STATISTICS AND PUBLIC POLICY
2017, VOL. 4, NO. 1, 1–8
http://dx.doi.org/10.1080/2330443X.2016.1277966

Taylor & Francis
Taylor & Francis Group

The 2008 Election: A Preregistered Replication Analysis

Rayleigh Lei[a], Andrew Gelman[b], and Yair Ghitza[c]

[a]University of Michigan, Ann Arbor, MI; [b]Columbia University, New York, NY; [c]Catalyst, Chicago, IL

ABSTRACT

We present an increasingly stringent set of replications, a multilevel regression and poststratification analysis of polls from the 2008 U.S. presidential election campaign, focusing on a set of plots showing the estimated Republican vote share for whites and for all voters, as a function of income level in each of the states.

We start with a nearly exact duplication that uses the posted code and changes only the model-fitting algorithm; we then replicate using already-analyzed data from 2004; and finally we set up preregistered replications using two surveys from 2008 that we had not previously looked at. We have already learned from our preliminary, nonpreregistered replication, which has revealed a potential problem with the earlier published analysis; it appears that our model may not sufficiently account for nonsampling error, and that some of the patterns presented in that earlier article may simply reflect noise.

In addition to the substantive interest in validating earlier findings about demographics, geography, and voting, the present project serves as a demonstration of preregistration in a setting where the subject matter is historical (and thus the replication data exist before the preregistration plan is written) and where the analysis is exploratory (and thus a replication cannot be simply deemed successful or unsuccessful based on the statistical significance of some particular comparison).

Our replication analysis produced graphs that showed the same general pattern of income and voting as we had found in our earlier published work, but with some differences in particular states that we cannot easily explain and which seem too large to be explained by sampling variation. This process thus demonstrates how replication can raise concerns about an earlier published result.

ARTICLE HISTORY
Received February 2016
Accepted July 2016

KEYWORDS
Bayesian methods; Other; Process improvement

1. Background

Replication is central to scientific objectivity and is increasingly recognized to be important in social science as well. In social science, there has also been a movement toward *preregistration*, the specification of protocols for data collection, data analysis, and data processing ahead of time (see, e.g., Gonzales and Cunningham 2015), as a way of eliminating selection bias that has called into question inference from individual studies and entire literatures.

A much-discussed cautionary example of such bias from psychology is the study of "embodied cognition," in particular an article by Bargh, Chen, and Burrows (1996), which has been cited over 3600 times but has recently been called into question. After a series of failed replications (Doyen et al. 2012; Wagenmakers et al. 2015), it seems possible that the empirical results of that entire subfield can entirely be explained by a series of optimistic researchers capitalizing on noise. In a more humble example, Nosek, Spies, and Motyl (2012) recounted the story of their "50 shades of gray" experiment in which they obtained a striking result relating political extremism to color perception, a correlation that seemed to be strongly backed up by statistically significant p-values but which yielded null results under a careful preregistered replication.

It has been increasingly recognized in recent years that nonpreregistered studies can have problems arising from "researcher degrees of freedom" (Simmons, Nelson, and Simonsohn 2011) or the "garden of forking paths" (Gelman and Loken 2014). When a study is conducted in an open-ended fashion, researchers have many degrees of freedom to decide what data to collect, what data to exclude from their analysis, and what comparisons to perform, while the data are still coming in. It becomes easy to get statistically significant p-values even if underlying effects are zero (or, more realistically, in settings with low signal-to-noise ratio so that statistically significant findings are likely to be highly exaggerated and often in the wrong direction; Gelman and Carlin (2014). As with the use of a hold-out sample in machine learning, external replication offers the promise of reducing such biases.

In political science, the term "replication" has traditionally been applied to the simple act of reproducing a published result using the identical data and code as used in the original analysis. Anyone who works with real data will realize that this exercise is

CONTACT Andrew Gelman gelman@stat.columbia.edu Columbia University, 1016 Social Work Building, New York, NY 10027.

valuable and can catch problems with sloppy data analysis (e.g., the Excel error of Reinhart and Rogoff 2010, or the "gremlins" article of Tol 2009, which required nearly as many corrections as the number of points in its dataset; see Gelman 2014). Reexamination of raw data can also expose mistakes, such as the survey data of LaCour and Green (2014); see Gelman (2015).

But procedural replication does not address researcher degrees of freedom or forking paths. To address these concerns it is helpful to have true replication with new data and a preregistered data-processing and analysis plan. Humphreys, de la Sierra, and van der Windt (2013) discussed how formal preregistration can work with laboratory or field experiments and this seems like a promising approach: We will not want every analysis to be preregistered but it is useful as an option, especially in studies that attempt to replicate controversial previously published claims.

Preregistered replication is more challenging in observational settings. For one thing, observational data can sometimes not be replicated at all. We cannot, for example, replicate an international relations study on a new sample of wars or recessions. The other challenge is that some datasets are so well understood that it would be meaningless to talk about a preregistered data collection and analysis protocol. Consider, for example, the much analyzed and much debated time series of economic growth and the party of the president (Bartels 2008; Campbell 2011; Comiskey and Marsh 2012), a problem that can never again be virgin territory for statistical analysis.

All this has led to the awkward situation that we applaud the calls for preregistration of others' work but have never conducted a preregistered replication of our own (Gelman 2013).

Recently, however, we have come across an opportunity to perform a preregistered replication of our own work. In Ghitza and Gelman (2013), we reported the results of a statistical analysis of poll data from the Pew Research Center in the lead-up to the 2008 U.S. presidential election. Ghitza and Gelman performed several analyses; in the present article we replicate one of them, an estimate of John McCain's share of the two-party by income, ethnicity, and state, as summarized in Figure 2 of that earlier article, which displays raw data and estimated McCain support as a function of five income categories for white voters and all voters in each of the 50 states.

We perform four replications of this analysis. Because we have already performed the first two replications and will describe them below, they are *not* preregistered:

1. A nearly exact duplication, using the same data and model, just changing the statistical analysis slightly by fitting a fully Bayesian analysis in Stan in place of the marginal maximum likelihood estimate presented in Ghitza and Gelman (2013).
2. A replication of the fully Bayesian analysis on a slightly different problem, the 2004 presidential election, using the Annenberg preelection poll from that year.

The above replications help us build trust in our method and smoke out any problems before then setting up the protocol for our two preregistered replications:

(3) A replication using the fully Bayesian analysis on the *telephone sample* from the 2008 Annenberg preelection survey.

(4) A replication using the fully Bayesian analysis on the *internet sample* from the 2008 Annenberg preelection survey.

In Sections 1–4 of this article, we give the results from analyses 1 and 2 above and set up our preregistration plan for analyses 3 and 4. We will time-stamp our article up to that point and post on the internet. Section 5 reports the results of replications 3 and 4 and compares them to the earlier findings obtained from the Pew survey.

Replications in psychology are often performed because of suspicion or controversy about published findings, and the goal of such replications is often to resolve the controversy in some way. The present replication is different. The findings of Ghitza and Gelman (2013) have not been controversial; indeed that article is largely methodological with no particular headline findings to confirm or reject. Rather, the role of the present article is to demonstrate the challenges of preregistration in a setting in which data are observational and analysis is complex. In this case, as we suspect in many statistically intensive problems, the design of a replication requires a surprising amount of effort. This suggests why such replications are not performed more often and, we hope, motivates us to a future in which statistical workflows are specified in a more replicable fashion.

That said, there are some substantive implications that we would like our replication to address. From a practical standpoint, the message of Ghitza and Gelman (2013) is that researchers can use multilevel regression and poststratification to make inferences about small subgroups of the population, for example, the voting patterns of whites at different income levels within a state. Our substantive focus will be to examine the results on income and voting from that published article and see how they replicate with new data. Unlike many replications, we are not checking a particular comparison or coefficient to see if it remains statistically significant. We hope this work is helpful to researchers in demonstrating replication in a more diffuse setting which, we believe, is characteristic of much social science research.

2. Duplications and Replications Using Existing Data

Ghitza and Gelman (2013) analyzed voter turnout and vote choice using a set of multilevel models predicting individual survey responses given income (divided into 5 categories), age (4 categories), ethnicity (white, black, hispanic, and other), and 51 states (including the District of Columbia). For each binary outcome y (e.g., vote intention for the Republican or Democratic candidate, excluding respondents who are undecided or express other preferences), a logistic regression is fit, $\Pr(y_i = 1) = \text{logit}^{-1}(X_i\beta)$, where X includes indicators for the demographic and geographic variables listed above, along with certain interactions of these main effects. The coefficients β are given a hierarchical prior distribution, in which batches of coefficients ("random effects," also called "varying intercepts and slopes;" Gelman and Hill 2007) are assigned normal distributions with variances that are estimated from the data. The model was used to create a predicted probability for each of the $5 \times 4 \times 4 \times 51$ *poststratification cells* j corresponding to combinations of the demographic and geographic factors in the model. Cells were then combined using Census data for the number of people

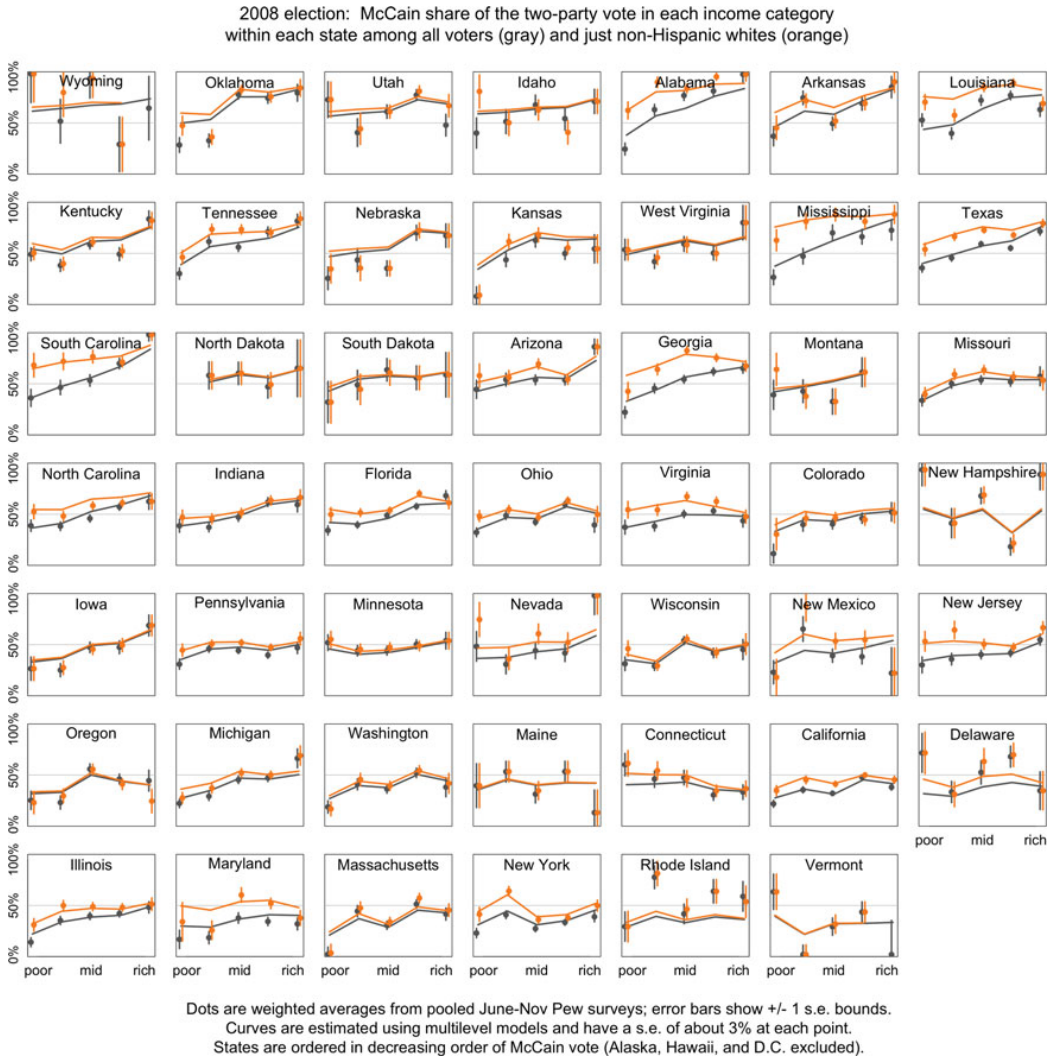

Figure 1. Results from using Stan to fit the multilevel model to the 2008 Pew data, showing estimated McCain share of the two-party vote by state and income category for all voters (black) and just white voters (brown). This change to fully Bayesian analysis had little effect on the inferences, as can be seen by comparing to the earlier version, Figure 2 from Ghitza and Gelman (2013).

or voters in each category. This poststratification step was performed using simple weighted averaging, for any subset S of the population, computing $\hat{\theta}_j = \sum_{j \in S} N_j \hat{\theta}_j / \sum_{j \in S} N_j$, where $\hat{\theta}_j$ represents the fitted $\Pr(y = 1)$ for people in cell j, and N_j is the Census estimate of the population in the cell.

We performed our first replication, using the same data and altering the code only to fit a fully Bayesian version of our model in Stan (Stan Development Team 2015). Compared to the earlier-fit model that performed marginal maximum likelihood, the only difference we noticed was that when the earlier point estimate calculated zero for a variable's terms, the Stan-calculated terms were nonzero. The practical differences between the two estimates were tiny, though, because when the marginal maximum likelihood estimate was zero for hierarchical variance parameters, the full Bayes estimates were small as well, and the differences were essentially nil when it came to prediction of turnout and vote choice proportions.

Figure 1 shows the results from the replications on the 2008 Pew data, with the only change being the switch to fully Bayesian inference. The result is essentially the same as Figure 2 of Ghitza and Gelman (2013), which is no surprise given the large sample size and large number of groups in the multilevel model.

For our second replication, we applied our newly cleaned code to the 2004 Annenberg preelection poll and produced Figure 2, which displays the raw data and Bayesian estimates for George W. Bush's vote share, by income, ethnicity, and state.

These results look reasonable, but we were struck by some differences as compared to the Pew 2008 analyses shown in Figure 1. It should be no surprise to see changes in individual states, as the two elections were different, most notably among African Americans throughout the country and white voters in the South. But we also notice a systematic difference: the lines in Figure 2 are much smoother than those in Figure 1. Our analysis from Annenberg 2004 shows a much more regular and monotonic pattern of income and voting by state, compared to our analysis from Pew 2008. It is *possible* that this represents a real change but we think it is more likely a statistical artifact.

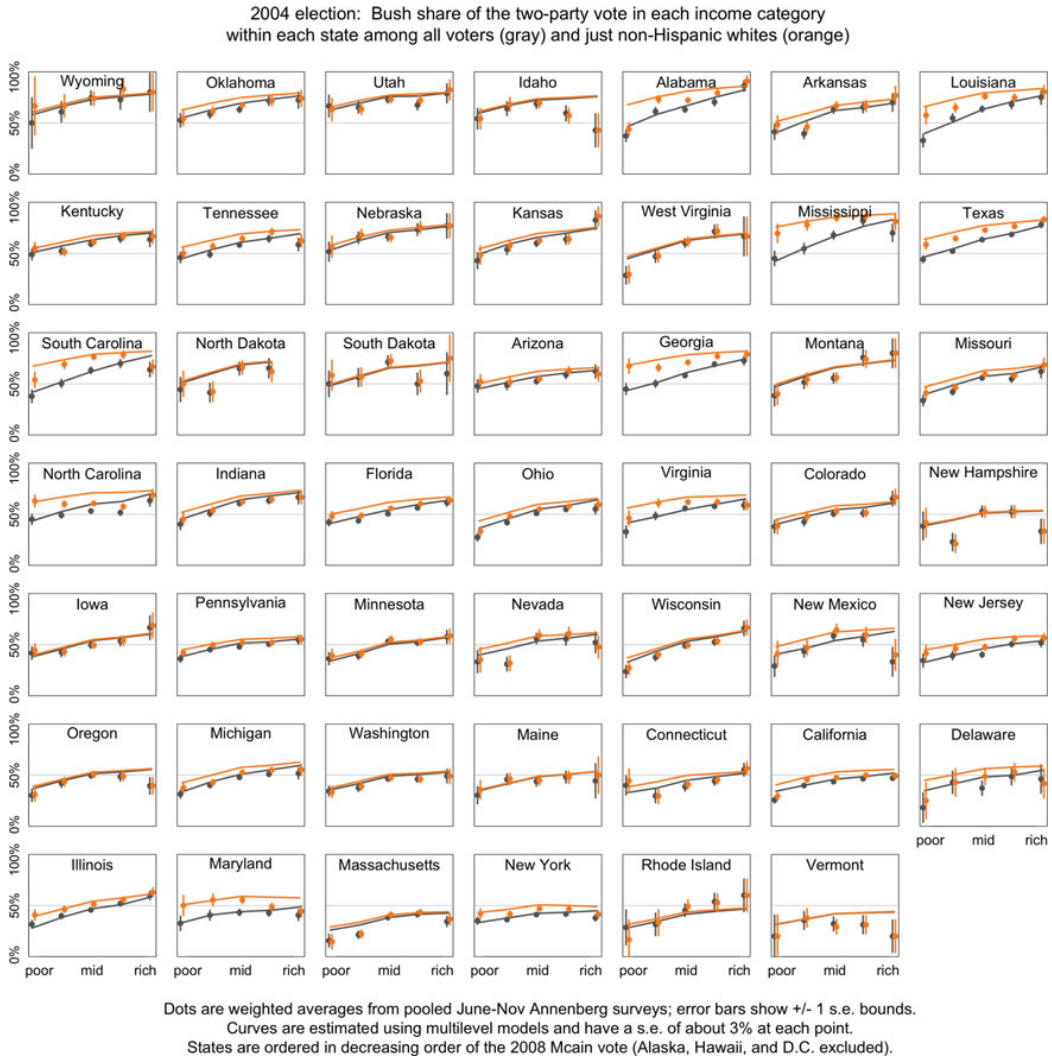

Figure 2. Replication of Stan analysis using 2004 Annenberg survey, showing Bayesian estimates of George W. Bush's share of the two-party vote. Compare to Figure 1, which shows the corresponding estimated Republican votes for 2008.

But what sort of artifact? The sample size from Annenberg 2004 is 43,970, whereas Pew 2008 is based on only 19,170 respondents. Based on this difference in sample size, we would expect the Pew 2008 analysis to yield the smoother graphs: with its smaller sample, we would expect more pooling toward the logistic regression model, hence less jumpy curves. Actually, though, as we see in Figures 1 and 2, the curves from Pew (smaller sample) are jumpier than those from Annenberg (larger sample).

What is going on? Again, it is possible that income was a less consistent predictor of vote in 2008 than in 2004, and one could come up with explanations for specific patterns. For example, consider the bump up in McCain support among whites in the second-lowest income category in Massachusetts and New York (see the bottom row of Figure 2). Perhaps this can be understood as a disinclination of some voters in this group toward voting for an African American. But it also seems plausible that many of the discordant patterns in Figure 2 simply represent noise—nonsampling error—that has not been accounted for the model. The Pew and Annenberg surveys were conducted in different ways, and it is conceivable that Annenberg, which was entirely focused on the election campaign, could have more consistent responses during the months of data collection, and a more representative sample than the Pew polls, which were designed with multiple purposes.

In any case, this discrepancy gives us another reason to perform a preregistered replication. We will be able to compare different survey organizations using data from the same time period.

3. Preregistering the New Replications

In preparation for the preregistered replications, we prepared the following files:

1. An R script to process the new data from the 2008 Annenberg surveys and also to load in the other information—Census summaries and state-by-state election results—needed for the multilevel regression and poststratification;
2. A Stan program to fit the multilevel model;

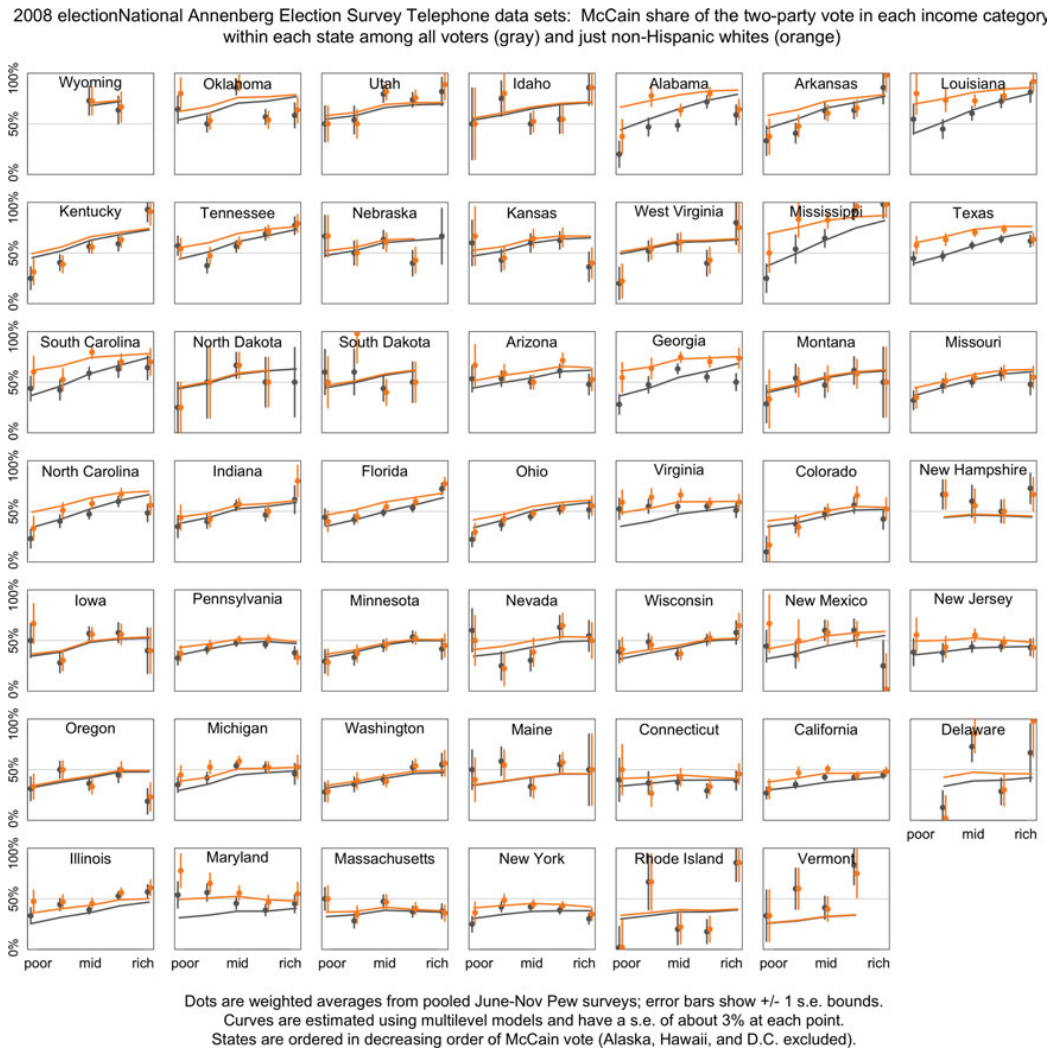

Figure 3. Preregistered replication on the 2008 Annenberg telephone survey. Compare to Figure 1 that performed the identical analysis on the 2008 Pew survey. The overall results are similar, but the lines in this graph are generally smoother, as a result of the Annenberg data being less noisy than the Pew data. We do not understand why the numbers from the surveys differ in this way.

3. An R script to run the Stan program;
4. An R script to produce the equivalent of Figure 1.

We plan to time stamp the present article and then perform the planned replications. Having done so, we will compare the resulting graphs to Figure 1, which shows the estimates from the Pew data.

4. Discussion

4.1. Value of a Preregistered Replication Plan

In many of the most publicized examples of replication, the goal is to confirm or debunk some controversial existing research claim. This present example is a bit different in that we began this study with no particular concern about the Ghitza and Gelman (2013) results, but a bit of replication would, we believe, give us a better sense of uncertainty about the details. In addition, one can always be worried about opportunistic interpretations of statistical results.

The main point of the present article is to demonstrate that preregistered replication can be done, and it can be useful, even in a setting where the subject matter is historical (and thus the replication data exist before the preregistration plan is written) and where the analysis is exploratory (and thus a replication cannot be simply deemed successful or unsuccessful based on the statistical significance of some particular comparison).

Laying out the details of this replication were instructive. In practice it is not easy to replicate an existing analysis from several years ago. Even in a case such as this where the data and code are accessible and so all the results can be reproduced, it can be a challenge to alter the analysis. In this case, a series of adaptations were required to move to a fully Bayesian analysis. In addition, it took some effort to prepare the Annenberg datasets for our new analysis. And when we did the 2004 replication, we found interesting differences, which suggested that there may indeed be problems with our published results (see Section 2).

Our replication plan is preregistered and time-stamped but is not intended to be of forensic quality. For example, there was

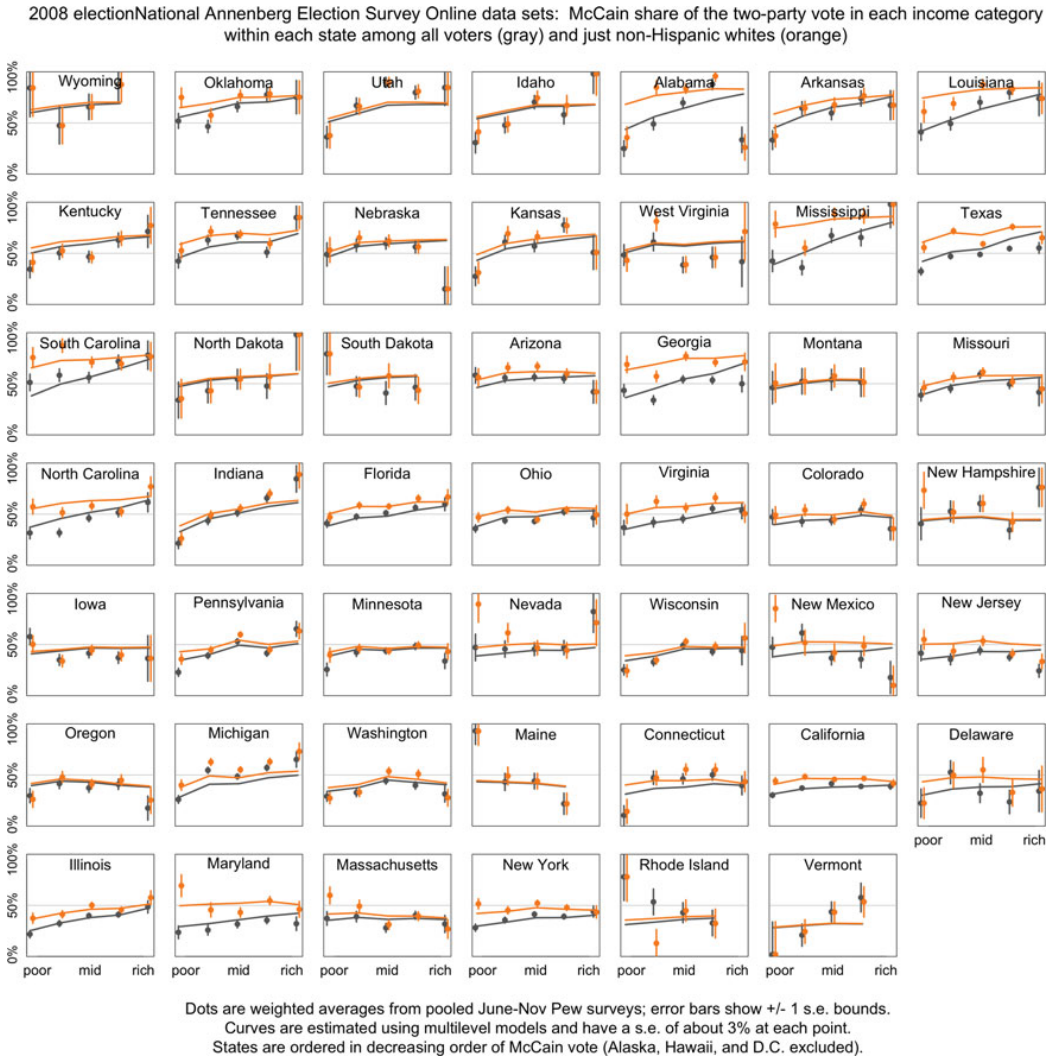

Figure 4. Preregistered replication on the 2008 Annenberg internet survey. Results are similar to those from the Annenberg telephone survey shown in Figure 3.

nothing to stop us from secretly performing various analyses on the Annenberg data and using them to decide on the details of our purportedly preregistered design. We did indeed have to crack open the new data to make coding decisions, and we can only offer our word that we did not make these choices based on outcomes.

Our next step in this project is to analyze the Annenberg datasets from 2008 using our preregistered replication plan and compare to our earlier published graphs. We plan to report these results in a follow-up article.

4.2. How to Interpret the Replication Results?

If all goes according to plan, performing the replications should take just a few minutes as they should merely involve running existing code on the two new, cleaned Annenberg 2008 preelection polls, resulting in inferences and graphs for each dataset. We will end up with two replications of Figure 2 from this article (a figure that is itself essentially a duplication of Figure 2 from Ghitza and Gelman 2013), showing data and estimated McCain vote share among whites and among all voters, for each of five income categories within each of the 48 contiguous states.

At that point, two challenges will arise. The first is that, yes, the two new surveys represent independent data collections, but that all three surveys share biases: they are all preelection polls, and they all omit people who do not respond to surveys. Our replications will be valid *as replications* but it is possible that any patterns confirmed by the replications are still just artifacts driven by nonsampling error. Thus, we cannot automatically interpret patterns in our data, even if confirmed by replications, as representing truths about voters: such a conclusion requires an extra-data assumption about representativeness of the surveys.

The other issue is that we have no predetermined criteria for "success" or "failure" of the replication. Our plan is to remake and compare it to what we got from the earlier survey, and many comparisons can be made. For example, voters in the second-lowest income category in several states (notably New York, Massachusetts, and California) appear to be noticeably more McCain-voting than those in the first and third income categories. Is this something we should believe—it does, after all, appear to be reflected in the Bayesian estimates as well as in the raw data—or is it some artifact of the sample? This is the kind of pattern we would like to check in a replication. Confirmation of

this pattern with the new surveys would not only increase our belief that this particular pattern from 2008 is real; it would also give us greater confidence in our inferential process. Conversely, if the pattern does *not* replicate, we would be inclined to feel that Figure 2 overstates our certainty about the vote.

As the aforementioned example indicates, we plan our inspection of the replications to be somewhat open-ended, and we recognize that researcher degrees of freedom and forking paths will arise in our analysis. Depending on what we see, we might well follow up with a larger Bayesian analysis including data from all three surveys, perhaps with survey-level error terms to capture differences between the samples beyond what could be explained by random sampling from a common population. Another option would be to include the date of interview in the model, thus allowing for public-opinion shifts during the months in which these surveys were in the field.

4.3. Presenting the Replication Results

Once the present article (excluding Section 5) has been accepted for publication by *Statistics and Public Policy*, we will publicly post it online on the date shown on the first page. We will then run the replication, produce the two graphs as planned, and put those graphs and our discussion in Section 5, a section that will be blank in the initially posted version of the article. *Statistics and Public Policy* will then publish the entire article.

5. Results from the Preregistered Analysis

The article up to this point was accepted by the journal and then we performed the planned replication. This section and the last paragraph of the abstract were added after that; the only other change we made in the article was to rewrite one of the figure captions for clarity.

The preregistrated replication worked but was not quite so simple as we envisioned. As planned, we took the code described earlier in this article and ran it on the 2008 Annenberg telephone and online samples. In the replication, one step was required that we had not entered in our preregistration plan. There was, of course, missing data in the Annenberg survey. When the outcome variable, vote preference, was missing, we simply excluded those respondents from the analysis. But when predictors in the model were missing, we followed standard statistical practice and multiply imputed the missing values (in this case using our `mi` package in R). The difficulty is that in our earlier analysis we had used a single one of the multiple imputed datasets to avoid the trouble of keeping track of multiple analyses. So in our replication we did the same thing. We used the first of the imputed datasets, which since the imputations were created randomly, was equivalent to choosing one of the imputations at random. This is what we had done with the Pew 2008 data in our earlier article but we had not remembered to include this as part of the replication instructions.

Figures 3 and 4 show the results of our replication analysis. Compared to Figure 1 that came from our earlier analysis, these new graphs are similar in their broad outlines but show some systematic differences. In particular, the patterns look a bit smoother within each state. Compare, for example, Massachusetts and New York (in the middle of the bottom row of plots), whose lines are much jumpier in Figure 1 than in the new Figures 3 and 4. The inferences from Annenberg phone and online surveys are gratifyingly similar, considering that they were estimated independently from different datasets.

Why are the estimates from the 2008 Pew and Annenberg surveys so different? The sample sizes are of the same order of magnitude (19,170 for Pew, 10,671 for Annenberg phone, 24,423 for Annenberg internet) and none of these surveys were conducted using cluster sampling (which could lower the effective sample size), so we suppose the difference must come from data collection or processing. The Pew sample comes with weights, which we use in the analysis within each poststratification cell, whereas the Annenberg samples are self-weighting. There may be a problem in how we handle the weighting in our analysis, if the Pew estimates are being pulled too closely to the raw data (the dots in Figure 1). But we reran the Pew analysis removing the weights and obtained a similar picture with jagged lines for states such as Massachusetts and New York.

At this point we remain unsure. Perhaps one way to pursue this would be to more fully poststratify, modeling the Pew data conditional on all the variables used in the weighting and then poststratifying on the resulting larger number of cells. Another starting point might be to reanalyze the Pew data ignoring the weights, just so see what the resulting inference looks like. Yet another tack would be to focus on a single large state and try to understand why the raw numbers from Pew (for example, the dots in the New York plot in Figure 1) are so jumpy.

In any case, this exercise demonstrates what can be learned from a replication analysis of survey data. Even in constructing the preregistration plan, we clarified some issues in the original analysis, and then the replication confirmed a suspicion we had, to not trust the jumpy estimates of vote choice as a function of income. At this point it is not clear what to trust, and further data analysis is needed to track down the systematic differences between original and replication analysis. Unlike many studies in the experimental-replication literature, in this case the discrepancies cannot be attributable to mere chance variation or sampling error, as the sample sizes are just too large for that. This process thus shows how replication can raise concerns about an earlier published result; it also demonstrates, as with Nosek, Spies, and Motyl (2012), that this process can be valuable even in a nonadversarial context in which researchers are attempting to replicate their own work.

Acknowledgments

The authors thank Catalist, the National Science Foundation, and the Office of Naval Research for partial support of this work, Macartan Humphreys for helpful comments, and Richard McElreath, Rob Trangucci, Jonah Gabry, Bob Carpenter, Ben Goodrich, and Daniel Lee for preparing the R and Stan functions that allowed them to easily fit their models. Further details on their replication procedures appear in their online appendix, https://github.com/rayleigh/election_stan_analysis/blob/master/appendix.pdf, and all their code can be found at https://github.com/rayleigh/election_stan_analysis.

Supplementary Material

Supplemental data for this article can be accessed on the publisher's website.

References

Bargh, J. A., Chen, M., and Burrows, L. (1996), "Automaticity of Social Behavior: Direct Effects of Trait Construct and Stereotype-Activation on Action," *Journal of Personality and Social Psychology*, 71, 230–244. [1]

Bartels, L. M. (2008), *Unequal Democracy: The Political Economy of the New Gilded Age*, Princeton, NJ: Princeton University Press. [2]

Campbell, J. E. (2011), "The Economic Records of the Presidents: Party Differences and Inherited Economic Conditions," *The Forum*, 9, 1–29. [2]

Comiskey, M., and Marsh, L. C. (2012), "Presidents, Parties, and the Business Cycle, 1949–2009," *Presidential Studies Quarterly*, 42, 40–59. [2]

Doyen, S., Klein, O., Pichon, C. L., and Cleeremans, A. (2012), "Behavioral Priming: It's All in the Mind, But Whose Mind?" *PLoS ONE*, 7, e29081. [1]

Gelman, A. (2013), "Preregistration of Studies and Mock Reports," *Political Analysis*, 21, 40–41. [2]

——— (2014, May 27), "A Whole Fleet of Gremlins: Looking More Carefully at Richard Tol's Twice-Corrected Paper, The Economic Effects of Climate Change," *Statistical Modeling, Causal Inference, and Social Science blog*, Available at *http://andrewgelman.com/2014/05/27/whole-fleet-gremlins-looking-carefully-richard-tols-twice-corrected-paper-economic-effects-climate-change/* [2]

——— (2015, May 20), "Fake Study on Changing Attitudes: Sometimes a Claim That is too Good to be True, isn't," *Monkey Cage blog*, Available at *http://www.washingtonpost.com/blogs/monkey-cage/wp/2015/05/20/fake-study-on-changing-attitudes-sometimes-a-claim-that-is-too-good-to-be-true-isnt/* [2]

Gelman, A., and Carlin, J. (2014), "Beyond Power Calculations: Assessing Type S (Sign) and Type M (Magnitude) Errors," *Perspectives on Psychological Science*, 9, 641–651. [1]

Gelman, A., and Hill, J. (2007), *Data Analysis Using Regression and Multilevel/HIerarchical Models*, Cambridge: Cambridge University Press. [2]

Gelman, A., and Loken, E. (2014), "The Statistical Crisis in Science," *American Scientist*, 102, 460–465. [1]

Ghitza, Y., and Gelman, A. (2013), "Deep Interactions With MRP: Election Turnout and Voting Patterns Among Small Electoral Subgroups," *American Journal of Political Science*, 57, 762–776. [2,3,5,6]

Gonzales, J. E., and Cunningham, C. A. (2015), "The Promise of Pre-Registration in Psychological Research," *Psychological Science Agenda*. Available at *http://www.apa.org/science/about/psa/2015/08/pre-registration.aspx* [1]

Humphreys, M., de la Sierra, R. S., and van der Windt, P. (2013), "Fishing, Commitment, and Communication: A Proposal for Comprehensive Nonbinding Research Registration," *Political Analysis*, 21, 1–20. [2]

LaCour, M. J., and Green, D. P. (2014), "When Contact Changes Minds: An Experiment on Transmission of Support for Gay Equality," *Science*, 346, 1366–1369. [2]

Nosek, B. A., Spies, J. R., and Motyl, M. (2012), "Scientific Utopia II. Restructuring Incentives and Practices to Promote Truth Over Publishability," *Perspectives on Psychological Science*, 7, 615–631. [1,7]

Reinhart, C. M., and Rogoff, K. S. (2010), "Growth in a Time of Debt," *American Economic Review: Papers & Proceedings*, 100, 573–578. [2]

Simmons, J., Nelson, L., and Simonsohn, U. (2011), "False-Positive Psychology: Undisclosed Flexibility in Data Collection and Analysis Allow Presenting Anything as Significant," *Psychological Science*, 22, 1359–1366. [1]

Stan Development Team (2015), *Stan Modeling Language: User's Guide and Reference Manual* (2.7 ed.), [Computer software manual]. Available at *http://mc-stan.org* [3]

Tol, R. (2009), "The Economic Effects of Climate Change," *Journal of Economic Perspectives*, 23, 29–51. [2]

Wagenmakers, E. J., Wetzels, R., Borsboom, D., Kievit, R., and van der Maas, H. L. J. (2015), "A Skeptical Eye on psi," in *Extrasensory Perception: Support, Skepticism, and Science*, eds. Edwin C. May and Sonali Bhatt Marwaha, Westport, CT: Praeger, 153–176. [1]