Expanding the scope of reproducibility research through data analysis replications

Jake M. Hofman
jmh@microsoft.com
Microsoft Research
New York, NY

Daniel G. Goldstein
dgg@microsoft.com
Microsoft Research
New York, NY

Siddhartha Sen
sidsen@microsoft.com
Microsoft Research
New York, NY

Forough Poursabzi-Sangdeh
forough.poursabzi@microsoft.com
Microsoft Research
New York, NY

ABSTRACT

In recent years, researchers in several scientific disciplines have become concerned with published studies replicating less often than expected. A positive side effect of this concern is an increased appreciation for replicating other researchers' work as a vital part of the scientific process. To date, many such efforts have come from the experimental sciences, where replication entails running new experiments, generating new data, and analyzing it. In this article, we emphasize not experimental replication but *data analysis replication*. We do so for three reasons. First, experimental replication excludes entire classes of publications that do not run experiments or even collect original data (for example, papers that make use of economic data, census data, municipal data, and the like). Second, experimental replication may in some cases be a needlessly high bar: there is great value in replicating the data analyses of published experimental work. As analytical replications require a lower investment of time and money than experimental replications, their adoption should expand the number and variety of scientific reproducibility studies undertaken. Third, we propose educating the undergraduate students to perform data analysis replications, which has scalable benefits for both the students themselves and the broader research community. In our talk we will provide details of a pilot program we created to teach undergraduates the skills necessary to conduct data analysis replications, and include a case study of the first set of students who completed this program and attempted to replicate the data analyses in a widely-cited social science paper on policing.

KEYWORDS

reproducibility, replication, robustness, education, data analysis

, ,

ACM Reference Format:
Jake M. Hofman, Daniel G. Goldstein, Siddhartha Sen, and Forough Poursabzi-Sangdeh. 2020. Expanding the scope of reproducibility research through data analysis replications. In . ACM, New York, NY, USA, 5 pages.

1 INTRODUCTION

Researchers in several scientific disciplines are concerned with a replication crisis in which the results of published studies replicate less often than expected [2, 13, 19]. This alarming realization presents the scientific community with both the challenge and the opportunity to improve how research is done. A good deal of progress has already been made in this direction in terms of increasing the reliability and verifiability of published work.

For instance, many researchers have adopted the practice of pre-registration, which amounts to publicly declaring the design and analyses of a study (e.g., hypotheses to be tested, experimental manipulations to be studied, and statistical tests to be run) before conducting it [14]. Publicly declaring the details of a study forces researchers to think about these technicalities before any data are collected or analyzed, which reduces (and ideally eliminates) the type of data-dependent decision making that can otherwise lead to high false discovery rates [11, 21]. It also has the benefit of enabling reviewers and consumers of a study to easily check if the study was executed as planned, which helps to distinguish between exploratory and confirmatory research [6, 14].

Standards have also improved around how research results are shared with the community. For example, some journals now require authors to submit transparent research materials such as data and analysis code with their publications,[1] making it easier for others to check and verify their work and build upon it. Other outlets leave this as optional, but reward authors with badges or provide similar incentives for submitting reproducible work.[2] In addition, improvements in software engineering practices, open source software tools, and computational infrastructure have made it easier than ever for authors to share their work in a way that is convenient for others to consume.

[1] See, for instance, the data policies for PLOS One (https://journals.plos.org/plosone/s/data-availability) or the American Economic Review (https://www.aeaweb.org/journals/policies/data-code/).

[2] See, for instance, badges awarded for computer science work by the Association for Computing Machinery (https://www.acm.org/publications/policies/artifact-review-badging) and the Open Science Framework (https://osf.io/tvyxz/wiki/1.%20View%20the%20Badges/).

The hope is that these practices will eventually become commonplace, leading to more credible original findings and early identification of problematic results. In the meantime, however, there is a good deal of existing research that does not adhere to these standards, making it difficult to assess the reliability of previously published work. Often readers are only presented with claims without access to any of the data or code that produced them.

A natural solution to this problem is to independently repeat the entire procedure specified in the paper and check to see if similar results are obtained. There have been notable recent efforts to do so, mainly in fields such as experimental psychology where replications involve running entirely new versions of a previously described experiment [16]. This requires recruiting new subjects, collecting new data, and following the original analysis plan to test previously specified claims. These replication projects are impressive, but are also costly and relatively difficult to scale as they require the time and expertise of highly trained researchers who, for instance, have access to a physical laboratory and are experienced in running human subjects experiments.[3]

Less attention, however, has been paid to reproducing the results of non-experimental work, for instance from research that relies on publicly available surveys or observational data. There is an abundance of such research, and reproducing these results has a much lower barrier to entry compared to reproducing experimental work. Assuming the underlying data behind such studies are publicly available, in principle all one needs to reproduce data analyses and results is access to and training with standard software packages to program and run the stated methods in the original work.

It is our conviction that there should be more data analysis replication attempts. What would it take to get the scientific community to embrace them? In theory, such replications could come as a result of journals and research institutions rewarding for this type of work. There has been some progress in this direction, but many researchers have not rushed to prioritize (data analysis) replications over other research activities. At the same time, there is an alternative approach that relies on a large pool of individuals who could aid in this effort, and benefit from doing so in the process: undergraduate data science students.

Engaging undergraduates in data analyses replications is in some sense a natural match. There is a sizeable overlap in the skills needed to perform data analysis replications and the skills that we aim to teach students at the undergraduate level, specifically in statistics, the social sciences, and computer science. And whereas it might be difficult to incentivize established researchers to work on data analysis replications, it is relatively straightforward to incentivize undergraduates to do so, by simply assigning data analysis replications as homeworks or class projects. Not only would this be an effective way to reinforce the skills that students are already being taught, but it also offers students a unique perspective on research and encourages them to think critically about the scientific process. The result of such a program would be a scalable mechanism for vetting scientific studies with benefits for both researchers and students alike.

[3]There are, however, a few notable efforts to engage undergraduates in re-running entire experiments in lieu of more experienced researchers [4, 10].

In our talk we will discuss our experience in piloting such a program during a data science summer school for undergraduate students.[4] First we will define what we mean by data analysis replications and distinguish them from other efforts to replicate published work. Then we will give an overview of a training program we piloted to teach students the skills needed to perform data analysis replications. We will conclude with some insights from the students' attempt to replicate the analyses in a paper on racial differences in police use of force, discussing challenges faced along the way and lessons learned for generalizing the program to a larger audience.

2 DATA ANALYSIS REPLICATIONS

What do we mean by a data analysis replication? Before answering this question, we should note the point of this manuscript is not to debate the semantics of different terms used to categorize replication attempts, among which there is a good deal of confusion and disagreement [7, 15, 17, 18].[5] We also do not wish to suggest that data analysis replications are an entirely new concept. In fact, there has been growing interest in various kinds of data analysis replications over the past few years [9, 20]. However, because this category of work typically receives less attention than other kinds of replications, our purpose is primarily to promote data analysis replications as an effort worth undertaking, and to provide advice on carrying them out.

To clarify the terms we will use going forward, by a data analysis replication we mean an attempt to verify the claims of a paper by writing *new analysis code* that follows the methods in the paper with the *original data* used by the authors. As shown in Figure 1, this is more involved than a reproducibility check that simply amounts to having a third party run the author's *same analysis code* on the *original data* from the paper. It is also distinct from and less involved than experimental replications, which require running an entirely *new experiment*, collecting *new data*, and conducting a *new analysis* on this newly collected data. Data analysis replications, in contrast, focus on *only the last step* of conducting a new analysis with existing data.

As a result, data analysis replications involve much less work than experimental replications while simultaneously applying to a broader range of scenarios than both reproducibility checks and experimental replications. For instance, data analysis replications apply to work that relies on surveys, observational data, or publicly available data of any sort used in a research paper. They also apply to existing datasets generated from experiments. In all of these cases one can ask whether, given the data and the description of the analysis in the paper, the claims of the paper can be replicated. Data analysis replications are important when the focus is not on the data generating components of a study, but rather on the analyses which treat the data as given.

Figure 2 is a simple flowchart to help determine whether a data analysis replication is possible. The main requirement for a data analysis replication is an existing dataset. This can come in two

[4]http://ds3.research.microsoft.com

[5]For instance, we use the definitions of reproducibility and experimental replication provided by the American Statistical Association [3] that have become commonplace, but these differ slightly from the definitions used by the National Science Foundation [1].

forms. The first is a well-documented, interpretable dataset from the authors themselves. If this is not available—or if one wants to check any decisions the authors may have made in deriving their own version of the dataset—it may be the case that well-documented data are available from another source. For instance, the paper might rely on publicly available census data from the government or from data that can be obtained through other online databases or APIs. From here, if there is interpretable code available from the authors that runs in a new environment, an exact data analysis replication is not necessary; one can simply re-run the existing code to see if results are reproduced, or look to the code to understand any details of the analysis in more depth than might be described in the paper. In all other cases a data analysis replication is possible.

Ideally, all papers would include well-documented data and interpretable, easy-to-run code, making data analysis replications largely unnecessary. Unfortunately, however, it is often the case that neither data nor code are made available, and most publication outlets do not require them. The next most common case is that the data used by a paper are available, but that the corresponding code is either unavailable or difficult to re-run or understand due to broken software dependencies.[6] The case we are concerned with is that in which one must write independent code, based on the methods described in the paper.

Data analysis replications are primarily focused on verifying past claims, but also leave room for critical thinking and robustness checks. It may be of interest to examine how sensitive a previous result is to the set of analysis choices made in arriving at that claim [6, 8]. For instance, perhaps the authors used a particular statistical method to test a hypothesis, but upon re-implementing this analysis it becomes apparent that the data do not adhere to certain criteria required for the test (e.g., an ANOVA was done with non-normally distributed residuals). Likewise, it could be the case that changing the way a particular concept is operationalized—for example by changing how a continuous variable is discretized, or modifying a model specification [22]—leads to qualitatively different findings than the original paper. These scenarios are certainly relevant and within scope of a data analysis replication, but our recommendation is that data analysis replications should focus first on doing exact replications of previous claims—following the methodology specified in a paper—and only then check the robustness of claims to various choices made in the analysis process.

3 SUMMARY

We promote the practice of conducting data analysis replications. Compared to experimental replications, data analysis replications invite more types of publications to be replicated because they are not limited to studies that collect new data. In addition, data analysis replications can be applied to the data gathered from experimental papers and be of great value. Because they require less time and money than experimental replications, data analysis replications should broaden the set of people who can participate in replication work, increase the number of replication projects overall, and give more visibility to papers that have undergone replication attempts.

[6]For an interesting example of this, see [12], where 12 papers were submitted to a special issue that used the same dataset—which was agreed upon in advance—and only 7 could ultimately be run by the organizers due to problems with software dependencies and package versions, even after a considerable time investment in resolving these issues. Even when one can run the code, it is often the case that the code is poorly documented and difficult to understand, leading to little additional insight over reading the manuscript alone.

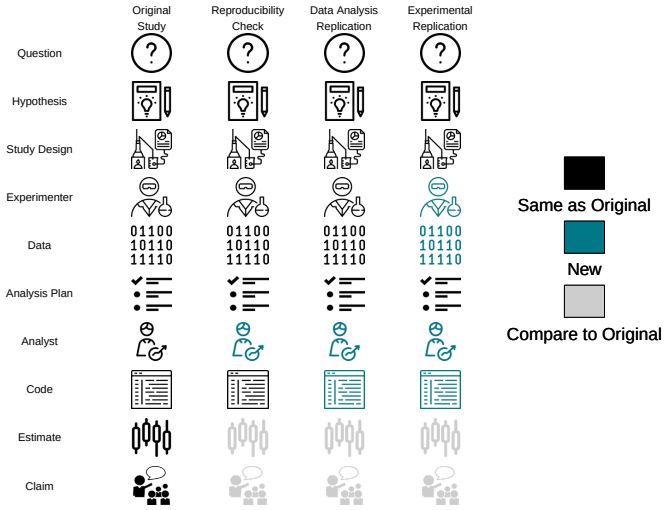

Figure 1: A figure following [17] to define what we mean by a data analysis replication. The first column depicts the stages of an original study. The second column defines a reproducibility check, where nearly everything is identical to the original study, but a third party analyst runs the original code provided by the authors on the original data to check results. The fourth column depicts an experimental replication, which requires an entirely new experiment, new data collection, and new analysis. The third column defines a data analysis replication, which sits between a reproducibility check and experimental replication in terms of effort because it leverages the original data but requires a new analyst to write new code to check the original claims in the paper. Note that an "experimenter" is depicted in the reproducibility check and data analysis replication columns, but is not strictly necessary, as these apply to non-experimental as well as experimental work.

While data analysis replications are simpler than experimental replications, they are nonetheless substantial research projects that should be valued by the scientific community. Though it might seem at first glance that data analysis replications can be carried out quickly, our case study of a month-long replication of a well-documented recent paper [5] demonstrated that many obstacles can stand in the way of such efforts. Although we do not have space to provide full details of the case study here, we will do so in our talk. Here we provide a brief summary of the case study and the challenges we faced in doing it.

Eight undergraduate students spent four weeks replicating and extending the analyses in "An Empirical Analysis of Racial Differences in Police Use of Force" [5]. We selected this paper because it was a widely-read paper that was also an ideal candidate for a data analysis replication. It not only met all of the requirements for a data analysis replication (see Figure 2), but also used relatively simple methodology that seemed straightforward to implement

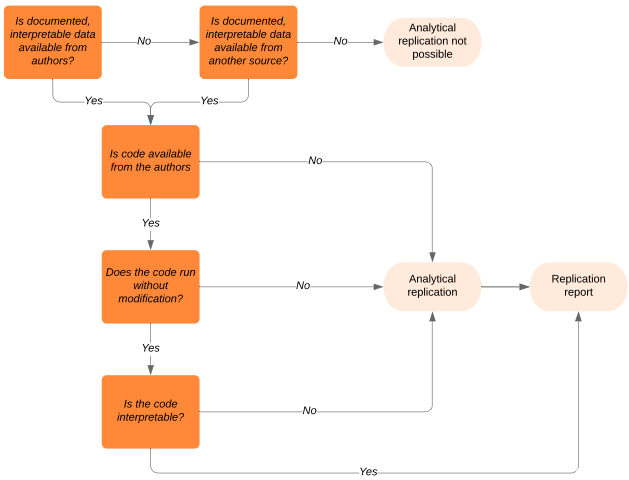

Figure 2: A flow chart to determine if a data analysis replication is possible given information about the original data and code used in a paper. If well-documented data are available from either the authors or another source, and the authors provide code, and the code runs without modification, and the code is interpretable, a data analysis replication is not strictly necessary but can still be done. In all other cases where data are available, a data analysis replication is possible.

and check, relied on two publicly available datasets,[7] and contained more than 100 pages between the main text and extensive appendices. Importantly, it also included code which enabled us to attempt the replication before and after looking at the author's code.

In short, the data analysis replication amounted to obtaining, cleaning, and recoding publicly available datasets, checking descriptive statistics on these datasets, and using them to perform a series of logistic regressions on features derived from them. This seemed deceptively simple, and the students estimated that they would complete the replication within a few days, after which they planned to spend several weeks working on robustness checks and extending the paper's original results.

In practice, completing the data analysis replication turned out to be much more complicated than expected and took several weeks itself, mainly for reasons that centered around how the original data were cleaned and featurized. These challenges came despite the extensive documentation in the paper and its appendix but also uncovered issues that might not have been clear without undertaking a data analysis replication. It was only after the students gained access to the original authors' code that they were able to resolve some of these issues.

One obstacle was having to speculate how variables were coded according to the author's description in the text. Another was getting basic counts (e.g., row counts, type counts) to match published tables, sometimes because of missing data. A third, and perhaps most important obstacle was getting fitted model coefficients to match those in published tables. Without looking at the author's code, we were able to get numbers that came close to the published ones, but which did not match exactly.

Performing these analyses led to insights that would not be apparent to a typical reader of the published work. For instance, some of the public data (upon which the original analysis was based) were incomplete in certain years, and this affected model estimates. In addition, some of the public data were inconsistently coded across years, which required subjective judgments in the analysis phase and also impacted estimates. Furthermore, we learned that a key conclusion of the original work—that blacks and hispanics were 50% more likely to experience use of force in interactions with the police—was based on a non-standard way of communicating risk. We found this number to be only 42% in the replication. The difference was attributable to the author writing about what is more likely in terms of odds ratios instead of probability ratios.

While we could have arrived at some of these insights by inspecting the author's code, reaching others—for instance, issues with the public data—were only made in the process of trying to reproduce the analysis *without* looking at the author's code. We endorse the practice we undertook here of attempting to reproduce the analysis without simply re-running the author's code. For many papers this will be the only option, as code is often not provided.

In closing, we would like to emphasize that data analysis replications are not just important for assessing the reproducibility of published work, but also useful for training future generations of researchers. The undergraduates who carried out the case study felt that engaging in a data analysis replication gave them valuable exposure to aspects of the scientific process that they would not have encountered otherwise for years to come. Typically, conducting a full data analysis as comprehensive as the one in a published paper is something that a student would not experience until after they have been admitted to graduate school, helped formulate a research and helped collect original data. We feel that bypassing these steps and going straight into data analysis replications shows undergraduates more aspects of what researchers do and helps them make better career decisions. We also believe it nicely complements more traditional textbook-based curricula in statistics by not only teaching students how to carry out statistical analyses themselves, but also encouraging them to think critically about analyses carried out by others in previously published research.

ACKNOWLEDGMENTS

We would like to thank Gabe Perez-Giz for his helpful ideas and conversations about this project. We would also like to thank the 2019 Microsoft Data Science Summer School students who we worked with on the data analysis replication case study: Brenda Fried, Harpreet Gaur, Adnan Hoq, Emeka Mbazor, Naomi Moreira, Cindy Muso, Etta Rapp, and Roymil Terrero.

[7] The paper contains analyses that rely on two additional datasets, but these datasets were not publicly available, so we could not attempt a data analysis replication with them.

REFERENCES

[1] 2018. Companion Guidelines on Replication and Reproducibility in Education Research. https://www.nsf.gov/pubs/2019/nsf19022/nsf19022.pdf

[2] C Glenn Begley and John PA Ioannidis. 2015. Reproducibility in science: improving the standard for basic and preclinical research. *Circulation research* 116, 1 (2015), 116–126.

[3] Karl Broman, Mine Cetinkaya-Rundel, Amy Nussbaum, Christopher Paciorek, Roger Peng, Daniel Turek, and Hadley Wickham. 2017. Recommendations to funding agencies for supporting reproducible research. In *American Statistical Association*, Vol. 2.

[4] Katherine Button. 2018. Reboot undergraduate courses for reproducibility. *Nature* 561, 7723 (2018), 287–288.

[5] Roland G. Fryer. 2019. An Empirical Analysis of Racial Differences in Police Use of Force. *Journal of Political Economy* 127, 3 (2019), 1210–1261. https://doi.org/10.1086/701423

[6] Andrew Gelman and Eric Loken. 2014. The Statistical Crisis in Science. *Am Sci* 102, 6 (2014), 460.

[7] Steven N. Goodman, Daniele Fanelli, and John P. A. Ioannidis. 2016. What does research reproducibility mean? *Science Translational Medicine* 8, 341 (2016), 341ps12–341ps12. https://doi.org/10.1126/scitranslmed.aaf5027 arXiv:https://stm.sciencemag.org/content/8/341/341ps12.full.pdf

[8] Jake M Hofman, Amit Sharma, and Duncan J Watts. 2017. Prediction and explanation in social systems. *Science* 355, 6324 (2017), 486–488.

[9] Jan H. Höffler and Thomas Kneib. 2013. Economics Needs Replication. https://www.ineteconomics.org/perspectives/blog/economics-needs-replication.

[10] Hans IJzerman, Mark J Brandt, and Jon E Grahe. 2018. How to make replications mainstream. https://doi.org/10.31234/osf.io/rwufg

[11] Norbert L Kerr. 1998. HARKing: Hypothesizing after the results are known. *Personality and Social Psychology Review* 2, 3 (1998), 196–217.

[12] David Liu and Matthew Salganik. 2019. Successes and struggles with computational reproducibility: Lessons from the fragile families challenge. (2019).

[13] Zacharias Maniadis, Fabio Tufano, and John A List. 2017. To replicate or not to replicate? Exploring reproducibility in economics through the lens of a model and a pilot study.

[14] Brian A Nosek, Charles R Ebersole, Alexander C DeHaven, and David T Mellor. 2018. The preregistration revolution. *Proceedings of the National Academy of Sciences* 115, 11 (March 2018), 2600–2606.

[15] Brian A Nosek and Timothy M Errington. 2019. What is replication? https://doi.org/10.31222/osf.io/u4g6t

[16] Open Science Collaboration. 2015. Estimating the reproducibility of psychological science. *Science* 349, 6251 (2015). https://doi.org/10.1126/science.aac4716 arXiv:https://science.sciencemag.org/content/349/6251/aac4716.full.pdf

[17] Prasad Patil, Roger D Peng, and Jeffrey T Leek. 2019. A visual tool for defining reproducibility and replicability. *Nature human behaviour* (2019), 1.

[18] Hans E Plesser. 2018. Reproducibility vs. replicability: a brief history of a confused terminology. *Frontiers in neuroinformatics* 11 (2018), 76.

[19] Patrick E Shrout and Joseph L Rodgers. 2018. Psychology, science, and knowledge construction: Broadening perspectives from the replication crisis. *Annual review of psychology* 69 (2018), 487–510.

[20] Joseph P Simmons and Leif D Nelson. 2019. Data Replicada. http://datacolada.org/81.

[21] Joseph P Simmons, Leif D Nelson, and Uri Simonsohn. 2011. False-Positive Psychology Undisclosed Flexibility in Data Collection and Analysis Allows Presenting Anything as Significant. *Psychological Science* 22, 11 (Oct. 2011), 0956797611417632–1366.

[22] Uri Simonsohn, Joseph P Simmons, and Leif D Nelson. 2015. Specification Curve: Descriptive and Inferential Statistics on All Reasonable Specifications. *SSRN Electronic Journal* (2015).