[23.10.2012-3:35pm] [1-6] Paper: OP-POLA120017
Copyedited by: PSB MANUSCRIPT CATEGORY: COMMENT

Political Analysis (2012) pp. 1–6
doi:10.1093/pan/mps033

Fisheries Management

Q1

David D. Laitin
Department of Political Science, 423 Encina Central, Stanford University, Stanford, CA 94305-6044
e-mail: dlaitin@stanford.edu

Edited by R. Michael Alvarez

Both papers in this volume on which I was asked to comment (James E. Monogan III, "A Case for Registering Studies of Political Outcomes: An Application in the 2010 House Elections"; and Macartan Humphreys, Raul Sanchez de la Sierra, and Peter van der Windt, "Fishing, Commitment, and Communication: A Proposal for Comprehensive Nonbinding Research Registration") advocate registration regimes for our discipline. The recommendations in both are incremental [promoting, as Lindblom (1965) might have said it, the "intelligence of research"][1] and cognizant of the costs in scientific learning from such a regime if rigidly enforced. Moreover, both papers cite studies by Gerber and various co-authors (e.g., Gerber, Green, and Nickerson 2001) demonstrating publication bias in political science, incentivizing researchers to manipulate their regression models until they can show a z-statistic ≥ 1.96, and thereby reaching standard levels of significance. I fully accept that Gerber et al.'s papers have detected a serious flaw in our scientific practices; there is a problem to be solved. The Monogan and Humphreys et al. proposals are therefore worthy of consideration.[2]

Nonetheless, I raise here several issues that hopefully will prevent more zealous proponents of registration to take the articles published herein as having established registration as a practice *sine qua non* for empirical work in political science. I also suggest an alternative approach for the management of fishing practices through incentivizing replications and the publication of null results. Of course, these can be complementary strategies for the improvement of scientific practices, but given the high costs of implementation (see Humphreys et al., section 4), redundancies should be limited.

1 The Analogy with Medical Research

Both papers see procedures developed in medical research (going back to requirements set by the FDA Modernization Act of 1997, with a new penalty regime for noncompliance legislated in 2007) as a model for political science.[3] This analogy assumes that the incentives faced by laboratories putting new treatments to the test are of the same type as we face as political scientists. But let us explore this analogy a bit more carefully. In medical research putting new treatments to the test,

Author's note: Thanks to Douglas Dion, James Fearon, Donald Green, Macartan Humphreys, Nzeera Ketter, Neil Malhotra, James Monogan III, Amanda Robinson, and Uri Simonsohn for their advice on earlier versions of this comment.

[1] As readers of this comment will see, I am less sanguine that incremental changes will be self-correcting in our discipline than Lindblom was for democratic governance.

[2] See Casey, Glennerster, and Miguel (forthcoming) for an exemplary step in setting a new standard for what they call a "pre-analysis plan" (PAP) that involved negotiations with funders, dated registration of specifications of the dependent variables and statistical models, and guidelines for handling deviations. To the extent that peer reviewers will expect similar care by future experimental studies, the practice will evolve without institutional reform.

[3] While there is mention in the two papers under review here of incomplete compliance with registration in the medical world, neither of them gives evidence or a citation showing that the registration regime in medical research has in fact reduced false positives.

[23.10.2012-3:35pm] [1-6] Paper: OP-POLA120017
Copyedited by: PSB MANUSCRIPT CATEGORY: COMMENT

there is a single supplier (say, of a drug) with an enormous financial stake in showing positive results for at least some segment of the population. The eager supplier either performs the necessary tests in-house, or contracts with an outside lab (usually from a prestigious medical school) to perform the stipulated trials. With results collected and analyzed, the FDA has the responsibility for examining whether the research results merit treatment approval. Because the data are proprietary, counter-labs with an interest in disconfirmation (e.g., representing firms with a competing drug) cannot replicate. Nor can basic scientists. And because the FDA is an agent of Congress, and the political process is more attuned to Type 1 errors (in which harmful drugs are approved) than Type 2 errors (in which beneficial drugs are denied), the overall incentive of the FDA is to compromise on innovation to eliminate the possibility of harmful effects.

In political science, the incentives are not precisely the same. First, unlike in medical research where a protocol is largely context independent, in political research the same protocol cannot normally be employed in even slightly different contexts. For lack of phone coverage, you cannot do GOTV robo-calls in Uganda, so the protocol to learn how best to improve voter turnout will not be the same as one conducted in the United States. Micro-adjustments are often needed to reflect local realities, and these local realities often become apparent after the project hits the field. It would be hard for any field experiment in political science to comply with current FDA rules that would not permit rapid adjustments to procedures once the test hits the field.

Second, even if labs are funded by NGOs (who, like Pharma, may have an interest in a particular outcome), lab survival (and the livelihoods of their soft-money post-docs) is rarely dependent on producing favorable results for their NGO. The worry about bias to one's contractor is less strong in political science. Third, political scientists can get equal professional status by showing that a treatment works as the NGO hoped or by showing that a treatment widely believed to be important (in theory or in conventional wisdom) produces null results. Fourth, the costs of Type 1 errors are usually less consequential in domains of political science as compared with the benefits of avoiding Type 2 errors. We can and should value innovation more than does the typical FDA panel. Finally, political science does not have an administrative review board capable of putting thumbs-up or thumbs-down with analysis of proprietary data. Rather, we have an army of graduate students and serious senior policemen (Neal Beck, Gary King, and the late David Freedman) who have received high professional recognition replicating reported results and sometimes undermining others' reputations! Replication is an alternative form of policing not available in the medial context with proprietary data.

All this is not to say that political science research is immune from the same problems that infect medical research (we are not); rather, it is to say that given different structures and incentives, the remedy imposed by the FDA on medical research may not be the optimal one in a different institutional environment. Indeed, neither paper notes that in public health research (linking smoking to cancer; finding the cause of what we now call "bird flu") fishing, supplemented by a vigorous replication regime, is the *modus operandi* with considerable public benefit.[4] The focus of both papers on the FDA obscures the affinity that much empirical work in political science shares with the Centers for Disease Control (the CDC), an issue I will raise in the next section addressing the inductive element in political science research.

2 The Distinction between Deductive and Inductive Research Programs

Both papers see registration as appropriate for research in which there are strong theoretical predictions that employ an empirical strategy to put those predictions to test. As Humphreys et al. write, their proposal "is for studies . . . that claim to be engaging in hypothesis testing." But this is not a precise view of cumulative knowledge in political science. The lay of our land is that we have for any research question a plethora of theories, all pointing to results in somewhat different directions, and research is conducted to decide which among those theories has the stronger

[4]For a recent example involving diabetes research, see http://med.stanford.edu/ism/2012/april/diabetes.html.

evidence, at least within a set of parameter values.[5] Knowledge accumulates as we learn over the course of a set of related studies the conditions under which any of the theories is correct.

Consider the case of democratic onsets, in which country wealth, social equality, divisions within the ruling apparatus in authoritarian regimes, and international facilitating conditions are only a few of many predicting variables. We have learned (from Przeworski, Alvarez, and Cheibub 2000) that in the era since 1945 wealth is not a good predictor of democratic onsets but (from Boix and Stokes 2003) in an earlier period it was. With these two findings in some tension, a recent paper (Boix 2011) provides data showing the international conditions that favor democratic onsets differentiate the two eras. Meanwhile, we remain without consistent empirical support identifying the effects of social equality or elite splits on democracy. Empirical work in this field has not been biased by a commitment to confirming a theory; rather, it is in deciding among a set of theories. Researchers are rewarded for showing any (significant) results; and, in the case of Przeworski, Alvarez, and Cheibub (2000), a study not organized as a test of any hypothesis, but rather an inductive exploration, a null result was their *coup de grace*. Is this research program deductive (testing a variety of theories out there) or inductive (the research community constructing data sets, and then seeking to draw blood from onions)? It is usually a pragmatic combination of both.

Nearly all research in political science, as with the example above, goes back and forth between a set of theories and an organically growing data archive with befuddled researchers trying to make sense of the variance on their dependent variables. Most regressions we report are descriptive, with the p values (implicitly) reflecting measures of strength of partial correlations. They are not, as is often explicitly (and falsely) claimed, tests of hypotheses. These partial correlations, especially if they challenge expectations from our quite loosely formulated theories, serve as a foundation for further theoretical exploration. Given the enormously complex and contingent world of politics, few of our theories have been tight enough to allow for "pure" deductively based tests.

King, Keohane, and Verba (1994) are often portrayed as methodologists seeking to encourage the rigorous testing of well-formulated theories. But they acknowledge (1994: 22) the reality of data exploration. They reasonably advocate, once we have a pattern in our data that we think is right, testing its observable implications with a different set of data. Going back and forth between exploration and testing is our disciplinary mode. Perhaps in this quasi-deductive and quasi-inductive research environment what is needed more than bias-reduction mechanisms for single investigations is incentives for meta-analyses that draw out from a plethora of quasi-comparable investigations (observational and experimental)[6] the relationships that hold and the parameter values that support these relationships. For this to work, of course, we need an archive of null results, something that preregistration advocates believe will be available to the scientific community only after their recommendations have been implemented. I will address this in the conclusion.

3 Credible Counterfactuals

Both papers compare the merit of the preregistration regimes they advocate with a research regime in which an isolated single author (or team) engages in a practice that is fundamentally biased. In the Humphreys et al. paper, the gains of registration are compared in simulations to the result of an investigator having no community oversight about model specification, and therefore the preferred approach is not being compared with the status quo in our discipline. As King, Keohane, and Verba (1994: 9) emphasize, science is a "social enterprise"; the study of bias without building in community incentives is a biased view of disciplinary practice. In the Monogan paper, the counterfactual is a research report on returns to a Republican congressional candidate for upgrading

[5]Nor is it the lay of the land in the world of structural equations in economics, in which collection of data on all parameters is followed by model manipulation until computer simulations report a correct fit.

[6]The two papers are less certain that registration is appropriate for research in which data are already publicly available. But my general point here holds for experimental research, viz. that in political science, there are invariably competing theories that are compared rather than an established theory that is tested.

[23.10.2012-3:35pm] [1-6] Paper: QP-POLA120017
Copyedited by: PSB MANUSCRIPT CATEGORY: COMMENT

anti-immigrant rhetoric in which the researcher omits from his/her model the possible confound of district ideology. Monogan admits that "Ideally, reviewers would have caught this omission." However, if our research community is so abysmally blind to missing-variable bias, I wonder if an archived registration regime would be more effective in enforcing our scientific norms. In other words, if scholars know that enforcement is weak in our scientific community, what would stop them from loading their preregistration plans with theoretical justifications for every possible functional form? This of course would subvert the very purpose of the registration regime. Or what is to stop an enterprising researcher from uploading replication data from a preregistered study that failed to correctly guess the set of covariates that would lend support to a theory and to publish the significant results of data analysis of the now publicly available data (a procedure, according to Monogan, that should have looser registration rules)? The answer is a community of scholars willing to police, and this is the principal cavity in our enterprise that needs to be filled.

4 Measuring the Cost/Benefit Ratio of a Preregistration Regime

The benefits seem clear! Our discipline suffers from model uncertainty: choosing the best model to test our hypotheses. The principal payoff to preregistration is that any ex-ante model is unbiased, and the reader can be assured of this lack of bias through referral to a registration archive. I do not dispute this important argument, but only note that model uncertainty can also be reduced with robustness tests, increasingly demanded as supplementary material by peer reviewers in leading journals. The advantage of a single preregistered model without robustness over a fishing expedition in which a particular explanatory variable retains significance across a range of specifications is not obvious.

What about the costs? Having commented on an initial draft of this commentary, Donald Green wrote that he saw some potential gain to preregistration, but could not see if it would make anyone worse off. Given the incrementalist proposals on which I am commenting, Green's point appears unimpeachable. However, I foresee costs that need to be reckoned. First, given my belief that it is the Bolsheviks who make history, the Mensheviki proposals (properly incremental, and fully cognizant of the costs) under review, if instituted, are likely to be captured by the ideological purists, refusing to consider the value of findings that do not meet the most rigorous (à la the FDA) preregistration standards. Second, to the extent that a radical fisheries management regime is imposed, empiricists may be discouraged from the data exploration that many researchers report as the basis of their deepest insights. This would be especially the case if those researchers saw "reject" down the equilibrium path with the expectations that the median reviewer would "red line" experimental reports that acknowledged post-plan fishing expeditions.

5 Conclusion

Registration without a community of scholars interested and incentivized to challenge findings is worthless. The bigger challenge for our discipline—and this is especially true for work in the CDC rather than the FDA mode—is not to deter misrepresentation but rather to incentivize the reverse engineering of work conducted in competing labs.

How might this be done most efficiently given the research challenges and publication incentives that political scientists face? I direct readers to Simmons, Nelson, and Simonsohn (2011). While they demonstrate that in psychology false positives are rife, they see preregistration as a "nonsolution." As Simonsohn elaborates (personal communication), even in medicine "people change design, don't report all variables, change number of participants, don't post data or results... etc. Enforcement is just too costly." Furthermore, he points out, "I also think of science as a process of discovery... Every paper I have [written] has some really interesting robustness, extensions, follow-ups that I would have never thought about at the beginning."

But this does not lead Simmons, Nelson, and Simonsohn (2011) to accept the status quo in their discipline. Instead, they provide a six-point set of guidelines for reports of experimental papers; the most relevant one for us is that all variables collected in the study should be part of the replication

[23.10.2012–3:35pm] [1–6] Paper: OP-POLA120017
Copyedited by: PSB MANUSCRIPT CATEGORY: COMMENT

data set and not just the ones that were eventually modeled.[7] And they have a four-point set of guidelines for peer reviewers; the most relevant for us is that they should be more tolerant of imperfections in results and thereby lowering incentives for paper-writers to obscure results that do not go precisely as predicted.

I would go three steps further. First, to the extent that a field experiment involves a contract with a party interested in a particular result, a PAP is recommended to prevent undue pressure from the funders to modify specifications in order to get their desired result. Second, no matter whether we have a formal registration regime, we ought to encourage a disciplinary practice of telling readers at what point in the research process (and for what reason) a particular model emerged, and to keep careful notebooks allowing us to report this accurately.

Third, we need to incentivize the publication of replications and null results. This might entail pressures on the already influential *Annual Review of Political Science* for more meta-analyses especially of experimental findings. But I would prefer the following: that one issue per year of the *APSR* be devoted to both replication of previous work and theoretically important null results from new research. This annual issue, as I envisage it, would entail the peer review and subsequent publication of papers no longer than six published pages.[8] With null results on well-established expectations appearing in our leading disciplinary review, the current publication bias would be reduced. Even for replications, with p values both below and above standard significance levels appearing, either increasing or decreasing confidence in previously reported results, this too should on average reduce publication bias. This annual issue would have three positive implications for scientific advance: (1) it would provide high disciplinary rewards for reporting on null results, what today we call "failure"; (2) it would through replications act as a police patrol on common misdemeanors such as omitted variables and results presented that were infected by cherry-picked covariates; and (3) it would provide a foundational (though admittedly not complete) archive of failed experiments for future meta-analyses.

By no means do I wish to discredit the proposals in the papers on which I was asked to comment. They are an important and sophisticated set of recommendations to raise our disciplinary standards. But the idea of registration is a new entry into our methodological medicine cabinet, and institutional implementation without broader understanding of its costs and benefits, and of the false analogies that may be driving proposals for institutional change, is risky.[9] I therefore wish only to open a disciplinary dialogue on the most efficient next step in transparency—and indeed preregistration might well be that step—that will deter bad practices while at the same time encourage us to learn from our data.

References

Boix, Carles. 2011. Democracy, development, and the international system. *American Political Science Review* 105(4):809–28.

Boix, Carles, and Susan Stokes. 2003. Endogenous democratization. *World Politics* 55:517–49.

Casey, Katherine, Rachel Glennerster, and Edward Miguel. Forthcoming. Reshaping institutions: Evidence on aid impacts using a pre-analysis plan. *Quarterly Journal of Economics.*

Gerber, Alan S., and Donald P. Green. 2012. *Field experiments.* New York: Norton.

Q3

[7] This would be impossible to enforce with data from field experiments, as the research community would not know if an uncomfortable variable (for the robustness of the preferred analysis) were deleted from the replication data set. If one believes that misrepresentation is less the problem than lack of incentives to replicate, as I do, this is a second-order issue.

[8] *Proceedings of the National Academy of Sciences (PNAS)* is the leading journal in generating citations. Its six-page limit forces researchers to report on what they found, why it is important, and why they are confident in their results. Of course, supplementary materials are available to reviewers and future readers.

[9] The idea is so new that in their recently published text Gerber and Green (2012) do not give sustained attention to the best procedures for a pre-analysis plan. In fact, in a few scattered mentions of the possibility for such planning, they celebrate the possibilities of fishing (p. 290) and make two suggestions that can correct for not using such a plan: (a) including difference-of-means estimates to accompany regressions in order to delimit the strategic use of covariates (p. 105) and (b) doing Bonferroni corrections to assure readers that the researcher isn't cherry-picking the one-in-twenty likely significant result (pp. 433–34).

[23.10.2012-3:35pm] [1-6] Paper: OP-POLA120017
Copyedited by: PSB MANUSCRIPT CATEGORY: COMMENT

Gerber, Alan S., Donald P. Green, and David Nickerson. 2001. Testing for publication bias in political science. *Political Analysis* 9(4):385–92.

King, Gary, Robert Keohane, and Sidney Verba. 1994. *Designing social inquiry*. Princeton, NJ: Princeton University Press.

Lindblom, Charles. 1965. *The intelligence of democracy*. New York: Free Press.

Przeworski, Adam, Michael E. Alvarez, and José Antonio Cheibub. 2000. *Democracy and development*. Cambridge, UK: Cambridge University Press.

Simmons, Joseph P., Leif D. Nelson, and Uri Simonsohn. 2011. Anything as significant false-positive psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychological Science* 22:1359.