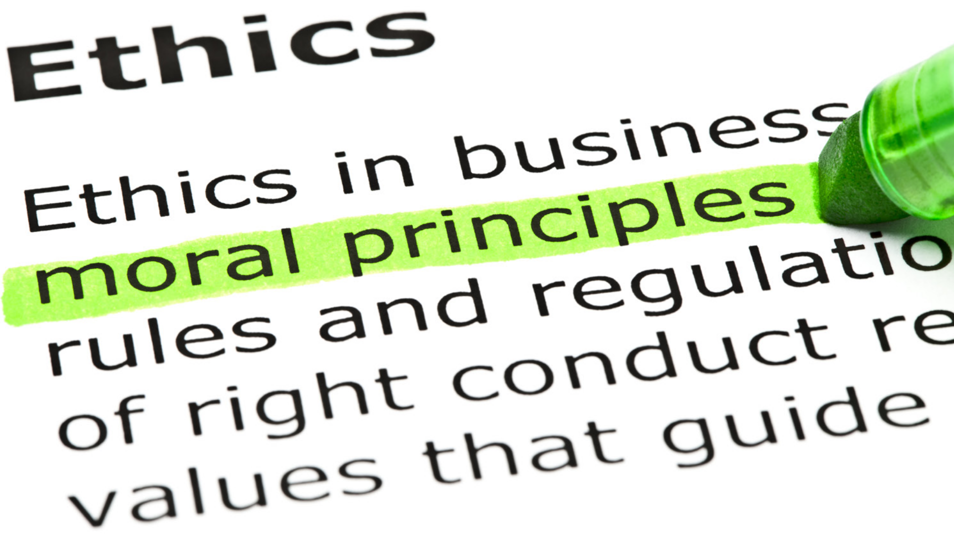
# [Ethics and Statistics]

Andrew Gelman
*Column Editor*

# They'd Rather Be Rigorous Than Right

Most of this particular column will be devoted to a scientific controversy that does not have a direct ethical dimension. The ethics question will come up at the end, when I argue that there are systematic features of our scholarly publication system that may encourage behavior that is counterproductive to science. Thus, the ethical failing is not of any individual, but of our institutions.



## A Controversy in Economics and Anthropology

Here's the story. Two economics professors, Quamrul Ashraf and Oded Galor, wrote a paper, "The Out of Africa Hypothesis, Human Genetic Diversity, and Comparative Economic Development," that is scheduled to appear in the *American Economic Review*. In their article, Ashraf and Galor claim to have an explanation for the wealth of Europe, the United States, and other former European settler nations; the moderate income levels of east Asian countries; and the poverty of Africa and Latin America. Their explanation, backed by statistical analysis, is that, as Ewen Calloway put it in a *Nature* article:

> High genetic diversity in a country's population is linked with greater innovation, the paper says, because diverse populations have a greater range of cognitive abilities and styles. By contrast, low genetic diversity tends to produce societies with greater interpersonal trust, because there are fewer differences between populations. Countries with intermediate levels of diversity, such as the United States, balance these factors and have the most productive economies as a result, the economists conclude.

Calloway continues, "The manuscript had been circulating on the Internet for more than two years, garnering little attention outside economics—until last month, when *Science* published a summary of the paper in its section on new research in other journals. This sparked a sharp response from a long list of prominent scientists" criticizing the claims.

The paper looked pretty silly to me, and I was surprised it was accepted in such a top journal. I wasn't the only person to read the paper and be unimpressed. In fact, I heard about it via an email from Kyle Peyton, who wrote:

> I am very skeptical of the reductive approach taken by the economics profession generally, and the normative implications this kind of research generates. For example, p. 7 of the working paper states: "… [according to our model,] decreasing the diversity of the most diverse country in the sample (Ethiopia) by 1 percentage point would raise its income per capita by 21 percent." Understandably, this piece is couched in assumptions that would take hours to pick apart, but their discussion of the approach belies the uncertainty involved. The main response by the authors in defense is that genetic diversity is a 'proxy variable.' This is a common assertion, but I find it really infuriating. I happen to drink coffee most days, which correlates with my happiness. So coffee consumption is a 'proxy' for my happiness. Therefore, I can put it in a regression and predict the relationship between my happiness and the amount of times I go to the bathroom. Ergo universal conclusions: 'Relieving yourself improves mental well-being.' New policy—you should relieve yourself at least two times per day in order to maintain high levels of emotional well-being.

I agree with Peyton's skepticism. Social scientists can be credulous, but I'd expect better from economists writing on economic development, which is one of their central topics. Ashraf and Galor have, however, been somewhat lucky in their enemies, in that they've been attacked by a bunch of anthropologists who have criticized them on political and scientific grounds. This gives the pair of economists the scientific and even moral high ground, in that they can feel that, unlike their antagonists, they are the true scholars, the ones pursuing truth wherever it leads them, letting the chips fall where they may.

The real issue for me is that the chips aren't quite falling the way Ashraf and Galor think they are. Let's start with the claims on Page 7 of their paper:

> Once institutional, cultural, and geographical factors are accounted for, [the fitted regression] indicates that: (i) increasing the diversity of the most homogenous country in the sample (Bolivia) by 1 percentage point would raise its income per capita in the year 2000 CE by 41 percent, (ii) decreasing the diversity of the most diverse country in the sample (Ethiopia) by 1 percentage point would raise its income per capita by 21 percent.

If taken literally, the above bit is not a claim at all; it's just an interpretation of some regression coefficients. But it clearly is a claim, in that the authors want us to take these examples seriously.

So let's take them seriously. What would it mean to increase Bolivia's diversity by 1 percentage point? I assume that would mean adding some white people to the country. What kind of white person would go to Bolivia? Probably someone rich enough to increase the country's income per capita. Hey, it works! What if some poor people from Ethiopia were taken to Bolivia? They'd increase the country's ethnic diversity too, but I don't see them increasing its per-capita income by 41%. But that's okay; nobody's suggesting filling Bolivia with poor Africans.

What about Ethiopia? How do you make it less diverse? I guess you'd have to break it up into a bunch of little countries, each of which is ethnically pure. Is that possible? I don't actually know. If you can't do that, you'd need to throw in lots of people with less genetic diversity. Maybe, hmmm, I dunno, a bunch of whites or Asians? What sort of whites or Asians might go to Ethiopia? Not the poorest ones, certainly. Why would they want to go to a poor country in the first place? Maybe some middle-income or rich ones (if the country could be safe enough, or if there's a sense there's money to be made). And, there you go; per-capita income goes up again.

So I don't see it. It's true that later on Page 7 the authors try to wriggle out of this one:

> Reassuringly, the highly significant and stable hump-shaped effect of genetic diversity on income per capita in the year 2000 CE is not an artifact of postcolonial migrations towards prosperous countries and the concomitant increase in ethnic diversity in these economies. The hump-shaped effect of genetic diversity remains highly significant and the optimal diversity estimate remains virtually intact if the regression sample is restricted to (i) non-OECD economies (i.e., economies that were less attractive to migrants), (ii) non-Neo-European countries (i.e., excluding the U.S., Canada, Australia, and New Zealand), (iii) non-Latin American countries, (iv) non-Sub-Saharan African countries, and, perhaps most importantly, (v) countries whose indigenous population is larger than 97 percent of the entire population (i.e., under conditions that virtually eliminate the role of migration in contributing to diversity).

I don't buy it. I'm not saying their central point is wrong—it's basically a twist on the classic "why are some countries so poor" question—but the extrapolations they give reveal the problems with their interpretation of the regression model. The way you make Bolivia more diverse is by adding more white people. It's fine to study these things, but you have to think about what your models mean.

Everybody wants to be Jared Diamond, that's the problem.

When I posted the above remark on my blog, there was vigorous discussion, including much on the details of Ashraf and Galor's genetic measures. I am not so interested in exactly how these are defined (even though they are central to the paper and perhaps central to its failings) because I am looking at the work from the outside, as a consumer. And from that perspective, it seems pretty clear that I am intended to think of "diversity" in something close to its usual English-language meaning, not merely something cooked up in the lab, but something related to social phenomena such as cooperation and trust. So I am taking diversity at face value in my discussion, while recognizing that it would be necessary to understand the underlying genetic calculations for a fuller understanding of this work.

## What Went Wrong?

If all (or even some) of these criticisms are valid, then, what went wrong, and how could Ashraf and Galor have done better? I would start with their big pattern: The most genetically diverse countries (according to their measure) are in east Africa, and they're poor. The least genetically diverse countries are remote undeveloped places like Bolivia and are pretty poor. Industrialized countries are not so remote (thus they have some diversity), but they're not filled with east Africans (thus they're not extremely genetically diverse). From there, you can look at various subsets of the data and perform various side analyses, as the authors indeed do for much of their paper.

This is how good social science commonly proceeds: You identify a pattern that is general enough to be considered a "stylized fact" (in this expression, "stylized" is not a negative term; it just refers to the inevitable smoothing-around-the-edges that takes place when a complex pattern is summarized). You then examine the data and establish this fact. You develop causal theories to explain it, and you try to identify, estimate, and isolate effect that place the stylized fact in a causal framework. The paper under discussion largely followed this pattern, but I think it failed in not clearly identifying its descriptive findings and then jumping from a regression to a series of implausible causal claims.

One commenter wrote in defense of Ashraf and Galor:

> This paper passed to scrutiny of five referees, and it was presented in about 50 leading universities in the

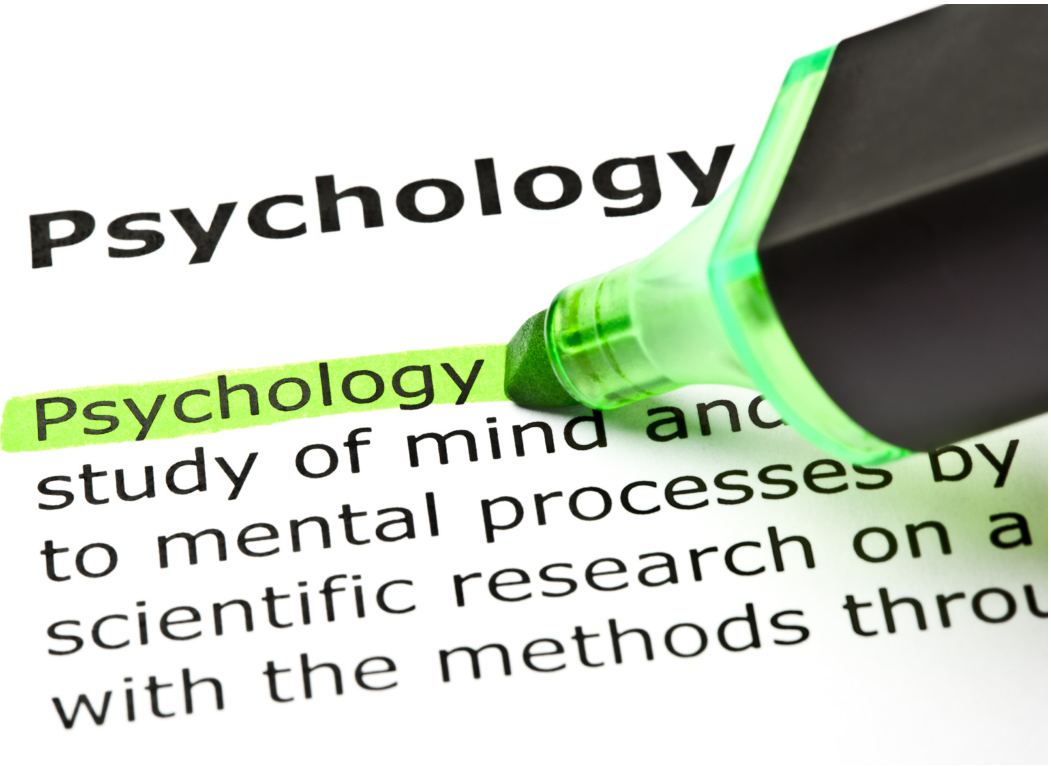
world. With all due respect, a priori, the likelihood that your comments are off-mark is higher than the likelihood that 1000s of people who heard the paper in seminar and conferences, and referees and editors who read it very carefully as part of your editorial duties."

On the other hand, mistakes are published, even in top journals. I should know—I once published a false theorem in a top statistics journal. And none of the referees noticed it! Years later, someone pointed out to me that we had made a mistake.

Back to the paper that passed the scrutiny of five referees. It was rebutted in a paper written by 18 authors, including nine Harvard professors. Could nine Harvard professors all be wrong? As you can see, this argument from authority is getting us nowhere. There's real disagreement here. It is as reasonable (or unreasonable) to agree with two non-Harvard professors who happen to be economists as to agree with nine Harvard professors who happen to be anthropologists. Ashraf and Galor explored an interesting pattern. That's a great thing to do. And then they overstated their claims, which I don't think is so great, but it seems to be what it takes to be published in a top journal.

I do understand they are talking about a thought experiment, not a policy directive. The problem is that the thought experiment is not well defined. "Increasing Bolivia's diversity" could be done in many ways, and these would have many effects. An unclearly defined thought experiment is not much of a thought experiment at all. The issue is not the time scale of any potential intervention, but rather that it is not defined.

In that case, what should the authors have done? I think they should have just written, "Using this measure of genetic diversity, we see an interesting nonmonotonic pattern between country-level diversity and country-level income," and gone on to explore this without the causal language. Or, if they wanted to talk about causality (which would be fine), I think they should be specific about potential interventions, whether set in the present era or many years in the past. They don't have to be actual interventions that anyone's done, but they should be clearly defined. "Increasing the genetic diversity of Bolivia" doesn't count as a potential intervention for me, because there are many ways that this could be done and I'd think these different interventions would have many different effects on Bolivia's per-capita income.



## What Is the Ethics Question Here?

So far, we have talked a lot about causal reasoning as it applies in a particular research paper, but nothing really about ethics. The anthropologists' criticism of Ashraf and Galor did have an ethics angle—they claimed the paper could be used to support racist attitudes—but that's not my concern here. As a statistician, I am familiar with the issue that any method my colleagues and I might develop can be used for immoral purposes, and if I considered that a strong general ethical argument, I might as well just close up shop right now. In some cases, I agree that the use of one's statistical tools is relevant, but in this case, I am more worried about problems in the causal claims under discussion than about the uses to which they may be put.

So where do the ethics come in? I claim that the fundamental problem is in the incentives that have contributed to Ashraf and Galor's paper being framed the way it was, which suited it to publication in a top journal. The way I see this work, the authors have an interesting idea and want to explore it. But exploration won't get you published in the *American Economic Review.* Instead of the explore-and-study paradigm, Ashraf and Galor went with assert-and-defend. They made a strong claim and kept banging on it, defending it with a bunch of analyses to demonstrate its robustness. I have no problem with robustness studies (for example, I was upset about some claims about age and happiness a while back because I had difficulty replicating them with new data), but I don't think this lets you off the hook of having to think carefully about causal claims. And presenting tables of numbers to three (meaningless) decimal places doesn't help either.

High-profile social science research aims for proof, not for understanding—and that's a problem. The incentives favor bold thinking and innovative analysis, and that part is great. But the incentives also favor silly causal claims. In many social sciences, it's not enough to notice an interesting pattern and explore it (as my colleagues and I did in our *Red State Blue State* book). Instead, you are supposed to make a strong causal claim, even in a context where it makes little sense.

In summary, I see the ethical problem in our publication system, in which the appearance of a definitive argument is valued over open exploration. Authors are encouraged to see potential criticisms not as open questions to pursue, but as "threats to validity" to be quashed. Now, I don't want to go too far on this. Of course, if you have a hypothesis, it makes sense to gather and analyze new data, or perform new analyses of existing data, to rule out alternative explanations. A research paper is more than a series of hypotheses and a data dump. But some hypotheses are necessarily speculative, and it might not work to try to fit them into a conventional framework of identification of causality. This is where the expectations of top journals might be contributing to a problem, where researchers first are incentivized to make and defend dramatic claims and then find it difficult to moderate their stances under criticism.

That said, I recognize the recursive difficulty in my analysis: I am criticizing the authors of the paper under discussion for implausible and ill-defined causal claims, and here I am making a causal speculation about the effects of a system of publication without considering clear alternatives or potential interventions. All I can say in my defense is my argument here is openly speculative, and I encourage subject-matter experts to think seriously about the incentives for publication in their fields in light of the evidence available for the claims being made (as in the celebrated paper by John Ioannidis speculating on the frequent publication of false claims in medical research, or the work by Gregory Francis, Uri Simonsohn, and others on $p$-values in psychology).

## Further Reading

Ashraf, Q., and O. Galor. 2013. The out of Africa hypothesis, human genetic diversity, and comparative economic development. *American Economic Review* 103(1):1–46.

Calloway, E. 2012. Economics and genetics meet in uneasy union. *Nature* 490:154–155.

Francis, G. 2013. Replication, statistical consistency, and publication bias (with discussion). *Journal of Mathematical Psychology*.

Gelman, A., and T.P. Speed. 1993. Characterizing a joint probability distribution by conditionals. *Journal of the Royal Statistical Society* B 55:185–188.

Gelman, A., and T.P. Speed. 1999. Corrigendum: Characterizing a joint probability distribution by conditionals. *Journal of the Royal Statistical Society* B 61:483.

Guedes, J., et al. 2013. Is poverty in our genes? A critique of Ashraf and Galor, "The out of Africa hypothesis, human genetic diversity, and comparative economic development." *Current Anthropology* 54:71–79.

Ioannidis, J. 2005. Why most published research findings are false. *PLOS Medicine* 2(8), e124.

Simmons, J., L. Nelson, and U. Simonsohn. 2011. False-positive psychology: Undisclosed flexibility in data collection and analysis allow presenting anything as significant. *Psychological Science* 22:1359–1366.

## 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.*