Andrew Gelman
Column Editor

It's Too Hard to Publish Criticisms and Obtain Data for Replication

A few months ago, I was pointed to a paper by Laura Hamilton in the *American Sociological Review* arguing that "the more money that parents provide for higher education, the lower the grades their children earn" and that "parental investments create a disincentive for student achievement." This claim was reported uncritically in *The New York Times* and was spread widely, perhaps in part because the conclusion was attractive both to the left (as an exposé of the high cost of college education and a poke at the rich and privileged) and to the right (in that it criticizes universities, which are one of the pillars of leftism in American culture).

But when I looked at the paper, I noticed a serious problem: The central analysis selects on observed cases, which is sometimes called "survivorship bias."

If you are not doing so well in college, but it's being paid for, you might stay. But if you are paying for it yourself, you might just quit. This would induce the negative correlation in the data, but not at all through the causal story asserted, that students who are parentally supported "dial down their academic efforts." The regression results are consistent with the alternative hypothesis that these students are working as hard as they ever would, and the parental support just makes them more likely to stay in school.

To say it another way: It seems completely plausible to me that (a) there could be a negative correlation between parental support and student grades, conditional on students being in college, but (b) paying more of a college student's tuition would not (on average) lower his or her grades.

By analogy, consider a study of a medical treatment that keeps the sickest patients alive. If an analysis were performed only on the survivors, it would appear that the treatment actually harms overall health: The control looks better because the patients who died were not counted!

And, indeed, Hamilton finds, "even net of sociodemographics and parental SES, parental aid significantly increases the likelihood of obtaining a bachelor's degree." It is reasonable to suppose the subset of students without parental financial support who did not complete their degree would have had lower-than-average grades, had they received the aid and stayed in school.

The paper does not address this possible selection bias in a satisfactory way. The closest is this passage: "It is also possible that by the final year of college, many students with little parental funding and low GPAs have simply dropped out, creating the appearance that lower levels of parental aid lead to higher student GPAs." Hamilton attempts to address this concern with a supplementary analysis, a regression fit for a different data set that allows within-student comparisons where aid levels vary over time. It is reasonable to do such a within-student analysis, but it has the same problem as already noted: students do not get counted after they drop out. Again, an observed correlation does not necessarily correspond to a causal effect—and I say this not in a vague "correlation does not equal causation" sense but more specifically in that there is a clear survivorship bias here.

Hamilton writes, "These results provide strong evidence that selectivity processes are not driving the negative relationship between parental aid and GPA," but does not address the concerns presented here. In addition, even if this supplementary analysis had not been subject to its own survivorship bias, the problem would remain that the results from the main regression have not been corrected for selection and thus should not be taken at face value. An appropriate analysis would have to correct the raw regression results in some way before giving them causal interpretations.

Even before reading the article, the brief description of the study I'd seen in a news report had made me suspicious. I'd expect the effect of parent's financial contributions to be positive (as they free the student from the need to get a job during college), but not negative. Hamilton argues that "parental investments create a disincentive for student achievement," which may be—but I'm generally suspicious of arguments in which the rebound is bigger than the main effect.

I mention these reactions not because my preconception should be considered more important than a published paper, but just to point out a general problem that can arise with high-profile journals such as the *American Sociological Review*, which is generally considered the top journal in its field. Exploratory research, inconclusive research, and research that confirms existing beliefs—all these can be difficult to get published in a top journal. Instead there is a premium on the surprising and counterintuitive—especially if such claims can be demonstrated in a seemingly conclusive way via statistical significance. I worry that the pressure to produce innovation leads to the sort of mistakes that led to this manuscript getting through the refereeing process. Another form of survivorship bias, perhaps?

Difficulty in Publishing a Criticism or Reanalyzing the Data

After reading Hamilton's article, noting its problems, and blogging about it, I asked a colleague in the sociology department if there was anything else I could do. He suggested two things: write a letter to the editor of the journal and reanalyze the data myself so as to provide an example for other researchers in the field.

So I proceeded on two tracks. I submitted a short comment to the *American Sociological Review*, including the criticisms described above along with a discussion of some other methodological flaws in the paper, and I asked a student to get the data so we could try our own analysis, probably some sort of latent-data model to account for drop-out.

Getting the data wasn't easy; in fact, I never got there. The data were public and obtainable from the National Center for Education Statistics, but we could not simply get the data directly. We had to file a request, and then we were told that we needed approval by the local institutional review board (IRB) before we could send a request to the data provider. So my student and I filed out two forms, one for the IRB and one for the data provider. After a bit, we heard back from the IRB that we needed more information. We did a bit more, and then after a month or so we had the local IRB approval. But after another month, we still had heard from the National Center for Education Statistics. Maybe there is some other set of hoops we need to jump through. But I don't know, because the student graduated in the meantime.

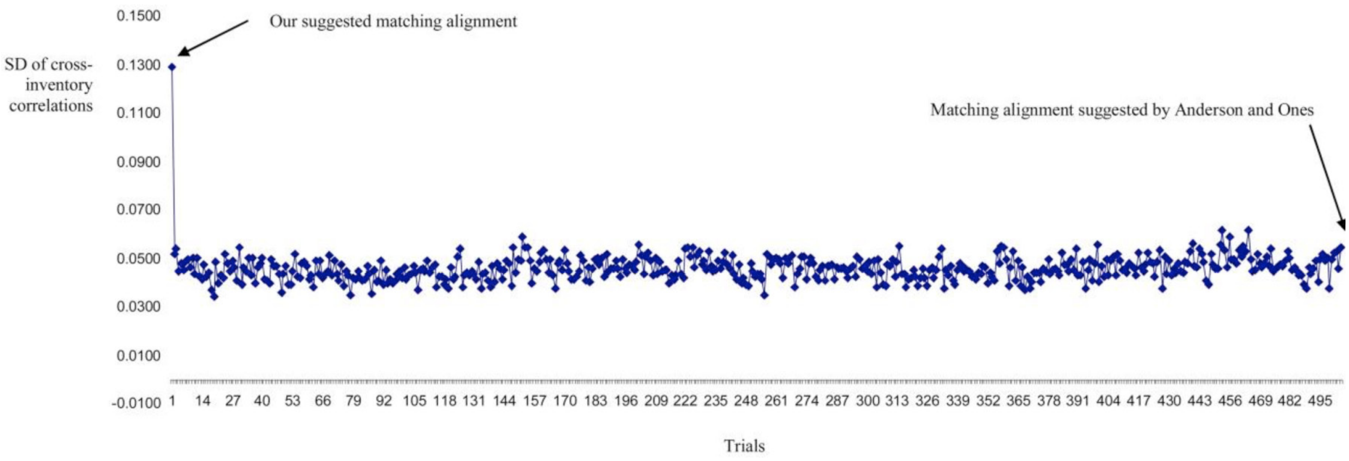
Meanwhile, I heard from the *American Sociological Review* about my letter to the editor. This news was not good, either. The letter was sent my letter to three reviewers, none of whom disagreed on the merits of my criticisms, but the editors declined to publish the letter because they judged it to not be in the top 10% of submissions to the journal. I'm sure my letter was indeed not in the top 10% of submissions, but the journal's attitude presents a serious problem, if the bar to publication of a correction is so high. That's a disincentive for the journal to publish corrections, a disincentive for outsiders such as me to write corrections, and a disincentive for researchers to be careful in the first place. Just to be clear: I'm not complaining how I was treated here; rather, I'm griping about the system in which a known error can stand uncorrected in a top journal, just because nobody managed to send in a correction that's in the top 10% of journal submissions.

The editors concluded their letter to me as follows: "We do genuinely share with you an enthusiasm for your topic and its importance to sociology. We hope that the reviewer comments on this particular manuscript provide useful guidance for that future work." I appreciate their encouragement, but again the difficulty here is that selection bias is not a major topic of my research. The journal does not seem to have a good system for handling short critical comments by outsiders who would like to contribute without it being part of a major research project.

Publishing drive-by criticism such as mine would, I believe, serve the following three benefits:

- Correct the record so future researchers, when they encounter this research paper, are not misled by the biases in the research methods
- Provide an incentive for journal editors to be careful about accepting papers with serious mistakes, and it takes some of the pressure off the review process, if there is a chance to catch problems in the post-publication stage
- Spur an expert (such as Hamilton or someone else working in that field) to revisit the problem, in this case, aiming toward a better understanding of student motivation and performance.

I do not fault Hamilton here. She overlooked a possible bias in her data analysis. But we all make mistakes. And, as she was working with public data under use restrictions, it would not necessarily have been appropriate for her to share these data with an outside researcher such as me. And it is certainly not her fault that it is so difficult to access the public data that she used. It is also hard for me to blame the *American Sociology Review* for accepting the paper:



journal editors and reviewers make mistakes, and, given that sociology is not a highly quantitative field, it is understandable that a subtle statistical error might not get caught.

The problem is that, for whatever reasons of bureaucracy or confidentiality, a splashy claim was made that was able to be published in a leading scholarly journal and widely disseminated in the news media—but not so easily criticized or reanalyzed.

The asymmetry is as follows: Hamilton's paper represents a major research effort, whereas my criticism took very little effort (given my existing understanding of selection bias and causal inference). The journal would have had no problem handling my criticisms, had they appeared in the pre-publication review process. Indeed, I am pretty sure the original paper would have needed serious revision and would have been required to fix the problem. But once the paper has been published, it is placed on a pedestal and criticism is held to a much higher standard.

One of the anonymous referees of my letter wrote, "We need voices such as the voice of the present comment-writer to caution those who may be over-interpreting or mis-specifying models (or failing to acknowledge potential bias-inducing aspects of samples and models). But we also need room in our social science endeavor to reveal intriguing empirical patterns and invoke sociological imagination (that is, theorizing) around what might be driving them." This seems completely reasonable to me, and I find it unfortunate that the current system does not allow the exploration and criticism to appear in the same place. Instead, the exploration appears in the form of settled science in the journal (and is cited uncritically in *The New York Times* and elsewhere), while the criticism appears on blogs and in articles such as this one.

Researchers Who Refuse to Admit They Made a Mistake

My next story is about Neil Anderson, a professor of human resources management, and Deniz Ones, a professor of psychology, who (inadvertently, I assume) made a data coding error and were eventually moved to issue a correction notice, but even then refused to fully admit their error. As psychologist Sanjay Srivastava puts it, the story "ended up with Lew [Goldberg] and colleagues [Kibeom Lee and Michael Ashton] publishing a comment on an erratum—the only time I've ever heard of that happening in a scientific journal."

In their "erratum and addendum," Anderson and Ones (this issue) explained that we had brought their attention to the "potential" of a "possible" misalignment and described the results computed from realigned data as being based on a "post-hoc" analysis of a "proposed" data shift. That is, Anderson and Ones did not plainly describe the mismatch problem as an error; instead, they presented the new results merely as an alternative, supplementary reanalysis.

I was annoyed by Anderson and Ones's unusual rejoinder to the comment on the correction. To the end, they refuse to admit their mistake, instead referring to "clerical errors as those alleged by Goldberg et al." and concluding:

> When any call is made for the retraction of two peer-reviewed and published articles, the onus of proof is on the claimant and the duty of scientific care and caution is manifestly high. Yet, Goldberg et al. (2008) have offered only circumstantial and superficial explanations … As detailed above, Goldberg et al. do not and cannot provide irrefutable proof of the alleged clerical errors. To call for the retraction of peer-reviewed, published papers on the basis of alleged clerical errors in data handling is sanctimoniously misguided. We continue to stand by the analyses, findings and conclusions reported in our earlier publications.

That's the best they can do "no irrefutable proof"?? Scientists should be aiming for the truth, not acting like defendants in a criminal case. Anderson and Ones were not accused of any crime; they just happened to make a mistake in a data analysis.

The "sanctimonious" graph from Goldberg et al. featuring the "no irrefutable proof" appears above.

And here is an excerpt from an interview with the University of Minnesota research bulletin: "When Ones discovers a piece of knowledge that's new, she says she gets a literal rush. 'In my career, this has happened around a dozen times or so. When you know you're looking at a set of results that will change the pool of knowledge for your field, it's exhilarating.'" It would be appropriate for her to apply that attitude in this case.

This particular story has a happy ending in that the correction was published in the original journal. But I find it sad that the authors of the error are holding on so tenaciously. They should welcome corrections, not repulse them.

In What Sense Is Science Self-Correcting?

Science is said to be self-correcting, with the mechanism being a dynamic process in which theoretical developments are criticized; experiments are designed to refute theories, distinguish among hypotheses, and (not least) to suggest new ideas; empirical data and analyses are published and can then be replicated (or fail to be replicated); and theories and hypothesized relationships are altered in light of new information.

There has been much discussion in recent years over the merits of open-access journals and post-publication peer review as compared to the traditional refereeing process. One of the difficulties, though, as discussed above, is that it is difficult to publish post-publication peer review and it can be difficult to obtain the data required to perform replication analyses.

A recent example arose with Carmen Reinhardt and Kenneth Rogoff, who made serious mistakes handling their time-series cross-sectional data in an influential paper on economic growth and government debt. It was over two years before those economists shared the data that allowed people to find the problems in their study. I'm not suggesting anything unethical on their part; rather, I see an ethical flaw in the system, so that it is considered acceptable practice not to share the data and analysis underlying a published report.

The revelation of all these problems (along with others such as the fabrications of disgraced primatologist Marc Hauser and the much-mocked ESP experiments of Daryl Bem) can be taken as evidence that things are getting better. Psychology researcher Gary Marcus writes:

> There is something positive that has come out of the crisis of replicability—something vitally important for all experimental sciences. For years, it was extremely difficult to publish a direct replication, or a failure to replicate an experiment, in a good journal. . . . Now, happily, the scientific culture has changed.

And sociologist Fabio Rojas writes:

> People may sneer at the social sciences, but they hold up as well. Recently, a well-known study in economics was found to be in error. People may laugh because it was an Excel error, but there's a deeper point. There was data, it could be obtained, and it could be replicated. Fixing errors and looking for mistakes is the hallmark of science. …

I agree with Marcus and Rojas that attention to problems of replication is a good thing. It's bad that people are making mistakes in their analysis and even faking data at all, but it's good that these problems are being aired. And, to the extent that scientific practices are improving to help detect error and fraud, and to reduce the incentives for publishing erroneous and fraudulent results in the first place, that's good too.

But I worry about a sense of complacency. The common thread in all these stories is the disproportionate effort required to get criticisms and replications into the standard channels of scientific publication. In case after case, outside researchers who notice problems with published articles find many obstacles in the way of post-publication replication and review.

Economist Steven Levitt recognizes the problem:

> Is it surprising that scientists would try to keep work that disagrees with their findings out of journals? … Within the field of economics, academics work behind the scenes constantly trying to undermine each other. I've seen economists do far worse things than pulling tricks in figures. When economists get mixed up in public policy, things get messier.

Following this idea, it could be desirable to change the incentives associated with scholarly publication. If post-publication criticism and replication were encouraged rather than discouraged, this could motivate a new wave of research evaluating existing published claims and also motivate published researchers to be more accepting about problems with their work. The larger goal is to move science away from a collection of isolated studies that are brittle (in the terminology of financial critic Nassim Taleb) to a system of communication that better matches the continually revising process of science itself. C

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*.