Andrew Gelman
Column Editor

Honesty and Transparency Are Not Enough

There has been a replication crisis in applied statistics (most notably in psychology, but also in social science and medical research), in which studies published in top scientific journals "fail to replicate"; that is, outside research teams are unable to reproduce published claims.

Sometimes replications are difficult because the underlying data and code are inaccessible. Data can be withheld for confidentiality reasons, because a researcher does not want to admit other interpretations or codings (as when primatologist Marc Hauser refused to let his research assistants view his monkey videos), or because the data simply are not there (as with political science student Michael Lacour's fake survey on attitudes toward same-sex marriage, or the survey on gun control that economist John Lott says was entirely lost in a computer crash), or for reasons that remain unclear (as in a much-publicized survey published in 2006 of mortality in Iraq; Spagat, 2014).

In political science and economics, it is common practice to work with data that are publicly available but can take a bit of effort to obtain and clean, effort that is hidden in the public record. This leads to problems such as the famous "Excel error" of economists Reinhart and Rogoff, whose conclusions from a much-publicized paper in 2010 turned out to depend on an embarrassing data-processing error that was not caught for years, in part because data sharing is not the scientific norm. Any outside researchers who questioned Reinhart and Rogoff's claims had to go get the data themselves and reconstruct the analyses from scratch, so it took time for the failure to be discovered.

These examples are notorious, but problems with data access are pervasive in published empirical work, including my own. When an outsider requests data from a published paper, the authors typically will not post or send their data files and code, but instead will just point to their sources, so replicators have to figure out what exactly to do from there. End-to-end replicability is not the norm, even among scholars who actively advocate for the principles of open science.

For example, our 2008 book *Red State Blue State* featured dozens of analyses of political data, and we reported where the numbers came from, but that was it. We did not supply data or code, which made our analyses difficult to check and, perhaps more importantly, inhibited the ability of others to extend our work. Our motivation was not to make things hard for people; we just did not feel like putting in the effort to construct a set of replication files. Perhaps if my colleagues and I had made a habit of providing replication materials for all our papers, we would have avoided the data coding error that invalidated all the empirical claims from one of my papers.

In experimental sciences such as psychology, challenges arise in not just reproducing published work but replicating via new experiments. Conditions for data collection are often unclear (for example, survey forms and complete descriptions of experimental protocols are often not available,

even in supplementary materials), which leads to potentially endless ways that a replication cannot be trusted.

Biologist Simone Schnall points to a report claiming that "even in studies with mice seemingly irrelevant factors such as the gender of the experimenter can make a difference," and it is presumably not a requirement of published papers to supply full demographic information in all lab assistance (Schnall, 2014; Gelman, 2014).

As a result, scientific claims can seem to be taken as the personal property of their promoters. Psychologist Daniel Kahneman wrote in 2014 that if replicators make no attempts to work with authors of the original work, "this behavior should be prohibited, not only because it is uncollegial but because it is bad science. A good-faith effort to consult with the original author should be viewed as essential to a valid replication."

I believe that Kahneman's suggestions are in good faith, but it seems clear that they give a privileged role to authors of published or publicized papers.

Various reforms have been suggested to resolve the replication crisis. I am in favor of more active post-publication review as part of a larger effort to make the publication process more continuous, so researchers can release preliminary and exploratory findings without the requirement that published results be presented as being certain.

Any published paper is just part of a larger flow of data collection and analysis, and should be treated as such. Related proposed institutional reforms include reducing the role of high-profile journals in academic hiring and promotion, and making it easier to publish criticisms and replications.

This works well with the norm or expectation that papers come with replication materials attached: raw data, analysis code, and precise descriptions of the data collection process. This is already required by some journals, such as the *Quarterly Journal of Political Science*. I can attest that the process can be a hassle, but I still think it is a good idea. Along with providing openness in data, researchers should be transparent about their choices in analysis (Steegen, et al., 2016; Wicherts, et al., 2016).

Honesty and transparency are not enough, though; I worry that the push toward various desirable *procedural* goals can make people neglect the fundamental *scientific* and *statistical* problems that, ultimately, have driven the replication crisis.

In recent years, we have seen many published and publicized papers that were essentially dead on arrival because they were attempting to identify small effects in the presence of noisy and often biased data. In such research projects, even findings that are statistically significant are likely to be in the wrong direction and can massively exaggerate effect sizes (Button, et al., 2013; Gelman and Carlin, 2014).

Honest reporting is important, even necessary, but design is important, too. Dishonest reporting can mislead, but honest reporting is not enough to save a poor design. It does not matter how honest and careful the researcher is, if the data are just too noisy to learn anything from them, from an exploratory or a confirmatory perspective.

Sociologist Satoshi Kanazawa published a series of papers several years ago claiming to have found various social factors that were predictive of the sex of children. Setting aside any concerns with the biological theory that motivated this work, the statistics made the problem simply hopeless (Gelman and Weakliem, 2009). Any underlying effect sizes were almost certainly less than 0.005 (that is, the probability of girl births was highly unlikely to vary more than that among the subpopulation being compared), but the sample size of these studies was in the low thousands. From the binomial distribution, we could know ahead of time that any estimate would have a standard error much larger than the effect being studied.

For example, in a study of N=4000, the standard error of a comparison between two equally sized groups is sqrt(0.5^2/2000 + 0.5^2/2000) = 0.016.

This point—that transparency is not enough—is important for two reasons.

First, consider the practical consequences for a researcher who eagerly accepts the message of ethical and practical values of sharing and openness, but does not learn about the importance of data quality. He or she could then just be driving very carefully and very efficiently into a brick wall, conducting transparent experiment after transparent experiment and continuing to produce and publish noise. The openness of the work may make it easier for a later researcher to attempt—and fail—to replicate the resulting published claims, but little if any useful empirical science will be done by anyone concerned. I do not think we are doing anybody any favors by having them work more openly using data that are inadequate to the task.

The second reason that honesty and transparency do not suffice

is that, even with the best of intentions and scholarly practices, researchers will fail. When transparency is touted as a solution to the replication crisis, I worry that the reasoning will go in both directions, so replication failure will be taken as a sort of moral failing of the original experimenter.

Yes, sometimes unreplicable research can be associated with ethical violations—even if researchers are innocently unaware of statistical errors in their published work, it is poor behavior for them to refuse to admit they could ever have been wrong—but we can also make mistakes in good faith. Making errors is inevitable; learning from them is not. To learn from errors, we want a system of science that facilitates and provides incentives for such learning; we don't want an attitude that automatically links error to secrecy or dishonesty.

This discussion echoes similar points in ethics more generally. For instance, doctors should first do no harm and should care for their patients, but there are times when all the care in the world is no substitute for a good antibiotic. We want medical researchers to convey their uncertainty appropriately, but more important is for them to have useful results to talk about. This is facilitated not just by openness but by close links between data, measurement, and substantive theory.

Further Reading

Button, K.S., J.P.A. Ioannidis, C. Mokrysz, B.A. Nosek, J. Flint, E.S.J. Robinson, and M.R. Munafo. 2013. *Nature Reviews Neuroscience* 14, 365–376.

Gelman, A. 2013. Correction on Should the Democrats move to the left on economic policy? *Annals of Applied Statistics* 7, 1248.

Gelman, A. 2014. "An experience with a registered replication project." Statistical Modeling, Causal Inference, and Social Science blog, Jul. 25. *http://andrewgelman.com/2014/07/25/experience-registered-replication-project.*

Gelman, A., and J.B. Carlin. 2014. Beyond power calculations: Assessing Type S (sign) and Type M (magnitude) errors. *Perspectives on Psychological Science* 9, 641– 651.

Gelman, A., and D. Weakliem. 2009. Of beauty, sex, and power: statistical challenges in estimating small effects. *American Scientist* 97, 310–316.

Kahneman, D. 2014. A new etiquette for replication. *Social Psychology* 45, 299–311.

Schnall, S. 2014. Social media and the crowd-sourcing of social psychology. Cambridge Embodied Cognition and Emotion Laboratory Blog, Mar. 4. *www.psychol.cam.ac.uk/cece/blog.*

Spagat, M. 2014. Questioning the *Lancet*, PLOS, and other surveys on Iraqi deaths. Musings on Iraq blog, Jan. 27. *http://musingsoniraq.blogspot.co.uk/2014/01/questioning-lancet-plos-and-other.html.*

Steegen, S., F. Tuerlinckx, A. Gelman, and W. Vanpaemel. 2016. Increasing transparency through a multiverse analysis. *Perspectives on Psychological Science 11*, 702–712.

Wicherts, J.M., C.L.S. Veldkamp, H.E.M. Augusteijn, M. Bakker, R.C.M. van Aert, and M.A.L.M van Assen. 2016. Degrees of freedom in planning, running, analyzing, and reporting psychological studies: A checklist to avoid *p*-hacking. *Frontiers in Psychology* 7, 1832.

About the Author

Andrew Gelman is a professor of statistics and political science and director of the Applied Statistics Center at Columbia University. He has received many awards, including the Outstanding Statistical Application Award from the American Statistical Association and the award for best article published in the *American Political Science Review*. He has coauthored many books; his most recent is *Red State, Blue State, Rich State, Poor State: Why Americans Vote the Way They Do* (Princeton University Press).