Failure to Reproduce: The Replication Crisis in Research – Can Librarians Help?

Andrée Rathemacher, Amanda Izenstark, & Harrison Dekker
University of Rhode Island Libraries

May 4, 2018
ACRL New England Annual Conference

Failing Forward: Experimentation and Creativity in Libraries
ACRL/NEC Conference 2018
2018 ACRL New England Chapter Annual Conference
Friday, May 4, 2018
Hotel 1620 Plymouth Harbor
Plymouth, Massachusetts

We often talk at conferences about projects that went well. In contrast, we rarely discuss initiatives that failed, or unexpected obstacles that forced us to find another route to success. In our 2018 conference, the ACRL New England chapter is highlighting experimentation and creativity in college and research libraries by acknowledging that missteps and roadblocks are all part of the process. Join us in Plymouth, Massachusetts in May 2018 to talk about 'failing forward.'

Official conference hashtag #acrlnec18

Conference materials https://scholarworks.umass.edu/acrl_nec_conf/2018/

Your session, "Failure to Reproduce: The Replication Crisis in Research - Can Librarians Help?", will be held during Breakout Session IV held from 3:40 - 4:30pm, in the Carver room. Please remember to plan for a 40-minute session, followed by a 10-minute Q&A.
1. The Reproducibility Crisis

2. Reproducible Workflows

3. Introduction to the Open Science Framework

1. Speaker introductions:

Andrée Rathemacher, Head, Acquisitions
Amanda Izenstark, Reference & Instructional Design Librarian
Harrison Dekker, Data Services Librarian

University of Rhode Island

2. Talk outline [see slide]
The Reproducibility Crisis
“It can be proven that most claimed research findings are false.”

– John P. A. Ioannidis, 2005

Those are the words of John Ioannidis (yo-NEE-dees) in a highly-cited article from 2005.

Now based at Stanford University, Ioannidis is a meta-scientist who conducts “research on research” with the goal of making improvements.

Sources:


“Reproducibility crisis”
(aka “replication crisis”)

“A methodological crisis in science in which scientists have found that the results of many scientific experiments are difficult or impossible to replicate on subsequent investigation, either by independent researchers or by the original researchers themselves.”

— Wikipedia

A recent survey by Nature found that more than 70% of researchers have tried and failed to reproduce another scientist’s experiments and more than half have failed to reproduce their own experiments.

90% of respondents agreed there is a reproducibility crisis.

Just a brief note on terminology: Many experts use reproducibility and replication more or less interchangeably, while others make distinctions between the two terms.

For the purposes of this presentation, it may help to distinguish between two concepts: 1) a broad problem in research in general, in which a large portion of research findings do not turn out to be true as initially reported and, 2) the necessary conditions for any given experiment to be repeated with consistent results.

Source:


Psychology is one of the disciplines where the crisis has received the most attention.

It is also a field where 91.5% of all published studies found positive results, that is, they supported the outcome predicted by researchers.

In 2015, the results of the Reproducibility Project: Psychology were published.

This was a four-year project lead by Brian Nosek, a professor of psychology at the University of Virginia and the co-founder & executive director of the Center for Open Science.

In the project, 270 authors replicated 100 social and cognitive psychology studies that had been published in three top psychology journals in 2008.

While 97% of the original studies produced significant results, only 36% of the replications did. Also, the effect sizes in the replication studies were only about half that of the original studies.

Sources:

Economics is another discipline with a reputation for non-reproducibility.

A 2015 paper by two researchers from the Federal Reserve and the Department of the Treasury tried to replicate results from 67 papers published in 13 prestigious economics journals. Even after contacting the authors of the original studies when necessary, they were only able to replicate 49% of the results. They concluded, “Because we were able to replicate less than half of the papers in our sample even with help from the authors, we assert that economics research is usually not replicable.”

Sources:


“Homeless man in Vancouver” by Jay Black is licensed under CC BY-SA 2.0.
Reproducibility has also been a big problem in biomedical research, or so-called “preclinical research.”

Preclinical research describes the type of exploratory research that goes on at universities (as opposed to clinical research and testing in humans that may be conducted by drug companies, for example).

In 2011, a team from Bayer reported that in only 20-25% of 67 seminal studies they tried to reproduce did they come to results “completely in line” with those of the original publications.

In 2012, Glenn Begley, the former head of cancer research at Amgen, reported that over a 10 year stretch, Amgen’s scientists had tried to replicate the findings of 53 landmark studies in cancer biology. They obtained positive results in only 6, a rate of just over 11%.

Ioannidis studied 49 of the most highly-regarded research findings in medicine published between 1990-2003, each cited more than 1,000 times. These were articles that helped lead to the widespread popularity of treatments such as the use of hormone-replacement therapy for menopausal women, vitamin E to reduce the risk of heart disease, coronary stents to ward off heart attacks, and daily low-dose aspirin to control blood pressure and prevent heart attacks and strokes. Of the 49 claims, 34 had been re-tested, and 14, or 41% had been convincingly shown to be wrong or significantly exaggerated.
One analysis of past studies indicates that the total prevalence of irreproducible preclinical research exceeds 50%, resulting in approximately $28 billion a year spent on research that is not reproducible, in the United States alone.

Sources:


Why? “File-drawer problem”

Researchers do not bother to write up experiments with negative / null results or the results of replication studies. Instead of submitting them to journals, they file them away.

“Filing” by Jeff Youngstrom is licensed under CC BY-NC 2.0.

File-drawer problem: Researchers do not bother to write up experiments with negative / null results or the results of replication studies.

Instead of submitting them to journals, they file them away.

They believe journals are not interested in publishing negative results or replications.

Replication studies are viewed as lacking prestige, originality, or excitement. They can be seen as a challenge to authority in the field.

Researchers fear that replications and negative results will not get them published, promoted, or even hired.
Why? Publication bias

"...the small proportion of results chosen for publication are unrepresentative of scientists’ repeated samplings of the real world."

— Neal S. Young, John P. A. Ioannidis, and Omar Al-Ubaydli, 2008

Not only are researchers biased against writing up and submitting negative results, but journals are biased toward publishing positive results.

A 2012 study by Danielle Fanelli of the University of Edinburgh analyzed over 4,600 papers published in all disciplines between 1990 and 2007. She found that the proportion of positive results increased from 70% in 1990-1991 to 86% in 2007.

Sources:


Another factor in the reproducibility crisis is so-called “questionable research practices.”

These are problems with experimental design and data analysis that include:

- Using sample sizes that are too small -- while results might be positive, their statistical power is low
- HARK-ing (hypothesizing after the results are known): instead of stating the hypothesis in advance of the study, researchers come up with a hypothesis after they have conducted their experiment, making the hypothesis fit the data collected.
- “P-hacking”
  - Viewing experimental data are they are coming in and stopping the experiment once a positive result is obtained
  - Performing hundreds of tests on dozens of variables and only reporting those that produced positive results
- In biological sciences, not validating reagents, antibodies, cell lines, and other lab materials
- Not attempting to replicate results before publication
- Not documenting experimental methods adequately, making it impossible for others to replicate the study
- Not openly sharing data and code underlying the experiment.
Source:

Why? Incentive structure

“Today I wouldn’t get an academic job. It’s as simple as that. I don’t think I would be regarded as productive enough.”

— Peter Higgs, 2013 (winner of the 2013 Nobel Prize in Physics)

Perhaps the biggest (and most intractable) reason for the reproducibility crisis is the incentive structure for academic researchers.

Researchers know that they need to publish regularly in the most prestigious journals possible in order to get tenure and receive grants.

● One piece of evidence showing the growing competition for academic research positions is that there have been persistent increases in the average number of publications of researchers at their time of hiring.

Because high-prestige journals like to publish exciting, surprising, or “sexy,” results, there is incentive to test hypotheses that are unlikely.

● So, for example, in the years between 1974 and 2014, the frequency of the words “innovative,” “groundbreaking,” and “novel” in PubMed abstracts increased by 2500% or more.

To make sure a hypothesis has firm theoretical grounding and an experimental design is well powered slows down the rate of production, so sound science is discouraged.

Shai Silberberg of the NIH notes that, “As long as universities think that the way for investigators to get money is to publish in Nature and Science, then that’s what they’re going to look for in investigators. They want that money.”
Sources:


What about peer review?

“The need to get away from the notion, proven wrong on a daily basis, that peer review of any kind at any journal means that a work of science is correct.”

— Michael Eisen, 2014

What about peer review?

The Cochrane Collaboration reviewed the evidence on peer review of manuscripts and grant proposals. They concluded, “At present, little empirical evidence is available to support the use of editorial peer review as a mechanism to ensure quality of biomedical research.”

The British Medical Journal took a study that was about to be published and inserted 8 errors in it, then sent it out to 300 reviewers. The median number of errors spotted was two, and 20% of the reviewers did not spot any.

And, one of the most well-known failures of peer review was the article by Andrew Wakefield, et al., published in The Lancet in 1998, finding that the MMR vaccine caused autism. That article was reviewed by 6 reviewers, none of whom found any of the problems with it that were later identified.

In short, peer review fails to detect false research findings primarily because:

- Reviewers don’t re-run experiments or examine the underlying data. Even if they did, they could have no knowledge as to what data investigators have chosen to exclude or of any questionable research practices investigators have engaged in.
- And again, incentives: Reviewing is not highly valued for career advancement — thus diligence in reviewing and time spent are not rewarded.
[Alternatives: Publish everything and then let the world decide what is important; post-publication peer review; increased transparency; credit for reviewers.]

Another quote: Long-time editor Irene Hames: “I don’t think that saying something is ‘peer reviewed’ can any longer be considered a badge of quality or rigour.” [Retractionwatch interview]

Sources:


Reproducible Workflows

Now that we’ve had an introduction to what is meant by “reproducibility crisis,” I’m going to talk about opportunities for Librarians to help train researchers on more transparent and reproducible practices. For full disclosures, all of my academic library work has been as a data librarian -- so the opportunities I’ve had when consulting with researchers often veer towards discussions of data manipulation and coding. That said, helping researchers to learn and adopt reproducible practices isn’t limited to teaching them to code and as we hope to show you in the remainder of this talk, there are opportunities to contribute for librarians of all backgrounds.
Communicating computational results

Modern data analysis typically involves dozens, if not hundreds of steps, each of which can be performed by numerous algorithms that are nominally identical but differ in detail, and each of which involves at least some ad hoc choices. If researchers do not make their code available, there is little hope of ever knowing what was done to the data, much less assessing whether it was the right thing to do.

Stark, 2018

Consider this quote from Philip Stark, a UC Berkeley Statistics professor and advocate for reproducible research:

Modern data analysis typically involves dozens, if not hundreds of steps, each of which can be performed by numerous algorithms that are nominally identical but differ in detail, and each of which involves at least some ad hoc choices. If researchers do not make their code available, there is little hope of ever knowing what was done to the data, much less assessing whether it was the right thing to do.

When we consider that more researchers from more disciplines are doing research that’s based on computation, then think about all the reviewers, collaborators, and future “consumers” with an interest in the validity of the research, I would argue that, to a significant extent, reproducibility is a scholarly communication challenge. As such, it seems appropriate that librarians are “at the table” when it comes to developing approaches for communicating the specific details of the decisions made by researchers as they prepare and analyze data.

Source:
Scholarship or advertising?

An article about computational science in a scientific publication is not the scholarship itself, it is merely advertising of the scholarship. The actual scholarship is the complete software development environment and the complete set of instructions which generate the figures.

- Jonathan Buckheit and David Donoho, 1995

In a 1995 publication, Stanford scientists Buckheit and Donoho famously paraphrased a their colleague geophysicist Jon Claerbout, an early advocate of reproducible research methods:
An article about computational science in a scientific publication is not the scholarship itself, it is merely advertising of the scholarship. The actual scholarship is the complete software development environment and the complete set of instructions which generate the figures.
------
This reinforces the case for framing reproducibility as a scholarly communication challenge, but there’s more going on here.

A lot has changed in the 20+ years since these remarks were made. As I’ve already mentioned, computational science has expanded, to more disciplines, more data and better hardware and software tools. In particular, scientists have access to programming languages like R and Python, which are open source, reliable, and have free documentation and learning materials and benefit from vast and supportive online communities that span many disciplines.

These languages are also, by design, extensible, thus making it possible for scientists to leverage the underlying language to build new tools to solve the specific needs of their disciplines. In other words, scientists face fewer constraints when they develop and share new tools. Furthermore, networks of software repositories exist making it easier for others to find and use these tools.
One of the side effects of this more open and democratized computing ecosystem is that it can be challenging, particularly for practicing researchers, to keep on top of change. This challenge can manifest itself in various ways. An obvious example is the need to quickly ramp up skills of research assistants who’s prior training might not meet a particular lab’s needs. Another example is knowing when a tool already exists to satisfy a particular need in the data workflow.

In my work as a data librarian, I’ve often been able to save a researcher time and effort of building a tool from scratch, by teaching them to use something that already exists. And I’d argue that this type of intervention is analogous to how librarians have traditionally intervened in research by, for instance, teaching student or researcher to use an existing reference tool rather than, say “browsing the stacks”.

Source:
The reproducible science movement has called attention to the notion of the analytical workflow and the importance of capturing all steps in code. This general model breaks the workflow into three distinct stages - and what takes place in each stage will vary by discipline. For instance, lab science might involve a very complex data acquisition stage involving experiments and instruments and in involve numerous files and formats, whereas social science might involve something as simple as a spreadsheet containing the results of a survey.

Even though data analysts frequently cite that they spend upwards of 80 percent of their time in acquisition and processing activities, it’s the final data analysis stage that has traditionally received the greatest emphasis in the curriculum. One strategy that I’ve employed in working with researchers is to advise them on how to use languages like Python, R, SQL, shell scripts to automate the acquisition and processing steps. (I should point out that with the growth of interest in reproducible research, there has been a concurrent growth in tools and practices for solving common challenges encountered in the acquisition and processing stages. I’ll introduce a few of these in a moment)

To help flesh out the idea of a data analysis workflow, I’ve provided a real world example from a book of case studies that was recently published. In this particular example, almost all steps from data ingestion to the creation of a final manuscript, complete with tables, plots, and bibliography have been automated. The implication of this model is that when the researcher needs to make a change, whether it be to the text, the underlying data, the code that generates a figure, a bibliography item, all that he must do to incorporate this change into the manuscript is to give a single command, which will invoke the various scripts that pull in the data, do the analysis, and assemble the outputs into a formatted document.

Contrast this with the manual processes we’ve probably all employed at one time or another to copy and paste figures into a document, carefully format table and bibliographies. I also suspect that many of us in this room can recall having to communicate a set of steps using screen captures or menu navigation narratives. It should be clear that this is not the best way to do reproducible science given how easy it is to leave out important details, or given the transient nature of user interfaces.

Source:
Workflow skills and tools

<table>
<thead>
<tr>
<th>Skill type</th>
<th>Description</th>
<th>Tools</th>
</tr>
</thead>
<tbody>
<tr>
<td>Literate computing</td>
<td>Enable writing self-contained documents combining text and code</td>
<td>Rstudio : Markdown : LaTeX : Jupyter</td>
</tr>
<tr>
<td>Version control</td>
<td>Track file changes over time. Revert to earlier versions. Branch/fork</td>
<td>Git : GitHub : BitBucket : Open Science Framework</td>
</tr>
<tr>
<td>Tracking provenance</td>
<td>Capture complex workflows involving multiple research objects/tools</td>
<td>VisTrails : Kepler : Taverna</td>
</tr>
<tr>
<td>Automation</td>
<td>Automate workflows using time-tested and ubiquitous command line tools</td>
<td>Unix command line : shell scripts : make</td>
</tr>
<tr>
<td>Virtual environments</td>
<td>Capture complex computation environments and configurations</td>
<td>VirtualBox : VMWare : Docker</td>
</tr>
</tbody>
</table>

https://ropensci.github.io/reproducibility-guide/sections/introduction/, 2018

This chart provides a sampling of specific skills and tools that can play a role in reproducible workflow and represent potential workshop opportunities for librarians to coordinate or teach.

[summarize the table - point out what i’ve taught, what’s in the Carpentry curriculum, etc.]

No one should feel like they need to master all of these in order to be effective in supporting reproducible research, but to have some familiarity with what these skills and tools are and why they are important and whether there are resources on your campus to meet the needs of your scientists. One of the advantages of basing training in the Library is that it can provide a discipline neutral and potentially less judgemental environment for learners.

Source:
Learning incentives

The first step to making science reproducible is to build good habits. Your most important collaborator is your future self. It’s important to make a workflow that you can use time and time again, and even pass on to others in such a way that you don’t have to be there to walk them through it.

Culich, 2014

In a 2017 Nature article called “A manifesto for reproducible science,” a group of prominent scientists argue for “the adoption of measures to optimize key elements of the scientific process: methods, reporting and dissemination, reproducibility, evaluation and incentives” by researchers, institutions, funders, and journals. These changes will improve the reliability and efficiency of research and the credibility of scientific literature.

In a 2014 interview, Berkeley research technology guru Aaron Culich stated:

The first step to making science reproducible is to build good habits. Your most important collaborator is your future self. It’s important to make a workflow that you can use time and time again, and even pass on to others in such a way that you don’t have to be there to walk them through it.

But I think, as all of us are aware from our outreach and teaching efforts, advocating for change is usually an uphill battle. The reality is that Learning and adopting new practices has a steep learning curve. Not all researchers are subject to funder or publisher mandates to share code, data or willing to invest the time to learn these skills just on the basis of being told that reproducibility is ethically “the right thing to do”

Given this reality, I’ve found that an effective argument is to remind scientists that the most immediate beneficiary of reproducible practices will be “future you.” Almost
everyone can relate to anecdotes about how easy it is to forget some step in an analytical process (even when it seemed obvious initially) and how common it is to need to repeat a complicated set of steps in order to accommodate a data revision or a reviewer suggestion. Having the technical skills to automate a workflow can literally eliminate these types of scenarios which would otherwise consume minutes, hours, or even days of a researcher's time. Similar arguments can be made about the benefits to collaboration, e.g. onboarding a research assistant, explaining analysis logic to a collaborator and so on.

Source:
More information

Training:
Data/Software Carpentry -- https://carpentries.org/
Library Carpentry -- https://librarycarpentry.github.io/

Case Studies:

Teaching materials:
Project TIER -- https://www.projecttier.org/
BITSS -- https://www.bitss.org/resources/
Amanda

I'm going to talk a bit about the Open Science Framework.

If we have time, I'll do a quick demo.
Why the Open Science Framework?

Project of the Center for Open Science, a nonprofit based in Charlottesville, VA

Funded by a variety of grants and sponsors, including DARPA, the NSF, NIH, and others.

https://osf.io/

Amanda

Why use the Open Science Framework? In addition to being a nonprofit funded by these grants and sponsors, it’s an interdisciplinary repository - well suited for much current research that includes two or more disciplines.

If a user heads to a discipline-specific archive, it’s likely s/he will miss related materials that happen to be hosted elsewhere.
What it does

Connects various parts of your workflow, wherever they are
- Google Drive
- Dropbox
- Mendeley
- FigShare
- GitHub...

Share other non-project files individually as well (new feature)

Amanda

So you (or maybe researchers you're working with) have some data, code, maybe some documents you've shared with colleagues, in a bunch of places.
What it does

Supports versioning

Allows date-stamped registration of research projects

Provides an additional backup of research materials

Amanda

You can register your hypothesis in advance, so you will show you haven't been HARKing.

There's a movement to provide DOIs for registrations so that it's easier to cite and track.
What it does

Centralizes access to research information

Provides granular sharing of elements with collaborators

Provides access for others who can provide feedback at any stage of the research process

Amanda

Researchers can share their work, documentation, methodology, and more.
Additional Related Project - OSF Preprints

Not just for science - includes the Arts & Humanities, Business, Education, Law, and more.

* Once research is published, encourage researchers to post their final manuscripts your institutional repository for increased visibility!

Amanda
Closing thoughts

“As readers of scientific work, all we can do is be more skeptical of everything that is published.”

– Cristobal Young, Assistant Professor of Sociology, Stanford University, 2015

“I want to adopt a stance of humility and assume that there are errors and that’s why I need to be cautious in my conclusions.”

– Brian Nosek, Professor of Psychology, University of Virginia and co-founder and director of the Center for Open Science, 2016

Amanda

Be critical: We can view research claims with a critical eye, educate ourselves on basic ways of evaluating research quality.

Be humble: Don’t arrogantly assume that we have figured everything out and have all the answers, because we usually don’t.

Be willing to revise our own personal opinions: We should be able to disconnect our positions from our identities, not stake our sense of self on a particular position we hold or practice that we’ve embraced.

Push for openness and transparency in the production and dissemination of knowledge.

Sources:


Closing thoughts

Sharing research at various stages of the process for feedback and input from others can improve researchers’ visibility, the actual research, and the final product.

(and in case you need additional talking points...)

Amanda

And just in case this is overwhelming...
Amanda

Courtesy of Jeff Leek, Biostatistics researcher @ Johns Hopkins:

“A few things that would reduce stress around reproducibility/replicability in science”

https://simplystatistics.org/2017/11/21/rr-sress/
From “A few things...”

2. We can remember that replication is statistical, not deterministic

3. We can remember that there is a difference between exploratory and confirmatory research

6. We can be persistent and private as long as possible

7. We can make the realization that data is valuable but in science you don’t own it

Amanda
Thank you!

Amanda -

Any questions?
Andrée Rathemacher     andree@uri.edu
Professor, Head of Acquisitions

Amanda Izenstark       amanda@uri.edu
Professor, Reference & Instructional Design Librarian

Harrison Dekker        hdekker@uri.edu
Associate Professor, Data Services Librarian

THINK BIG. WE DO™