Reducing Political Bias in Political Science Estimates

L. J. Zigerell, *Illinois State University*

ABSTRACT Political science researchers have flexibility in how to analyze data, how to report data, and whether to report on data. A review of examples of reporting flexibility from the race and sex discrimination literature illustrates how research design choices can influence estimates and inferences. This reporting flexibility—coupled with the political imbalance among political scientists—creates the potential for political bias in reported political science estimates. These biases can be reduced or eliminated through preregistration and preacceptance, with researchers committing to a research design before completing data collection. Removing the potential for reporting flexibility can raise the credibility of political science research.

Political science is an important tool for understanding and improving the world (see Sides 2013); but several recent events have raised questions about political science research and social science more generally. The LaCour and Green (2014) *Science* article was retracted after Broockman, Kalla, and Aronow (2015) reported irregularities in the raw data for that article (Singal 2015). Moreover, results from a project replicating a large number of psychology studies reported that the mean estimated effect size of the original studies was twice the mean estimated effect size of the replication studies, with substantially fewer effect size estimates in the replications reaching statistical significance (Open Science Collaboration 2015). The variation between estimates reported in the original studies and estimates reported in the replications likely involves flexibility in data reporting (Simmons, Nelson, and Simonsohn 2011), in which only selected studies are reported or only selected outcome variables and experimental conditions are reported.[1]

Many leading political science journals have adopted requirements that data to reproduce newly published analyses be publicly available (DA-RT 2015). These requirements can plausibly be expected to improve inferences through an increase in the likelihood of uncovering irregularities or errors; however, leading political science journals have not yet adopted policies to improve inferences by eliminating flexibility in data reporting, even though such flexibility permits post hoc research design choices that can produce biased inferences.

In the next section I review examples from the race and sex discrimination literature to illustrate reporting flexibility and how this flexibility can be used to produce unrepresentative estimates that provide more support for the policy preferences of a particular political party or ideology. Subsequent sections identify potential negative consequences of reporting flexibility for political science and discuss mechanisms that can be adopted to eliminate this flexibility.

EXAMPLES OF REPORTING FLEXIBILITY IN THE RACE AND SEX DISCRIMINATION LITERATURE

Reporting flexibility includes which outcome variables to report. For example, the questionnaire for the survey experiment reported on in Rattan et al. (2012) about the effect of a treatment manipulating the race of a juvenile offender contained four items that could be used as outcome variables, but Rattan et al. (2012) mentioned only two of these items, both of which could be used to detect a statistically significant difference between experimental conditions; standardized estimates of antiblack discrimination for these items were 0.17 and 0.19[2]; however, using the same research design and the Rattan et al. (2010) data, neither of the unreported items detected a statistically significant difference between experimental conditions, with standardized estimates of –0.01 and 0.09. Such selective reporting thus created an overestimate of antiblack discrimination detected in the experiment.

Reporting flexibility includes coding of outcome variables. The survey experiment reported on in Banks (2014) measured attitudes about President Barack Obama and the Democrats' health care reform proposal with multiple items: one item measured approval for the way that Obama is handling health care, on a scale from 1 (*approve strongly*) to 5 (*disapprove strongly*); another set of items measured support for the health care bill with a dichotomous favor-or-oppose item and with a follow-up item

L. J. Zigerell *is an assistant professor of politics and government at Illinois State University. He can be reached at ljzigerell@ilstu.edu.*

doi:10.1017/S1049096516002389

about the strength of favoring or opposing. Banks (2014) reported results for only the dichotomous favor-or-oppose item, which permitted an inference that "anger uniquely increases the impact of racial attitudes on health care opinions" (p. 508). However, based on the Banks (2010) data, the inference about the unique influence of anger is not supported when predicting the approval item or when predicting a favor-or-oppose item coded to include the follow-up measure of strength of favoring or opposing (see Zigerell 2016).

among black respondents: blacks rated black targets 0.61 standard deviations higher than white targets (two-tailed p value of 0.018). Moreover, straightforward analysis of the data permitted an inference of sex discrimination that favored women: across all respondents, female targets were rated 0.22 standard deviations higher than male targets (two-tailed p value less than 0.001). The absence of this study from the literature produces an underestimation of detected discrimination against whites and men.

These requirements can plausibly be expected to improve inferences through an increase in the likelihood of uncovering irregularities or errors; however, leading political science journals have not yet adopted policies to improve inferences by eliminating flexibility in data reporting, even though such flexibility permits post hoc research design choices that can produce biased inferences.

Reporting flexibility for cross-time data includes selection of endpoints for the analysis. Maliniak, Powers, and Walter (2013) reported that female-authored international relations articles have received fewer citations than equivalent male-authored international relations articles, based on an analysis restricted to selected articles published between 1980 and 2006; for this selection, the two-tailed p value from the model with the full set of control variables was 0.093 for the citation difference between female-authored articles and male-authored articles. However, data on articles from 2007 were available in the dataset, and reproduction code for the article indicated that models for the main table contained a control for articles published in 2007. When articles from 2007 were included in the analysis, the two-tailed p value inflated to 0.223 for the citation difference between female-authored articles and male-authored articles (Zigerell 2015b).[3]

Previously discussed examples are consistent with a preference for statistically significant results, but reporting flexibility also permits statistically significant results to be unreported. Discussing an Associated Press report of an implicit association test estimating that 56% of Americans had antiblack sentiments (Agiesta and Ross 2012), Moore (2012) disclosed that the report failed to mention that the same implicit association test also produced an estimate of 33% of Americans with antiwhite sentiments. This criticism about a failure to report estimated discrimination against whites when reporting estimated discrimination against blacks can be extended to the manuscript on which the Associated Press report was based (Pasek, Krosnick, and Tompson 2012) and to other political science publications (see Zigerell 2015a).

The Benard (2005) survey experiment is an example of an entire study whose results have not yet been reported on in the literature, even though the study permitted detection of race and sex discrimination. The survey experiment tested for race and sex discrimination in ratings of target workers whose names signaled race and sex: Brad, Kareem, Kristen, and Tamika. Straightforward analysis of the data did not permit detection of racial discrimination in overall ratings among white respondents: whites rated black targets 0.09 standard deviations higher than white targets (two-tailed p value of 0.204). However, straightforward analysis of the data did permit detection of racial discrimination

Certain fields have relatively unique types of reporting flexibility, such as the measurement of racial discrimination. Researchers testing for racial discrimination using the 2012 American National Election Studies (ANES) Time Series study can measure antiblack discrimination with symbolic racism items, feeling thermometer ratings for blacks and/or for whites, stereotype scales for blacks and/or whites being hardworking and/or intelligent, an item about whether blacks have too much influence in American politics, and an item about how much discrimination blacks face in the United States today. Researchers failing to find racial discrimination in the ANES dataset can then check racial discrimination items in the General Social Survey or other surveys.

Reporting flexibility in the use of multiple measures of racial discrimination can be illustrated with the survey experiment reported on in Krupnikov and Piston (2015), which included items measuring blacks' and whites' perceived intelligence, laziness, and amount of influence in American politics. Results predicting a difference in candidate thermometer ratings were reported in that article with the "black influence" measure used as the measure of racial attitudes in model 1 of table 2 of that article, which permitted detection of racial discrimination at a conventional level of statistical significance in a triple interaction of racial attitudes, black ad sponsor, and negative ad. But reanalysis of the Krupnikov (2012) data indicated that use of the stereotype items as the measure of racial attitudes did not permit detection of racial discrimination in that triple interaction term at a conventional level of statistical significance.

Another area of relatively unique reporting flexibility in the study of racial discrimination is sample restriction. Researchers testing for antiblack discrimination in a survey can analyze data for all respondents or can restrict the sample to groups such as nonblack respondents, white respondents, or non-Hispanic white respondents; it is also possible to disaggregate samples by sex (e.g., Hutchings, Walton, and Benjamin 2010) and to mix sample restrictions within a study. Baker (2015) used a sample of non-Hispanic whites for one survey experiment and a sample of nonblacks for another survey experiment reported on in the same article. Using the Baker (2012) data and code posted for the analysis with the nonblack sample, restricting the sample to white

respondents raised the two-tailed *p* value from 0.080 to 0.285 for the test of whether respondents supported more aid to persons in Guyana or Armenia. In this case, the choice to use a nonblack sample or a white sample determined whether an inference of discrimination in favor of Africans was supported.[4]

THE POTENTIAL FOR POLITICAL BIAS IN POLITICAL SCIENCE ESTIMATES, AND HOW TO REDUCE IT

The aforementioned studies permitted flexibility to report a range of estimates for race or sex discrimination or to not report an estimate at all. Reporting choices could have been influenced by many factors, such as sincere researcher theoretical preference for one research design over another or requests from journal peer reviewers or journal editors to report certain results or to analyze the data in certain ways. The reasons for these reporting choices are less important than the presence of these choices and the fact that political science currently has no mechanism to guarantee that reported results represent the full range of estimates from reasonable research designs. This reporting flexibility might not be much of a problem if selective reporting of estimates canceled out in the aggregate, but it is unreasonable to assume such balancing, given the political imbalance in political science, with surveys indicating that Democrats outnumber Republicans by a ratio of 5.6:1 (Klein and Stern 2005) or 7.25:1 (Rothman, Lichter, and Nevitte 2005) and liberals outnumbering conservatives 81:2 (Rothman et al. 2005).

Given the combination of reporting flexibility and the substantial political imbalance within political science, it would not be unreasonable to suspect that political science estimates have been and are biased toward the preferred inferences of the political Left.

Given the combination of reporting flexibility and the substantial political imbalance within political science, it would not be unreasonable to suspect that political science estimates have been and are biased toward the preferred inferences of the political Left. Duarte et al. (2015) noted that greater ideological diversity in social psychology can produce benefits for the field, such as combatting confirmation bias. Greater ideological diversity in the political science peer reviewer pool might help reduce reporting flexibility if peer reviewers skeptical of a claim require reporting of additional models or require declarations about the presence of unreported analyses.

But more straightforward—and easier to accomplish—mechanisms for eliminating reporting flexibility are preregistration (see Monogan 2015) and preacceptance (see Nyhan 2015). In a preregistered study, researchers announce their planned research design choices before completing data collection, and then report results from this preannounced research design; readers can thus have more confidence that reported results have not been influenced by reporting bias. Preregistration has an advantage that researchers can announce planned research design choices unilaterally, but a shortcoming of preregistration is that there is no guarantee that preregistered studies will enter the literature. This shortcoming can be eliminated through preacceptance of articles in which journals agree to publish an article based on a preregistered research design no matter what the results are, as was done in a special issue of *Comparative Political Studies* (Findley et al. 2016).[5]

Monogan (2013) is an exemplar of a preregistered study. The preregistered Monogan (2010) file describes the planned research design to assess whether a Republican House candidate's percentage of the major-party vote in 2010 was influenced by the tone of the candidate's website statement on immigration. The preregistered file identified the planned outcome variable, explained how the explanatory variable would be measured, identified which control variables would be included in the model and how these control variables would be measured, and explained how the data would be analyzed. Analysis of the vote data collected after the research design had been made public did not provide support for the expectation that Republican candidates' stance on immigration influenced constituent vote choice.[6]

CONCLUSION

Political science has made improvements to increase transparency, such as leading journals adopting requirements regarding the public posting of data. Providing other researchers access to data can lead to discovery of irregularities (Broockman et al. 2015), correction of errors (Herndon, Ash, and Pollin 2014), and debate about how best to interpret the data (Bashir 2015). This can foster the production of improved inferences through reanalysis of the data to find irregularities, errors, or imperfections, and by encouraging more honest and representative reporting in the original studies if researchers and others are aware that the data will eventually be public and subject to reanalysis.

However, public posting of data is not sufficient for eliminating selective reporting of results. Some political science studies have unreported outcome variables and unreported experimental conditions even when the data were planned to be publicly released (Franco et al. 2015); moreover, many of the political science survey experiments in the archives of the Time-sharing Experiments for the Social Sciences program have not been reported on in the academic literature, suggesting that political science has a file drawer of unreported studies (see the data from Franco et al. 2014). Thus, publicly posting data is insufficient for fostering a representative reporting of results and is inefficient for eliminating selective reporting if other researchers must spend time checking whether reported results are representative of the range of inferences that could be produced by reasonable research designs.

Preregistration and preacceptance are better mechanisms for eliminating reporting flexibility. Data can often produce a range of estimates, and the full extent of this range is not always reported, as indicated by the aforementioned examples from the race and sex discrimination literature. Preregistration of research designs reduces this range of estimates to only those estimates derived from analyses announced in advance, and preacceptance guarantees that these estimates enter the literature.[7] Readers of

https://doi.org/10.1017/S1049096516002389 Published online by Cambridge University Press

preregistered studies can thus have confidence that the reported results reflect a researcher's preferred research design and not the researcher's preferred inferences.[8]

Public policy should be informed by data, and political scientists have an important role in collecting, analyzing, and reporting on these data. But public officials and the public cannot be expected to update their preferences to reflect political science research if results from political science studies can be selectively reported. Many political science estimates cannot be checked against real-world outcomes, as the estimates of election forecasters can; therefore, much of the deference that political science research receives reflects only a reader's trust in the research process and the reputations of the researchers and publishing journals.

Perhaps critics of political science can be convinced that an overwhelmingly liberal set of researchers can be trusted to not exploit flexibility in data reporting, but a better solution is to eliminate this reporting flexibility when possible so that political science estimates can be accurately represented as untainted by post hoc research design choices made to favor one set of policy preferences over another. In a time of risk that federal funding for political science will be cut or eliminated, political scientists should adopt mechanisms to help ensure that the estimates that the field produces have higher credibility (see Nyhan 2015, 81) and are thus more deserving of funding. Preregistration and preacceptance can be used to remove reporting flexibility and make political science research more credible.

ACKNOWLEDGMENTS

Thanks to Andy Baker, Antoine Banks, Stephen Benard, Daniel Maliniak, Spencer Piston, Ryan Powers, and Aneeta Rattan for providing data and/or information about their study discussed in the manuscript. Thanks to the anonymous reviewers and the *PS* editors. Data from the Time-sharing Experiments for the Social Sciences (TESS) program described in this manuscript were collected from the inception of TESS to August 31, 2008, under NSF Grant 0094964 (Diana C. Mutz and Arthur Lupia, Principal Investigators), and from September 1, 2008, to August 31, 2012, under NSF Grant 0818839 (Jeremy Freese and Penny Visser, Principal Investigators). ■

NOTES

1. For evidence of selective reporting in social science, see Franco, Malhotra, and Simonovits (2014), Franco et al. (2015), and Franco et al. (2016).
2. These effect size estimates differ slightly from the estimates in Rattan et al. (2012) due to the Rattan et al. (2012) analysis using frequency weights in the base module of SPSS and not probability weights.
3. The reproduction datasaet for Maliniak, Powers, and Walter (2013) appeared to have a sample of articles from 2007 that would permit use of articles from 2007 in the analysis: the dataset contained 124 articles from 2005, 133 articles from 2006, and 146 articles from 2007, all with a value of 1 for the in_analysis variable.
4. Inclusion of nonwhite respondents in the nonblack sample does not alter the key inference from Baker (2015) about the interaction of paternalism and race of the recipients: the effect size and two-tailed p value for the interaction term are 0.26 and p=0.005 for the nonblack sample and 0.25 and p=0.014 for the white sample.
5. See http://www.ipdutexas.org/cps-transparency-special-issue.html
6. Null results for preregistered studies are not uncommon. According to Kaplan and Irvin (2015), after the FDA Modernization Act of 1997 required preregistration of research designs for certain clinical trials, positive results for these clinical trials declined from 57% to 8%.
7. Reports of preregistered and preaccepted studies can include analyses that were not preregistered, but these non-preregistered analyses should be identified in the report as non-preregistered.
8. It is possible that reports of preregistered and preaccepted studies fail to reflect the preregistered research design, but in such cases the preregistration file is available for comparison so that inconsistencies can be detected.

REFERENCES

Agiesta, Jennifer and Sonya Ross. 2012. "AP Poll: Majority Harbor Prejudice against Blacks." Associated Press, October 27. http://news.yahoo.com/ap-poll-majority-harbor-prejudice-against-blacks-073551680–election.html.

Baker, Andy. 2012. Replication data for: "Race, Paternalism, and Foreign Aid: Evidence from US Public Opinion." http://spot.colorado.edu/~bakerab/data.html.

———. 2015. "Race, Paternalism, and Foreign Aid: Evidence from US Public Opinion." *American Political Science Review* 109 (1): 93–109.

Banks, Antoine J. 2010. Replication data for: "The Public's Anger: White Racial Attitudes and Opinions toward Health Care Reform." http://www.tessexperiments.org/data/banks019.html.

———. 2014. "The Public's Anger: White Racial Attitudes and Opinions toward Health Care Reform." *Political Behavior* 36 (3): 493–514.

Bashir, Omar S. 2015. "Testing Inferences about American Politics: A Review of the 'Oligarchy' Result." *Research & Politics* 2 (4). doi: 10.1177/2053168015608896.

Benard, Stephen W. 2005. Reproduction data for: "Statistical Discrimination, Stereotyping, and Evaluations of Worker Productivity." http://www.tessexperiments.org/data/benard241.html.

Broockman, David, Joshua Kalla, and Peter Aronow. 2015. "Irregularities in LaCour (2014)." http://web.stanford.edu/~dbroock/broockman_kalla_aronow_lg_irregularities.pdf.

DA-RT. 2015. "The Journal Editors' Transparency Statement (JETS)." http://www.dartstatement.org/#!blank/c22sl.

Duarte, José L., Jarret T. Crawford, Charlotta Stern, Jonathan Haidt, Lee Jussim, and Philip E. Tetlock. 2015. "Political Diversity Will Improve Social Psychological Science." *Behavioral and Brain Sciences* 38: e130.

Findley, Michael G., Nathan M. Jensen, Edmund J. Malesky, and Thomas B. Pepinsky. 2016. "Can Results-Free Review Reduce Publication Bias? The Results and Implications of a Pilot Study." *Comparative Political Studies* doi: 10.1177/0010414016655539.

Franco, Annie, Neil Malhotra, and Gabor Simonovits. 2014. "Publication Bias in the Social Sciences: Unlocking the File Drawer." *Science* 345 (6203): 1502–05.

———. 2015. "Underreporting in Political Science Survey Experiments: Comparing Questionnaires to Published Results." *Political Analysis* 23 (2): 306–12.

———. 2016. "Underreporting in Psychology Experiments: Evidence from a Study Registry." *Social Psychological and Personality Science* 7 (1): 8–12.

Herndon, Thomas, Michael Ash, and Robert Pollin. 2014. "Does High Public Debt Consistently Stifle Economic Growth? A Critique of Reinhart and Rogoff." *Cambridge Journal of Economics* 38 (2): 257–79.

Hutchings, Vincent L., Hanes Walton, and Andrea Benjamin. 2010. "The Impact of Explicit Racial Cues on Gender Differences in Support for Confederate Symbols and Partisanship." *Journal of Politics* 72 (4): 1175–88.

Kaplan, Robert M. and Veronica L. Irvin. 2015. "Likelihood of Null Effects of Large NHLBI Clinical Trials Has Increased Over Time." *PloS ONE* 10 (8): e0132382.

Klein, Daniel B. and Charlotta Stern. 2005. "Professors and Their Politics: The Policy Views of Social Scientists." *Critical Review* 17 (3–4): 257–303.

Krupnikov, Yanna. 2012. Replication data for: "Accentuating the Negative: Candidate Race and Campaign Strategy." http://www.tessexperiments.org/data/krupnikov245.html.

Krupnikov, Yanna and Spencer Piston. 2015. "Accentuating the Negative: Candidate Race and Campaign Strategy." *Political Communication* 32 (1): 152–73.

LaCour, Michael J. and Donald P. Green. 2014. "When Contact Changes Minds: An Experiment on Transmission of Support for Gay Equality." *Science* 346 (6215): 1366–1369.

Maliniak, Daniel, Ryan Powers, and Barbara F. Walter. 2013. "The Gender Citation Gap in International Relations." *International Organization* 67 (4): 889–922.

Monogan, James E., III. 2010. "The Immigration Issue and the 2010 House Elections: A Research Design." Political Science Registered Studies Dataverse at the University of Georgia, November 1. http://hdl.handle.net/1902.1/16470.

———. 2013. "A Case for Registering Studies of Political Outcomes: An Application in the 2010 House Elections." *Political Analysis* 21 (1): 21–37.

———. 2015. "Research Preregistration in Political Science: The Case, Counterarguments, and a Response to Critiques." *PS: Political Science & Politics* 48 (3): 425–29.

Moore, David W. 2012. "What the AP Poll on Racial Attitudes Really Tells Us, Part 1." iMediaEthics, Nov. 2. http://www.imediaethics.org/News/3555/What_the_ap_poll_on_racial_attitudes_really_tells_us__part_1_.php.

Nyhan, Brendan. 2015. "Increasing the Credibility of Political Science Research: A Proposal for Journal Reforms." *PS: Political Science & Politics* 48 (SI): 78–83.

Open Science Collaboration. 2015. "Estimating the Reproducibility of Psychological Science." *Science* 349 (6251): aac4716.

Pasek, Josh, Jon A. Krosnick, and Trevor Tompson. 2012. "The Impact of Anti-Black Racism on Approval of Barack Obama's Job Performance and on Voting in the 2012 Presidential Election." Unpublished report. http://joshpasek.com/wp-content/uploads/2012/10/2012-Voting-and-Racism.pdf.

Rattan, Aneeta, Cynthia S. Levine, Carol S. Dweck, and Jennifer L. Eberhardt. 2010. Replication data for: "Race and the Fragility of the Legal Distinction between Juveniles and Adults." http://www.tessexperiments.org/data/eberhardt033.html.

———. 2012. "Race and the Fragility of the Legal Distinction between Juveniles and Adults." *PloS One* 7 (5): e36680.

Rothman, Stanley, S. Robert Lichter, and Neil Nevitte. 2005. "Politics and Professional Advancement among College Faculty." *The Forum* 3 (1), article 2.

Sides, John. 2013. "Why Study Social Science." *The Monkey Cage*, February 5. http://themonkeycage.org/2013/02/why-study-social-science.

Simmons, Joseph P., Leif D. Nelson, and Uri Simonsohn. 2011. "False-Positive Psychology Undisclosed Flexibility in Data Collection and Analysis Allows Presenting Anything as Significant." *Psychological Science* 22 (11): 1359–66.

Singal, Jesse. 2015. "The Case of the Amazing Gay-Marriage Data: How a Graduate Student Reluctantly Uncovered a Huge Scientific Fraud." *New York Magazine*, May 29. http://nymag.com/scienceofus/2015/05/how-a-grad-student-uncovered-a-huge-fraud.html.

Zigerell, L. J. 2015a. "Inferential Selection Bias in a Study of Racial Bias: Revisiting 'Working Twice as Hard to Get Half as Far.'" *Research & Politics* 2 (1). doi: 10.1177/2053168015570996.

———. 2015b. "Is the Gender Citation Gap in International Relations Driven by Elite Papers?" *Research & Politics* 2 (2). doi: 10.1177/2053168015585192.

———. 2016. "The Public's Fear: Revisiting the Emotional Influences on White Racial Attitudes about Health Care Reform." Unpublished manuscript.

https://doi.org/10.1017/S1049096516002389 Published online by Cambridge University Press