

Guest Editorial

Perspectives on Peer Review of Data: Framing Standards and Questions

Morten Wendelbo

*Policy Sciences Lecturer, Policy Sciences & Economics Library
Texas A&M University*

Peer review serves two major purposes: verification and collaboration. A peer reviewer's role is to ensure, to a reasonable degree, that what journals publish is objective, correct and relevant. Further, a peer reviewer functions to give authors feedback on their writing, methods and presentation. Both aspects play their own part in ensuring the quality of the research published and so, by extension, peer review plays a part in moving the research in a field forward. However, when it comes to data, peer review is often inconsistent or outright absent, and that is hurting every field where data is not subject to the same rigorous standards as the manuscript itself.

Data is a powerful tool in research and academic publications. Its use is widespread in most disciplines, not least in Library and Information Sciences; the tangible nature of data adds to its strength, allowing it to be applied to test a hypothesis. Not only can data help show a change over time, but it can quantify the extent of that change. By extending data-measured trends, it can also be used to predict future events.

Data is also so powerful because it is difficult to dispute. Readers and peer reviewers rarely have the access and resources necessary to verify a manuscript's results, much less to review the integrity of the data itself. Thus far, peer review of a manuscript has often functioned as a proxy for the quality of the underlying data: a well-written, well-reasoned and well-supported manuscript is likely to be based on qualified data and established data practices. In other words, a good manuscript acts as a "signal" that the underlying data and practices are reliable. However, with the increasing influence of data, it is pertinent to ask whether that assumption is adequate. A few examples will illustrate why it is not.

It is a trite observation that statistics can be whatever you desire them to be. By displaying only certain isolated summary statistics of a larger set of data, nearly any dataset can be made to tell nearly any story. In a similar vein, by asking questions in certain ways, or measuring variables in a dataset in a particular fashion, data can be made to fit a hypothesis, rather than to objectively measure a phenomenon. Without access to the underlying dataset and data work, there are very few ways to tell if a dataset is sufficiently objective to warrant the findings that a manuscript claims.

Thankfully, academic dishonesty of that kind is, to my knowledge, quite rare. Instead, data shortcomings often happen because data users lack the skills to understand the limits of data, and the limits of the models the data at hand can support. Data literacy, a subject I teach, has not kept up with the expansion of data use in academics—or in business, for that matter. Peer review of data serves a dual purpose in this case, because it prevents publication of results that the data do not in reality support, *and* it acts as a screen for those readers who do not have the requisite skills to evaluate data and methods themselves.

Bad data practices can have severe, even fatal, consequences depending on the discipline. One of the worst examples is the work behind the now infamous and

disproven connection between the MMR vaccine (a vaccine used to prevent measles, mumps and rubella) and the incidence of autism. Subsequent scrutiny of the data and methods employed by the team behind the work revealed that the data had been collected selectively and had otherwise been subject to fraudulent practices. The falsified evidence, along with irresponsible media practices and celebrity endorsements of the research team¹ provided a near perfect storm, resulting in the epidemic outbreak of diseases that had been under control for decades. The results derived from the falsified data significantly lowered vaccination rates and subsequently led to the unnecessary deaths of children from preventable diseases.² It is not certain that access to the data would have prevented the initial publication of the work, but without peer review of the data itself the authors were free to tweak the data without impunity. With an established norm for the quality of data work, it is significantly less likely that work based on fraudulent data would make it into respectable and influential journals.

Another study that would have been unlikely to see the light of day was conducted by Lacour and Greene and claimed that opinions about controversial topics could be changed with short conversations. The study was influential, but when two graduate students at U.C. Berkeley wanted to replicate and expand those findings, the original authors could not prove the authenticity of the data and the journal was forced to retract it.³

The true value in reviewing the underlying data of submitted manuscripts is that it will increase the quality of research by providing an incentive for authors to work through, clean and verify data. This not only prevents the rare case of fraud, but also lowers the chance of results being skewed by improperly specified or formatted data, by helping to prevent mistakes. With access to the data, reviewers are able to verify that the methods described in the manuscript do in fact produce the results presented.

A paper published in 2010 by Kenneth Rogoff and Carmen Reinhart both of Harvard University, presents a fascinating case of mistakes in research.⁴ Their work in development economics showed that countries burdened with a very specific level of debt grow at a much slower rate than other countries. The paper became highly influential and was widely cited in academics and politics both. The allocation of millions of dollars was affected by their findings. A few years later, Thomas Herndon, a graduate student at the University of Massachusetts, found multiple mistakes when he tried to replicate the findings from the original article for a homework assignment.⁵ The mistakes included a simple excel formula mistake without which one of the central findings of the article would no longer be fully supported by the data.⁶ The mistakes were no doubt accidental; Rogoff and Reinhart provided Herndon with the data in the first place. Access to the data, and the appropriate expertise and due diligence by the reviewers in the manuscript submission phase, would have uncovered these mistakes and prevented the publication of faulty results⁷.

Moreover, whether intentional or not, submitted manuscripts do not always represent the true contents of the underlying data. Data are not always collected in a manner that justify using the results as evidence for a hypothesis. It is not sufficient to merely review the statistical methodology—peer review must include a review of the way in which the data was collected, collated and cleaned for analysis because improperly sourced data can invalidate it as evidence, *even when the statistical methodology is sound*.

*PS: This presents an interesting dilemma: Any mention of an article—with a subsequent citation and reference—adds to an article's citation count, regardless of whether it was mentioned for its merits or lack thereof. In an academic system where citation counts are increasingly important for rankings and tenure decisions, a “negative” citation is a valuable citation nonetheless. I choose to cite it because the authors readily admitted their mistake, and because many of their findings still stand, despite the error.

Without access to the data—and documentation of its collection—the content and manner of collection of data are not verifiable by the peer reviewer.

For these and other reasons, it is essential that reviewers be more critical of the data used in manuscripts, even if that means simply verifying the data. Academic work, in almost every discipline, is too often dismissed as “lies, damn lies” because of examples including the ones mentioned here. However, the vast majority of data used in academia is honest and precise and as reviewers of academic work, we have a responsibility to make sure that such integrity and good work actually matter.

How Would Peer Review of Data Work?

There is no way around the fact that reviewing the data work behind a manuscript makes peer review an even more onerous task than it already is. This is especially true if it is done without a pre-determined framework and good standards. However, the vast majority of mistakes and fraudulent practices are actually relatively easy to uncover for an experienced reviewer with a background in data work.

By virtue of being fabricated, fraudulent data may at first appear difficult to uncover, particularly because having access to data does not by itself allow a reviewer to verify how that data was collected and calculated. Nevertheless, while it may be hard for a reviewer to spot cleverly fabricated data, it is even harder for a writer to create comprehensive data that leaves no trail behind in the first place. The data will often be too perfect. In such cases, it is less a case of what is in the data, but rather what is *not* there that should cause suspicion. Despite the difficulty of exposing fraudulent data, the requirement to submit data along with manuscripts means that authors cannot fabricate information without the risk of getting caught.

However, the gains of data review are most pertinent for the non-fraudulent cases. The framework and standards necessary for systematic peer review of data will take time and experience to develop. Here I present an overall outline in the hopes that it will spur a much needed conversation, and course of action, surrounding peer review of data in the field of Library Science.

Joanna Lahey, a colleague who teaches graduate students quantitative methods at The Bush School of Government and Public Service at Texas A&M University, uses the humorous but apt metaphor of taking the data out for coffee. The first step in dealing with data at any level is to get to know the data that is in front of you. That means becoming familiar with the overall attributes of the data, including ranges, averages and other summary statistics of every variable. Whatever stands out requires a closer look, and if nothing stands out an even closer look is warranted, because no dataset is perfect.

At a bare minimum, data review requires access to a clean data file and to a “do” file—a file containing the commands necessary to accurately replicate the exact analysis done by the author. With these two resources, a reviewer with data expertise has the ability to easily ensure the vast majority of mistakes and oversights do not make it into publication. Beyond allowing for peer review, if these are shared with the reader, it provides a more complete academic study that can be replicated and augmented to move the field forward at a greater pace. With the “do”-file, the reviewer can verify the reported results. More than that, an experienced reviewer can verify that the method described in the manuscript is actually representative of the analysis done.

Access to the data also allows the reviewer to suggest further work to be done that is specifically based on the data available. Many authors will no doubt recognize the frustration of receiving a notice from the editor (citing the peer reviewers) to include work and results that the data cannot possibly support in order to improve the prospects for acceptance; with access to the data, the reviewers will have a deeper understanding of what the data can and cannot support and a sense of what it would

take to augment the data with new or better variables if the initial submission is inadequate for publication.

Much of the benefit to reviewing data may even accrue before the article and the data are submitted. With the knowledge that reviewers will be more likely to understand and accept the data when it is properly cleaned, labeled and presented, authors are likely to review the data in greater depth themselves to preempt issues during the peer review process. Authors, themselves, will discover more unintentional mistakes and discrepancies. Since authors have to carefully clean and label the data for the review process, little additional effort is necessary to publish the data, either in the journal or elsewhere. This will also improve the quality of the data and increase the potential for data-sharing in the academic and public domains.

Where Do We Go from Here?

There may be push back against peer reviewing data. As stated earlier, data review will increase the burden on both authors and reviewers. However, in reality, and as a course of ensuring academic integrity, authors should already be reviewing their data and data analyses closely.

Naturally, reviewing data requires some expertise. I will argue that it asks less of the reviewer's time than it may appear at first. The vast majority of the skills necessary to review the data and the manuscript itself—keep in mind that mere summary statistics go a long way—are necessary to be a good consumer of academic scholarship in the first place. In order for the review to move beyond the easy-to-spot mistakes, reviewers will need a dedicated data skillset. Some reviewers will already have this skillset while others will need to acquire it. When I teach data literacy to graduate students, who only have two introductory statistics classes under their belt, they acquire the vast majority of necessary skills in a few weeks through periodic workshops. For a frequent reviewer, time spent training would pay off several times over in time saved reviewing.

It is important that we must develop a set of standards and practices that we expect authors to abide by to ensure the integrity of the manuscripts they submit. It is also a clear priority to ensure that authors are protected from misuse of their (often hard earned) data by those who are privy to the data in the review process, much as authors must be protected from having reviewers use the manuscripts themselves for their own gain already. It is also important that we mandate that authors publish the data upon which their manuscripts are based—especially after the data is no longer a source for forthcoming manuscripts. This is already a long held practice within economics and political science, for example, where data must often be published within a certain timeframe from the data of the manuscript's publication. Further, many funding agencies and foundations now make financial support dependent on data review and publication of data.

I recognize that this will spell a substantial change for many, whether author or reviewer, but there is no way around it if we intend to uphold the integrity of research in the field of Library Science. Our colleagues in other fields underwent this change already and we can learn from their tribulations and successes to design a framework and standards to guide peer review of data.

Notes

1. Liza Gross, "A Broken Trust: Lessons from the Vaccine–Autism Wars," *PLoS Biol*, 7, no. 5 (May 2009), e1000114.
2. DK Flaherty, "The Vaccine–Autism Connection: A Public Health Crisis Caused by Unethical Medical Practices and Fraudulent Science," *Annals of Pharmacotherapy*, 45, no. 10 (October 2011):1302–04.

3. Steve Kolowich, "The Researchers Who Sank a Bogus Canvassing Study Have Replicated Some of Its Findings," *Chronicle of Higher Education*, (2016), <http://chronicle.com/article/The-Researchers-Who-Sank-a/236024>.
4. Carmen M. Reinhart and Kenneth S. Rogoff, "Growth in a Time of Debt," *American Economic Review*, 100, no. 2 (May 2010): 573–8.
5. Ruth Alexander, "Reinhart, Rogoff... and Herndon: The student who caught out the profs," BBC (April 20, 2013). Retrieved from <http://www.bbc.com/news/magazine-22223190>.
6. Thomas Herndon, Michael Ash, Robert Pollin, "Does High Public Debt Consistently Stifle Economic Growth? A Critique of Reinhart and Rogoff," *Cambridge Journal of Economics*, 38 no. 2 (2014): 257–279.