METHOD ARTICLE

REVISED Internal replication of computational workflows in scientific research [version 2; peer review: 2 approved]

Jade Benjamin-Chung[1], John M. Colford, Jr.[1], Andrew Mertens[1], Alan E. Hubbard[1], Benjamin F. Arnold[1,2]

[1]Division of Epidemiology & Biostatistics, University of California, Berkeley, Berkeley, CA, 94720, USA
[2]Francis I. Proctor Foundation, University of California, San Francisco, San Francisco, CA, 94122, USA

v2 **First published:** 05 Feb 2020, **4**:17
https://doi.org/10.12688/gatesopenres.13108.1
Latest published: 17 Jun 2020, **4**:17
https://doi.org/10.12688/gatesopenres.13108.2

Abstract

Failures to reproduce research findings across scientific disciplines from psychology to physics have garnered increasing attention in recent years. External replication of published findings by outside investigators has emerged as a method to detect errors and bias in the published literature. However, some studies influence policy and practice before external replication efforts can confirm or challenge the original contributions. Uncovering and resolving errors before publication would increase the efficiency of the scientific process by increasing the accuracy of published evidence. Here we summarize the rationale and best practices for internal replication, a process in which multiple independent data analysts replicate an analysis and correct errors prior to publication. We explain how internal replication should reduce errors and bias that arise during data analyses and argue that it will be most effective when coupled with pre-specified hypotheses and analysis plans and performed with data analysts masked to experimental group assignments. By improving the reproducibility of published evidence, internal replication should contribute to more rapid scientific advances.

Keywords

replication, reproducibility, masking, blinding, computational workflow

Open Peer Review

Reviewer Status ✓ ✓

	Invited Reviewers	
	1	**2**
version 2 (revision) 17 Jun 2020	✓ report	✓ report
version 1 05 Feb 2020	? report	

1. **Garret Christensen**, U.S. Census Bureau, Washington, USA

2. **Lifeng Lin**, Florida State University, Tallahassee, USA

Any reports and responses or comments on the article can be found at the end of the article.

Corresponding author: Jade Benjamin-Chung (jadebc@berkeley.edu)

Author roles: Benjamin-Chung J: Conceptualization, Data Curation, Formal Analysis, Investigation, Methodology, Project Administration, Resources, Software, Supervision, Validation, Visualization, Writing – Original Draft Preparation, Writing – Review & Editing; **Colford, Jr. JM**: Conceptualization, Funding Acquisition, Supervision, Writing – Review & Editing; **Mertens A**: Data Curation, Formal Analysis, Investigation, Resources, Software, Validation, Writing – Review & Editing; **Hubbard AE**: Supervision, Writing – Review & Editing; **Arnold BF**: Conceptualization, Data Curation, Formal Analysis, Investigation, Methodology, Project Administration, Resources, Software, Supervision, Validation, Writing – Review & Editing

Competing interests: No competing interests were disclosed.

Grant information: This research was supported by Bill & Melinda Gates Foundation through a Global Development grant to the University of California, Berkeley, CA, USA [OPPGD759].
The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

How to cite this article: Benjamin-Chung J, Colford, Jr. JM, Mertens A *et al.* **Internal replication of computational workflows in scientific research [version 2; peer review: 2 approved]** Gates Open Research 2020, **4**:17
https://doi.org/10.12688/gatesopenres.13108.2

First published: 05 Feb 2020, **4**:17 https://doi.org/10.12688/gatesopenres.13108.1

REVISED **Amendments from Version 1**

In this revised version of our manuscript, we have addressed the comments of our reviewer Dr Garret Christensen. The key differences compared to the initial version are as follows:

- We have added a discussion of how internal replication compares to pre-publication review and pair programming.
- We also discuss the importance of internal replication in the case when studies cannot publicly share data and made some clarifications to our programming tips for internal replication.

Any further responses from the reviewers can be found at the end of the article

Introduction

A growing body of research has highlighted failures to replicate original study findings across disciplines[1–3]. Studies fail to be replicated for reasons ranging from unintentional coding errors to fraud[4]. Even in the absence of errors or fraud, researchers' own confirmation bias may impact the reproducibility of their findings[5]. In response to mounting concerns, interest in methods to improve transparency and reproducibility in research has skyrocketed: researchers across disciplines have published recommended practices to improve reproducibility[6–11] and detect lapses in publication integrity[12]. In addition, certain funders have announced grant review criteria for rigor and reproducibility and dedicated funding for replication studies[9,13].

A growing practice to diagnose reproducibility of the published literature is external independent replication, in which investigators outside an original study team attempt to replicate published results using the original dataset[14]. More importantly, external replication may occur too late to prevent funding or policymaking based on erroneous results. For example, a landmark social science study was recently retracted due to a coding error[15,16], yet the lead investigator had already received millions of dollars in funding based in large part on the initial study's erroneous findings[15,17]. In addition, external replication may create an incentive for replicators to overturn original study findings that introduces bias and undermines constructive scientific discourse[16,17]. An alternative approach that does not create such an incentive is for journals to conduct "pre-publication review", in which they attempt to replicate study findings using data and analytic code submitted prior to publication[16,17].

Here we describe "internal replication", a process through which investigators from an original study team independently replicate a computational workflow in order to identify and resolve errors and help thwart biases that occur during computational analyses prior to publication[10]. This practice is a natural complement to the growing practice of replicating experiments in different laboratories prior to publication in preclinical studies and has been recommended as a standard practice for computational workflows used by biologists[18,19]. In this article, we argue that adopting internal replication in scientific studies with a computational workflow should reduce errors and bias prior to publication. If adopted as a common practice, internal replication should improve the reliability of published evidence, increase the proportion of published findings that can be externally replicated, and increase the efficiency of the scientific process[5]. Below we describe the workflow for internal replication and how the process can reduce errors and confirmation bias during computational analysis.

The status quo workflow

In many disciplines, a typical computational workflow proceeds as follows: investigators conduct an experiment and/or collect data (Figure 1). Afterwards, they often make decisions about computational analyses with full knowledge of experimental group assignment, and often a single analyst performs computation and error checking without independent replication prior to publication. Researchers naturally tend to confirm their own beliefs and are prone to making choices that – consciously or not – lead them to a statistically significant finding[5]. In addition, it is common for researchers to thoroughly check unexpected results while errors in expected results may go unnoticed, introducing "disconfirmation bias"[5]. Yet another threat to validity lies in human error. A typical computational workflow requires thousands of lines of code, and it is inevitable that some will include mistakes. Though only some mistakes will ultimately alter a study's findings, occasionally a small error can

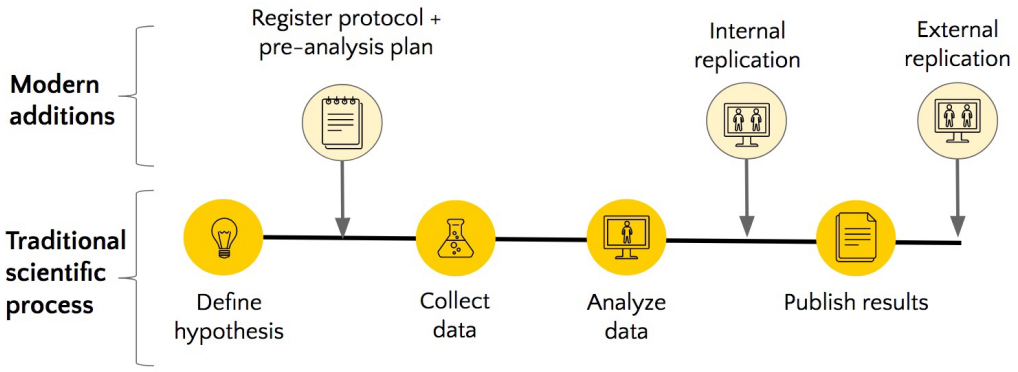


Figure 1. Modern additions to the traditional scientific process to increase rigor and reproducibility. Dark yellow circles indicate components of the traditional scientific process. Light yellow circles indicate modern additions to the scientific process.

amplify through an analysis like a genetic mutation, ultimately yielding vastly different results and policy implications. For example, a recent external replication[20] of a highly influential study that found externalities of school-based deworming identified a coding error in a variable used in a regression model; when corrected, one of the study's most novel, policy-relevant findings (that worm infections were lower in control schools 3–6 km away from intervention schools) was closer to the null and no longer statistically significant[21,22]. These recent examples highlight the urgent need to improve computational analysis practices prior to publication to improve the accuracy of published literature. We argue that human error and confirmation bias are inevitable. Scientists need computational workflows that anticipate and minimize these cognitive bias traps.

Here, we discuss procedures for internal replication and explain how the practice can reduce unintentional errors and thwart bias during data analysis[5]. Then we review internal replication practices we developed in eight internal replications we conducted for 32 outcomes in two large, cluster-randomized trials evaluating public health interventions in Bangladesh and Kenya[23–25].

Methods

The internal replication workflow

The internal replication workflow is a best practice that embraces confirmation bias and errors as an inevitable feature of scientific computation and reduces the likelihood that they will ultimately influence study findings. It complements other modern additions to traditional scientific practice, such as study registration and pre-analysis plans (Figure 1). At a high level, this workflow consists of pre-specification of computational analyses before a study commences, masking of analysts to experimental group assignment, and internal replication of key results before publication.

Including internal replication in pre-analysis plans. Pre-specification of analysis plans prior to study commencement is an increasingly common best practice that should reduce confirmation bias. Pre-analysis plans define the study hypotheses and objectives, experimental groups or exposures, outcomes, and statistical analysis methods[10]. Pre-analysis plans can also include plans for internal replication, including details about which analysis components will be replicated as well as the minimum allowable difference between independent analysts' results (i.e., the tolerance level). For example, a tolerance level could be chosen so that any differences in results between replicators are small enough that they would not appear in published manuscripts. Pre-specifying internal replication procedures prior to study commencement minimizes the chance of confirmation bias during the replication process.

Internal replication of key results before publication. Following an experiment and/or data collection, the internal replication process begins when analysts independently prepare computational datasets by cleaning and merging raw data and generating variables (Figure 2). During this step, they do not share their code with each other. Once each prepares an analysis dataset, analysts compare the values and summaries of each variable (e.g., the range and mean) between their datasets. If discrepancies exist, analysts work independently to resolve them and then iteratively compare and revise until their datasets are functionally identical (i.e., they have the same number of study units in each dataset and same values in each column). Once analysis datasets are replicated, the same process guides replication of computational analyses: analysts independently perform analyses, compare results, identify and resolve discrepancies between their results, and repeat the process until the difference in their results is less than the pre-determined tolerance level. Throughout the process, analysts may share their datasets and results with each other, but they do not share their code or analysis scripts until results are fully replicated. Code templates for comparing results while attempting to replicate are available from Zenodo[26], and programming tips for internal replication are available in Box 1. In studies with a high degree of repetition (e.g., multiple sites and outcomes) another tool to increase reproducibility is to create a software package using replicated code.

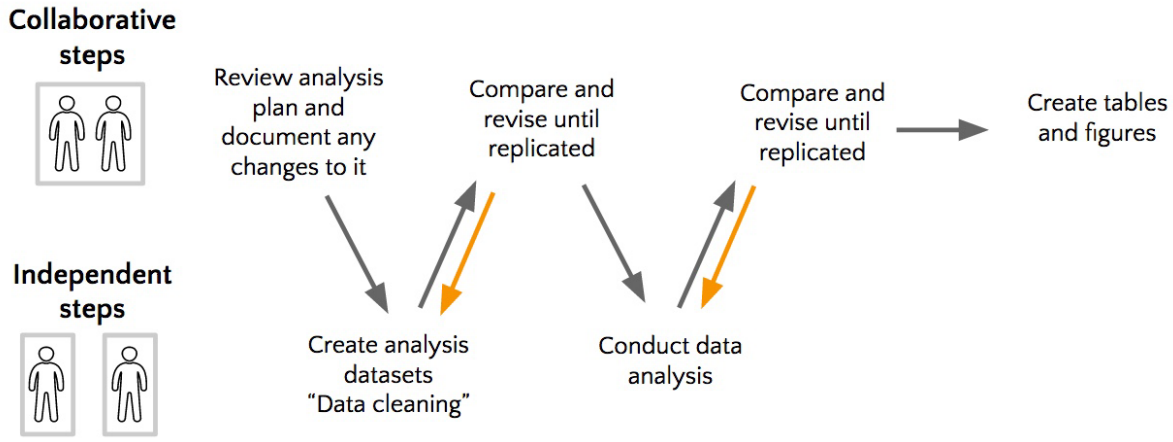


Figure 2. Internal Replication Workflow.

Box 1. Programming tips for internal replication

- **Decision log**: We recommend that independent data analysts keep a log of decisions they make together that are not covered by the pre-analysis plan. This includes how to handle unexpected outliers, discrepant identification numbers, and erroneous variable values. The log provides a thorough, transparent record of minor decisions made during the analysis.
- **Software**: Using statistical software such as R or Python that allows data analysts to efficiently load a large number of objects of differing dimensions (e.g., scalars, vectors, and matrices of different dimensions from different data sources), take the difference between them, and identify which are replicated facilitates replication. Other languages, such as Stata or SAS, allow objects of different dimensions to be loaded simultaneously, but the default is to work with a particular dataset with specific dimensions. As a result, while replicating, it may be more difficult to efficiently compare large numbers of matrices or other objects generated by each analyst to check for replication when using these languages. If analysts use the same software, we recommend that they use the same version of the software to ensure that differences in their results are not due to differences in software versions.
- **Version control**: We recommend using a version control system, such as Git, to track changes made during replication facilitates collaboration of data analysts during and after replication.
- **Variable type**: Agreeing upon the variable type, particularly for continuous variables, facilitates smooth replication. Some software truncates the number of significant figures when saving numeric variables in different formats. For example, since Stata stores numeric variables in binary, the value 0.1, which does not have a perfect binary representation, is stored differently for float and double variables types. These differences can carry forward and prevent replication.
- **Sorting and seed**: For analyses utilizing any kind of resampling (e.g. bootstrapping) or cross-validation, sorting and seed matter. Agreeing on variable sort order, sorting data at the same location in your script (e.g., right before analysis), and using the same seeds at the same locations facilitate replication.
- **Modular scripts**: Writing modular scripts that perform a limited number of discrete analyses makes it easier to diagnose failures to replicate. For analyses with thousands of objects, saving results in relatively small batches speeds replication by allowing data analysts to diagnose and resolve failures to replicate without having to re-run the entire analysis.
- **Bash scripts**: Bash scripts are plain text files that list a series of commands across different software packages; for instance, they could delete previous results objects and analysis logs from their stored location and then re-run analytic scripts in R or Stata. Separate bash scripts can be written for data management and different components of the analysis. Including code to remove all previously saved objects each time a script is re-run ensures that old versions of objects aren't compared by accident when assessing whether replication was achieved.

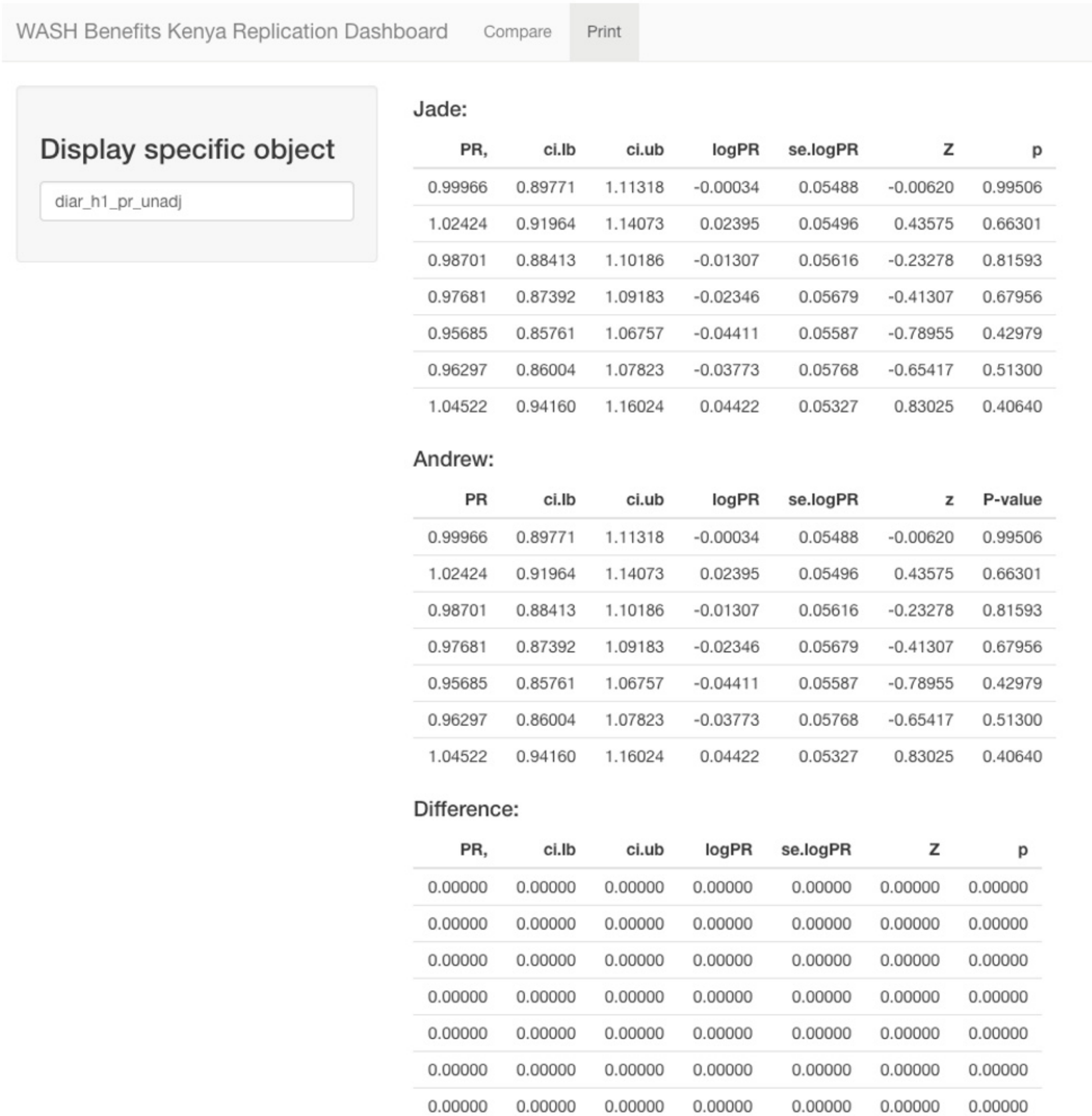
For example, the pre-analysis plan for the WASH Benefits Kenya trial included estimation of unadjusted prevalence ratios for diarrhea. Each analyst separately wrote the code in accordance with the pre-analysis plan using analysis datasets that they had ensured were functionally identical. Analyses were performed using *R* version 3.2.3. They saved estimates to a shared directory that could be accessed by both analysts (link to analyst 1 code[27], link to analyst 2 code[28]). We then compared each analyst's estimates using a dashboard created with Shiny R to determine whether results were replicated. The dashboard displayed each analysts' results, including the estimated prevalence ratio, confidence interval, log prevalence ratio, log of the standard error, Z-statistic, and p-value for the analysis in the columns, and each row displayed these estimates comparing each intervention arm to the control arm (Figure 3). In addition, the dashboard showed the difference between each analyst's estimates, which are all equal to 0, indicating that this analysis was internally replicated. We used this overall internal replication process for each component of the statistical analysis in this study (e.g., unadjusted analyses of other outcomes and adjusted analyses).

Masked computational analyses. Masking analysts to experimental group assignment, or in theory any other key variable during data preparation and analysis, can further reduce bias during internal replication. In a masked analysis, prior to working with data, an independent analyst re-randomizes the experimental group assignment variable. Subsequently, analysts use the re-randomized experimental group variable during dataset preparation and computational analysis. Results viewed during internal replication are scrambled, preventing any judgments that could produce favorable findings. Following internal replication, analysis scripts are re-run with the true experimental group assignment variable to obtain unmasked, final results. Though masked analyses are common in some fields, such as clinical trials[29] and particle and nuclear physics[30], to our knowledge the approach has not been widely adopted in other types of biomedical studies or in other fields, such as biology, psychology, or social sciences.

Comparing internal replication to alternative approaches

Internal replication vs. pair programming. At first glance, internal replication may resemble pair programming, a practice in which one analyst writes code while the other simultaneously reads and comments on the code to suggest coding strategies and improvements. Costs associated with pair programming and internal replication are likely to be similar since both approaches require two analysts to complete a single analysis. Unlike pair programming, in internal replication the vast majority of coding is done independently with minimal communication until data or results are compared. The advantage of this approach over pair programming is that it allows analysts to pursue completely different coding strategies which may be subject to differing sources of error and bias. Pair programming may be more subject to "group think" in which shared biases or judgment calls are reinforced or amplified. Thus, we believe that internal replication is more likely to identify failures to replicate results due to coding errors and biases than pair programming.

Internal replication vs. pre-publication review. One strategy to increase reproducibility that is complementary to internal



Jade:

PR,	ci.lb	ci.ub	logPR	se.logPR	Z	p
0.99966	0.89771	1.11318	-0.00034	0.05488	-0.00620	0.99506
1.02424	0.91964	1.14073	0.02395	0.05496	0.43575	0.66301
0.98701	0.88413	1.10186	-0.01307	0.05616	-0.23278	0.81593
0.97681	0.87392	1.09183	-0.02346	0.05679	-0.41307	0.67956
0.95685	0.85761	1.06757	-0.04411	0.05587	-0.78955	0.42979
0.96297	0.86004	1.07823	-0.03773	0.05768	-0.65417	0.51300
1.04522	0.94160	1.16024	0.04422	0.05327	0.83025	0.40640

Andrew:

PR	ci.lb	ci.ub	logPR	se.logPR	z	P-value
0.99966	0.89771	1.11318	-0.00034	0.05488	-0.00620	0.99506
1.02424	0.91964	1.14073	0.02395	0.05496	0.43575	0.66301
0.98701	0.88413	1.10186	-0.01307	0.05616	-0.23278	0.81593
0.97681	0.87392	1.09183	-0.02346	0.05679	-0.41307	0.67956
0.95685	0.85761	1.06757	-0.04411	0.05587	-0.78955	0.42979
0.96297	0.86004	1.07823	-0.03773	0.05768	-0.65417	0.51300
1.04522	0.94160	1.16024	0.04422	0.05327	0.83025	0.40640

Difference:

PR,	ci.lb	ci.ub	logPR	se.logPR	Z	p
0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000
0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000
0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000
0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000
0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000
0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000
0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000

Figure 3. Screenshot of a Shiny R dashboard indicating that the diarrhea unadjusted prevalence ratios in WASH Benefits Kenya were replicated.

replication is pre-publication review in which journals replicate study findings using replication scripts and datasets prior to accepting a manuscript for publication[16,17]. The *American Journal of Political Science* is one example of a journal that uses this approach. Internal replication before submission to peer review would catch errors internally before peer reviewers and journal editors consider a manuscript. Detecting errors after submission or after peer review is less efficient because it may require another round of peer review and revision. In addition, pre-publication review may be sufficient to ensure that the results generated by analytic code match those in the manuscript but may miss coding errors.

How internal replication reduces bias

The act of comparing independently generated results during internal replication should reduce errors, confirmation bias, and disconfirmation bias. Analysts are unlikely to make identical mistakes, and discrepant results must be corrected in order to achieve replication. By requiring analysts to compare all of their results, internal replication improves the quality and extent of error checking, reducing disconfirmation bias.

The process of resolving discrepancies in independently generated results also reduces confirmation bias by illuminating judgment calls and decisions that may influence study results

but are outside the scope of the pre-analysis plan. For example, decisions about how to handle erroneous responses to a coded survey question or missing responses for individual variables used to generate a composite variable cannot feasibly be included in pre-analysis plans. These types of small, seemingly inconsequential decisions often cannot be predicted before seeing the data. When differences in analysts' decisions lead to discrepant results, internal replication provides a platform for transparent analytic choices outside the scope of pre-analysis plans.

Investigators must balance the need to reduce errors and bias with the significant costs required to perform internal replication. Internal replication can double the person-time required to complete an analysis and puts the burden of replication on the original study team. Yet, our view is that, overall, internal replication is far more efficient than external replication because the original study team is most knowledgeable about a study; external replication efforts require significant investment from the original investigator team because external replicators are not familiar with study materials[16,17].

Results

We developed best practices for internal replication while internally replicating data analyses for two randomized trials conducted in Bangladesh and Kenya named "WASH Benefits". These trials measured the effect of single interventions (water (W), sanitation (S), handwashing (H), nutrition (N)) and combined interventions (combined W+S+H, combined W+S+H+N) on over 32 outcomes including child growth, diarrhea, parasite infection, and child development[23–25]. In addition, each trial tested 3 core hypotheses related to the effects of single interventions vs. combinations of interventions. The WASH Benefits trials were unusually complex and had a large number of interventions, hypotheses, and outcomes across two countries. Yet, it was exactly this complexity combined with the global importance of the results that motivated the study team to embrace internal replication. Our internal replication of these trials led us to uncover and resolve errors at every stage of the data analysis and brought to light numerous small judgment calls and assumptions made by each analyst. Correcting errors and transparently discussing data analysts' assumptions helped us reduce bias in our study findings prior to publication. Trial data is available as underlying data[31].

Including internal replication in pre-analysis plans

The WASH Benefits team published a protocol that described the study's rationale, design, and analysis near the beginning of the study[23] (Table 1). At the time of analysis but before working with the data, we updated the analysis plan for each country with additional details pertinent to specific analyses, and we registered the updated plans through the Open Science Framework (e.g., https://osf.io/63mna/). Having detailed pre-analysis plans in place improved the efficiency of internal replication by providing a clear roadmap for individual data analysts.

Masked computational analyses & internal replication of key results before publication

To reduce potential bias, analysts were masked to treatment assignments; we performed analyses using scrambled treatment assignment labels instead of real ones. During masked analyses, we encountered numerous differences in judgment calls that initially prevented replication. For example, two data analysts calculated age in months by dividing age in days using different numbers for average days per month. When age was used to calculate the height-for-age Z-score, the small differences in age in months had ripple effects that produced different results, particularly for effects that were borderline significant. We developed a workflow in which analysts kept notes about their judgement calls in a shared document, and they used these to reconcile differences and to make joint decisions in advance of future analyses to reduce replication time.

Another difference between analysts that initially prevented replication was the use of different software (Stata vs. R) and different data structures. For example, in an analysis of child growth data collected at two time points, one analyst happened to create a single dataset including both time points, and another included a separate dataset for each time point. The process of merging datasets with additional covariates against these two differing data structures created discrepancies in results that the replicators had to discuss and resolve. A concrete example of a mistake we caught that affected results was in the coding of the variable for the month of data collection. The variable was intended to be coded as a set of indicator variables when it was included in adjusted statistical models. One analyst accidentally coded it as a numeric variable, (e.g., 1 for January, 2 for February, etc.). The analysts were unable to replicate results until this discrepancy had been resolved. These examples illustrate the value of internal replication for detecting and resolving errors.

After completing our first replication for WASH Benefits, we developed a R software package[32] for the trials for internal use based on the replicated code. The package streamlined the analysis across additional outcomes in the trials by providing a consistent template for analyses and standardizing output into a single, coherent format. Beyond the benefits of efficiency, the package's consistent interface and internal data handling reduced the number of steps that each analysis needed to replicate. The time required to create the software package was justified since WASH Benefits was conducted in two countries and measured numerous outcomes. For studies with fewer iterations of the same type of analysis, the time investment required to develop a software package would likely outweigh its benefits. For any internal replication, creating a dashboard, such as the one we created with Shiny R (https://osf.io/xbyrn/), to compare analysts' estimates greatly increases the efficiency of internal replication by rapidly identifying estimates that failed to replicate. Advances in data science have made package and application development much easier, and we anticipate that in future

Table 1. Internal replication in the WASH Benefits trials.

Internal replication workflow steps	How each internal replication step reduces:			Examples from internal replication of WASH Benefits
	Confirmation bias	Disconfirmation bias	Human error	
1) Pre-specify computational analyses before study commences	Prevents p-hacking and analytic choices that produce favorable results by requiring investigators to make analytic decisions before seeing the data	Since all analyses – including secondary, subgroup, and sensitivity analyses – are pre-specified rather than selected post hoc, analysts incorporate them into the computational workflow and check them for errors in a systematic way	May indirectly reduce human error during analysis by reducing the number of decisions analysts must make during analysis, decreasing cognitive load	Prior to primary outcome data collection, the study investigators published a description of the study rationale, design, and analysis plan[23]. After data collection but before analysis, investigators published minor modifications to the pre-analysis plan on the Open Science Framework (https://osf.io/krezy).
2) Mask analysts to experimental group assignment	Prevents analysts from seeing study results during analysis	Results viewed during analysis are not meaningful because of re-randomized labels, so all results must be reviewed rather than only those that do not confirm expectations	May indirectly reduce human error during analysis by shifting attention from interpreting study findings to ensuring that the computational workflow is error-free	Prior to analysis, an independent analyst created a treatment variable that was randomly permuted within the trial's randomized blocks. This scrambled treatment variable was used during data cleaning and analysis and was replaced by the real treatment assignments only after step 3 (below) was complete.
3) Internal replication of key results before publication	If there are any discrepancies in analytic decisions outside the scope of the pre-analysis plan between independent analysts, these are likely to prevent internal replication, requiring analysts to transparently discuss and agree upon major and minor analytic decisions	Requires every result to be compared and replicated, not only those that fail to confirm expectations	Catches and resolves numerous potential errors during data cleaning and analysis since such errors are likely to prevent internal replication	Investigators first internally replicated the analysis of primary outcomes at one site. Then using that internally replicated code, they developed a software package using internally replicated for use in analyses of secondary and tertiary outcomes that standardized output into a standard format (https://github.com/ben-arnold/washb).

Caption: The WASH Benefits trials were two randomized, controlled, epidemiologic field trials conducted in Bangladesh and Kenya that measured the effect of single and combined interventions water, sanitation, handwashing, and nutrition interventions on over 32 outcomes including child growth, diarrhea, parasite infection, and child development[23–25]. Each trial tested 3 core hypotheses in 6 intervention arms related to the effects of single interventions vs. combinations of interventions. The complexity of the trials and anticipation of the trials' results in the global health sector motivated the study team to perform internal replication.

years the process will become even more streamlined, making this approach feasible for studies with limited funding for internal replication.

Conclusions

The internal replication tools we presented range from relatively easy (masking analysts to treatment assignment) to more resource intensive (developing an analytic software package). If resources are limited, replicating the analysis steps that are most error-prone and require the most judgment calls is a good place to start. For example, unglamorous data cleaning and processing steps are likely to be more error-prone since they require significantly more arbitrary decisions and complex programming steps than the computational analysis, especially when a pre-analysis plan is used. Alternatively, investigators could replicate a subset of outcomes or comparisons or a key portion of the analysis. In some disciplines research is conducted alone, and in this case a single analyst could partially replicate their own work by programming the analysis in alternative software packages (e.g., R vs. Stata) or by writing the same error-prone section of code twice.

Internal replication is one of many tools available to researchers to increase the reproducibility of their work, including publication of pre-analysis plans and publishing analytic datasets. A large proportion of studies are unable to publish their datasets due to human subjects protections or other privacy restrictions. In these cases, since external replication may not be possible, internal replication is even more valuable.

A limitation of internal replication is that it is performed by analysts from the same study team, who may make the same judgment calls or mistakes due to "group think". Internal replication cannot detect identical errors or judgment calls made by each analyst. Nevertheless, our view is that internal replication can still prevent the majority of errors and biases, and publicly posting analysis plans and complete replication files allows external investigators to vet judgement calls.

Internal replication should increase the accuracy of published scientific results, thereby increasing the efficiency of the scientific process[33]. Furthermore, it can reduce public controversies about how to interpret externally replicated results that differ from original results. These broader benefits should motivate funders to consider dedicated financial support for internal replication and spur journals to incentivize internal replication[7,16,17]. For example, completion of internal replication could be a criterion editors use to assess studies during scientific journal peer review, and internally replicated studies could receive a reproducibility kite-mark or badge, such as those instituted by the journals *Psychological Science* and *Biostatistics*[34–36]. Investigators who provide details of their plans for internal replication could be prioritized during grant proposal review. Including internal replication in the modern computational workflow allows scientists to embrace the fact that errors and bias are inevitable—a critical step towards advancing science and strengthening the culture of reproducibility.

Data availability

Underlying data

Open Science Framework: WASH Benefits Kenya Primary Analysis. https://doi.org/10.17605/OSF.IO/KREZY[31]

This repository contains the following underlying data:

- washb-kenya-tr (This file contains the randomized treatment assignment for each cluster in the study. Available as in dta and csv format with codebook)
- washb-kenya-tracking (This file provides tracking information for the 8,246 households enrolled at the year 1 and year 2 visits. Available as in dta and csv format with codebook)
- washb-kenya-uptake-baseline (This file includes intervention adherence indicators collected at baseline in each household. Available as in dta and csv format with codebook)
- washb-kenya-uptake-midline (This file includes intervention adherence indicators collected at midline in each household. Available as in dta and csv format with codebook)
- washb-kenya-uptake-endline (This file includes intervention adherence indicators collected at endline in each household. Available as in dta and csv format with codebook)
- washb-kenya-midline-anthro (This file includes anthropometry measurements collected at midline in index children. Available as in dta and csv format with codebook)
- washb-kenya-endline-anthro (This file includes anthropometry measurements collected at endline in index children. Available as in dta and csv format with codebook)
- washb-kenya-diar (This file includes diarrhea illness symptoms collected at midline and endline in index children and children < 36 months in the compound. Available as in dta and csv format with codebook)

Software availability

Source code for the internal replication dashboard is available from: https://github.com/jadebc/replicate

Archived source code at the time of publication: https://doi.org/10.5281/zenodo.3626134[26]

License: Apache License 2.0

Source code for the WASH Benefits software package is available from: https://github.com/ben-arnold/washb

Archived source code at time of publication: https://doi.org/10.5281/zenodo.3626168[32]

License: GNU General Public License v3.0 or later

Source code for Analyst 1 of the WASH Benefits Kenya primary outcome analysis is available from: https://github.com/jadebc/WBK-primary-outcomes

Archived source code at the time of publication: https://doi.org/10.5281/zenodo.3627316[27]

License: Apache License 2.0

Source code for Analyst 2 of the WASH Benefits Kenya primary outcome analysis is available from: https://github.com/amertens/Wash-Benefits-Kenya

Archived source code at the time of publication: https://doi.org/10.5281/zenodo.3627359[28]

License: Apache License 2.0

Acknowledgements

The authors would like to thank Matthew Akamatsu for his helpful comments on this manuscript.

References

1. Open Science Collaboration: **Estimating the reproducibility of psychological science.** *Science.* 2015; **349**(6251): aac4716.
PubMed Abstract | Publisher Full Text

2. Camerer CF, Dreber A, Forsell E, *et al.*: **Evaluating replicability of laboratory experiments in economics.** *Science.* 2016; **351**(6280): 1433–1436.
PubMed Abstract | Publisher Full Text

3. Ioannidis JP: **Acknowledging and Overcoming Nonreproducibility in Basic and Preclinical Research.** *JAMA.* 2017; **317**(10): 1019–1020.
PubMed Abstract | Publisher Full Text

4. Crocker J, Cooper ML: **Addressing scientific fraud.** *Science.* 2011; **334**(6060): 1182.
PubMed Abstract | Publisher Full Text

5. Nuzzo R: **Fooling ourselves.** *Nat Lond.* 2011; **526**(7572): 182–185.
Publisher Full Text

6. Open Science Collaboration: **An Open, Large-Scale, Collaborative Effort to Estimate the Reproducibility of Psychological Science.** *Perspect Psychol Sci.* 2012; **7**(6): 657–660.
PubMed Abstract | Publisher Full Text

7. Nosek BA, Alter G, Banks GC, *et al.*: **Promoting an open research culture.** *Science.* 2015; **348**(6242): 1422–1425.
PubMed Abstract | Publisher Full Text | Free Full Text

8. Miguel E, Camerer C, Casey K, *et al.*: **Promoting transparency in social science research.** *Science.* 2014; **343**(6166): 30–31.
PubMed Abstract | Publisher Full Text | Free Full Text

9. Munafò MR, Nosek BA, Bishop DVM, *et al.*: **A manifesto for reproducible science.** *Nat Hum Behav.* 2017; **1**(1): 0021.
Publisher Full Text

10. DeMets DL, Cook TD, Buhr KA: **Guidelines for Statistical Analysis Plans.** *JAMA.* 2017; **318**(23): 2301–2303.
PubMed Abstract | Publisher Full Text

11. Aczel B, Szaszi B, Sarafoglou A, *et al.*: **A consensus-based transparency checklist.** *Nat Hum Behav.* 2020; **4**(1): 4–6.
PubMed Abstract | Publisher Full Text

12. Grey A, Bolland MJ, Avenell A, *et al.*: **Check for publication integrity before misconduct.** *Nature.* 2020; **577**(7789): 167–169.
PubMed Abstract | Publisher Full Text

13. **National Institutes of Health Rigor and Reproducibility**. 2017; Accessed 19 Dec 2017.
Reference Source

14. Brandt MJ, IJzerman H, Dijksterhuis A, *et al.*: **The Replication Recipe: What makes for a convincing replication?** *J Exp Soc Psychol.* 2014; **50**(1): 217–224.
Publisher Full Text

15. George RP: **Opinion | Confirmation Bias Hurts Social Science.** *Wall Str J.* 2019.
Reference Source

16. Regnerus M: **Is structural stigma's effect on the mortality of sexual minorities robust? A failure to replicate the results of a published study.** *Soc Sci Med.* 2017; **188**: 157–165.
PubMed Abstract | Publisher Full Text

17. Gertler P, Galiani S, Romero M: **How to make replication the norm.** *Nature.* 2018; **554**(7693): 417–419.
PubMed Abstract | Publisher Full Text

18. Voelkl B, Vogt L, Sena ES, *et al.*: **Reproducibility of preclinical animal research improves with heterogeneity of study samples.** *PLoS Biol.* 2018; **16**(2): e2003693.
PubMed Abstract | Publisher Full Text | Free Full Text

19. Shade A, Teal TK: **Computing Workflows for Biologists: A Roadmap.** *PLoS Biol.* 2015; **13**(11): e1002303.
PubMed Abstract | Publisher Full Text | Free Full Text

20. Aiken AM, Davey C, Hargreaves JR, *et al.*: **Re-analysis of health and educational impacts of a school-based deworming programme in western Kenya: a pure replication.** *Int J Epidemiol.* 2015; **44**(5): 1572–1580.
PubMed Abstract | Publisher Full Text | Free Full Text

21. Miguel E, Kremer M: **Worms: Identifying Impacts on Education and Health in the Presence of Treatment Externalities.** *Econometrica.* 2004; **72**(1): 159–217.
Publisher Full Text

22. Hicks JH, Kremer M, Miguel E, *et al.*: **Commentary: Deworming externalities and schooling impacts in Kenya: a comment on Aiken *et al.* (2015) and Davey *et al.* (2015).** *Int J Epidemiol.* 2015; **44**(5): 1593–1596.
PubMed Abstract | Publisher Full Text

23. Arnold BF, Null C, Luby SP, *et al.*: **Cluster-randomised controlled trials of individual and combined water, sanitation, hygiene and nutritional interventions in rural Bangladesh and Kenya: the WASH Benefits study design and rationale.** *BMJ Open.* 2013; **3**(8): e003476.
PubMed Abstract | Publisher Full Text | Free Full Text

24. Luby SP, Rahman M, Arnold BF, *et al.*: **Effects of water quality, sanitation, handwashing and nutritional interventions on diarrhoea and child growth in rural Bangladesh: a cluster randomised trial.** *Lancet Glob Health.* 2018; **6**(3): e302–e315.
PubMed Abstract | Publisher Full Text | Free Full Text

25. Null C, Stewart CP, Pickering AJ, *et al.*: **Effects of water quality, sanitation, handwashing and nutritional interventions on diarrhoea and child growth in rural Kenya: a cluster-randomised controlled trial.** *Lancet Glob Health.* 2018; **6**(3): e316–e329.
PubMed Abstract | Publisher Full Text | Free Full Text

26. Benjamin-Chung J: **jadebc/replicate: v1. Zenodo archive (Version 1).** *Zenodo.* 2020.
http://www.doi.org/10.5281/zenodo.3626134

27. Benjamin-Chung J, Mertens A: **jadebc/WBK-primary-outcomes: Version associated with the internal replication and the primary outcomes manuscripts (Version v1).** *Zenodo.* 2020.
http://www.doi.org/10.5281/zenodo.3627316

28. Mertens A: **amertens/Wash-Benefits-Kenya: Initial release (Version 1.0.0).** *Zenodo.* 2020.
http://www.doi.org/10.5281/zenodo.3627359

29. Schulz KF, Grimes DA: **Blinding in randomised trials: hiding who got what.** *Lancet.* 2002; **359**(9307): 696–700.
PubMed Abstract | Publisher Full Text

30. MacCoun R, Perlmutter S: **Blind analysis: Hide results to seek the truth.** *Nature.* 2015; **526**(7572): 187–189.
PubMed Abstract | Publisher Full Text

31. Benjamin-Chung J, Arnold BF, Mertens A: **WASH Benefits Kenya Primary Analysis**. 2020.
http://www.doi.org/10.17605/OSF.IO/KREZY

32. Nguyen A, Arnold B, Mertens A: **ben-arnold/washb: Version 0.2.2 (Version v0.2.2).** *Zenodo.* 2020.
http://www.doi.org/10.5281/zenodo.3626168

33. Ebersole CR, Axt JR, Nosek BA: **Scientists' Reputations Are Based on Getting It Right, Not Being Right.** *PLoS Biol.* 2016; **14**(5): e1002460.
PubMed Abstract | Publisher Full Text | Free Full Text

34. Peng RD: **Reproducible research in computational science.** *Science.* 2011; **334**(6060): 1226–1227.
PubMed Abstract | Publisher Full Text | Free Full Text

35. Kidwell MC, Lazarević LB, Baranski E, *et al.*: **Badges to Acknowledge Open Practices: A Simple, Low-Cost, Effective Method for Increasing Transparency.** *PLoS Biol.* 2016; **14**(5): e1002456.
PubMed Abstract | Publisher Full Text | Free Full Text

36. Rowhani-Farid A, Barnett AG: **Badges for sharing data and code at *Biostatistics*: an observational study [version 2; peer review: 2 approved].** *F1000Res.* 2018; **7**: 90.
PubMed Abstract | Publisher Full Text | Free Full Text

Open Peer Review

Current Peer Review Status: ✔ ✔

Version 2

Reviewer Report 04 August 2020

https://doi.org/10.21956/gatesopenres.14341.r29139

✔ **Lifeng Lin**

Department of Statistics, Florida State University, Tallahassee, Florida, USA

This manuscript illustrates the idea of internal replication for dealing with the crisis of research reproducibility and replicability in many fields. I like the framework and examples presented in this manuscript. I have the following specific comments:

- I think it might be helpful to use a few sentences to distinguish research reproducibility and research replicability in the introduction. It seems that this manuscript has used these two words in an exchangeable way, but I think the former may refer to the process that the research findings are exactly reproduced by re-performing the analyses, and the latter may refer to other studies conducted by other research teams (and thus based on different data) that produce consistent findings.

- In the second paragraph in the introduction section, the authors provided an example of coding error in a social science study. Similar issues also occurred in medical research. For example, in a recent clinical trial published in JAMA (Aboumatar *et al.*, 2018[1]), the treatment labels were erroneously switched in the analyses, leading to reversed findings, and the article was subsequently retracted (Aboumatar *et al.*, 2019[2]). It seems that such errors could be avoided by using the internal replication approach described by the authors.

- In the subsection of “Including internal replication in pre-analysis plans”, the authors mentioned that: “a tolerance level could be chosen so that any differences in results between replicators are small enough that they would not appear in published manuscripts.” I think the authors may further clarify how such differences may arise and how the tolerance level should be set. It seems that the authors referred to potential computational inaccuracies (e.g., due to rounding errors) or differences between methods used to implement statistical analyses. Although the authors provided an example of the WASH study (Figure 3), this example gave identical results by two analysts. I was wondering how researchers should address the discrepancies that substantially exceed the tolerance level. For example, multiple statistical software programs can be used to implement generalized linear mixed models, but different programs could lead to noticeable

differences (Zhang *et al.*, 2011[3]).

- I like Figures 1 and 2. In Figure 2, I was wondering if the last step, "create tables and figures," should be also conducted by both/multiple analysts independently or at least cross-validated by the analysts. Errors may occur when generating tables and figures (which could even occur during the typesetting process by publishers), while tables and figures are important content for readers to understand study findings.

- In the subsection of "Internal replication vs. pre-publication review," it might be helpful to elaborate how the pre-publication review can be feasibly implemented. To my knowledge, few journals in the field of statistics and biostatistics actually replicate study findings using replication scripts and datasets prior to accepting a manuscript for publication. Many statistical projects involve extensive computations (e.g., simulation studies, very high-dimensional data analyses), and some require high-performance computers to perform the analyses. It is not very feasible to ask reviewers to run replication scripts, especially given that most review tasks are volunteer. For medical projects, patient-level data may be also needed for replicating findings, and they may not be shared without proper de-identification processes.

- In the results section, the authors discussed about masked computational analyses. "To reduce potential bias, analysts were masked to treatment assignments; we performed analyses using scrambled treatment assignment labels instead of real ones." This is not new in clinical trials. Many trials are double blinded; treatment assignments are masked not only for analysts, but also for patients and healthcare providers.

- At the end of this manuscript, I think the authors may consider adding some discussion for time-sensitive research. This topic may be particularly important during the covid-19 pandemic. There have been thousands of studies on covid-19; some provide timely evidence, but some have been found to be misleading (e.g., Mehra *et al.*, 2020[4]). The internal replication is clearly able to reduce error and bias; however, it also requires much more amount of time and effort paid by research teams, and it may delay the publication of important time-sensitive evidence.

References

1. Aboumatar H, Naqibuddin M, Chung S, Chaudhry H, et al.: Effect of a Program Combining Transitional Care and Long-term Self-management Support on Outcomes of Hospitalized Patients With Chronic Obstructive Pulmonary Disease. *JAMA*. 2018; **320** (22). Publisher Full Text
2. Aboumatar H, Wise R: Notice of Retraction. Aboumatar et al. Effect of a Program Combining Transitional Care and Long-term Self-management Support on Outcomes of Hospitalized Patients With Chronic Obstructive Pulmonary Disease: A Randomized Clinical Trial.JAMA . 2018;320(22):2335-2343.*JAMA*. 2019; **322** (14). Publisher Full Text
3. Zhang H, Lu N, Feng C, Thurston SW, et al.: On fitting generalized linear mixed-effects models for binary responses using different statistical packages.*Stat Med*. 2011; **30** (20): 2562-72 PubMed Abstract | Publisher Full Text
4. Mehra M, Desai S, Ruschitzka F, Patel A: RETRACTED: Hydroxychloroquine or chloroquine with or without a macrolide for treatment of COVID-19: a multinational registry analysis. *The Lancet*. 2020. Publisher Full Text

Is the rationale for developing the new method (or application) clearly explained?
Yes

Is the description of the method technically sound?
Yes

Are sufficient details provided to allow replication of the method development and its use by others?
Yes

If any results are presented, are all the source data underlying the results available to ensure full reproducibility?
Yes

Are the conclusions about the method and its performance adequately supported by the findings presented in the article?
Yes

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: Statistical methods

I confirm that I have read this submission and believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard.

Reviewer Report 13 July 2020

https://doi.org/10.21956/gatesopenres.14341.r28939

Garret Christensen
U.S. Census Bureau, Washington, DC, USA

The authors did a good job addressing my initial concerns. At this point I have no remaining concerns and give it my full approval.

Is the rationale for developing the new method (or application) clearly explained?
Yes

Is the description of the method technically sound?
Yes

Are sufficient details provided to allow replication of the method development and its use by others?
Yes

If any results are presented, are all the source data underlying the results available to ensure full reproducibility?
Yes

Are the conclusions about the method and its performance adequately supported by the findings presented in the article?
Yes

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: Reproducibility, research transparency, poverty programs, applied econometrics, labor economics

I confirm that I have read this submission and believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard.

Version 1

Reviewer Report 19 February 2020

https://doi.org/10.21956/gatesopenres.14274.r28542

?

Garret Christensen
U.S. Census Bureau, Washington, DC, USA

This article summarizes the internal replication (pre-publication double coding of data cleaning and analysis) of the WASH Benefits study, and describes the process so other projects can emulate the methods widely. I find it to be mostly well-written and its argument sound. I have mostly minor comments and suggestions, which are below. I do have a few larger topics that might be addressed:

A. How does your method compare to journal post-acceptance pre-publication review, as currently done in several political science journals?

B. How does your method work for solo researchers?

C. How does your method work in a world of proprietary or confidential data that cannot be

publicly released? Something like 40% of articles in a top economics journal use data that can't be released, and the Federal Statistical Research Data Centers produce important research from data that can't be publicly released. (This is probably a point in your favor, since it's harder to do external replications of non-sharable data.)

D. How does your method compare in costs and benefits to pair programming?

Details in order of appearance:

1. Introduction, paragraph 2: I'm not sure citation 15 argues for external replication, at least not exclusively. Doesn't it make the case for quasi-internal, or at least pre-publication review (by the journal itself, after acceptance but before publication)? Political science journals such as AJPS do this now. It's just a "does your code actually produce your tables" and not as thorough as your method, but I think you should clarify this and possibly compare or contrast this method of journal pre-publication review to your method.

2. "a landmark study was recently retracted...": In which field/on what topic?

3. Status quo workflow paragraph: I think the use of citation 21 as an example needs to be more nuanced. Which "one of the study's most novel, policy-relevant findings"? I believe the original study authors would disagree with that claim. For example, from the Commentary response (which could also be cited) to cite 21: "The argument in Davey *et al*., that deworming impacts on school participation are not robust to different statistical approaches, is based on several analytical errors." Anyway, the point that an external replication led to disagreement or even controversy is certainly valid.

4. Methods including internal replication paragraph: "regardless of units" is a little unclear. I think I get the point about ultimate display in a journal, but often coefficients will be scaled. For example, population or income will be divided by 10K or 1m to avoid having to display a bunch of zeroes.

5. Box 1, Software: Using non-object-oriented languages. What exactly do you mean by "makes it more difficult to load objects of different dimensions and compare them simultaneously"? An example could help.

6. Variable type: truncating the number of sig figs: I assume you are talking about character vs. numeric. Is this not just a machine precision/binary representation thing that all languages/computers do? Please clarify.

7. Bottom of page 4/just after Box 1: add a link here to the dashboard. It appears later, but this is the first mention.

8. Last paragraph before results: You have not discussed the large world of solo investigators. How do they fit into this scheme?

9. Results, making computational analysis...paragraph: "we kept notes about their judgement calls...": This seems close to pair programming. Software teaching organizations seem to advocate for this, and say it reduces errors. How would your method compare in terms of

cost and reliability? Did you consider pair programming?

10. Conclusions: It seems kite-marks are very similar to badges, as adopted in psychology? (https://fivethirtyeight.com/features/even-psychologists-respond-to-meaningless-rewards/) Are there differences worth mentioning, or maybe just cite both?

Any opinions and conclusions expressed herein are those of the authors and do not necessarily reflect the views of the U.S. Census Bureau.

Is the rationale for developing the new method (or application) clearly explained?
Yes

Is the description of the method technically sound?
Partly

Are sufficient details provided to allow replication of the method development and its use by others?
Yes

If any results are presented, are all the source data underlying the results available to ensure full reproducibility?
Yes

Are the conclusions about the method and its performance adequately supported by the findings presented in the article?
Partly

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: Reproducibility, research transparency, poverty programs, applied econometrics, labor economics

I confirm that I have read this submission and believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard, however I have significant reservations, as outlined above.

Author Response 11 Jun 2020

Jade Benjamin-Chung, University of California, Berkeley, Berkeley, USA

Thank you for these constructive comments. Below we describe how we have addressed them in the revised manuscript. We feel that the revisions in response to these comments have strengthened the manuscript.

A. How does your method compare to journal post-acceptance pre-publication review, as currently done in several political science journals?

Response: Thank you for raising this question. We have added the following section that compares internal replication to pre-publication review in the Methods section.

"Internal replication vs. pre-publication review

One strategy to increase reproducibility that is complementary to internal replication is pre-publication review in which journals replicate study findings using replication scripts and datasets prior to accepting a manuscript for publication [Gertler et al., 2018]. The American Journal of Political Science is one example of a journal that uses this approach. Internal replication before submission to peer review would catch errors internally before peer reviewers and journal editors consider a manuscript. Detecting errors after submission or after peer review is less efficient because it may require another round of peer review and revision."

Gertler P, Galiani S, Romero M (2018) How to make replication the norm. In: *Nature*. http://www.nature.com/articles/d41586-018-02108-9. Accessed 9 Mar 2018

B. How does your method work for solo researchers?

Response: We have added the following text in the first paragraph of the Conclusions:

"In some disciplines research is conducted alone, and in this case a single analyst could partially replicate their own work by programming the analysis in alternative software packages (e.g., R vs. Stata) or by writing the same error-prone section of code twice."

C. How does your method work in a world of proprietary or confidential data that cannot be publicly released? Something like 40% of articles in a top economics journal use data that can't be released, and the Federal Statistical Research Data Centers produce important research from data that can't be publicly released. (This is probably a point in your favor, since it's harder to do external replications of non-sharable data.)

Response: The internal replication process is even more important for papers that cannot publish their data. We have added the following to the Conclusions:

"Internal replication is one of many tools available to researchers to increase the reproducibility of their work, including publication of pre-analysis plans and publishing analytic datasets. A large proportion of studies are unable to publish their datasets due to human subjects protections or other privacy restrictions. In these cases, since external replication may not be possible, internal replication is even more valuable."

D. How does your method compare in costs and benefits to pair programming?

Response: We have added the following text to the Methods section:

"Internal replication vs. pair programming

At first glance, internal replication may resemble pair programming, a practice in which one analyst writes code while the other simultaneously reads and comments on the code to suggest coding strategies and improvements. Costs associated with pair programming and internal replication are likely to be similar since both approaches require two analysts to complete a single analysis. Unlike pair programming, in internal replication the vast majority of coding is done independently with minimal communication until data or results are compared. The advantage of this approach over pair programming is that it allows analysts to pursue completely different coding strategies which may be subject to differing sources of error and bias. Pair programming may be more subject to "group think" in which shared biases or judgment calls are reinforced or amplified. Thus, we believe that internal replication is more likely to identify failures to replicate results due to coding errors and biases than pair programming."

Details in order of appearance:

1. Introduction, paragraph 2: I'm not sure citation 15 argues for external replication, at least not exclusively. Doesn't it make the case for quasi-internal, or at least pre-publication review (by the journal itself, after acceptance but before publication)? Political science journals such as AJPS do this now. It's just a "does your code actually produce your tables" and not as thorough as your method, but I think you should clarify this and possibly compare or contrast this method of journal pre-publication review to your method.

Response: We agree, and we have revised the sentences citing this paper in the Introduction as follows:

"In addition, external replication may create an incentive for replicators to overturn original study findings that introduces bias and undermines constructive scientific discourse [Gertler et al., 2018]. An alternative approach that does not create such an incentive is for journals to conduct "pre-publication review" in which they attempt to replicate study findings using data and analytic code submitted prior to publication [Gertler et al., 2018]."

Gertler P, Galiani S, Romero M (2018) How to make replication the norm. In: *Nature*. http://www.nature.com/articles/d41586-018-02108-9. Accessed 9 Mar 2018

Regarding pre-publication review, please see our response to item A above.

2. "a landmark study was recently retracted...": In which field/on what topic?

Response: The authors of the original study were in both sociology and public health departments, so we have revised the text to refer to the study as a *"landmark social science study".*

3. Status quo workflow paragraph: I think the use of citation 21 as an example needs to be more nuanced. Which "one of the study's most novel, policy-relevant findings"? I believe the original study authors would disagree with that claim. For example, from the Commentary response (which could also be cited) to cite 21: "The argument in Davey *et al*., that deworming impacts on school participation are not robust to different statistical approaches, is based on several analytical errors." Anyway, the point that an external replication led to disagreement or even controversy is certainly valid.

Response: We have added additional details about the specific result we referred to in that section as shown below. We were referring to the finding of lower worm infections in control schools 3-6 km away from intervention schools, not to school participation results. We believe that our sentence stating that "one of the study's [...] findings" reflects the fact that the "pure" replication effort identified errors that only changed some, not all, of the original study's findings when corrected. We have also cited the Hicks et al. commentary responding to Aiken et al. We have not added the Davey et al. citation because it is not strictly a replication, but included re-analyses using different statistical approaches than the original paper.

"For example, a recent external replication [Aiken et al., 2015] of a highly influential study that found externalities of school-based deworming identified a coding error in a variable used in a regression model; when corrected, one of the study's most novel, policy-relevant findings (that worm infections were lower in control schools 3-6 km away from intervention schools) was closer to the null and no longer statistically significant [Miguel & Kremer, 2004; Hicks et al., 2015]."

Aiken AM, Davey C, Hargreaves JR, Hayes RJ (2015) Re-analysis of health and educational impacts of a school-based deworming programme in western Kenya: a pure replication. *Int J Epidemiol* 44:1572–1580 . https://doi.org/10.1093/ije/dyv127

Miguel E, Kremer M (2004) Worms: Identifying Impacts on Education and Health in the Presence of Treatment Externalities. *Econometrica* 72:159–217 . https://doi.org/10.1111/j.1468-0262.2004.00481.x

Hicks JH, Kremer M, Miguel E (2015) Commentary: Deworming externalities and schooling impacts in Kenya: a comment on Aiken et al. (2015) and Davey et al. (2015*). Int J Epidemiol* 44:1593–1596 . https://doi.org/10.1093/ije/dyv129

4. Methods including internal replication paragraph: "regardless of units" is a little unclear. I think I get the point about ultimate display in a journal, but often coefficients will be scaled. For example, population or income will be divided by 10K or 1m to avoid having to display a bunch of zeroes.

Response: We have revised this sentence as follows:

"*For example, a tolerance level could be chosen so that any differences in results between replicators are small enough that they would not appear in published manuscripts."*

5. Box 1, Software: Using non-object-oriented languages. What exactly do you mean by "makes it more difficult to load objects of different dimensions and compare them simultaneously"? An example could help.

Response: We have revised this part of Box 1 accordingly:

*"**Software:** Using statistical software such as R or Python that allows data analysts to efficiently load a large number of objects of differing dimensions (e.g., scalars, vectors, and matrices of differing dimensions from different data sources), take the difference between them, and identify which are replicated facilitates replication. Other languages such as Stata or SAS, allow objects of different dimensions to be loaded simultaneously, but the default is to work with a particular dataset with specific dimensions. As a result, while replicating, it may be more difficult to efficiently compare large numbers of matrices or other objects generated by each analyst to check for replication when using these languages. If analysts use the same software, we recommend that they use the same version of the software to ensure that differences in their results are not due to differences in software versions."*

6. Variable type: truncating the number of sig figs: I assume you are talking about character vs. numeric. Is this not just a machine precision/binary representation thing that all languages/computers do? Please clarify.

Response: This is a good point, and we have revised this part of Box 1 accordingly:

*"**Variable type**: Agreeing upon the variable type, particularly for continuous variables, facilitates smooth replication. Some software truncates the number of significant figures when saving numeric variables in different formats. For example, since Stata stores numeric variables in binary, the value 0.1, which does not have a perfect binary representation, is stored differently for float and double variable types. These differences can carry forward and prevent replication."*

7. Bottom of page 4/just after Box 1: add a link here to the dashboard. It appears later, but this is the first mention.

Response: We have done so.

8. Last paragraph before results: You have not discussed the large world of solo investigators. How do they fit into this scheme?

Response: We have added the following text in the first paragraph of the Conclusions:

"In some disciplines research is conducted alone, and in this case a single analyst could partially replicate their own work by programming the analysis in alternative software packages (e.g., R vs. Stata) or by writing the same error-prone section of code twice."

9. Results, making computational analysis...paragraph: "we kept notes about their judgement calls...": This seems close to pair programming. Software teaching organizations seem to advocate for this, and say it reduces errors. How would your method compare in terms of cost and reliability? Did you consider pair programming?

Response: We have added the following paragraph comparing internal replication and pair programming in the Methods section.

"Internal replication vs. pair programming

At first glance, internal replication may resemble pair programming, a practice in which one analyst writes code while the other simultaneously reads and comments on the code to suggest coding strategies and improvements. Costs associated with pair programming and internal replication are likely to be similar since both approaches require two analysts to complete a single analysis. Unlike pair programming, in internal replication the vast majority of coding is done independently with minimal communication until data or results are compared. The advantage of this approach over pair programming is that it allows analysts to pursue completely different coding strategies which may be subject to differing sources of error and bias. Pair programming may be more subject to "group think" in which shared biases or judgment calls are reinforced or amplified. Thus, we believe that internal replication is more likely to identify failures to replicate results due to coding errors and biases than pair programming. "

10. Conclusions: It seems kite-marks are very similar to badges, as adopted in psychology? (https://fivethirtyeight.com/features/even-psychologists-respond-to-meaningless-rewards/) Are there differences worth mentioning, or maybe just cite both?

Response: Yes, our understanding is that kite-marks and badges are essentially the same (see the Rowhani-Farid citation below). We have revised the text to use both terms and included new citations as well.

Kidwell MC, Lazarević LB, Baranski E, Hardwicke TE, Piechowski S, Falkenberg L-S, Kennett C, Slowik A, Sonnleitner C, Hess-Holden C, Errington TM, Fiedler S, Nosek BA (2016) Badges to Acknowledge Open Practices: A Simple, Low-Cost, Effective Method for Increasing Transparency. PLoS Biol 14:e1002456 . https://doi.org/10.1371/journal.pbio.1002456

Rowhani-Farid A, Barnett AG. Badges for sharing data and code at Biostatistics: an observational study. *F1000Res* 2018; **7**. DOI:10.12688/f1000research.13477.2.

Competing Interests: The authors have no competing interests to declare.