Organizational Behavior and Human Decision Processes 164 (2021) 192–202

Contents lists available at ScienceDirect

Organizational Behavior and Human Decision Processes

journal homepage: www.elsevier.com/locate/obhdp

Expanding the scope of reproducibility research through data analysis replications☆

Jake M. Hofman*, Daniel G. Goldstein, Siddhartha Sen, Forough Poursabzi-Sangdeh, Jennifer Allen, Ling Liang Dong, Brenda Fried, Harpreet Gaur, Adnan Hoq, Emeka Mbazor, Naomi Moreira, Cindy Muso, Etta Rapp, Roymil Terrero

Microsoft Research New York City, United States

ARTICLE INFO

Keywords:
Reproducibility
Replication
Robustness
Education
Data analysis

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 appreciation that replicating other researchers' work is an essential 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 replications but *data analysis replications*. We do so for three reasons. First, experimental replication excludes entire classes of publications that do not run experiments or even collect original data (e.g., archival data analysis). Second, experimental replication may in some cases be a needlessly high bar: there is great value in replicating just the data analyses of published experimental work. As data analysis replications require a lower investment of resources than experimental replications, their adoption should expand the number and variety of scientific reproducibility studies undertaken. Third, we propose that teaching undergraduate students to perform data analysis replications will greatly increase the number of replications done while providing them with research experience that should inform their decisions to pursue research or to attend graduate school. Towards this end, we 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. In addition, we present a summary of ten additional data analysis replications carried out entirely by students in a university course.

1. Introduction

Recently, researchers across the sciences have been concerned that the results of published studies replicate less often than expected (Begley & Ioannidis, 2015; Maniadis, Tufano, & List, 2017; Shrout & Rodgers, 2018). This realization has presented the scientific community with both the challenge and the opportunity of improving how reproducible science should be 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 preregistration, 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 (Nosek, Ebersole, DeHaven, & Mellor, 2018). Publicly declaring the plans for a study forces researchers to think about these technicalities before any data are collected or analyzed, which can reduce (and ideally eliminates) the type of data-dependent decision making that can otherwise lead to high false discovery rates (Kerr, 1998; Simmons, Nelson, & Simonsohn, 2011). 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 (Gelman & Loken, 2014; Nosek et al., 2018).

Standards have also improved around how research results are shared with the community. For example, some journals now require

☆ This article is an invited submission. It is part of the special issue "Best Practices in Open Science," Edited by Don Moore and Stefan Thau.

* Corresponding author.
E-mail address: jmh@microsoft.com (J.M. Hofman).

https://doi.org/10.1016/j.obhdp.2020.11.003
Received 18 December 2019; Received in revised form 16 November 2020; Accepted 17 November 2020
Available online 13 April 2021

authors to submit research materials such as data and analysis code with their publications, making it easier for others to check, verify, and expand upon their work.[1] 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.

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 and not 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 previously described experiments (Open Science Collaboration, 2015). Experimental replications require recruiting new participants, collecting new data, and following the original analysis plan. 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 funded participant pool and are experienced in running 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 empirical, non-experimental papers, and reproducing their results has a much lower barrier to entry compared to reproducing experimental work. Given publicly available data, in principle all one needs to reproduce data analyses and results is access to and training with standard software packages to recreate the analysis plan in the original work. So in theory there should be a large number of data analysis replications—many more so than experimental replications—but in practice there are many fewer. Why is this the case? The blessing and curse of data analysis replications is that they require *less* skill and expertise than experimental replications, and ironically this may be exactly why they are not encouraged or rewarded by the academic research community.

It is our conviction that there should be more attempts at replicating data analyses. What would it take to get the scientific community to embrace them? Replications could come as a result of journals and research institutions rewarding this type of work, and there has been some progress in this direction, but it is safe to say that venues currently prioritize other research activities over data analysis replications. Here we suggest an alternative approach that recognizes that, given the current incentives in academic research, it may be difficult to get established researchers to undertake data analysis replications. Instead we propose a solution that relies on a large pool of individuals who could aid in this effort, and who would benefit from doing so in the process: undergraduate students.

Data analysis replications seem particularly well suited to undergraduate instruction. 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 class projects. Not only is this 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.

The plan of this paper is as follows. First, we more precisely define what we mean by data analysis replications and distinguish them from similar efforts to replicate published work. Next, we give an overview of a training program we piloted to teach students the skills needed to perform data analysis replications. As a case study, we report on the students' attempt to replicate a paper on disparities across racial groups in police use of force, discussing challenges faced along the way and lessons learned for generalizing the program to a larger audience. To show that our proposal is applicable to a wide set of studies, we present an overview of ten additional data analysis replications undertaken as course projects. Finally, we propose a best practice of "checkpointing" for data analyses, which we borrow from the field of computer systems research and conclude with some of the insights we gained from observing students conduct data analysis replications.

2. Data analysis replications

What do we mean by a data analysis replication? Before answering this question, we should note the point of this article 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 (Christensen, Freese, & Miguel, 2019; Goodman, Fanelli, & Ioannidis, 2016; Nosek & Errington, 2019; Patil, Peng, & Leek, 2019; Plesser, 2018).[4] 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 (Homer & Kneib, 2013; Simmons & Nelson, 2019). 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 Fig. 1, this is more involved than a reproducibility check that simply amounts to having a third party run the authors' *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 are an instance of "verification" according to the terminology of Christensen et al. (2019), who define verification as using the same data and "focusing on repeating procedures". Since the latter is underspecified—one could "verify" many different aspects of a paper (e.g., data collection, experimental protocol, etc.), a data analysis replication specifically refers to repeating the data analysis of a paper.

[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/).

[3] There are, however, a few notable efforts to engage undergraduates in re-running entire experiments in lieu of more experienced researchers (Button, 2018; IJzerman, Brandt, & Grahe, 2018).

[4] Here we use the definitions of reproducibility and experimental replication provided by the American Statistical Association (Broman et al., 2017) that have become commonplace, but these differ slightly from the definitions used by the National Science Foundation (Companion Guidelines on Replication & Reproducibility in Education Research, 2018).

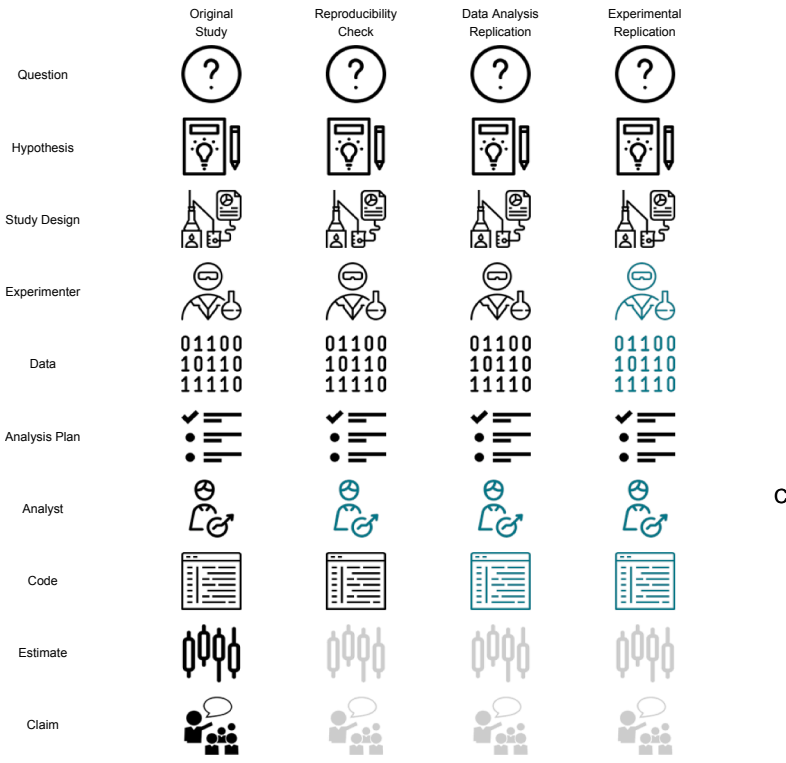

Fig. 1. A figure following Patil et al. (2019) to define what we mean by a data analysis replication. The first column depicts the stages of an original study. The second column defines a reproducibility check, where nearly everything is identical to the original study, but a third party analyst runs 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.

Data analysis replications apply to a broader range of scenarios than both reproducibility checks and experimental replications. For instance, data analysis replications apply to work that relies on archival data in addition to datasets generated from experiments. 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 source data as given. As we demonstrate with the case studies that we present later in this paper, undertaking a data analysis replication and writing new analysis code to follow an existing analysis plan can expose discrepancies and other issues that simply re-running existing analysis code might not reveal.

Another important distinction is between data analysis replications that start with source data and those which start with summary statistics. For instance, Bergh and colleagues (Bergh, Sharp, Aguinis, & Li, 2017) attempted to replicate the empirical findings of 88 papers but started with summary statistics published in papers (e.g., means, standard deviations, and correlations), as opposed to starting with source data and attempting to reproduce these summary statistics. While theirs is a valuable approach, and the only viable approach when source data are not available, our approach allows one to discover mistakes that might have occurred earlier in the analysis process. Clearly, different approaches to data analysis replication focus on different stages of the scientific process.

Fig. 2 is a simple flowchart to help determine whether a data analysis replication is possible in a given situation. The main requirement for a data analysis replication is an existing source dataset. This can come in two 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, every paper would include well-documented data and interpretable, easy-to-run, and correct code, making data analysis replications largely unnecessary. Unfortunately, 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 source data are available, but that the corresponding code is either unavailable or difficult to understand or re-run due, for example, to broken software dependencies.[5] We are concerned with the case 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 (Gelman & Loken, 2014; Hofman, Sharma, & Watts, 2017; Silberzahn et al., 2018). 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 (Simonsohn, Simmons, & Nelson, 2015)—leads to qualitatively different findings than the original paper. Finally, one can ask if the independent, dependent, and proxy variables recruited to make an argument about a construct are

[5] For an interesting example of this, see (Liu & Salganik, 2019), where 12 papers were submitted to a special issue that used the same dataset, 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.

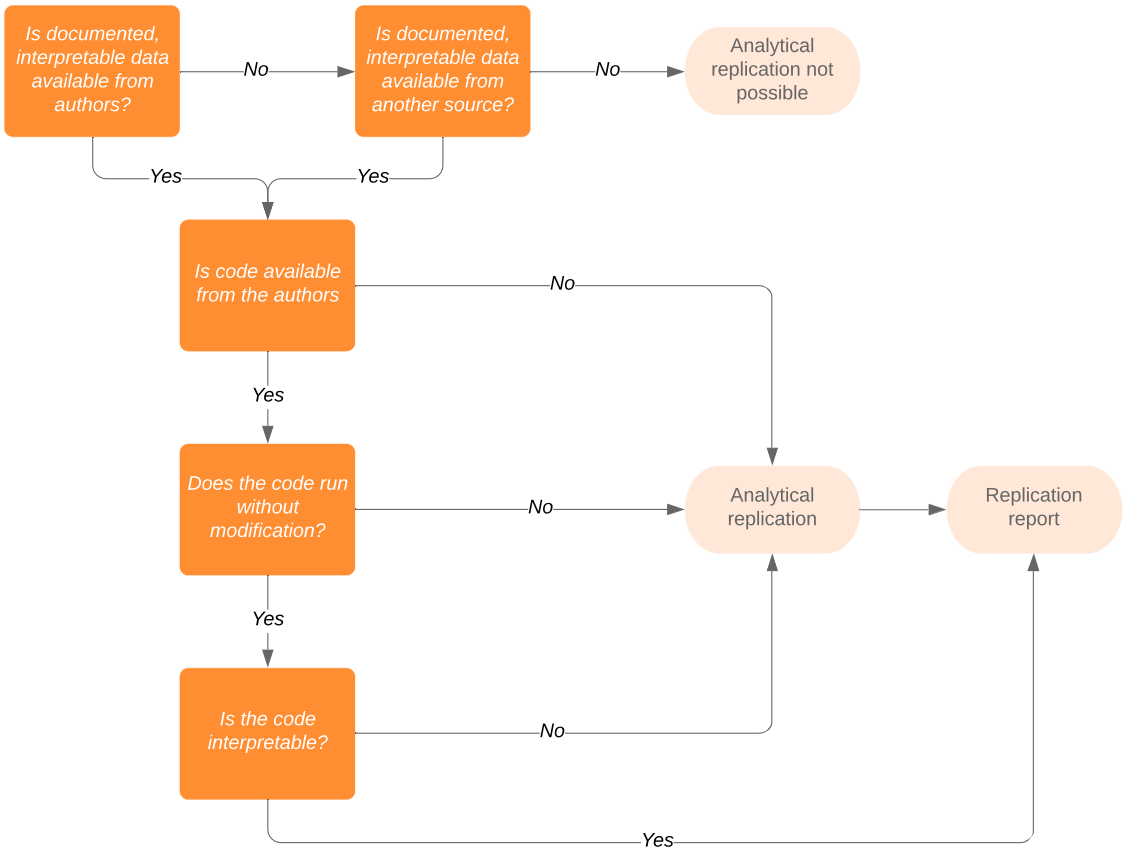

Fig. 2. A flow chart to determine if a data analysis replication is possible given information about the original data and code used in a paper. If well-documented data are available from either the authors or another source, the authors provide code, the code runs without modification, and the code 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.

appropriate (Dalton & Aguinis, 2013; Ketchen, Ireland, & Baker, 2013) and could imagine coding new analyses with more suitable variables. These scenarios are certainly relevant 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. Having laid out what we mean by data analysis replications, we will now relate experiences working with students to carry them out.

3. Piloting a training program with undergraduates

In this section, we discuss our experience piloting a substantial data analysis replication project with undergraduates. The project was part of the Microsoft Research Data Science Summer School (DS3),[6] which is aimed at increasing diversity in computer science and related fields. DS3 consists of four weeks of coursework followed by four weeks of work on an applied research project. The students were eight undergraduates who had a background in programming and familiarity with introductory probability and statistics. They had an interest in research before joining DS3, but most of them had not completed an original research project before doing so. This manuscript is jointly co-authored by both the instructors and the students from the 2019 DS3 program.

The first four weeks of the program were used to ensure the students had the skills that they would need to conduct a data analysis replication. This included tools for doing collaborative, reproducible research such as: the Unix command line for automated scripting; git and Github for version control; R for data analysis and statistics; and R Markdown for literate programming. To reinforce collaborative coding and good software engineering practices, the instructors used the repository hosting service Github to organize the course material, and the students used Github to submit all of their work as well.[7]

The students also learned concepts from statistics and machine learning ranging from introductory topics like expectations, variance, and statistical tests to more advanced concepts such as regression, classification, overfitting, and regularization. The statistics curriculum was based on simulations rather than asymptotic tests to emphasize conceptual understanding over rote memorization or procedural execution (Diez, Barr, & Cetinkaya-Rundel, 2014; Yakir, 2011). Students were exposed to the ongoing replication crisis in several fields, highlighting prominent examples of how flawed practices have led to unreliable results.

During these first four weeks, the students completed assignments that reinforced these topics while also gradually building their abilities to do data analysis replications. For instance, after they learned exploratory data analysis in R—i.e., how to tabulate and plot large datasets—they were asked to replicate the results of a published paper that required only these skills, with results that were known to replicate.[8] Later on the students conducted a more involved replication assignment that required more effort in terms of obtaining and cleaning the data, and for which the conclusions of the paper were more sensitive

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

[7] https://github.com/msr-ds3/coursework.

[8] https://github.com/msr-ds3/coursework/tree/2019/week2#the-anatomy-of-the-long-tail.

to the particular analysis choices made by the authors.[9] The students read about past replication attempts and the types of critiques that have been made around brittle results in the past (Coupe, 2018; Open Science Collaboration, 2015). All of this gave them a chance to think critically about published work, which was a substantial change from the textbook-based approach to statistics and statistical thinking they had seen in the past.

Upon successfully completing these assignments, the students not only had a thorough understanding of the underlying material, but it was also clear that they were capable of carrying out data analysis replications. This was achieved during the course of four weeks, with approximately two hours of lecture each weekday morning and several hours of independent exercises each afternoon. All of this is to say that it is entirely feasible to teach undergraduate students to do data analysis replications in a relatively short amount of time (e.g., a one semester course).

4. Case study: Replicating "An Empirical Analysis of Racial Differences in Police Use of Force"

The students spent the last four weeks of the summer program replicating and extending the analyses in an academic paper—"An Empirical Analysis of Racial Differences in Police Use of Force" (Fryer, 2019)—referred to as the "Policing Study" hereafter. 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 Fig. 1), but also used relatively simple methodology that seemed straightforward to implement and check, relied on two publicly available datasets,[10] 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, however, 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 transformed into independent variables. These challenges came despite the extensive documentation in the paper and its appendices 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.

In what follows, we organize the challenges and discrepancies that the students faced into four categories, based on the stage of analysis they pertain to: raw data, featurization, statistical modeling, and interpretation. Raw data refers to the content and summary properties of the original data used by the study, prior to any manipulation or synthesis. Featurization refers to the manipulation of this data (e.g., encoding categorical values) to generate features that can be used for statistical modeling. Statistical modeling refers to the methods used to fit statistical models (e.g., logistic regression) on the featurized data. Interpretation refers to how the results of the fitted models are interpreted and communicated in the paper.

Raw data. The Policing Study analyzed two publicly-available datasets. The first are police-reported encounters from New York City's Stop, Question, and Frisk program (Stop and Frisk hereafter), a program running from 2003–2019 that allowed NYPD officers to stop and question a pedestrian, and then possibly frisk them for weapons or contraband. The dataset includes information about civilian race, characteristics of the encounter, and the degree and type of force used, ranging from placing hands on a civilian to hitting them with a baton.[11] The second dataset is the Police-Public Contact Survey (PPCS hereafter), a survey of a nationally representative sample of civilians conducted by the Bureau of Justice Statistics every three years from 1996–2015. The dataset includes civilian reports of interactions with the police, some of which also involved force.[12] Both datasets are available for download from the corresponding program's website.

Since the datasets spanned multiple years, the students had to download and combine multiple data files, one for each year. In this very simple act alone, discrepancies began to arise. The students counted a total of 4,982,825 datapoints in the Stop and Frisk dataset and a total of 409,678 datapoints in the PPCS dataset; in contrast, the Policing Study counted 4,982,925 datapoints (0.002% more) and 426,000 datapoints (3.8% more), respectively. The data files of some years had missing columns and mismatched column names compared to the data of other years; several columns had values that were nonsensical (e.g., civilian ages greater than 200 or civilian race labeled as "male"); and some values were inconsistent across the years.

Though seemingly trivial, these discrepancies added significant overhead and uncertainty to the replication process. The Policing Study did not provide details about how they handled these discrepancies or cleaned the data; as a result, the students spent a considerable amount of time experimenting with different cleaning methods to match the number of datapoints in the original study. The students contacted the respective government programs to inquire about possible changes to the raw data, but these contacts confirmed that the data had not been changed and the totals computed by the students were accurate. The consequence of this was that every downstream discrepancy encountered during the replication process carried with it the uncertainty in the possibility of being caused by discrepancies in the raw data. This added substantial *debugging overhead* to the replication process.

Although the students did not have access to the cleaning process or the cleaned data, the Policing Study provided summary statistics in the appendix of various categorical values such as civilian race, characteristics of the encounter, and type of force used. Due to the tiny fraction of missing datapoints in the Stop and Frisk dataset, the students were able to replicate all but a few of the summary statistics. The PPCS dataset suffered from a larger fraction of missing data, causing challenges in replicating a significant number of summary statistics.

Featurization. After cleaning the raw data to the best of their abilities, the students used their cleaned data to derive the features (independent variables) used by the Policing Study to fit statistical models. Some of these features mapped directly to the raw data values, but others had to be interpreted and encoded based on the raw values, leading to another source of discrepancies. In the Stop and Frisk dataset, the students had trouble encoding civilian age and race. For example, the Policing Study encoded "White Hispanic" as "Hispanic" but encoded "Black Hispanic" as "Black", which was not explicitly documented in the manuscript.

Featurizing the PPCS dataset was more problematic. The dataset used a varying number of columns to describe race across the years: for example, the 2005 dataset used 2 columns while the 2011 dataset used 12. No details were provided in the manuscript for how to encode race consistently across the years. Another feature used by the study was a

[9] https://github.com/msr-ds3/coursework/tree/2019/week2/ngrams, following Chapter 2 Exercise 6 in Salganik (2017).

[10] 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.

[11] https://www1.nyc.gov/site/nypd/stats/reports-analysis/stopfrisk.page.

[12] https://www.bjs.gov/index.cfm?ty=dcdetail&iid=251.

categorical variable encoding stops as either "traffic stop" or "street stop" based on the reason for stopping. The raw data had 15 different reasons for stopping, but no mapping was provided between them and the two categories. The reasons ranged from ones that were clearly traffic stops (e.g., "involved in traffic accident"), clearly street stops (e.g., "jaywalking"), either/or (e.g., "bike violation"), or seemingly orthogonal (e.g., "seeking information"). The consequence of these discrepancies is that the students remained unconfident about several of the features they used in the next stage of the analysis, statistical modeling.

Statistical modeling. The Policing Study performed a series of logistic regressions to relate police use of force (encoded as a yes/no variable) to different sets of features, showing the effect of controlling for these sets of features on whether force was used. The series of regressions are summarized in the rows of Fig. 3, beginning with race as the only control variable, and then adding civilian demographics, encounter characteristics, civilian behavior, and so on. The regression procedure was sufficiently well defined that the students were able to perform it without much difficulty.

The columns of Fig. 3 show the estimates from the Policing Study for the Stop and Frisk dataset (top) and the estimates derived by the students (bottom). Although none of the estimates replicated exactly, the trends observed were similar to those found in the study. The first column ("white mean") represents the percentage of stops of white civilians that involved use of force (the study calculated 15.3%, the students calculated 16.1%). The remaining columns report the odds ratios for different races relative to the white mean. As more controls are added, the discrepancy between races diminishes but does not disappear completely. This replicates the Policing Study's primary finding, that black and hispanic civilians are more likely to experience police use of force than white civilians, even when controlling for different context variables. The students obtained similar data analysis replication results for the PPCS data.

As discussed earlier, the discrepancies in the regression estimates can partly be explained by discrepancies in the raw data, the data cleaning process, and the featurization process. However, another source of error could arise from the specific implementation of logistic regression used by the students versus the original study. Accounting for implementation differences in modeling packages can be resolved by providing access to the source code and the *same code execution environment* as the original study. Containerization technologies like Docker (Liu & Salganik, 2019) can be used to facilitate this kind of replication.

Interpretation. Having approximately reproduced the modeling results in the Policing Study, the students next thought about interpreting these results. In particular, one oddity in the table is that the columns show two very different types of quantities related to probabilities, which can lead to confusion. Specifically, the first column of the first row shows the average *probability* that a white civilian in the dataset experienced use of force—given as 0.153, or 15.3%—while the remaining columns give the *odds ratios* for civilians of other races to experience use of force (e.g., 1.534 for black civilians, from the second column of the first row of Table 2).[13] The paper uses these odds ratios to make statements about the relative likelihood of use of force—e.g., "Blacks are 53 percent more likely to experience any use of force relative to a white mean of 15.3 percent"—derived by comparing the estimated odds of 1.534 for black civilians to a hypothetical odds ratio of 1.0, which would indicate that white and black civilians have the same odds of experience use of force.

In discussing this with the students, it became clear that the paper compared *odds ratios* for black and white civilians, but presented the result as if these were *risk ratios* (i.e. ratios of probabilities). This is a common issue in communicating the output of logistic regression analyses, and can result in potentially misleading statements about relative probabilities, especially as the absolute value of the underlying probabilities becomes large (Davies et al., 1998; McNutt, Wu, Xue, & Hafner, 2003). To determine the actual probability ratio, the students used the fitted model to compute the probabilities that civilians of each race would experience use of force. For black civilians, this probability is 21.8%. Comparing this to the average of 15.3% for white civilians yields a ratio of 1.42, or a 42% higher probability for black civilians to experience use of force compared to white civilians. While a 42% higher probability for black civilians compared to white civlians is still an alarmingly large difference, we were quite surprised to learn of this discrepancy in what became the main abstract-level finding of the paper and was subsequently quoted in several major news outlets (Ehrenfreund & Guo, 2016).

4.1. Comparing to the authors' code

After the students completed their replication attempt we gained access to the authors' original source code, which was located behind a paywall that was previously inaccessible to us. Despite having the full manuscript and appendix that totaled more than 100 pages, we were unable to resolve some of the discrepancies between the results of our replication attempt and the original paper until we gained access to this code. Surprisingly, even then some differences remained.

For instance, having access to the code helped in resolving the interpretation of odds ratios mentioned above. Without the code, it was unclear if it was raw odds ratios that were reported in the text or if these were mean marginal effects, in which case the interpretation would have been different. This was a source of confusion that slowed down the replication study, but it showed the students that the interpretation of results is an important part of assessing a paper's replicability.

Access to the code also resolved the sizeable difference in the number of observations in the PPCS data between our replication attempt and the original paper. Upon seeing the code it became clear why the Policing Study had 27,000 more observations than ours did, which came down to a difference in how files were formatted across years. In particular, in 1999, people were asked about both traffic stops and other stops, whereas in other years they were asked only about their most recent stop. Counting traffic stops and other stops separately in 1999 resulted in our dataset having 425,903 observations, which is much closer to the approximately 426,000 observations reported in the published paper.

At the same time, if we had access to the authors' code in the first place and did a reproducibility check by simply re-running the code, we might not have uncovered some of the issues that we found by doing our own data analysis replication. This is an important benefit of data analysis replications over more standard reproducibility checks, even when the original code for a paper is available.

4.2. Reproducible replication results

In an effort to make our own work reproducible, we have created a Github repository with all of the materials necessary to repeat our analyses.[14] This includes scripts to download the raw data from its original sources, code to clean the raw data according to the methodology used in the Policing Study, and code to fit and analyze the corresponding models. Our code is documented and contains a "Makefile" that runs all of these analyses.

[13] As a reminder, odds are computed by comparing the probability that an event occurs to the probability that it does not occur (i.e., $o(\text{event}) = \frac{p(\text{event})}{1-p(\text{event})}$), and odds ratios compare the odds for one event to another (e.g., $\frac{o(\text{force}|\text{black})}{o(\text{force}|\text{white})}$). This is not the same as the corresponding probability ratio (e.g., $\frac{p(\text{force}|\text{black})}{p(\text{force}|\text{white})}$). In particular, odds ratios provide an overestimate of probability ratios (Davies, Crombie, & Tavakoli, 1998).

[14] https://github.com/msr-ds3/stop-question-frisk.

Original study results

	White mean	Black	Hispanic	Asian	Other race
No controls	0.153	1.534	1.582	1.044	1.392
+Civilian Demographics		1.480	1.517	1.010	1.392
+Encounter Characteristics		1.655	1.641	1.059	1.452
+Civilian Behavior		1.462	1.516	1.051	1.372
+Precinct and Year Fixed Effects		1.178	1.122	0.953	1.060

Our data analysis replication results

	White mean	Black	Hispanic	Asian	Other race
No controls	0.161	1.500	1.548	1.047	1.322
+Civilian Demographics		1.461	1.503	1.021	1.303
+Encounter Characteristics		1.644	1.611	1.060	1.415
+Civilian Behavior		1.483	1.519	1.073	1.374
+Precinct and Year Fixed Effects		1.203	1.140	0.975	1.074

Fig. 3. Results of the logistic regression from Table 2 of the Policing Study (left) and our attempt to replicate these results (right). The first column of the first row in each table gives the estimated unconditional average proportion of white civilians in the Stop, Question, and Frisk dataset that experienced use of force from a police officer. Following the presentation in Fryer (2019), the remaining cells give the odds ratios for civilians of different races to experience use of force compared to white civilians. Columns give these odds ratios for different races, and rows correspond to increasingly complex sets of control variables to try to eliminate confounds. We note that our estimates are roughly in line with the original estimates, and discuss the subtle interpretation of these results i.n the main text.

To facilitate reproducibility, our code specifies all software packages required to run the code, along with version numbers of the packages that were used at the time this manuscript was written. In the future, we plan to containerize the code in an environment that will automatically provide these packages, to make reproducibility checks even easier for future researchers (Liu & Salganik, 2019). Although we have not implemented all of the best practices prescribed by prior work, there currently is no universal standard for which (subset of) practices should be implemented in which scenarios. The goal of our work is not to set such a standard, but instead to emphasize the value of data analysis replications and convey what we have learned from them.

A particularly useful aspect of our code is that it creates "checkpoints" of the datasets and results after each stage of the data analysis replication. This allows the analyst to check their work after each stage of the replication before proceeding to the next. It also has other important benefits, motivating us to propose it as a best practice for reproducibility later on in the paper.

5. Data analysis replications of ten additional papers

Complementing the data analysis replications discussed above, we present here a summary of ten additional data analysis replications performed by undergraduates and masters students as the final project in a semester-long course on computational social science at Columbia University in 2019. Replications were done in groups of three or four students. Each group covered a different paper chosen from disciplines ranging from economics, political science and law to computer science. The outcome of these replication attempts confirmed that undergraduate students are capable of carrying out independent data analysis replications. Table 1 provides information about each paper along with a summary of the challenges faced by each group, with further details of the replication attempts provided in Table 2.[15]

For each of these papers we document issues that the students faced in replicating the results of the corresponding paper, showing similar challenges to those faced in replicating the Policing Study. For instance, several teams had difficulty getting access to the original datasets used in a paper, as they were either no longer available or were available in some modified or less detailed format. The most common issue that arose across replication attempts was dealing with what data to include or exclude in the analysis. In some cases, this issue was due to unspecified procedures for dealing with outliers or dataset imbalance, while in other cases it involved observations being implicitly dropped during regression analyses because of missing values for some covariates. Once datasets were obtained, the process of cleaning and coding them presented challenges as well. For example, the students were uncertain about how exactly the original authors operationalized time-based variables in mixed-effects models.

Table 1
Challenges faced by students in ten other data analysis replications as part of an undergraduate course.

Paper	Data availability	Missing data and outliers	Data coding and cleaning	Statistical methods	Interpretation
Depken (2000)	✘	✘	✘		
Fearon and Laitin (2003)			✘		✘
Collier and Hoeffler (2004)		✘	✘		✘
Leskovec et al. (2010)		✘			
Choi and Varian (2012)	✘	✘			
Muchlinski et al. (2016)				✘	✘
Cattaneo et al. (2009)		✘		✘	✘
Davidson et al. (2017)	✘			✘	
Clauset et al. (2015)				✘	
Penney (2016)	✘	✘		✘	✘

Ultimately students were able to resolve the majority of these issues and replicate many of the results they sought to reproduce, at least qualitatively if not exactly. This, however, misses the even more important point that the students had the necessary skills to undertake these replication attempts in the first place, regardless of the outcome of each attempt. While each group faced challenges, nearly all of them had to do with the paper or dataset itself, as opposed to the students' own errors. In examining these replication reports in detail, we found only one group that clearly introduced their own error (by misunderstanding how to calculate Gini coefficients) as opposed to uncovering existing ambiguities or issues with the original paper or data.

As part of these replication attempts, students also spent time trying to extend and think critically about the results published in each paper,

[15] A full report of each replication attempt can be found here: http://modelingsocialdata.org/lectures/2019/05/17/final-project-reports-2019.html.

Table 2
Details of each replication attempt.

Paper	Main question addressed in paper	Insights from students' replication attempt
Depken (2000)	Should baseball teams invest in a few star players or instead focus on minimizing wage disparity across its players?	Exact data were not available, so a secondary data source was used; it was difficult to determine exactly which observations were included in analysis and unclear how certain covariates were operationalized (e.g., fixed effects for years); despite these issues, students reported qualitatively similar results; extension shows results are similar for alternative measures of wage disparity, but that the models have limited predictive accuracy
Fearon and Laitin (2003)	What factors explain if a country is at risk of civil war?	Students found an error in the coding of the dataset provided by the authors, but were able to address the issue and reported that the stated results replicated exactly; the fitted models were found to have relative low predictive power
Collier and Hoeffler (2004)	What causes civil wars?	It was unclear how certain variables were operationalized (e.g., fixed effects for time) and students found that observations were implicitly dropped as more and more variables were including in the modeling process, but they reported qualitatively similar results; predictive checks again yielded limited results
Leskovec et al. (2010)	How well can one infer whether a social relationship has a positive or negative affinity?	Datasets were provided by the authors, but summary statistics differed somewhat from what was reported in the paper; replicating revealed that some data were implicitly excluded from the analyses; predictive performance of models had the same qualitative trend, but did not match the original paper in overall level; extension showed decreased predictive performance over longer time periods
Choi and Varian (2012)	How good of a proxy is search data for real-world activity?	A version of the data were publicly available, but differed from what was used in the paper due to different normalization and time granularity; students reported qualitatively similar results but no exact numerical agreement; extension showed decreased predictive performance over longer time periods
Muchlinski et al. (2016)	How well do different models do in predicting civil war?	Impossible to determine train/test split for data because of missing random seeds and split ratio; original paper mistakenly evaluates performance on the training (instead of test) data; feature importance analysis showed the same variables, but in different order; both the type of model and the features used in them changed, making it difficult to isolate which accounts for performance differences
Cattaneo et al. (2009)	What were the effects of a government program to improve living conditions in low income areas on health and happiness?	Students found some issues with how missing values were dealt with, but results replicated quite closely once these were resolved; details of statistically methodology were not clear upon closer inspection, extension showed that some assumptions necessary for the natural experiment analysis may not hold; predictive power of the models was somewhat limited
Davidson et al. (2017)	How well can one automate detection of hate speech vs. offensive speech?	Some issues getting the original data (pointers to tweets that were subsequently deleted) but students reported that results mostly replicated; found that authors did not use a true held out test set in evaluating performance; extension showed that explicitly dealing with dataset imbalance improved results over original results
Clauset et al. (2015)	How does academic hiring vary with gender and institutional prestige?	Many results replicated successfully using author-provided data, students had conceptual difficulty calculating Gini coefficients; extension included predictive model with similar results to descriptives reported in the paper
Penney (2016)	Did the revelation of online surveillance by the US government result in fewer visits to sensitive topics on Wikipedia?	Original data source was no longer available, alternative was used; dealing with outlier removal affected results; students reported that some parts of the paper replicated while others did not; an extension looking at a longer timespan of the data shows subtleties in interpreting the results

which raised questions about the assumptions underlying statistical analyses, issues in how analyses were carried out, and concerns about the interpretation of results. For example, students found that a paper that claimed to do out-of-sample testing of predictive models—fitting a model on a training set and testing it on a held-out test set—mistakenly tested the model on the training data instead of the test set, as was independently reported by established researchers who sought to replicate the same paper (Neunhoeffer & Sternberg, 2019). Others had concerns surrounding the assumptions necessary for natural experiment analyses. For instance, one team found that authors had made an "as-if random" assumption in the assignment of different housing units to a government program, but a simple logistic regression could predict whether a housing unit would (or would not) be assigned to the program based on the features of the home significantly better than chance. Several teams also found differences between the (statistical) significance of results claimed in a paper and the practical importance of the findings. In some papers, despite statistically significant coefficient estimates, the corresponding models were found to have very limited predictive power (Ward, Greenhill, & Bakke, 2010), whereas in others, effects that held over the time period studied in the paper did not endure when examined over longer time frames.

Considering these challenges alongside those faced in the Policing Study, we see a fairly consistent pattern: undergraduates are more than capable of undertaking data analysis replications, and having them do so raises interesting questions about different parts of the data analysis pipeline, providing an opportunity to simultaneously educate students

and improve the quality of published research.

6. Proposed best practice: Checkpointing

We develop a technique for enhancing reproducibility based on our experience replicating the data analysis of the Policy study and the studies in Table 1. The idea is to create *checkpoints*, or "snapshots" of the data and derived results at various points during the replication process. Checkpoints allow an analyst to check their work incrementally, isolate and debug discrepancies effectively, and perform replications in parallel. Our use of checkpointing is inspired by a common technique in database systems for tolerating failures (Gray & Reuter, 1992), and is distinct from the practice of code versioning for workflow reproducibility as suggested by Christensen et al. (Christensen et al., 2019). As we noted earlier, data analysis replications involve writing new analysis code, not running the same analysis code written by the authors. Checkpointing facilitates the writing of this new code, by making the replication process more modular and efficient.

In a database system, checkpoints are used to increase the efficiency of fault tolerance and recovery. The system is modeled as a deterministic state machine that starts at an initial state and processes a sequence of commands in a particular order (Schneider, 1990). If a failure occurs, the system can be recovered by starting from the initial state and replaying the commands in the same exact order (assuming the commands have been persistently logged). If we periodically checkpoint (and log) the system's state, however, then recovery can be completed much more quickly, by initializing the system with the most recent checkpoint and only replaying the commands that occurred after that checkpoint. Checkpoints thus increase the efficiency of failure recovery. They can also be used to isolate and debug problems in the system: if the system is correct at checkpoint i but not at checkpoint $i + 1$, then a problem must have occurred while processing commands between i and $i + 1$.

In a data analysis replication, the original analysis is the "system", the analysis steps are the "commands", and the "state" is the data and derived results. Just like in a database system, the commands manipulate the system's state—in this case, the analysis steps manipulate the raw data to generate derived data and results. Thus, by adopting the practice of checkpointing, the authors of a study can periodically save the state of their analysis. For example, the raw data can be saved as the first checkpoint, the cleaned data can be the second checkpoint, the summary statistics and featurized data can be the third, and so on. By recording these checkpoints and including them *as part of a data/code release*, the authors can greatly facilitate data analysis replications of their work. In particular, checkpoints allow a data analyst to perform an analogous set of tasks to the failure recovery tasks described above; a "failure" in our setting is a discrepancy in the replication process. Suppose checkpoints $C_0, C_1, \ldots, C_n$ are provided as part of a study's release. Using these checkpoints, a data analyst can perform the following tasks:

1. Resume replication from the most recent checkpoint if/when a discrepancy occurs, rather than starting from the beginning (C_0). In general, replication can be performed between any pair of checkpoints C_i and C_j, e.g., if the analyst wishes to only replicate a subset of the analysis.
2. Isolate a discrepancy to a single checkpoint period, enabling faster debugging. If a replication is consistent up to C_i but not up to C_{i+1}, then one of the analysis steps between C_i and C_{i+1} is the likely cause of the discrepancy.
3. Replicate different parts of the analysis in *parallel*. By starting/ending at disjoint checkpoint periods, multiple analysts can simultaneously replicate disjoint parts of the analysis.

The third task is particularly conducive to a replication setting, where the goal is to replicate all parts of a data analysis: replicating disjoint checkpoint periods in parallel can greatly speed up the replication process. (In a database systems setting, the goal is typically to recover the system to the latest state, which typically requires only the most recent checkpoint.) Using checkpoints, we can parallelize a data analysis replication over multiple analysts (e.g., multiple undergraduate students in our program) as follows. The first student replicates the analysis from C_0 to C_1, the second student replicates from C_1 to C_2, the third from C_3 to C_4, and so on—all in parallel with each other. Fig. 4 illustrates this process. As the figure shows, the checkpoints act as the "interfaces" along which individual pieces of the data analysis replication connect to form the entire replication.

What information should the authors capture in a checkpoint, and how often should checkpoints be generated? A checkpoint should contain data and derived results that are independent of the analysis code written by the authors; in other words, it should be portable across different analysis scripts. This is because a data analysis replication involves writing new analysis code that attempts to replicate the checkpoint. Thus, besides formatting differences in the output of the new and original analysis codes, the contents of a checkpoint should be straightforward to compare across different analysis codes that generate it. A good candidate for the initial checkpoint is the raw data, which is typically provided by the authors (or publicly available) and hence replicable by definition. The next checkpoint could be a snapshot of the data after it has been cleaned; new analysis code is already required to replicate this checkpoint. The next checkpoint could be summary statistics and features encoded from the cleaned data, and so forth.

The granularity at which checkpoints are generated by the authors affects the ease and efficiency of data analysis replications. The finer the granularity, the closer a data analyst can check their replication work against the original study, the replication can be parallelized over more analysts, and discrepancies can be isolated to a smaller number of analysis steps. However, generating a checkpoint after every data transformation is clearly impractical for both the authors and the analyst performing a data analysis replication. Fig. 4 shows an example where two checkpoints are generated during the data cleaning stage instead of just one. This reflects our experience from replicating the Policing Study and the studies in Table 1, where the most common replication issue that arose was dealing with missing data, outliers, and other data cleaning artifacts. Consequently, one conclusion of our data analysis replications is that more checkpoints are needed during the data cleaning stage. Since checkpoints should be independent of the analysis code that generated them, a data analysis replication can contribute new checkpoints to a study in between the existing checkpoints provided by the authors (or by a previous data analysis replication). For example in Fig. 4, checkpoint 1 may have been contributed by a data analysis replication that found the leap between checkpoints 0 and 2 to be too large.

7. Conclusion

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 and increase the number of replication projects overall.

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 could be carried out quickly, we have shown through a case study of a month-long replication of a well-documented recent paper (along with ten additional data analysis replication examples) that many obstacles can stand in the way of such efforts. This was made particularly clear by the way the replication was

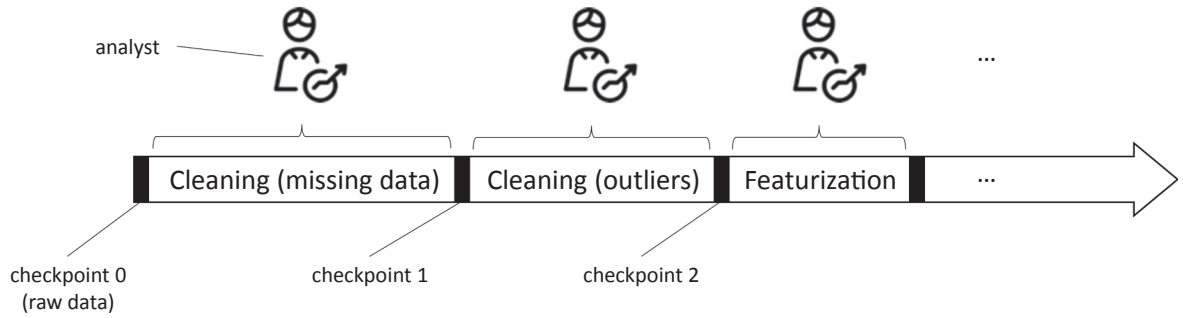

Fig. 4. Our proposed best practice: checkpointing. By providing checkpoints of the data and derived results at periodic stages of an analysis, authors can greatly improve the ease and efficiency of subsequent data analysis replications. In the figure, multiple analysts replicate different parts of the analysis in parallel; these efforts interface at the checkpoints to stitch together a complete replication.

attempted, both before and after looking at the author's published analysis code. For example, 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 obstacle was getting fitted model coefficients to match those in published tables and understanding how the authors interpreted these results. 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 was based on a non-standard way of communicating risk. Specifically, an event in the Policing Study that was stated to be 53% more likely was found to be only 42% more likely 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 revealed 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.

Based on our experience replicating several data analyses, we propose a best practice for authors called "checkpointing"—inspired by the eponymous practice in database systems—which entails taking snapshots of the data and derived results at periodic stages of an analysis. Checkpoints improve the ease and efficiency of a data analysis replication by allowing analysts to incrementally check their work, isolate discrepancies, and parallelize the replication effort.

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. Several of the co-authors of this article were undergraduates when this data analysis replication was conducted and 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 question 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.

In closing, it is our hope that this work will inspire data analysis replications across a wide range of fields. Specifically, we envision building an open, online platform that will enable anyone to learn the skills necessary for doing data analysis replications. This platform would serve as a reference for teachers, students, and the broader public. It would contain course material for training students to do replications that can be freely used in university courses. It would also contain a repository of research papers that are candidates for data analysis replications, along with the results of any attempts to replicate those papers. In the short term, this would give the research community and the broader public an easy way to assess the reliability of published results, and in the long term we hope it will lead to the publication of better, more reliable, and more robust research.

CRediT authorship contribution statement

Jake M. Hofman: Conceptualization, Validation, Resources, Writing - original draft, Writing - review & editing, Visualization, Supervision, Project administration. **Daniel G. Goldstein:** Conceptualization, Validation, Resources, Writing - original draft, Writing - review & editing, Visualization, Supervision, Project administration. **Siddhartha Sen:** Conceptualization, Validation, Resources, Writing - original draft, Writing - review & editing, Visualization, Supervision, Project administration. **Forough Poursabzi-Sangdeh:** Conceptualization, Validation, Resources, Writing - original draft, Writing - review & editing, Visualization, Supervision, Project administration. **Jennifer Allen:** Supervision, Project administration, Software, Validation. **Ling Liang Dong:** Supervision, Project administration, Software, Validation. **Brenda Fried:** Methodology, Software, Validation, Formal analysis, Investigation, Data curation, Visualization. **Harpreet Gaur:** Methodology, Software, Validation, Formal analysis, Investigation, Data curation, Visualization. **Adnan Hoq:** Methodology, Software, Validation, Formal analysis, Investigation, Data curation, Visualization. **Emeka Mbazor:** Methodology, Software, Validation, Formal analysis, Investigation, Data curation, Visualization. **Naomi Moreira:** Methodology, Software, Validation, Formal analysis, Investigation, Data curation, Visualization. **Cindy Muso:** Methodology, Software, Validation, Formal analysis, Investigation, Data curation, Visualization. **Etta Rapp:** Methodology, Software, Validation, Formal analysis, Investigation, Data curation, Visualization. **Roymil Terrero:** Methodology, Software, Validation, Formal analysis, Investigation, Data curation, Visualization.

Acknowledgements

We would like to thank Gabe Perez-Giz for his helpful ideas and conversations about this project. We also thank the anonymous reviewers of this journal for their insightful feedback and pointers to

related work.

Appendix A. Supplementary material

Supplementary data associated with this article can be found, in the online version, at https://doi.org/10.1016/j.obhdp.2020.11.003.

References

Begley, C. G., & Ioannidis, J. P. (2015). Reproducibility in science: Improving the standard for basic and preclinical research. *Circulation Research, 116*(1), 116–126.

Bergh, D. D., Sharp, B. M., Aguinis, H., & Li, M. (2017). Is there a credibility crisis in strategic management research? Evidence on the reproducibility of study findings. *Strategic Organization, 15*(3), 423–436.

Broman, K., Cetinkaya-Rundel, M., Nussbaum, A., Paciorek, C., Peng, R., Turek, D., & Wickham, H. (2017). Recommendations to funding agencies for supporting reproducible research. *American Statistical Association, 2*.

Button, K. (2018). Reboot undergraduate courses for reproducibility. *Nature, 561*(7723), 287–288.

Cattaneo, M. D., Galiani, S., Gertler, P. J., Martinez, S., & Titiunik, R. (2009). Housing, health, and happiness. *American Economic Journal: Economic Policy, 1*(1), 75–105.

Choi, H., & Varian, H. (2012). Predicting the present with google trends. *Economic Record, 88*, 2–9.

Christensen, G., Freese, J., & Miguel, E. (2019). *Transparent and reproducible social science research: How to do open science*. University of California Press.

Clauset, A., Arbesman, S., & Larremore, D. B. (2015). Systematic inequality and hierarchy in faculty hiring networks. *Science Advances, 1*(1), el400005.

Collier, P., & Hoeffler, A. (2004). Greed and grievance in civil war. *Oxford Economic Papers, 56*(A), 563–595.

Companion guidelines on replication and reproducibility in education research. (2018). https://www.nsf.gov/pubs/2019/nsfl9022/nsfl9022.pdf.

Coupe, T. (2018). Replicating Predicting the present with Google trends by Hyunyoung Choi and Hal Varian (The Economic Record, 2012). *Economics: The Open-Access, Open-Assessment E-Journal* 12(2018–34), 1–8.

Dalton, D. R., & Aguinis, H. (2013). Measurement malaise in strategic management studies: The case of corporate governance research. *Organizational Research Methods, 16*(1), 88–99.

Davidson, T., Warmsley, D., Macy, M., & Weber, I. (2017). Automated hate speech detection and the problem of offensive language. In *Eleventh international AAAI conference on web and social media*.

Davies, H. T. O., Crombie, I. K., & Tavakoli, M. (1998). When can odds ratios mislead? *BMJ, 316*(7136), 989–991. https://doi.org/10.1136/bmj.316.7136.989.

Depken, C. A., II (2000). Wage disparity and team productivity: Evidence from major league baseball. *Economics Letters, 67*(1), 87–92.

Diez, D. M., Barr, C. D., & Cetinkaya-Rundel, M. (2014). Introductory statistics with randomization and simulation. *Openlntro*. https://www.openintro.org/stat/textbook.php.

Ehrenfreund, M. & Guo, J. (2016). *How a controversial study found that police are more likely to shoot whites, not blacks*. The Washington Post. Retrieved July 13, 2016, from https://www.washingtonpost.com/news/wonk/wp/2016/07/13/why-a-massive-new-study-on-police-shootings-of-whites-and-blacks-is-so-controversial/.

Fearon, J. D., & Laitin, D. D. (2003). Ethnicity, insurgency, and civil war. *American Political Science Review, 97*(1), 75–90.

Fryer, R. G. (2019). An empirical analysis of racial differences in police use of force. *Journal of Political Economy, 127*(3), 1210–1261. https://doi.org/10.1086/701423.

Gelman, A., & Loken, E. (2014). The statistical crisis in science. *American Scientist, 102*(6), 460.

Goodman, S. N., Fanelli, D., & Ioannidis, J. P. A. (2016). What does research reproducibility mean? *Science Translational Medicine, 5*(341). https://doi.org/10.1126/scitranslmed.aaf5027, 341psl2-341psl2.

Gray, J., & Reuter, A. (1992). *Transaction processing: Concepts and techniques*. Elsevier.

Homer, J. H., & Kneib, T. (2013). *Economics needs replication*.

Hofman, J. M., Sharma, A., & Watts, D. J. (2017). Prediction and explanation in social systems. *Science, 555*(6324), 486–488.

IJzerman, H., Brandt, M. J. & Grahe, J. E. (2018). *How to make replications mainstream*. https://doi.org/10.31234/osf.io/rwufg.

Kerr, N. L. (1998). Harking: Hypothesizing after the results are known. *Personality and Social Psychology Review, 2*(3), 196–217.

Ketchen, D. J., Jr, Ireland, R. D., & Baker, L. T. (2013). The use of archival proxies in strategic management studies: Castles made of sand? *Organizational Research Methods, 16*(1), 32–42.

Leskovec, J., Huttenlocher, D., & Kleinberg, J. (2010). Predicting positive and negative links in online social networks. In *Proceedings of the 19th international conference on World wide web* (pp. 641–650).

Liu, D. & Salganik, M. (2019). *Successes and struggles with computational reproducibility: Lessons from the fragile families challenge*.

Maniadis, Z., Tufano, F. & List, J. A. (2017). *To replicate or not to replicate? exploring reproducibility in economics through the lens of a model and a pilot study*.

McNutt, L.-A., Wu, C., Xue, X., & Hafner, J. P. (2003). Estimating the relative risk in cohort studies and clinical trials of common outcomes. *American Journal of Epidemiology, 157*(10), 940–943. https://doi.org/10.1093/aje/kwg074.

Muchlinski, D., Siroky, D., He, J., & Kocher, M. (2016). Comparing random forest with logistic regression for predicting class-imbalanced civil war onset data. *Political Analysis, 24*(1), 87–103.

Neunhoeffer, M., & Sternberg, S. (2019). How cross-validation can go wrong and what to do about it. *Political Analysis, 27*(1), 101–106.

Nosek, B. A., Ebersole, C. R., DeHaven, A. C., & Mellor, D. T. (2018). The preregistration revolution. *Proceedings of the National Academy of Sciences, 115*(11), 2600–2606.

Nosek, B. A. & Errington, T. M. (2019). *What is replication*? https://doi.org/10.31222/osf.io/u4g6t.

Open Science Collaboration. (2015). Estimating the reproducibility of psychological science. *Science* 349(6251). https://doi.org/10.1126/science.aac4716.

Patil, P., Peng, R. D., & Leek, J. T. (2019). A visual tool for denning reproducibility and replicability. *Nature Human Behaviour, 1*.

Penney, J. W. (2016). Chilling effects: Online surveillance and wikipedia use. *Berkeley Technology Law Journal, 31*, 117.

Plesser, H. E. (2018). Reproducibility vs. replicability: A brief history of a confused terminology. *Frontiers in Neuroinformatics, 11*, 76.

Salganik, M. (2017). *Bit by bit: Social research in the digital age*. Princeton University Press.

Schneider, F. B. (1990). Implementing fault-tolerant services using the state machine approach: A tutorial. *ACM Computing Surveys, 22*(3), 299.

Shrout, P. E., & Rodgers, J. L. (2018). Psychology, science, and knowledge construction: Broadening perspectives from the replication crisis. *Annual Review of Psychology, 69*, 487–510.

Silberzahn, R., Uhlmann, E. L., Martin, D. P., Anselmi, P., Aust, F., Awtrey, E., Bahmk, S., Bai, F., Bannard, C., Bonnier, E., et al. (2018). Many analysts, one data set: Making transparent how variations in analytic choices affect results. *Advances in Methods and Practices in Psychological Science, 1*(3), 337–356.

Simmons, J. P. & Nelson, L. D. (2019). *Data replicada*.

Simmons, J. P., Nelson, L. D., & Simonsohn, U. (2011). False-positive psychology undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychological Science, 22*(11), 0956797611417632-1366.

Simonsohn, U., Simmons, J. P., & Nelson, L. D. (2015). Specification curve: Descriptive and inferential statistics on all reasonable specifications. *SSRN Electronic Journal*.

Ward, M. D., Greenhill, B. D., & Bakke, K. M. (2010). The perils of policy by p-value: Predicting civil conflicts. *Journal of Peace Research, 47*(4), 363–375.

Yakir, B. (2011). *Introduction to statistical thinking (with R, without calculus)*. http://pluto.huji.ac.il/msby/StatThink/index.html.