Andrew Gelman
Column Editor

Is It Possible to Be an Ethicist Without Being Mean to People?[1]

A few months ago, I published in *Slate* magazine an article called "Too Good to Be True," in which I discussed two seriously flawed articles that were published in a leading psychology journal.

Not long after my article was published, I received the following email from a well-known professor of psychology whom I'd never met:

> I was surprised to see your trashing of a recent paper in *Psych Science*, and even more surprised to discover that you didn't contact the authors of the paper before you publically leveled charges of sloppiness and stupidity (if not dishonesty) against them. It is really quite shameful. I was glad to see that they posted a thoughtful and temperate response on their website, though sad that it won't be read by all the people who first saw your attack. Some of your points were good, some were sophomoric, and it is hard to understand why you wouldn't have shared your complaints with the authors so that they could help you distinguish between the two. That's something scientists do when their goal is to learn the truth rather than to entertain readers. You hurt the feelings and reputations of some nice young people. I hope you're proud.

I replied that I think it's good for psychology researchers to be discussing these issues. I think that once a paper is published, it is not necessary to contact the authors before publishing criticisms. What motivated me in this case was disappointment in seeing a leading journal of psychology publish a paper on fertility that got the dates of peak fertility wrong with a study based on 100 Internet volunteers and 24 psychology students presented as having discovered the first-ever observable, objective behavior associated with women's ovulation. I have every reason to believe the authors of the papers I discussed have the goal of learning the truth. Unfortunately, I don't think their statistical methods are helping them; rather, I think these methods have led them (and their editor at *Psychological Science*) to jump to conclusions based on noise.

Regarding the authors of the study: Yes, I will willingly criticize work where I find mistakes, even if the authors happen to be nice and young. Daryl Bem is nice (I assume) and elderly; I will criticize his work, too. I am nice and middle-aged, and people can feel free to criticize my work. It's not personal. I think the best thing for all these people (including me) is for them to recognize when they have problems with their research methods.

I concluded my reply by asking the letter writer if he, having read the published study, believed women during their period of peak fertility are really three times more likely to wear red or pink shirts, compared to women in other parts of their menstrual cycle? I don't believe it, myself, or, to put it another way, I don't consider the paper under discussion to provide good evidence of that claim for all the reasons I discussed in my article.

The email exchange progressed for a few more iterations, but did not really move forward. I did not ever apologize for criticizing the research based only on information in the published article, nor did my correspondent answer the question of whether he believed women during their period of peak fertility are really three times more likely to wear red or pink shirts.

Moving beyond the specifics, though, I want to discuss a more general issue: Is it necessary for me, when writing about ethics, to be so negative? Contrary to what my correspondent suggested, when I criticize a study, I am trying to help—not hurt—its authors. In this case, I am hoping these two young scholars will learn more about statistics and avoid jumping to conclusions from small samples—but I can see how, in the short term, it can hurt to be told that one's published study is flawed.

But, beyond anything else, it is likely more difficult to persuade people with a negative message. Here's the paradox: Negativity gets attention (especially in a magazine such as *Slate*), but makes persuasion that much more difficult.

I have noticed that, of my eight ethics columns that have appeared so far in *CHANCE*, seven are largely negative (with, in one case, the negativity being directed back at the practices of other statisticians and me). And I've been thinking about using the title "Crimes Against Data" for my projected book on ethics and statistics.

So maybe I've gone too far in the critical direction. As balance, I will devote the rest of the present column to a positive example, one noted briefly in my response above. It's an article by Brian Nosek, Jeffrey Spies, and Matt Motyl that tells a refreshing story (which I shall excerpt here) about ambitious, but careful, research. It starts as follows:

> Two of the present authors, Motyl and Nosek, share interests in political ideology. We were inspired by the fast growing literature on embodiment that demonstrates surprising links between body and mind (Markman & Brendl, 2005; Proffitt, 2006) to investigate embodiment of political extremism. Participants from the political left, right and center (N = 1,979) completed a perceptual judgment task

[1] To appear in *CHANCE* as part of a regular column on ethics and statistics. We thank Eric Loken for helpful comments.

> in which words were presented in different shades of gray. Participants had to click along a gradient representing grays from near black to near white to select a shade that matched the shade of the word. We calculated accuracy: How close to the actual shade did participants get? The results were stunning. Moderates perceived the shades of gray more accurately than extremists on the left and right (p = .01). Our conclusion: political extremists perceive the world in black-and-white, figuratively and literally. Our design and follow-up analyses ruled out obvious alternative explanations such as time spent on task and a tendency to select extreme responses. Enthused about the result, we identified *Psychological Science* as our fall back journal after we toured the *Science*, *Nature*, and *PNAS* rejection mills. The ultimate publication, Motyl and Nosek (2012) served as one of Motyl's signature publications as he finished graduate school and entered the job market.

The authors continue:

> The story is all true, except for the last sentence; we did not publish the finding. Before writing and submitting, we paused. Two recent papers highlighted the possibility that research practices spuriously inflate the presence of positive results in the published literature (John, Loewenstein, & Prelec, 2012; Simmons, Nelson, & Simonsohn, 2011). Surely ours was not a case to worry about. We had hypothesized it, the effect was reliable. But, we had been discussing reproducibility, and we had declared to our lab mates the importance of replication for increasing certainty of research results. We also had an unusual laboratory situation. For studies that could be run through a web browser, data collection was very easy (Nosek et al., 2007). We could not justify skipping replication on the grounds of feasibility or resource constraints. Finally, the procedure had been created by someone else for another purpose, and we had not laid out our analysis strategy in advance. We could have made analysis decisions that increased the likelihood of obtaining results aligned with our hypothesis. These reasons made it difficult to avoid doing a replication. We conducted a direct replication while we prepared the manuscript. We ran 1,300 participants, giving us .995 power to detect an effect of the original effect size at alpha = .05.

And, then, the punch line:

> The effect vanished (p = .59).

Their paper is all about how to provide incentives for this sort of good behavior, in contrast to the ample incentives that researchers have to publish their tentative findings attached to grandiose claims.

This story also can be understood from a more formal statistical perspective, hypothesizing a true effect size and then considering what sort of experimental data are likely to arise.

Suppose we are studying a comparison whose true size is 0.05, on a scale in which 1 is the standard deviation of the measurements. An effect of 0.05 may sound too low but I think it is realistic or even optimistic in typical between-subject designs[2] given the large levels of measurement error and individual variation in psychology experiments. Now suppose data are gathered on 100 people in each of two groups to be compared. The standard error for the difference is sqrt(1/100+1/100) = 0.14. Any true effect will be lost in the noise—indeed, it is quite probable that any observed difference will go in the opposite direction of the average in the population.

In theory, the power of a study such as described above—the probability that the observed difference will be at least two standard errors away from zero—is easily computable using the normal distribution and comes to 0.06, which can be broken down into a 0.05 chance of finding a positive and statistically significant difference and a 0.01 chance of finding a negative and statistically significant difference. The theoretical Type S (sign) error rate is then 0.01/0.06 = 16% (Gelman and Tuerlinckx, 2000).[3]

In practice, though, the probability of finding statistical significance is much higher, given that even when a research hypothesis is prespecified, there will be many choices involved in the statistical analysis, choices such as which data (if any) to exclude, which categories of responses to combine, whether to look at interactions, and so forth. Simmons, Nelson, and Simosohn (2011) discuss the ubiquity of such "researcher degrees of freedom" in psychology research. Given such flexibility, it is extremely likely that *something* statistically significant will be found, and that this highlighted comparison will make sense in light of the researchers' hypotheses. This is what happened to Nosek, Spies, and Motyl in the first part of their story recounted above.

Now on to the next stage, in which a specific analysis is picked out and further data are gathered. Suppose the new study again has 100 participants: then it is likely that nothing statistically significant will turn up, because this time the theoretical assumptions hold and the analysis has been pre-chosen. At this point the story can take several paths. One possibility is that this study is taken as a negative replication, in which case it can be dismissed on grounds of lack of power, or crudely counted as a "-1" on a scale in which the original result counts as a +1. Another option is for the outcome criteria to be tweaked in some way so that the data to once again appear to be statistically significant.

In this case, however, Nosek, Spies, and Motyl replicated with a much larger study of 1300 people, for which the standard error for a simple comparison would be sqrt(1/650+1/650) = 0.055—smaller than the standard error for the hypothetical n = 200 study, but still leaving not much chance for statistical significance. From this perspective it is indeed no surprise that their attempted replication failed.

[2] In a *between-subject* design, two different groups of people are compared. This is distinguished from *within-subject* designs, in which two different measurements are compared within a single group of people. Within-subject designs typically have the advantage of lower variability—by taking the difference of two measurements taken on a single person, much of the between-person design is subtracted out. Between-subject designs have the advantage of simplicity of interpretation: there is no need to worry that the earlier measurement on a person is influencing the later measurements. In the experiments being discussed here, the losses from gathering between-subject data are large: there is so much variation between people that it is hard to learn much at all even from fairly large sample sizes.

[3] We define a *Type S error* as an estimated comparison that is made with confidence and has a sign opposite to the true value, and a *Type M error* as an estimate whose magnitude is far from that of the true value. For example, if the true value of a parameter is -0.1 and the estimate is 0.5 and statistically significant, this is a Type S error (wrong sign) and a Type M error (gross overestimate of magnitude). When sample sizes are small or measurement error is large compared to true effect sizes, Type S errors can be large; that is, it can be likely that statistically significant findings are in the wrong direction. This problem is exacerbated by the practice of multiple comparisons and analyses that are contingent on data. In addition, in these settings, statistically-significant point estimates tend to be much larger than true (population) comparisons; that is, Type M errors are common. I prefer the Type M and Type S framework to the traditional discussion of Type 1 and Type 2 errors because in the sorts of problems my colleagues and I work on, interest is in the sign, magnitude, and persistence of effects rather than the simple question of their presence or absence.

But what about the claim that a study of 1,300 people would given them ".995 power to detect an effect of the original effect size"? Given that their replication did not detect such an effect, or even come close, this suggests something went wrong in that power calculation. The researchers' mistake was to use in their design the estimate from the original study. That estimate is likely to be highly inflated (that is, a type M or magnitude error): the very fact that it was statistically significant in a small-sample study is evidence that it is, in whole or part, the product of noise. A more realistic effect size estimate would give a more realistic power calculation.

As we and others have discussed, the direct interpretation of "statistically significant" findings leads to three distinct problems. First, it is possible—indeed, common practice—to perform an analysis that is contingent on the particular data that appear, thus greatly increasing the probability that a pattern can be found to be statistically significant even if it occurs purely from chance (Simmons, Nelson, and Simosohn, 2011). Second, claims that have survived the "statistical significance filter" are probably overestimates of any true effects (Gelman and Weakliem, 2009, Button et al., 2013). Third, designs of future studies tend to be wildly optimistic if they are based on statistically-significant effect size estimates from previous studies, leading to a boom-and-bust cycle of hype and disappointment or, worse, an explaining-away of failed replications if too much trust is placed in the original finding.

It is to the credit of Nosek, Spies, and Motyl that they demonstrated all of this in the context of difficulties with their own work, thus giving other researchers a path forward. And it's been a pleasure for once to write an ethics column with a positive focus.

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. Power failure: Why small sample size undermines the reliability of neuroscience. *Nature Reviews Neuroscience* 14:365–376.

Gelman, A. 2013. Too good to be true. *Slate*, 24 Jul. *www.slate.com/articles/health_and_science/science/2013/07/statistics_and_psychology_multiple_comparisons_give_spurious_results.html*

Gelman, A., and F. Tuerlinckx. 2000. Type S error rates for classical and Bayesian single and multiple comparison procedures. *Computational Statistics* 15:373–390.

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

Harris, W. S., M. Gowda, J. W. Kolb, C. P. Strychacz, J. L. Vacek, P. G. Jones, A. Forker, J. H. O'Keefe, and B. D. McCallister. 1999. A randomized, controlled trial of the effects of remote, intercessory prayer on outcomes in patients admitted to the coronary care unit. *Archives of Internal Medicine* 159:2273–2278.

John, L., G. Loewenstein, and D. Prelec. 2012. Measuring the prevalence of questionable research practices with incentives for truth-telling. *Psychological Science* 23:524–532.

Markman, A. B., and C. M. Brendl. 2005. Constraining theories of embodied cognition. *Psychological Science* 16:6–10.

Motyl, M., and B. A. Nosek. 2012. Political extremists see the world in black-and-white, literally. Unpublished data.

Nosek, B. A., F. L. Smyth, J. J. Hansen, T. Devos, N. M. Lindner, K. A. Raganath, . . . M. R. Banaji. 2007. Persuasiveness and correlates of implicit attitudes and stereotypes. *European Review of Social Psychology* 18:36–88.

Nosek, B. A., J. R. Spies, and M. Motyl. 2013. Scientific utopia: II. Restructuring incentives and practices to promote truth over publishability. *Perspectives on Psychological Science* 7:615–631.

Proffitt, D. R. 2006. Embodied perception and the economy of action. *Perspectives on Psychological Science* 1:110–122.

Simmons, J. P., L. D. Nelson, and U. Simosohn. 2011. False-positive psychology: Undisclosed flexibility in data collection and analysis allows 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.*