sociologists need to be better at replication – a guest post by cristobal young

Cristobal Young is an assistant professor at Stanford's Department of Sociology. He works on quantitative methods, stratification, and economic sociology. In this post co-authored with Aaron Horvath, he reports on the attempt to replicate 53 sociological studies. Spoiler: we need to do better.

Do Sociologists Release Their Data and Code? Disappointing Results from a Field Experiment on Replication.

Replication packages – releasing the complete data and code for a published article – are a growing currency in 21st century social science, and for good reasons. Replication packages help to spread methodological innovations, facilitate understanding of methods, and show confidence in findings. Yet, we found that few sociologists are willing or able to share the exact details of their analysis.

We conducted a small field experiment as part of a graduate course in statistical analysis. Students selected sociological articles that they admired and wanted to learn from, and asked the authors for a replication package.

Out of the 53 sociologists contacted, only 15 of the authors (28 percent) provided a replication package. This is a missed opportunity for the learning and development of new sociologists, as well as an unfortunate marker of the state of open science within our field.

Some 19 percent of authors never replied to repeated requests, or first replied but never provided a package. More than half (56 percent) directly refused to release their data and code. Sometimes there were good reasons. Twelve authors (23 percent) cited legal or IRB limitations on their ability to share their data. But only one of these authors provided the statistical code to show how the confidential data were analyzed.

Why So Little Response?

A common reason for not releasing a replication package was because the author had lost the data – often due to reported computer/hard drive malfunctions. As well, many authors said they were too busy or felt that providing a replication package would be too complicated. One author said they had

never heard of a replication package. The solutions here are simple: compiling a replication package should be part of a journal article's final copy-editing and page-proofing process.

More troubling is that a few authors openly rejected the *principle* of replication, saying in effect, "read the paper and figure it out yourself." One articulated a deep opposition, on the grounds that replication packages break down the "barriers to entry" that protect researchers from scrutiny and intellectual competition from others.

The Case for Higher Standards

Methodology sections of research articles are, by necessity, broad and abstract descriptions of their procedures. However, in most quantitative analyses, the *exact* methods and code are on the author's computer. Readers should be able to download and run replication packages as easily as they can download and read published articles. The methodology section should not be a "barrier to entry," but rather an on-ramp to an open and shared scholarly enterprise.

When authors released replication packages, it was enlightening for students to look "under the hood" on research they admired, and see exactly how results were produced. Students finished the process with deeper understanding of – and greater confidence in – the research. Replication packages also serve as a research accelerator: their transparency instills practical insight and confidence – bridging the gap between chalkboard statistics and actual cutting-edge research – and invites younger scholars to build on the shoulders of success. As Gary King has emphasized, replications have become first publications for many students, and helped launched many careers – all while ramping up citations to the original articles.

In our small sample, little more than a quarter of sociologists released their data and code. Top journals in political science and economics now require on-line replication packages. Transparency is no less crucial in sociology for the accumulation of knowledge, methods, and capabilities among young scholars. Sociologists – and ultimately, sociology journals – should embrace replication packages as part of the lasting contribution of their research.

Table 1. Response to Replication Request

Response	**Frequency**	**Percent**
Yes: Released data and code for paper	15	28%
No: Did not release	38	72%
Reasons for "No"		
IRB / legal / confidentiality issue	12	23%
No response / no follow up	10	19%
Don't have data	6	11%
Don't have time / too complicated	6	11%
Still using the data	2	4%

'See the article and figure it out'	2	4%
Total	53	100%

Note: For replication and transparency, a blinded copy of the data is available on-line. Each author's identity is blinded, but the journal name, year of publication, and response code is available. Half of the requests addressed articles in the top three journals, and more than half were published in the last three years.

Figure 1: Illustrative Quotes from Student Correspondence with Authors:

Positive:

1. "Here is the data file and Stata .do file to reproduce [the] Tables…. Let me know if you have any questions."
2. "[Attached are] data and *R* code that does all regression models in the paper. Assuming that you know *R*, you could literally redo the entire paper in a few minutes."

Negative:

3. "While I applaud your efforts to replicate my research, the best guidance I can offer is that the details about the data and analysis strategies are in the paper."
4. "I don't keep or produce 'replication packages'… Data takes a significant amount of human capital and financial resources, and serves as a barrier-to-entry against other researchers… they can do it themselves."

50+ chapters of grad skool advice goodness: Grad Skool Rulz ($2!!!!)/From Black Power/Party in the Street

Written by fabiorojas

August 11, 2015 at 12:01 am

Posted in epistemology and methods, fabio, guest bloggers, research

36 Responses

Subscribe to comments with RSS.

> The majority of my articles are published from well-known and easily accessible public use datasets. I don't mind providing people with code for particular variables, but there is a bit of a "suck it up and do it yourself" that echoes in my own mind when I'm sending some kid a bunch of code that I have always had to write myself. "Read the god damned paper and figure it out yourself" is not an unreasonable response when one is using publicly available data. If I fucked something up, then you can point that out. I've always just sent people the code, but, frankly, for

papers using public dataset….maybe I shouldn't. Do your own work. And, I resent the fuck out of all of these fake experiments. I can't believe any of them pass human subjects. I'm not a subject. You should not be doing research to me to see how cooperative I am with your lazy worthless ass.

sherkat

August 11, 2015 at 1:00 am

Tell me how you *really* feel!

fabiorojas

August 11, 2015 at 1:07 am

I kinda agree with Sherkat here. Replication is important, but if I found out the replication request was part of an experiment to make someone else's bones, I'd be annoyed.

I've been asked for replication data twice. Once, it was a grad student who wanted a particular variable for their own work that I had compiled. Request granted and I was happy to see the study in print. The other time it was an undergrad who was working through a class assignment (experiment even?). Request granted again, as it should be.

But for those of us in the same boat as Jessica (https://scatter.wordpress.com/2015/07/20/help-wanted-managing-workflow/), we should all recognize that such request do take time. Which is actually a good argument for journals and sociology departments instituting replications standards so that it's just a matter of course rather than a hassle.

cwalken

August 11, 2015 at 3:28 am

One important clarification: the student assignment was to replicate a study that they admired, getting to work hands on with real world data, understand exactly how the analysis was done, and to run some alternative model specifications. It was a very successful assignment, in the sense students got a lot out of it.

A byproduct of that assignment is that we had this data which we are sharing here. I could not imagine asking anyone for a replication package 'just to see what they say'. Nobody asked anyone for a package just out of curiosity. Authors who shared their replication packages helped grad students better understand both their article and the methods they used.

Cristobal Young

August 11, 2015 at 4:02 am

I think replication assignments are a great idea and should be more widely used. Supporting replication work is only more work for people if they are not conducting their analysis in an organized way to begin with. it's just sloppy practice, the sort of thing that was maybe fine in the

1990s before ideas about the importance of maintaining good code became widespread, but now is a sign of somebody who is more slob than scientist and whose work should be evaluated accordingly.

jeremy

August 11, 2015 at 2:19 pm

A resource for those of you who want to get a head start muddling through some of more daunting publicly-available sets: http://www.asdfree.com

JNCohen

August 11, 2015 at 2:35 pm

Jeremy: I assume you'd support a push for all ASA journals to host replication files for public-use data as technical appendices? Would asking for these at the review stage be going too far?

micah

August 11, 2015 at 3:37 pm

It is interesting that all of the replication materials from sociology journals came from AJS (3 out of 10), ASR (4 out of 14) or Social Forces (6 out of 10). The other two came from APSR and Strategic Management. The other journals are an eclectic mix of places sociologists might publish, so it isn't simply a low/high status thing.

neal caren

August 11, 2015 at 3:45 pm

I thought this sounded familiar…sure enough, I found an email from early 2014 from a Stanford student asking me about providing data from a paper I wrote (with reference to Young's class). Interesting to be on the other side of the process for once. And Darren, I'd love to have you in the audience at our Audit Studies and Field Experiments special session at ASA. We'll be talking about IRB, ethics, and other interesting things, if you're so inclined.

S. Michael Gaddis (@smgaddis)

August 11, 2015 at 4:52 pm

"More than half (56 percent) directly refused to release their data and code." Why don't you publish the list of these people? Think of it as providing your replication dataset for others to see.

Gary King

August 11, 2015 at 6:42 pm

What were the distributions of years-since-publication of the "failed' vs. "successful" replications?

krippendorf

August 11, 2015 at 8:51 pm

I'd like to know more about how people were contacted: What was the text of the email? How many follow up emails were sent? Were they sent from .edu accounts or something else? While I'm not exactly in the Sherkat camp, a single email from some rando on a gmail account is unlikely to ever be read, let alone receive a response. In that light, an 81% response rate – even if to say "leave me alone" in various ways is pretty impressive.

Tom

August 11, 2015 at 9:01 pm

All the publications that the authors refused to provide replication data and codes should be retracted. Period. Aren't we doing science or producing BS?

John

August 11, 2015 at 9:26 pm

Great exercise. I do hope sociology moves in this direction. One issue that hasn't been mentioned though is that sometimes researchers deal with proprietary data that cannot be legally shared. I happen to work with twitter data and I had to sign a legal agreement to not disclose it to other parties to protect the privacy of twitter users. I assume other social media data are governed by the same rules. However, even with these restrictions, I do think that researchers like me could share the do file so others can learn more about our modeling strategies, etc.

Rene Flores

August 11, 2015 at 9:47 pm

I guess we just cannot do that. Because if we do, then probably not a small percentage of soc profs. will lose their jobs. It's not uncommon for biologists, chemists, physicians, and psychologists to cook up their data. Aren't we exceptions, especially since we even discourage our authors to make their data and codes publicly available? I've seen a couple of cases, in which even the reading of the questionnaire/variables from a database that I basically breathe with is totally wrong, but got published in high-impact journals. Once I requested data for an article published in one of our top three, one of the authors said he/she couldn't find it and the other simply ignored it. Sociology is probably the only discipline with a publication practice that has extremely long turn-around time without codes or data available to readers. I believe we are doomed to decline if this continues. We are not doing science; we are doing business. It's all about grants and quick-and-dirty publications. Tweak your data so that you have a good storyline (better to align well with the dominant discourse in our filed), put in some fancy models that most reviewers won't even understand but pretend to, have a couple of co-authors that are from top schools and rotate authorship, reciprocate promotion from your inner circle, and then viola you see a long list of junk that soon everyone forgets.

Josh

August 11, 2015 at 9:51 pm

Micah: Sorry for how jaded this will come off, but my view is that if there is any collective interest among quantitative sociologists in improving the quality or credibility of their craft, they need to engage in action that is not mediated by an organization that has enormous inertia and *very* little representation of quantitative expertise at any of its key decision-making levels. In other words, expecting that quantitative sociology will improve because of the leadership, or even followership, of its professional organization is waiting at a bus stop for a route that was discontinued long ago.

In my view, the more promising course is for other journals to show leadership, and for quantitatively-inclined readers to reward that leadership in how they think about the pecking order of different journals. Demography is one obvious candidate. Soc Science is perhaps another.

jeremy

August 12, 2015 at 12:38 am

I'm sure slob scientists are all of the problem. Yeah, right. People who use generally available data are fully in line for genuine replication. Availing people of code is actually less likely to result in the rejection of faulty findings, since the lazy motherfuckers who use the available odes are only going to find the same shit that the original author found. Duh! What needs to be scrutinized are the people who have "proprietary" data sources. The people who get grants and contract with bullshit marketing firms and claim to find whatever they claim to find. The real slobs are the people who protect their proprietary bullshit non-random data collected from some fly-by-night marketing firm…That is the shit we need to be monitoring. Whore social scientists using garbage on-line data to claim gods knows what. Regnerus, and shit.

sherkat

August 12, 2015 at 1:23 am

FYI, Professor Sherkat is Orgtheory's Official Anger Translator.

fabiorojas

August 12, 2015 at 2:19 am

It would be good if researchers started thinking about replication from the early start of their research projects. This would eliminate all reasons mentioned by researchers in your experiment (too difficult,no time, hardware problems,..). There are excellent tools available for making your projects reproducible. I have written a blog about one such tool after the wormswar controversy :

https://www.ifpri.org/blog/encouraging-transparency-agricultural-research

bvancampenhout

August 12, 2015 at 8:38 am

Josh, in his comment above, made a great point about the distinction between sociology as a science (the advancement of knowledge) and sociology as a business pursuit (getting publications). I agree – replication packages help keep in check the business pressures in our discipline.

Sociologists need to uphold the linkage between our individual aspirations to publish, and the collective need for sociology to advance knowledge and understanding about the world. Business incentives can lead to the decoupling of publication from knowledge creation. It is important for the discipline that the mandate "publish or perish" in practice also means "create valuable knowledge or perish". How confident are we that the pressure to publish means pressure to find valuable evidence and accurate knowledge?

I insist that my students make replication packages for their work. That may not be in their best business interest, because having some wiggle room with the results can make their papers sound more exciting. But, it is unquestionably in the best interests of science and of sociology – what matters for the discipline is the accumulation of knowledge that is robust and genuinely important.

Jeremy makes a great point that the journal Demography might be ready (with some nudging) to require replication packages. Until then, we should remember that providing code and shareable data is a signal of quality in research, and is the default expectation in our peer disciplines.

Cristobal Young

August 12, 2015 at 9:00 am

After I finish up a project–i.e., once it's been accepted–I go through and produce a replication dataset. I do this for my own sake, so I can reproduce everything later if I have to, and so I don't have to keep a ton of intermediate files on my computer. It just involves cleaning up the code that I wrote for the analyses. It takes maybe an hour or two per project. Maybe three if I was really messy while working on it.

Not all of those replications could be released–like some of the authors you mention, I have worked with, e.g., EEOC data that I cannot share. But I can still show the code that cleans up files and runs analyses if people are interested. This seems straightforward enough.

John-Paul Ferguson

August 12, 2015 at 2:18 pm

Closely related: Vines, et al. "The Availability of Research Data Declines Rapidly with Article Age." Current Biology 24, no. 1 (January 2014): 94–97. doi:10.1016/j.cub.2013.11.014.

bbolker

August 12, 2015 at 3:28 pm

With me, it is really difficult and hard! But almost I have understand more from your post

wikihoidap

August 12, 2015 at 3:35 pm

As much transparency as possible that doesn't violate IRB standards or inundate the reader with needless information should be the standard for research. Economists have always been very forthcoming with the statistical methods they use to arrive at their conclusions and build their models. Why haven't sociologists adopted the same standards? While an argument can be made that a trained investigator should be able to figure out for him or herself how the original investigators arrived at their conclusions, particularly if the data are publicly available, the fact of the matter is that some data sets, such as the PSID, are just plain messy to begin with and sometimes require complex preliminary manipulation before any actual inferential statistical tests are performed with the data.Not being transparent with this preparation is being obfuscatory to both the scholarly community and the review committee.

But a bigger problem, in my opinion, is journals' lack of statistical criticism where the investigator has shown that the tests meet their basic assumptions. It's not as big of a deal with more forgiving tests such as ANOVA, but I wonder how many published regression analyses are actually erroneous on their face, because they don't meet their own test assumptions. However, with journal pages at a premium, I don't foresee a big push for investigators being required to produce evidence that the data meets the multivariate models being reported.

Paul

August 12, 2015 at 3:43 pm

I knew I recognized this request. And I recognize my paper in the list (I was a "no" due to not having the data)–though I wasn't lead author, I wrote the code. The code was for a paper written while I was a grad student with little knowledge of the publishing and post-publishing world and was lost to a computer virus.

While I strongly support replication (having just struggled to try to replicate someone else's work for months), the failure to produce the data had nothing to do with sloppiness in my science (albeit sloppiness in my streaming soccer games while a grad student, I will admit to) or unwillingness to share. It was, in large part, due to ignorance of a relatively new interest in social science as a grad student now almost a decade ago from when I first started working on the paper. And much of the discussion here either explicitly claims or implicitly accuses non-replication as a sign of shoddy science or ethics. That's simply not the case and I stand behind our (extremely straightforward and easy to replicate sans code) regression. It is not a lie, it was science and it should not be asked to be retracted. I am admittedly embarrassed by my early-career mistake of not backing up that code from a project onto a separate computer, but my name does not need to be published publicly so I can be publicly shamed for that (as it's no longer the case). Assume good intent to gain allies in the push for more replication.

The email was from an edu account and mentioned the class. It also seemed to assume that replication was already a growing norm, which, i think clearly, is not the case in our field.

rk

August 12, 2015 at 10:24 pm

Jeremy, you do realize Demography is an association journal just like ASR, right? Just an association whose bus you assume hasn't stopped running? This doesn't make sense. Professional associations are going to be where publication standards are set. (Or, you know, trying to convince unaccountable individuals to change the policies at their private journals.)

Philip N. Cohen

August 12, 2015 at 11:38 pm

Phil: I think PAA is a more scientifically-minded organization and so I like its prospects for scientifically-minded change in its publications. The problem here isn't with associations per se, as many of the various changes that have been noted for other journals have happened at "association journals" of other disciplines.

jeremy

August 13, 2015 at 3:58 am

I sense that there is a bit of a backlash building against such requests. You never know what is going to happen when someone who doesn't work in your area and who you've never heard of requests a replication package. Increasingly, I hear people (in different disciplines) openly worrying: Am I going to get jacked by some tweeter or blogger with an agenda? Will an army of keyboard activists come after me? Will I wind up on Fox News or MSNBC? Will some nutcase start stalking me and hanging out in my building? I strongly support the idea of a blanket policy that rep packages be required as part of the publication process. I always provide code and data (when not prohibited), and will continue to do so, but the idea that we're all obviously duty bound to respond to essentially anonymous requests is at least questionable, especially in a place like the U.S. with very lax "protections" against defamation.

TR

August 13, 2015 at 2:03 pm

personally, my lesson learned is this: do not post on a sociology blog after a night of drinking. Apologies for my atrocious grammar above!

rk

August 13, 2015 at 3:05 pm

agreed w TR above. the recent debacle about Kremer and Miguel's worms paper in economics is a good example, where a journalist at Buzzfeed took a mediocre replication to a fairly extensive degree to critique data sharing.

ZC

August 14, 2015 at 12:40 am

[…] Lake May Not Be Great for Athletes' Health The worst piece of peer review I've ever received Sociologists need to be better at replication The biggest infectious disease threat we face isn't Ebola – it's our short attention span CDC […]

Links 8/14/15 | Mike the Mad Biologist

August 14, 2015 at 8:45 pm

[…] It was discovered that only about a quarter of sociologists are able, permitted, or willing to provi…. […]

replication and the future of sociology | orgtheory.net

August 17, 2015 at 12:01 am

I total agree with this: Josh, in his comment above, made a great point about the distinction between sociology as a science (the advancement of knowledge) and sociology as a business pursuit (getting publications). I agree – replication packages help keep in check the business pressures in our discipline.

Sociologists need to uphold the linkage between our individual aspirations to publish, and the collective need for sociology to advance knowledge and understanding about the world. Business incentives can lead to the decoupling of publication from knowledge creation. It is important for the discipline that the mandate "publish or perish" in practice also means "create valuable knowledge or perish". How confident are we that the pressure to publish means pressure to find valuable evidence and accurate knowledge?

I insist that my students make replication packages for their work. That may not be in their best business interest, because having some wiggle room with the results can make their papers sound more exciting. But, it is unquestionably in the best interests of science and of sociology – what matters for the discipline is the accumulation of knowledge that is robust and genuinely important.

Jeremy makes a great point that the journal Demography might be ready (with some nudging) to require replication packages. Until then, we should remember that providing code and shareable data is a signal of quality in research, and is the default expectation in our peer disciplines.

cach lam sua chua

August 17, 2015 at 3:43 pm

[…] contacted, only 15 of the authors (28 percent) provided a replication package." To read blog, click here. This compares to a recent study of economists that found that 44 percent provided data and code […]

At Least We're Better Than Sociologists? | The Replication Network

August 23, 2015 at 1:08 am

I do not see a problem with asking people whether they are prepared to share their code and data just to test whether access is granted. We also need to study how we do science, although I see that this was not the major purpose of the replication package requests. Besides all this, I do not understand why this is called a "field experiment".

ingorohlfing

August 23, 2015 at 8:34 pm

[…] to progress in science. Reproducibility requires the proper storage and sharing of data, which cannot be taken for granted, and a detailed, step-by-step description of the empirical analysis. Statistical research has no […]

Reproducible QCA studies with fs/QCA | Politics, Science, Political Science

August 27, 2015 at 8:57 pm

Comments are closed.