates within France, *cascad* collaborates with researchers, academic journals, and other users from all around the world. Its workforce comprises full-time reproducibility engineers, part-time graduate students, and a group of faculty oversees the operations and promotes the services offered.

The establishment of *cascad* was driven by two firm beliefs. First, we believe that for science to be taken seriously, there needs to be a serious commitment to reproducibility. Put it simply, if you want the chain of science to be strong and useful to society, you do not want reproducibility to be its weakest link. Second, we hold the conviction that merely making code and data publicly accessible does not fully address the reproducibility challenge. We have come to this resolute belief after engaging in several years of management at RunMyCode (www.runmycode.org), a repository for code and data used by various economics and management journals. In this capacity, we often saw researchers failing to share all the essential components (code, data, explanations) required to regenerate their results. This was frequently due to hurdles such as copyright issues, non-disclosure agreements (NDAs), or concerns related to data privacy. Moreover, even when all components were available, other researchers frequently struggled to execute them, and occasionally failed entirely (for consistent evidence, see Chang and Li, 2017, Gertler et al. 2018, Trisovic, 2022, and Pérignon et al. 2023).

We realized that a third party could be useful in this context. First, when all the required resources can be shared, a third party can run and regenerate all the results before uploading the code and data on an online repository. Second, when some data cannot be shared, the

third party can ask permission to access such data, to be able to run the code and reproduce the results (Pérignon et al. 2019). Finally, third-party verifiers can also be useful to academic journals (i) when the third party has permanent access to some restricted data, (ii) when it owns a license of, or expertise in, a software that the journals do not have, or (iii) when the journals do not have enough staff or computing power to verify all the newly accepted papers.

11.4 Examples of collaborations

Collaborations with economics journals: Since 2019, *cascad* has provided verification reports to the data editors of the American Economic Association and the Royal Economic Society. Such verifications concern conditionally accepted articles in one of the eleven journals managed by these two associations (e.g. *American Economic Review*, *American Economic Journal: Macroeconomics*, *Economics Journal*). To date, around 60 verifications have been conducted by *cascad* for these journals.

Collaboration with a restricted data access center: Since 2020, the *cascad* agency has partnered with the Centre d'Accès Sécurisé aux Données (CASD), a French public research infrastructure that enables researchers to access granular, individual data from the French Institute of Statistics and Economic Studies (INSEE) and from various French public administrations and ministries. In total, CASD hosts data from 378 sources and offers a data provider service to 742 user institutions. The CASD also gives access to restricted access data from the Banque de France, as well as individual health data, and environmental data. This example allows us to illustrate the economy of scale argument introduced earlier. Indeed, Colliard et al. (2023) found 134 articles on Google Scholar using CASD data, published in 91 different academic journals. To verify the reproducibility of all these articles, each journal would have had to go through a lengthy accreditation process to access the same original data. Instead, *cascad* offers a single point of entry to all academic journals seeking a reproducibility check for articles using restricted data accessed through CASD.

Collaboration with a scientific consortium: In 2021, *cascad* was tasked with assessing the reproducibility of the empirical results of 168 international research teams, gathered from more than 200 universities, who were participating in the Fincap project (Menkveld et al., 2023). Each team had to answer the same six research questions using the same dataset consisting of 720 million financial transactions. Pérignon et al. (2023) showed that running the original researchers' code on the same raw data regenerated exactly the same results only 52% of the time. Reproducibility was higher for researchers with better coding skills and for those who exerted more effort. It was lower for more technical research questions, those with more complex code, and for outlier results. Neither researcher seniority, nor peer-review ratings appeared to be related to the level of reproducibility.

The business model of third-party verifiers

Launching and operating a third-party reproducibility verification service is costly. Colliard et al. (2023) decomposed the total costs between the fixed costs corresponding to the IT infrastructure (including software) and the variables costs corresponding to labor costs, computing costs, and the costs of accessing data. They showed that exploiting economies of scale could lower the average cost per paper from $763 to $330.

Our experience at *cascad* suggests that in addition to accessing restricted data, the most challenging and time-consuming task is to reconstruct the computing environment used by the original authors (recall that dockers are not yet widespread in economics). Another challenge in practice is to be able to locate the results in the regenerated logfile because a surprisingly large fraction of code still does not automatically generate tables and figures (see Pérignon et al., 2023). These challenges suggest that one way to reduce verification costs is to increase automation in the verification process, raise awareness among researchers, and increase their coding skills.

The question of who should pay for the extra cost associated with reproducibility checks is also key. Should the readers pay, including nonacademic ones? Should the authors pay? Should only those who get their papers accepted or all authors who submit their manuscripts pay? Differently, should the costs be covered by research funding agencies or universities? Obviously, designing a sustainable business model is a prerequisite for third-party verifiers to scale up and operate efficiently.

11.5 Conclusion

We have shown in this paper that third-party verification services are useful actors in the reproducibility ecosystem. They complement the journals' verification efforts, especially when the research is based on restricted data or requires special skills or computing environments. We have shown that to prosper in the long term, third-party verifiers need to automate their labor-intensive process and clarify their business models.

Third-party verifiers could also be useful to systematically verify empirical findings based on *online experiments*, such as those conducted on Amazon Mechanical Turk (MTurk) and Qualtrics. While this kind of study often relies on pre-analysis plans and shares final datasets on public repositories (e.g. Open Science Foundations), the existence of several forensic investigations (http://datacolada.org/109) and subsequent paper retractions suggest that it is more important than ever to allow third-party verifiers to access the raw data collected on the online platforms.

11.6 References

Chang, A. C., & Li, P. (2017) A preanalysis plan to replicate sixty economics research papers that worked half of the time. *American Economic Review*, 107(5), 60–64.

Christensen, G., Miguel, E. (2018) Transparency, reproducibility, and the credibility of economics research. *Journal of Economic Literature*, 56(3), 920–980.

Colliard, J.-E., Hurlin, C., and Pérignon, C. (2023) The economics of computational reproducibility, Working Paper, HEC Paris.

Duflo, E., and Hoynes, H. (2018) "Report of the search committee to appoint a data editor for the AEA." *AEA Papers and Proceedings*, 108: 745.DOI: 10.1257/pandp.108.745

Gertler, P., Galiani, S., and Romero, M. (2018) How to make replication the norm. *Nature*, 554 (7693), 417–419.

Herbert, S., Kingi, H., Stanchi, F., & Vilhuber, L. (2021). The reproducibility of economics research: A case study. Banque de France Working Paper Series, WP #853.

Menkveld, A., Dreber, A., Holzmeister, F. Huber, J., Johannesson, M., Kirchler, M., Razen, M., Weitzel U., et al. (forthcoming) Non-standard errors, *Journal of Finance*.

National Academies of Sciences, Engineering, and Medicine. (2019) Reproducibility and replicability in science. The National Academies Press.

Pérignon, C., O. Akmansoy, C. Hurlin, A. Menkveld, Dreber, A., Holzmeister, F., Huber, J., Johannesson, M., Kirchler, M. Razen, M., Weitzel U. (2023) Computational reproducibility in finance: Evidence from 1,000 tests. Working Paper, HEC Paris.

Pérignon, C., Gadouche, K., Hurlin, C., Silberman, R. and Debonnel E. (2019) Certify reproducibility with confidential data, *Science*, 365 (2019), 127–128.

Trisovic, A., Lau, M. K., Pasquier T., and Crosas, M. (2022) A large-scale study on research code quality and execution. *Scientific Data*, volume 9, 60.

Vilhuber, L. (2020). Reproducibility and replicability in economics. *Harvard Data Science Review*, 2(4).

Vilhuber, L. (2021). Report by the AEA data editor. *AEA Papers and Proceedings*, 111, 808–817.

Part VII

Session 7: Why can or should research institutions publish replication packages?

This session brings various perspectives together on how research institutions think about publishing replication packages themselves, i.e., not a journal or generalist repository. Panelists come from a university with a specialized, university-centred data repository; from a Federal Reserve Bank with an active researcher community, and from a non-profit (non-academic) research institution. Each faces the requirements of varied internal researchers, external visibility, and differing audiences. The panelists can all speak to how a research institution makes decision about the degree of transparency, and how much of that to do with internal resources.

12 Open Data and Code at the Urban Institute

Abstract: As the leader of the Technology and Data Science team at the Urban Institute, I believe that the field of quantitative social science data analytics needs increased requirements and branding around reproducibility checks, open data, and helpful documentation. Doing so could help to separate high-quality science from less well supported arguments, and pave the way to increasing replication studies in the field. In this piece, I discuss my organization, the Urban Institute, our initiatives, successes, and challenges to date in this area, and my recommendations for next steps.

12.1 Background

The Urban Institute is an organization whose mission is to provide evidence, analysis, and tools to people who make change to ultimately empower communities and improve people's lives. We define "people who make change" broadly as policymakers, government agency employees, advocates, community leaders foundations, corporate leaders, and other similar actors.

Though Urban as an organization has a number of goals, I would categorize our primary drivers as 1) to make impact toward our mission and 2) to fundraise effectively to support that impact and the organizational supports that make it possible. This is important to consider later when I discuss organizational priorities around open data and code.

Similarly, Urban conducts work broadly across many policy areas, however I might summarize them succinctly as 1) conducting policy research and evaluations, 2) providing technical assistance on implementation, 3) producing data and data tools, 4) providing advisory services, 5) convening experts across sectors, and 6) translating research and communicating it to targeted audiences. In support of this work, Urban sometimes posts both the data and code powering the data on its website, Urban.org.

12.2 Main Thoughts

Existing Initiatives

Urban is home to a number of existing initiatives intended to make progress toward more open data and code. The first is Urban's Data Catalog (https://datacatalog.urban.org/), to which all researchers who wish to publish code on Urban's website must submit their data and document their submissions to a minimum extent. The second is Urban's central library of quality assurance materials and trainings, which promote open science standards, reproducibility checks, automation in programming, clear organization, and quality documentation throughout. The third is Urban's automated R (and soon Stata) themes, templates, and guides, which allow researchers to more easily automate the full research pipelines from data collection to publication in R. Urban has processes in place to comply with the requirements of third parties, such as the AEA Data Editor and ICPSR among others, to whom Urban is required to submit or may submit voluntarily. And finally, Urban has a central team that is available to conduct code reviews and reproducibility checks on demand.

Urban continues to make improvements in all of these areas, including adding supporting resources for quality assurance such as an internal code library, providing additional documentation and examples on Github for certain projects, improving our automation of publishing systems to extend to additional content on our website, and revamping and improving our data catalog experience.

Successes

As a result of these efforts, Urban has seen a number of successes that have led to substantial benefits to the organization. For our external users and partners, our publicly available data are now better documented, with a clear license for use, citations are clear and available, and our impact through open data is easier to see and track. I have seen better quality assurance materials and process automation lead to a more streamlined review process that saves time and allows for rapid iteration and even innovation under tight timelines. The processes and systems in place have also allowed for more redundancy and reduced stress on team members when they work well, especially when these efforts reduced in improved documentation, onboarding, communication, and collaboration across teams with diverse skill sets and backgrounds.

Challenges

Despite our efforts and successes, significant challenges remain. Our organization is decentralized and funded by many different parties across government, philanthropy, and the private sector. I have observed that many of these funders in recent years, especially in the philanthropic sector, have shifted their focus away from core research and more toward work that generates impact. I worry that this shift will lead open data and code efforts to be seen increasingly as "optional" when so many other "more impactful" activities are vying for the next marginal funding dollar. In any case, as a result of this landscape, these open practices are

only adopted on a voluntary basis or in certain cases where required by the funder or journal at Urban.

I also observe that researchers and organizational leadership are not directly incentivized, outside of the few funding requirements we do observe (such as the Sloan Foundation, the Arnold Foundation, the National Science Foundation, and others), by existing priorities to tackle these challenges. In the scope of our priority focuses on impact and fundraising, I have observed that while centralized quality assurance and open code are seen as a priority at Urban, they have been at times overwhelmed by even higher priorities, especially in light of my thoughts on funders' changing priorities in this space.

In the meantime, researchers continue to prioritize quality control in their own decentralized individual projects and efforts. However, in my experience the majority are not motivated by quality control arguments to adopt newer open data and code practices, and the strongest motivation remains funder and journal requirements. Most researchers I work with see their work and current processes as high quality, just defined differently across the organization. More importantly, in my view, is that open data and code efforts are often seen as an additional layer of bureaucracy and busy work on top of existing requirements, and thus they are perceived as reducing the agency and academic freedom that many researchers highly value.

12.3 Conclusion

It is hard for me to see openness and reproducibility change without increasing the requirements on researchers and institutions from those funding them and disseminating their work. Ultimately, despite the short-term perceived costs of increased bureaucracy, I believe these requirements will bring larger benefits and are worth considering.

I believe it would be wise for advocates for open data and reproducible research to call for funders and journals to require reproducibility checks at a minimum, and open data where possible as a next step. I would also be in favor of third-party reproducibility checks, and/or marks that certified that a third-party check has been passed and certain materials are available for reproduction.

I believe that these requirements would improve our clients' confidence in the quality of the complying institutions, and clearly help us to differentiate important policy signal from the noise. It would also pave the way toward a future where more replication studies are feasible. Urban currently has systems, processes, and materials in place to comply with these requirements, and the field now has sufficient examples from peer organizations and journals to enable the rapid spread of best practices once requirements are in place.

13 Prioritizing Transparency

The Federal Reserve Bank of Kansas City (FRB KC) operates with a mission to work in the public's interest by supporting economic and financial stability. Central to this mission is public and community engagement that emphasizes trust along with core values such as integrity and service. These characteristics point towards a strong organizational alignment with the foundational principles of replication, namely transparency.

When FRB KC first began exploring reproducibility in earnest, our researchers were sharing data on an infrequent, ad-hoc basis with full responsibility for preparing and disseminating their own replication packages. However, growing library services and evolving journal expectations drove us to reconsider our approach. An internal data preservation initiative posed questions about what types of data we should preserve and, importantly, why. As we considered the historical and administrative purposes of research data, we also examined preservation policies and practices in economic journals to better understand the state of the field[1]. There, we found a steadily increasing trend in journals requiring or strongly recommending the inclusion of replication packages as part of the publication process.

With both our mission and industry trends in mind, FRB KC ultimately decided to prioritize a high degree of transparency in our approach to sharing replication packages. Our researchers' findings inform important economic policies and decision-making, and we determined that a focus on openness was one way FRB KC could tangibly demonstrate our ongoing commitment to integrity and public trust. Thus, we elected to establish a formal process to enable researchers to share underlying data to the extent possible.

[1]Butler, C. R. & Currier, B. D. (2017). You can't replicate what you can't find: Data preservation policies in economic journals. Presentation to the International Association for Social Science Information Services & Technology (IASSIST) Conference, Lawrence, KS. Available at http://doi.org/10.17605/OSF.IO/HF3DS

14 The Data Release Process

Researchers initiate the data release process, and the Research Library coordinates the subsequent steps. These steps include legal reviews of contracts and terms of use, information security assessments, and the creation of a comprehensive ReadMe document. The ReadMe document contains vital information such as licenses, file descriptions, instructions for usage, data source references, and fixity details to ensure long-term usability.

Currently, our researchers are under no requirements to publish replication packages; it is up to their discretion to determine when to initiate the data release process. Those decisions are often driven by a range of motivations, including journal requirements, media inquiries, collegial requests, a desire to make research widely available, and more. However, not all data can be shared in every situation. Potential barriers include code and data complexity, contractual restrictions, privacy concerns, and many more.

Despite these challenges, FRB KC confronts obstacles to sharing head-on. While many journal policies provide exceptions for situations involving proprietary or confidential data, FRB KC will often seek permission from data vendors on an ad-hoc basis. We also work to incorporate explicit language during contract negotiations that allows for replication-focused sharing. In this way, we have successfully shared data from a wide range of public and proprietary sources. Appropriate sharing is also sometimes facilitated through methods such as aggregation and synthetic data. However, there are still times when sharing is not feasible, and legal and ethical matters always outweigh desires for transparency.

Once the review process is complete, researchers have a wide range of distribution options available. Replication packages may be posted on the FRB KC website, uploaded to an external repository, emailed to a requestor, submitted to a journal, or any other means of distribution needed. Research Library staff are available to help facilitate any desired publication method, but researchers can also manage distribution themselves.

The data release process currently has a turnaround time of approximately three weeks from initiation to publication, but it can also be expedited when needed and may take additional time if issues arise. Defining and operationalizing this timeline has required ongoing relationship-building and negotiations with key stakeholder functions such as legal and information security. These discussions have centered on clearly defining researchers' needs and balancing them within the organization's risk mitigation framework and available resources.

15 Resourcing Considerations

The curation process is completed entirely by internal staff since replication packages are considered public once they leave our internal systems. Whether posted publicly in a repository or emailed to a single well-known colleague, we recognize that we no longer have control over replication files once they have been distributed. As a result, our curation process ensures that all replication packages meet FRB KC's standards for sharing with external audiences before being distributed across any medium.

At this time, we have elected to place less emphasis on the mechanisms for distribution so that we can instead focus our limited resources on curation for the reasons outlined above. As we explore future opportunities to expand publication and preservation practices, scalability and capacity remain crucial considerations. Every replication package requires time-consuming context gathering, stakeholder coordination, and tailored decision-making. Growing demand for this service requires that we continue to be thoughtful and intentional in deploying available resources.

16 Conclusion

The Federal Reserve Bank of Kansas City's commitment to transparency and public trust has shaped our approach to publishing replication packages, and our transition from an informal, ad-hoc approach to a more structured process has helped promote secure, ethical, and effective data sharing. In this way, FRB KC research practices continue to contribute to the overall mission and values of the organization. This is especially important given that the Federal Reserve System is a long-lived institution. After more than 100 years of supporting economic and financial stability, trust in our research is intended to extend well into the future in addition to informing the audiences of today.

As other research institutions consider their role in the publication of replication packages, the overall approach should be influenced by factors such as organizational goals, risk assessments, known impediments, and available capacity. Comprehensive curation and publication services are not appropriate for all institutions, and there are many external partners and resources available to supplement where needed. However, replication best practices are maturing steadily in the field of economics and will continue to normalize with time. Research institutions that proactively consider their role and expectations in this space will be much better positioned to maintain ongoing trust and integrity in their findings.

Part VIII

Session 8: Should funders require reproducible archives?

Both private and government funders of academic research have been increasingly requiring that data collected or created as part of funded research be made openly available. However, it is still rare that this requirement extends to computational artifacts, such as code, and even more rare that computational reproducibility is required. The panelists all work for funders, and have experience with various funding models and approaches. But only one of them currently enforces computational reproducibility of funded research.

17 We Should Do More Direct Replications in Science

Despite an arguable reproducibility crisis in many scientific fields, some have questioned the value of performing or funding direct replications of prior studies. Their reasoning is that direct replications add little to our knowledge, and that we should focus on performing new studies. I argue, to the contrary, that direct replications are essential to scientific progress. Without direct replication, we have much less ability to know which prior scientific findings are actually worth trying to extend. As well, only direct replication can help us figure out puzzling anomalies about which contextual factors are actually important to a given scientific result.

18 Introduction

As we have seen over the past several years, there are problems with replicating the academic literature in many fields. The Reproducibility Project in Psychology found that only around 40% (give or take) of psychology experiments in top journals could truly be replicated. The Reproducibility Project in Cancer Biology similarly looked at studies from top journals in that field, and found that the replication effect was only about 15% as big as the original effect (for example, if an original study found that giving cancerous mice a particular drug made them live 20 days long, a typical replication experiment found that they lived 3 days longer). Many pharmaceutical companies have said that they can barely replicate the academic literature, despite the fact that they have a huge incentive to carry forward successful experiments into further drug development (see here and here).

Due to these results and many others, one current proposal is that science funders such as the National Institutes of Health (NIH) and the National Science Foundation (NSF) – which will spend nearly $60 billion this year, collectively – should dedicate at least 1/1000th of their budgets to doing more replication studies. Even just $50 million a year would be transformative, and would ensure that we can have higher confidence in which results are reliable and worth carrying forward into future work.

Oddly enough, not everyone agrees that directly replicating studies is a high-value activity. Indeed, when I was at a National Academies workshop recently, someone fairly high up at NIH told me that they weren't in favor of doing more replications (it was a personal conversation, so I won't name and shame the individual in question).

The gist of this person's view:

> "What do we really learn from trying to replicate experiments exactly? No experiment is ever going to be perfect, and we'll find some discrepancies, but who cares? What really matters is whether the finding is robust in different contexts, so instead of funding exact replications, we should just fund new work that extends it in a new direction."

This NIH official isn't the only one who is skeptical of the value of replication. Back when the Reproducibility Project in Psychology was finishing up in 2014, Jason Mitchell (of the Social, Cognitive and Affective Neuroscience Lab at Harvard) famously wrote a short piece in 2014 called "On the Evidentiary Emptiness of Failed Replications."

Mitchell's major claim is that it can be very hard to elicit a positive effect, and there are many more ways to mess up than to get things right. Moreover, there is a ton of tacit and unwritten knowledge in the fields of psychology and neuroscience (and, one presumes, other fields as well). By analogy, he says, if you take a recipe and follow it to the letter, but you don't actually know what "medium heat" means or how to thinly slice an onion, you might not get the same results as an expert cook. But that doesn't mean the recipe is wrong, it just means that you don't have enough tacit knowledge and skill. Thus, he suggests, unless the replicators do everything perfectly, a "failed replication" is uninformative to the readers.

These points are all well taken. Nonetheless, I think that direct replication of experiments in psychology, medicine, biology, economics, and many other fields, is highly useful and often essential to make progress. This is true for several reasons.

First, by doing direct replications (or at least *trying* to do so), at a minimum you learn how good a field is at disclosing its methods such that anyone else would be able to build upon a prior study.

With the Reproducibility Project in Cancer Biology (caveat: I funded that while in philanthropy), we saw that literally zero percent of the time was it even possible to ***try*** to replicate a study.

This wasn't because of tacit knowledge or because the original experimenters had some highly nuanced skill that the replicators lacked. Instead, it was because of obvious steps in the study that had to have happened, but that hadn't been documented very well at all.

For one example, "many original papers failed to report key descriptive and inferential statistics: the data needed to compute effect sizes and conduct power analyses was publicly accessible for just 4 of 193 experiments. Moreover, despite contacting the authors of the original papers, we were unable to obtain these data for 68% of the experiments." In other words, they couldn't even figure out the magnitude of the effect they were supposed to be replicating. This is utterly basic information that ought to be included in any study.

Perhaps worse, "none of the 193 experiments were described in sufficient detail in the original paper." In every single case, the team had to reach out to the original lab, which often was uncooperative or claimed not to recall what had actually happened in the study. For the 41% of the time that the original lab *was* cooperative, the answer was always, "You'll need more materials and reagents than we mentioned."

That's why the entire project took longer, cost more, and completed fewer experiments than the project investigators had originally proposed when I funded this work while at the Laura and John Arnold Foundation. The quality of the literature was so low that it was impossible for anyone to fathom just how much effort and expense it would take even to *try* to replicate studies.

Clearly, the scientific literature can do better than this. All the top scientific journals should commit to publishing a truly *comprehensive* description of methods for every relevant study

(including video as much as possible), so that others can more readily understand exactly how studies were conducted.

Second, if a study *is* successfully replicated, then you learn that you can have more confidence in that line of work. With so much irreproducibility and even fraud, it's good to know what to trust.

For example, last year *Science* published a lengthy story detailing how a prominent Alzheimer's study from 2006 was likely fraudulent. To quote from the *Science* article:

> The authors "appeared to have composed figures by piecing together parts of photos from different experiments," says Elisabeth Bik, a molecular biologist and well-known forensic image consultant. "The obtained experimental results might not have been the desired results, and that data might have been changed to … better fit a hypothesis."

Nobel Laureate Thomas Sudhof (a neuroscientist at Stanford) told *Science* that the "immediate, obvious damage is wasted NIH funding and wasted thinking in the field because people are using these results as a starting point for their own experiments."

A systematic replication project in Alzheimer's might have turned up that fact long before now. As a result, researchers in that field would have a better idea as to which studies to trust, and where to try to explore further.

Third, there's always the possibility that a study can't be replicated very well or at all. Let's take a specific example from the Repro. Project in Cancer Biology. The bottom-line results were that "50 replication experiments from 23 of the original papers were completed, generating data about the replicability of a total of 158 effects. . . . Replication effect sizes were 85% smaller on average than the original findings. 46% of effects replicated successfully on more criteria than they failed. Original positive results were half as likely to replicate successfully (40%) than original null results (80%)."

Contrary to Harvard's Jason Mitchell and to the NIH official who spoke with me, I do think you can learn a lot from "failed" replications. There are at least three possibilities.

a. Yes, it is possible that the replication team just isn't very good, > or doesn't have enough tacit knowledge, or made a simple mistake > somewhere. That is possible. But it doesn't seem likely to be true > in all cases. Indeed, the replicators might often be more skilled > than the original investigators. And when we know that so many > pharma companies can't replicate more than 1/3rd of the academic > literature -- despite highly-qualified teams who have every > incentive to come up with a successful replication so that the > program can move forward -- it seems like we have bigger problems > than "replicator incompetence."

a. Another possibility is that the original study can't be fully > trusted for any number of reasons. Perhaps there was improper > randomization, improper treatment of outliers, questionable use of > statistics, publication bias, outright fraud, or just a fluke. To > be sure, we don't *know* any of that just because of one failed > replication. But we do have a reason to suspect that further > investigation might turn up improper practices.

a. Perhaps the original study and the replication are *both* correct, > but there is some subtle difference in context, population, etc., > that explains the difference. Consider this classic > paper, > in which two labs on opposite coasts of the US tried to work > together on an experiment characterizing breast cancer cells, but > found themselves stymied for a year or so during which their > results were inconsistent. By traveling to each others' labs, they > finally figured out that, unbeknownst to anyone in the field, the > rate of stirring a tissue sample could change the ultimate > results. They would never have known that the rate of stirring was > important unless they had been trying to exactly duplicate each > other's results. Thus, it seems hugely important to know which > seemingly insignificant factors can make a difference--otherwise > someone trying to extend a prior study might easily attribute a > change in results to the wrong thing.

Thus, we have many reasons to think that direct replication of a scientific study (or of a company's data analysis) is actually important. A direct replication can expose flaws in how the original analysis was reported, can expose faulty practices (or even fraud), can help us know how to extend a prior study to new areas, and at a minimum can help us know which results are more robust and trustworthy.

As to federal funding, my conclusion is this: Let's say that we spend X on new science and R&D every year. A system that puts 99.9% of X towards new research, and 0.1% of X on replication studies, will be more reliable, productive, innovative, and will lead to more pharmaceutical cures, than a system that funds the whole 100% of X on new research.

Disclosure Statement: The author works for an organization (Good Science Project) dedicated to improving science, but has no financial conflicts of interest as to any topic discussed here.

Part IX

Session 9: Reproducibility, confidentiality, and open data mandates (at CEA)

Many granting agencies have adopted open data mandates. What is the interplay between reproducibility and those mandates? How can researchers be supported to meet those mandates, both in general, and specifically when data are confidential. At first glance, confidentiality and open data seem irreconcilable, but could we find practices that both respect confidentiality and provide enough information and transparency to foster reproducibility? *NOTE: Session was part of the annual CEA conference, online day.

19 Reproducibility, Confidentiality, and Open Data Mandates

There are a number of things to consider when contemplating reproducibility and open data mandates in the context of research uses of confidential and often highly-sensitive data. This note outlines a number of these considerations, and associated challenges, as well as a few potential responses to those.

20 The research context

The specific context of focus here is quantitative analysis of routinely-collected data at the intersection of health economics, health services and policy research, and population health. The context is important because it implies three important features that are relevant to reproducibility. First, routinely-collected data refer to administrative and other data sources that are highly relevant sources for applied and policy-relevant research. By definition, however, they are not collected by the researcher and therefore do not "belong to" the researcher or the researcher's institution. Second, these data sources can often be linked, which creates very rich information, and also means they can come through different data stewards; these stewards may not all have precisely the same rules and expectations around data use. Third, these data are highly sensitive, both because they are about sensitive subjects (e.g. health and socioeconomic status) and because they provide complex detail about individuals' interactions with health and social services.

There are many existing resources that can help researchers navigate access to and use of these complex linked data, including for example Health Data Research Network Canada, and the Canadian Research Data Centre Network. In these cases, and many other similar services, it is not possible to deposit data in an open access framework. Again, this has to do with ownership, but also because these data are broad resources that can support hundreds of projects annually. It would likely not be efficient or effective to archive all of the resulting permutations from the same base data resources.

Key definitions

Related to this general context is that this area of research often focuses on questions that are about the functioning of complex systems. This means that projects will use lots of variables and assess many different relationships among them. There is a need for deep knowledge of the policy systems from which data are generated and any dynamics in those across sites or geography and over time. In addition, many of the variables that might be of interest are measures of emergent properties, which are "...properties that manifest themselves as the result of various system components working together, not as a property of any individual component." In other words, these are measures that do not exist in the data in their original form, but need to be developed, often with complex algorithms through some sort of validation study. The implication is that the greater variety of data that are linkable, the more opportunity there is to develop and then study these kinds of emergent properties. Some simple examples might be aspects of health system performance (safety, efficiency, etc.) or the concept of resilience among children.

All of this might help shift focus from the deposit of data to the sharing of code, algorithms, metadata and other resources that might help with reproducibility, as well as with replicability, robustness and generalizability. As a quick overview, the definitions used here for these concepts are:

Reproducibility is using the same data and same analysis and getting the same result.

Replicability is using different data and the same analysis and getting the same result.

Robustness is using the same data and different analysis and getting the same result.

Generalizability is using different data and different analysis and getting the same result.

All of these are important, with the emphasis depending in part on how unique or new a research question is, the existing evidence, and the importance of, or difference in, context when studying the phenomenon of interest.

21 The challenges and possible solutions

Given the context and ambitions outlined above, there are three distinct challenges for research. First is accommodating sensitive or private data, second is the limited space in journal articles for complex methods for data development (i.e. data wrangling and shaping as opposed to statistical analytic methods), and the third is transferability of code across different systems and data generation processes.

Sensitive and private data: While some data cannot be placed in open access repositories, it is still possible to meet other principles of open science. Specifically, the data used can comply with the FAIR principles of being findable, accessible (as in, clarity on how they can be accessed), interoperable and reusable. Very much aligned with that, researchers can develop metadata for the final data set used for analysis. Good metadata will include both provenance information, such as the population covered, the time period, the purpose of the data set creation and so on, and specific details on the variables and the data sets and computational methods used to derive them. It can be helpful to think in terms of a "data set genealogy", meaning a figure or diagram to show original data sets and how they come together to create the analytic data set. This ideally would include information on the starting population (N=), inclusions and exclusions, and the resulting data set.

Limited space in journal articles for data development methods: The ability to reproduce or replicate studies relies on transparency. The required transparency includes both the process used to move from the real-world, routinely-collected data to an analysis-ready data set and the specific analysis or modelling approach used. Analysis and modelling is often the focus for scrutiny in peer review. The construction of the analytic-read data set, in contrast, is often described in enough detail to generally understand what was done but short of what is needed to replicate or reproduce the study.

One way to address this is to develop concept dictionaries such as the one run by the Manitoba Centre for Health Policy, and related code repositories that can help operationalize those concepts. Another possible response is to develop guidelines and a receptive publication venue for detailed methods, as a form of a "deposit paper" that could accompany new entries to a concept dictionary or new contributions to a code repository. This type of paper would provide transparency of the process used to create an analysis ready data set, and therefore the ability for other researchers to assess the process and the decisions made. A standard format and expectations for these methods papers would help researchers to structure the content and peer reviewers with their assessments of quality and completeness. Health Data Research Network

Canada is pursuing these options, so may be a source of more information about these ideas in the future.

Transferability of code: Code repositories provide a place to share, but it is also important to think about the transferability of code. For example, it may be important to document the context of analysis (lightly in the code, with direction to metadata) and specifically how the context might alter the analysis, the findings, and potential for generalizability. Another possibility is to use data harmonization or data standardization practices to develop common data models. The attraction of common data models is that they provide pre-specified definitions for variables of interest, and the ability to access readily-usable and analysis-ready data from multiple jurisdictions. In other words, common data models address data and concept transferability once, which then can be used many times by any researcher with access to those data. The drawback is that these models may not have all the variables or emergent properties of interest, but this opens the possibility for extensions to existing models. If these are in place, and are well-documented, they create transparency, efficiency and the potential for reproducibility and replicability (and potentially robustness and generalizability) in research.

22 Conclusion

Research using sensitive data raises some additional challenges for meeting the mandates of open science, including reproducibility. These challenges are solvable, but will require additional attention and work from research teams to ensure that the fundamental ingredients of transparency in data availability, access, manipulation and analysis are all present.

23 Reproducibility, Replicability and Open Science at the Canadian Research Data Centre Network

As an organization that exists at the nexus of academic research and public policy, the Canadian Research Data Centre Network (CRDCN) is committed to enabling projects at the forefront of the many disciplines represented among our users and to supporting robust analyses of public policy that can be used to improve outcomes for Canadians. In that regard, reproducibility and replicability figure as key criteria for assessing strength of research evidence. The CRDCN in partnership with Statistics Canada continues to advocate for and advance the ideal of open science within the constraints of providing access to confidential microdata.

23.1 Context

The CRDCN is a network of over 40 Canadian universities that support access to de-identified microdata from Statistics Canada at one of 33 university-based secure facilities (not dissimilar from the US RDC facilities). The data in the secure facilities are very detailed and include: question-by-question response data from nearly all social and industry surveys conducted by Statistics Canada; governmental program data from a wide variety of sources (Education, Health and Employment being the most prominent); administrative data from the Canada Revenue Agency and Immigration, Refugees and Citizenship Canada; and a large set of linkage keys through which many of these files can be used together. Academic researchers with suitable research questions can generally access the data at no cost to them following a review of a research proposal and obtaining a security clearance. Other researchers can similarly access the data on a cost-recovery basis. Core funding for the facilities and a small central staff are provided by federal grants, Statistics Canada, provincial support, and the partner universities. CRDCN researchers represent more than 25 disciplines, but are concentrated in economics, sociology, epidemiology, and public health.

The infrastructure that the CRDCN operates is currently undergoing a major transition. Whereas we now have site-specific servers that must be updated and maintained locally, two

advanced research computing facilities have been built at the University of Waterloo in Ontario and Simon Fraser University in British Columbia. Following security checks and technical pilots they will serve as the backbone computational and storage infrastructure for a new national IT platform. With this transition, the computing environment will be standardized for all users, meaning that software and storage will be managed in a way that is the same regardless of where the user is accessing the data. This environment will also allow users working on most data files to work remotely should they so choose.

As a network dedicated to providing access to confidential microdata in a way that ensures the confidentiality of the respondents, CRDCN is cognizant of our role as it relates to the open science movement. While open data is an important pillar of open science, there are other elements that are not affected by the sensitivity of the data being used. To ensure the credibility and strength of network research, we continue to push for more open methods and tools and for greater data accessibility within the confines of the legislation which governs access. The value of the confidential microdata accessible through CRDCN for answering both scientific and policy questions is unparalleled, and despite not being fully open these data remain an important tool for rigorous and robust analyses.

For several reasons the CRDCN is a fertile test bed for providing strong evidence to advance knowledge and inform policy. Foremost is the intrinsic advantage of microdata - de-identified individual records from surveys and administrative data – which de facto are not subject to the various biases inherent in aggregate records. Second, the breadth and depth of the repository afford a very diverse set of opportunities for analysis over time, space, and content domain. Third, several of the surveys are conducted over repeated cycles so enabling time series analysis. Fourth, the number of linked data files – both between surveys and between surveys and administrative data – further extends the scope for analyses which map on to the complex processes typical of many of the economic, social and health systems that are primary foci of investigation. And fifth, is the depth and breadth of expertise represented in the 2400 CRDCN researchers drawn from 42 partner and affiliated institutions across Canada and embracing a wide range of disciplines in the social and health sciences. Underpinning these strengths is the intrinsic robustness of Statistics Canada microdata collection, curation, and management which provides a strong foundation for conducting reproducibility and replicability analyses.

Despite these advantages of the CRDCN repository as a fertile test bed, replicability and reproducibility studies are overwhelmingly the exception not the rule and, in the case of the latter, conspicuous by their total absence from the Network's extensive bibliography of over 6000 publications. So, what are the challenges and/or disincentives that explain this apparent mismatch between the in-principle potential and yet the in-practice outcome? And how might those challenges be addressed?

23.2 Reproducibility and Replicability in a Secure Environment

Security provisions under the federal Statistics Act in Canada govern access to the confidential data made available through the CRDCN. As such, they cannot be open data, but this does not mean they cannot be fair [@fair]. Being able to find, access and use the data is, of course, a fundamental and necessary precondition to being able to either replicate or reproduce a study. Core functions of the CRDCN-StatCan partnership are to enable findability, accessibility, interoperability, and reuse given the constraints prescribed by the Statistics Act. In the context of confidential data this constitutes the most important activity since without it, no other steps towards reproduction or replication can proceed.

The code used to conduct the analysis becomes the second critical piece and it is here where our current efforts to foster reproducibility are targeted. Inside the secure facilities, statistical information compiled from the raw data are vetted by a staff person before the information is released to ensure that there are no residuals that might reveal an individual's information and that the release satisfies the guidelines for statistical information releases for that dataset. Part of this process involves providing some supporting documentation which most often includes statistical code files used to create the output. This means that, for most projects, the process of building a set of code files that can recreate output is automatic and researchers have only to request vetting of their already existing code.

Building on this operational advantage to support reproducibility analysis has been the subject of many discussions by a CRDCN working group. While the policies that relate to vetting favour reproducibility, there are currently several policies that do not. Most critically, updates to datasets (as a result of rebasing or errors later discovered) are uploaded without offering a way to use the original data. Therefore, while a researcher might know that the data needed to conduct a reproduction of a study have been updated, they cannot be requested (this does not preclude the possibility of verifying that the original study is replicable with the updated data). Next, there is no mechanism present by which code can be shared between projects.

The computational environment inside the secure facilities is tightly controlled, and there is no way to request or move code between research contracts (even for the same researcher). The implication here is that to re-use code it must be produced in such a way that it can also pass a vetting process for release from the facility. While this type of coding is consistent with best practices, many code files will not meet this standard. If the standard is met, the code can be vetted and released at any time during the archival period of five years.

Reproducing an RDC study would follow a three-part process. First, the reproducer would gather the code either through contact with the author or by accessing code that the author had released as part of a replication package. Next, they would apply for access to the data, including writing a full research proposal. Finally, they would undergo a security screening process with fingerprinting, and pay cost-recovery if they were not part of the network. None of the current incentive structures in academia for reproduction or replications accommodate this level of required effort. Indeed, if we search the CRDCN bibliography for "replication",

"reproduction" or "reproduce" we find a few studies on fertility rates, and some on social inequalities, but of course these research themes are very different to research work attempting to reproduce or replicate research done in the RDC facilities. In 2022 a replication workshop was hosted by the Canadian Journal of Economics Data Editor (Marie Connolly) with some replications taking place inside the secure facilities. Her takeaway was that the level of administrative burden required to set up access for participants was not worthwhile when there are so many other papers at the journal that could be replicated without requiring that level of effort (even though Statistics Canada and the RDC program were supportive of the mission).

We are unlikely to see more institutional support for reproducibility in the CRDCN in the short term given that this existing mechanism by which research can be reproduced is largely unused. Whatever reproducibility looks like in the future, it will almost certainly need to be led by academics and scholarly societies. However, should the demand to reproduce or replicate research arise, satisfying it will require a major shift in the way the program is operated. With well over 150 peer-reviewed articles produced by the Research Data Centres every year (not to mention policy reports, theses, and other research outputs), operational realities preclude reproduction or replication of any more than a small fraction of the body of research.

That said, influential work conducted in the RDC is replicated and extended in the normal course of advancing knowledge. The most high-profile example is the replication and extension of the research on work requirements in the Self-Sufficiency Project by Riddel & Riddel (2014) from Journal of Public Economics published in 2020 in the Journal of Labor Economics. The experimental design of the policy was theoretically well-suited to learn about the labour market effects of welfare program incentives, but the replication included statistical controls to account for some changes to the post-intervention policy environment which overturned many of the original findings. The data management and universality of access make this process easier for academics with access to the RDCs. Simply put, the rest of the world is in many ways catching up to where the CRDCN researchers have always been. Where data deposits have relatively recently become routine and/or required, researchers using the CRDCN facilities have always been able to request the raw data used by another project.

23.3 Conclusion

As a network, CRDCN's ambition is to encourage our researchers to move their work as far along the spectrum of open science as possible, ensuring that the data and research tools that can be made available are made available. At the same time, we advocate that Statistics Canada, as data provider, move as far along the spectrum of FAIR data as possible. We will continue to build capacity within our research community and particularly for early-career researchers in ways that will let them maximize the vision of "as open as possible". To accomplish this, we provide training and guidance on reproducibility in secure environments, and our Replicability and Reproducibility working group continues to look for ways to enable a

research culture at CRDCN that prioritizes open science including partnerships with journals and administrative support for reproducibility initiatives.

23.4 Bibliography