Andrew Gelman
Column Editor

Disagreements About the Strength of Evidence

In the world we live in, the scientific claims we notice have been selected through a complex filtering process involving funders, researchers, scientific journals, and the news media. The question to us, as statisticians and consumers of research, is how to interpret what we see. Our conclusions about the world—based on the research that we hear about—rely on certain selected data summaries rather than on all the potentially relevant data.

Interpretation of research claims can be contentious. For example, when Malcolm Gladwell credulously told the story of a mathematician who claimed to be able to predict marriage breakups with 90% accuracy—and reported on a research study on learning difficulties without noting that it failed to replicate on "a larger and more representative sample of subjects" (in the words of psychology researcher Christopher Chabris)—the two sides had completely different perspectives and different outlets. On one hand, Gladwell has an audience of millions who have never heard of Chabris and will never hear his criticisms. Meanwhile, among social scientists, Gladwell is known as a popularizer who makes frequent mistakes and cannot seriously counter any of Chabris's objections on the merits.

Any disagreement between them may start with a dispute on the science, but it quickly becomes a discussion of how much simplification and dramatic storytelling is appropriate in communicating scientific claims to general audiences.

Even among scientists, controversies often seem to take on asymmetric forms, with one side pointing out an error and another side furiously denying it. Unfortunately, people do not seem to like to admit their errors. And, as discussed in an earlier column, ethics questions arise regarding systemic incentives for exaggerating claims and denying errors. From a scientific perspective, though, these cases often seem pretty clear.

A Scientific Disagreement Without a Clean Resolution

Recently, though, I had an interesting dispute that did *not* follow the above pattern. It was a disagreement with Larry Bartels, a well-known political scientist who works in areas of public opinion and elections that are close to my own research topics in American politics. I have a lot of respect for Bartels's research and his scholarly judgment.

In this case, the disagreement involved the effects of subliminal stimuli on political attitudes in the context of interpretation of an experiment conducted a few years ago and recently published by political scientists Cengiz Erisen, Milton Lodge, and Charles Taber.

To put things in a positive way, Bartels was writing about some interesting recent research which I then constructively criticized. Or, to be more negative, Bartels was hyping some dramatic claims, and I was engaging in mindless criticism.

Before going on, let me emphasize that I am expressing no general opinion (and certainly no expertise) on the experiments performed by Erisen; the other work of his collaborators Lodge and Taber; the effects of subliminal stimuli; or priming studies in general.

The two main points I'm making here are, first, to discuss some statistical aspects of our disagreement regarding the strength of the evidence on effects of subliminal stimuli on attitudes; and, second, to tell the somewhat unhappy story about how my methodological discussion with Bartels was contentious. This second observation is relevant to a discussion of ethics

and statistics because it reflects the difficulties of communication even in this relatively clean example.

The balance between promotion and criticism is always worth discussing, but particularly so in this case because of two factors:

1. The research in question is on the borderline. The conclusions in question are not rock-solid—they depend on how you look at the data and are associated with p-values like 0.10 rather than 0.0001—but neither are they silly. Some of the findings definitely seem real, and the debate is more about how far to take it than whether there's anything there at all. Nobody in the debate is claiming that the findings are empty; there's only dispute about their implications.

2. The topic—effects of unnoticed stimuli on political attitudes—is important, and I recognize that Bartels has something valuable to say on the topic regarding both methods and public opinion.

I will go into detail about the example itself and then discuss some of the frustrations Bartels and I, in different ways, have expressed in our exchanges.

A Disputed Claim About the Effects of Subliminal Stimuli on Political Attitudes

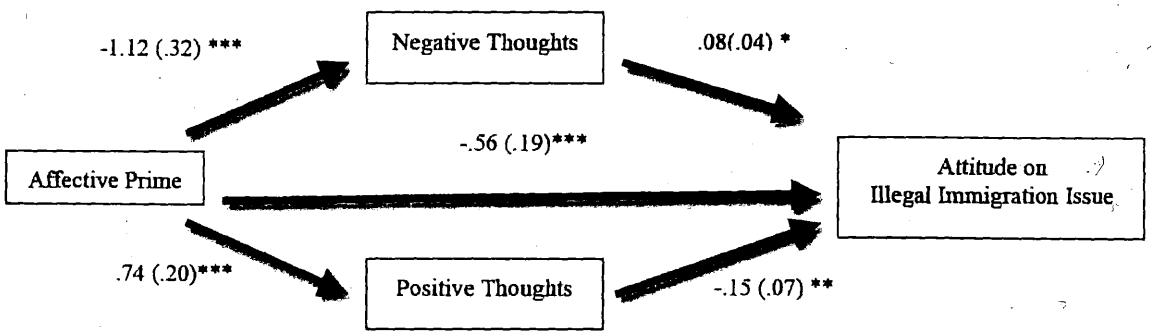
The story starts with a post by Larry Bartels in the Monkey Cage, a political science blog at *The Washington Post*, with the title, "Here's how a cartoon smiley face punched a big hole in democratic theory." The subtitle was "Fleeting exposure to 'irrelevant stimuli' powerfully shapes our assessments of policy arguments." Bartels wrote:

> What were these powerful "irrelevant stimuli" that were outweighing the impact of subjects' prior policy views? Before seeing each policy statement, each subject was subliminally exposed (for 39 milliseconds—well below the threshold of conscious awareness) to one of three images: a smiling cartoon face, a frowning cartoon face, or a neutral cartoon face. . . . [T]he subliminal cartoon faces substantially altered their assessments of the policy statements . . .

I followed up with a post expressing some skepticism:

> Unfortunately they don't give the data or any clear summary of the data from experiment No. 2, so I can't evaluate it. I respect Larry Bartels, and I see that he characterized the results as the "subliminal cartoon faces substantially altered their assessments of the policy statements—and the resulting negative and positive thoughts produced substantial changes in policy attitudes." But based on the evidence given in the paper, I can't evaluate this claim. I'm not saying it's wrong. I'm just saying that I can't express judgment on it, given the information provided.

Bartels then followed up with a post saying that further information was in Chapter 3 of Erisen's PhD dissertation and presented as evidence this path analysis:



Along with this summary:

> In this case, subliminal exposure to a smiley cartoon face reduced negative thoughts about illegal immigration, increased positive thoughts about illegal immigration, and (crucially for Gelman) substantially shifted policy attitudes.

Erisen also sent along a note with further explanation, the centerpiece of which was another path analysis.

Unfortunately, I still wasn't convinced. The trouble is, I get confused whenever I see these path diagrams. What I really want to see is a direct comparison of the political attitudes with and without the intervention. No amount of path analysis will convince me until I see the direct comparison.

However, as Bartels pointed out to me, I had not read the entire relevant chapter of Erisen's dissertation in detail. I'd looked at the graphs (which had results of path analyses and data summaries on positive and negative thoughts, but no direct data summaries of issue attitudes) and at some of the tables. It turns out there were some direct comparisons of issue attitudes in the text of the dissertation—but not in the tables and figures.

I'll get back to that in a bit, but first let me return to what I wrote at the time:

> I'm not saying that Erisen is wrong in his claims, just that the evidence he [and Bartels] have shown me is too abstract to convince me. I realize that he knows a lot more about his experiment and his data than I do and I'm pretty sure that he is much more informed on this literature than I am, so I respect that he feels he can draw certain strong conclusions from his data. But, for me, I have to go with what information is available to me.

Why do these claims from path analysis confuse me? An example is given in a blog comment by David Harris, who reported that Erisen et al. "seem to acknowledge that the effect of their priming on people's actual policy evaluations is nil," but that then follow up with a convoluted explanation involving a series of interactions.

Convoluted can be OK—real life is convoluted—but I'd like to see some simple comparisons. If someone wants to claim that "Fleeting exposure to 'irrelevant stimuli' powerfully shapes our assessments of policy arguments," I'd like to see if these fleeting exposures indeed have powerful effects. In an observational setting, such effects can be hard to tease out. But in this case, the researchers did a controlled experiment, and I'd like to see the direct comparison as a starting point.

Commenter Dean Eckles reported that "those effects are not significant at conventional levels in Exp 2," pointing to two passages from Erisen's dissertation. On illegal immigration:

> In the first step of the mediation model a simple regression shows the effect of affective prime on the attitude (beta=.34; p [less than] .07). Although not hypothesized, this confirms the direct influence of the affective prime on the illegal immigration attitude.

And on energy security:

> As before, the first step of the mediation model ought to present the effect of the prime on one's attitude. In this mediation model, however, the affective prime does not change energy security attitude directly (beta=-.10; p [greater than] .10. Yet, as discussed before, the first step of mediation analysis is not required to establish the model (Shrout & Bolger 2002; MacKinnon 2008).

From the standpoint of p-values, this would seem to pretty much cover it. The direct result was not statistically significant. When it went in the expected direction and was not statistically significant, it was taken as a confirmation of the hypothesis. When it went in the wrong direction and was not statistically significant, it was dismissed as not being required. There were several of this sort of comparison in the dissertation.

But the concerns about statistical significance should not be taken to imply a criticism of Erisen's dissertation or his research paper with Lodge and Taber, as the focus of that work was on the path analyses, not on the direct effects. Rather, it illustrates the difficulty of using a patchwork of comparisons to draw general conclusions—in this case, conclusions about direct effects that were not central to the original study.

Beyond this, as John Bullock pointed out in a blog comment, about 15-20% of the cases were excluded from these analyses, so "the reported estimates are not really estimates of the average treatment effects for the experimental sample." This does not mean the work in question is bad or that this selection had a material effect on the research conclusions—nor does it mean that Erisen's experiments were not worth doing. It just reveals the difficulty involved in identifying and interpreting even the simplest of comparisons.

Two Perspectives

So, here you have the story as I see it: Bartels learned of an interesting study regarding subliminal stimuli—a study that made a lot of sense, especially in light of some of his earlier work on ways voters can be

swayed by information that logically should be irrelevant to voting decisions or policy positions. (This is consistent with the work of Daniel Kahneman, Paul Slovic, and Amos Tversky regarding shortcuts and heuristics in decision making.) The work seemed solid and was supported by several statistical analyses. And there does seem to be something there; in particular, Erisen shows strong evidence of the stimulus affecting the numbers of positive and negative thoughts expressed by the students in his experiment. But the evidence for Bartels' headline claim—that subliminal smiley-faces affect political attitudes, not just positive and negative expressions—is not so clear.

That's my perspective. Now for Bartels' perspective. He wrote that my original post was sloppy in that I was not looking at all the evidence—that I reacted to the path analyses presented by him and Erisen and did not look carefully within Erisen's PhD thesis to find the direct comparisons. In particular, Bartels pointed to several comparisons of average issue attitudes among respondents—differences between the two treatment groups that were large and which had p-values in the range between 0.05 and 0.10. He considered the appearance of several positive aggregate differences in the data as representing good evidence, even if none of the single comparisons were statistically significant on their own.

This response by Bartels was fair, both in pointing out several comparisons that I had not noticed in the dissertation and in emphasizing that it would be wrong to say these results provide no evidence supporting his claims just because there happens to be no $p<0.05$.

My quick response is that the interpretation of these data depends a lot on your priors. If we accept the general quality of the measurements in this study (no big systematic errors, for example), then there's very clear evidence of the subliminal stimuli having effects on positive and negative expressions. Hence, it's completely reasonable to expect effects on other survey responses, including issue attitudes. That is, we're not in Bem territory here. Assuming the experiments are done competently, there are real effects here. Given that the stimuli can affect issue attitudes, it's reasonable to expect variation; to expect some positive and some negative effects; and for the effects to vary across people and across situations. So if I wanted to study these effects, I'd be inclined to fit a multilevel model, not excluding any data, to better estimate effects in the context of variation.

When it comes to specific effects, and to specific claims of large effects—recall the original claim that the stimulus "powerfully shapes our assessments of policy arguments," and the claims that it "substantially altered," and "punched a big hole in democratic theory"—I'd like to see some strong evidence. And this mixture of positive and negative comparisons does not look like strong evidence to me. I agree these results are consistent with some effect on issue attitudes, but I don't see the evidence for the large effects that have been claimed.

I respect the path analyses for what they are, and I'm not saying Erisen shouldn't have done them. But I think it's fair to say that these are the sorts of analyses used to understand large effects that exist. They don't directly address the question of the effects asserted by Bartels of the stimulus on policy attitudes (which is how we could end up with an explanation of large effects that cancel out).

As a Bayesian, I do accept Bartels' criticism that it was odd for me to claim there was no evidence just because p was not less than 0.05. Even weak evidence should shift my priors a bit, no? And I agree that weak evidence is not the same as zero evidence.

So let me clarify that, accepting the quality of Erisen's experimental protocols (which I have no reason to question), and assuming the results hold up if all the data are included in the analyses, I have no doubt that some effects are there. The question I am raising is about the size, consistency, and direction of the effects on policy attitudes.

An Unsatisfying Ending

In some sense, the post-publication review process worked well. Bartels gave the original work a wide audience by promoting it on the Monkey Cage; I read that post and offered my objection in a post of my own; and, in turn, Erisen and various commenters replied. And, eventually, after a couple of email exchanges, I finally got the point that Bartels had been trying to explain to me—that Erisen did have some version of the direct comparisons I'd been asking for; they were just in the text of his dissertation and not in the tables and figures.

In all of this, my contributions have been entirely methodological and administrative. I have made some remarks about what I consider strong, statistical evidence. And with my blog posts, I have provided a forum for several insightful commenters. My intention is not, nor has it ever been, to accuse either Erisen or Bartels of any ethical violations, and I recognize their expertise. Indeed, one of the most frustrating aspects of this entire episode has been the difficulty of raising a methodological objection (in this case, to Bartels' claim that the experiment demonstrated large effects on political attitudes) without such objections being taken as a personal attack. I am sure that some of this is my fault, that I have implicitly sent negative vibes without intending to—a bit of unintentional subliminal stimuli, I suppose. And this is regrettable enough that I think it's worth discussion. Again, criticism is central to science.

But when it is perceived as personal, it can be hard for the message to get through.

Our post-publication discussion was slow and frustrating for all concerned, I believe. But I still think it moved forward in a better way than it would have without the open exchange. It progressed better than if, for example, we'd had only a series of static, published articles presenting one position or another. At the least, we received a lot of helpful input from blog commenters.

Bartels posted on a research finding that he thought was important and perhaps had not received enough attention. I was skeptical. After all the dust has settled, I remain skeptical about any effects of the subliminal stimulus on political attitudes. I cannot be sure what Bartels' current view is, but I think he remains convinced that "subliminal exposure to a smiley cartoon face ... substantially shifted policy attitudes." Maybe our disagreement ultimately comes down to priors, which makes sense given that the evidence from the data is weak. In any case, it is not up to me or Bartels to adjudicate this issue; both of us can merely present and argue our views. Ultimately, the scientific consensus should and will be decided by further research in this area. I don't want to overplay the importance of the interaction between Bartels and me, which is more about how research such as Erisen's is presented and interpreted than about the research itself.

Our interactions have been difficult and at times painful, with Bartels expressing irritation at my expression of what he characterized as a "made-up disagreement." I think the disagreement is real, but from a statistical standpoint, this can be viewed as a sort of informational correlation by which one's priors rebound back and influence one's interpretation of evidence.

Meanwhile, new studies are published and neglected, hyped, or both. I offer no general solution to how to handle these; clearly, the standard system of scientific publishing has its limitations. I just want to raise some of these issues in a context where I see no easy answers.

I think social science can—and should—do better than we usually do. Consider the notorious economics paper by Reinhart and Rogoff that was published in a top journal with serious errors that were not corrected until several years after publication.

On one hand, the model of discourse I describe in this column is not at all scalable in that it relied on lots of effort from Bartels, Erisen, me, and several commenters, and all of us have only finite time available for this sort of thing. Beyond this, it involved the discomfort involved in open disagreement. On the other hand, consider the thousands of researchers who spend many hours refereeing papers for journals. Surely this effort could be channeled in a more useful way.

Further Reading

Achen, C. H., and L. M. Bartels. 2002. Blind retrospection: Electoral responses to drought, flu, and shark attacks. Prepared for presentation at the Annual Meeting of the American Political Science Association, Boston.

Bartels, L. 2014. Here's how a cartoon smiley face punched a big hole in democratic theory. Monkey Cage blog, *The Washington Post. www.washingtonpost.com/blogs/monkey-cage/wp/2014/09/04/heres-how-a-cartoon-smiley-face-punched-a-big-hole-in-democratic-theory*

Bartels, L. 2014. More on smiley-face democracy. Monkey Cage blog, *The Washington Post. www.washingtonpost.com/blogs/monkey-cage/wp/2014/09/08/more-on-smiley-face-democracy*

Chabris, C. 2013. The trouble with Malcolm Gladwell. *Slate. www.slate.com/articles/health_and_science/science/2013/10/malcolm_gladwell_critique_david_and_goliath_misrepresents_the_science.html*

Erisen, C. 2009. Affective contagion: The impact of subtle affective cues in political thinking. PhD dissertation, Department of Political Science, Stony Brook University.

Erisen, C., M. Lodge, and C. S. Taber. 2012. Affective contagion in effortful political thinking. *Political Psychology* 35, 187–206.

Gelman, A. 2014. Sailing between the Scylla of hyping of sexy research and the Charybdis of reflexive skepticism. Statistical Modeling, Causal Inference, and Social Science blog. *http://andrewgelman.com/2014/10/22/sailing-scylla-hyping-sexy-research-charybdis-reflexive-skepticism*

Gladwell, M. 2013. Christopher Chabris should calm down. *Slate*, 10 Oct. *www.slate.com/articles/health_and_science/science/2013/10/malcolm_gladwell_s_david_and_goliath_he_explains_why_christopher_chabris.html*

Kahneman, D., and A. Tversky. 1974. Judgment under uncertainty: Heuristics and biases. *Science* 185, 1124–1131.

Lodge, M., and C. S. Taber. 2013. *The rationalizing voter*. Cambridge University Press.

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*.