Trust, Transparency, and Replication in Political Science

David D. Laitin, *Stanford University*
Rob Reich, *Stanford University*

ABSTRACT Striving better to uncover causal effects, political science is amid a revolution in micro-empirical research designs and experimental methods. This methodological development—although quite promising in delivering new findings and discovering the mechanisms that underlie previously known associations—raises new and unnerving ethical issues that have yet to be confronted by our profession. We believe that addressing these issues proactively by generating strong, internal norms of disciplinary regulation is preferable to reactive measures, which often come in the wake of public exposés and can lead to externally imposed regulations or centrally imposed internal policing.

Amid a micro-revolution in empirical research designs and experimental methods in our discipline,[1] this article discusses the distinctive perils in monitoring unethical practices and errors in this new research tradition. Because political science needs to improve its capacity to address ethical dimensions of increasingly common research practices, we propose preventive procedures that should be considered as part of an already active disciplinary discussion on a wide range of ethical standards.[2] We aim to participate in an already broad conversation about norms and rules of oversight that do not sacrifice scientific freedom and creativity. We emphasize measures that we believe will improve the incentive structure for transparency and regulation in political science.

Why pay special attention to our micro-revolution? A major reason is that our findings have more immediate relevance to a wider audience than was previously the case. A classic study of an earlier era, Moore's *Social Origins of Dictatorship and Democracy* (1966), for example, was a landmark contribution to the study of democracy. However, almost nothing would have changed in the world outside of social science if errors were found in one of his analyses. We compare that to a recent paper on democracy that purported to show that transparency in parliamentary debates in partial democracies tends to reduce participation in debates and threatens reelection (Malesky, Schuler, and Tran 2012). In this case, the implications for democracy-promotion programs in the world of international aid are immediate and the costs of error for wider society much higher.

It is not that political science in the past has been politically irrelevant. Rather, the micro-findings with clear causal attributions for well-identified treatments tend to have more immediate and obvious applications. For example, the findings on how best to "get out the vote" are adopted rather quickly by electoral campaigns and the findings on how to address voter fraud in new democracies affect aid-program design. We compare this with the more indirect effects on budgets and strategies when political scientists in the past served as expert witnesses in voting-rights trials, or as advisers and consultants for state and federal campaigns, or in high-level positions in the American foreign-policy establishment.

Immediate relevance increases not only the costs of error but also the incentives to cheat. Several factors are involved. First, the greater the immediate impact of our field on wider society, the greater the incentive to present results that advance one's political or moral interests. Second, with an increasing number of well-designed studies all attacking the same problems (e.g., conditions favorable to democratic stability) using the same tools, the space for a truly original finding is narrowing. To make a mark on the discipline today is a genuine challenge; falsifying data or misportraying findings can ensure originality. Third, in part because of a diminution of public funding for social science, recent studies are increasingly funded by private donors who have an interest in the findings and who put considerable pressure on researchers (who hope for continued funding) to tailor questions and adjust findings consistent with the funder's goals.

Despite stronger incentives to cheat—or at least powerful incentives to exaggerate marginal findings and/or hide nonfindings—our discipline has not matured in a parallel manner to curb the unethical behavior that may follow. Ethical research practices are not a core—or even a peripheral—element of graduate training; furthermore, we have weak or nonexistent institutionalized policing mechanisms. Indeed, the American Political Science Association (APSA) lacks the personnel and—until now—the will to budget for disciplinary oversight that is becoming increasingly important. We hope that the widely publicized allegations of data

David D. Laitin *is Watkins Professor of Political Science and co-director of the Immigration Policy Lab at Stanford University. He can be reached at dlaitin@stanford.edu.*
Rob Reich *is professor of political science and director of the Center for Ethics in Society at Stanford University. He can be reached at reich@stanford.edu.*

doi:10.1017/S1049096516002365

falsification in a 2015 experimental design will spur concerted and proactive efforts for a disciplinary remedy (Singal 2015). Political science should not wait for what happened in economics: a popular documentary, *Inside Job*, exposed unsettling ethical transgressions at the heart of the discipline.

Despite stronger incentives to cheat—or at least powerful incentives to exaggerate marginal findings and/or hide nonfindings—our discipline has not matured in a parallel manner to curb the unethical behavior that may follow.

To be sure, several innovations from within our research community have helped us to monitor ourselves. In 1995, Gary King published "Replication, Replication" in this journal in which he advocated, at a minimum, a replication footnote for all data-driven contributions to our scientific journals. This idea at the time was so radical that it was voted down at the annual business meeting of the APSA Comparative Politics Section. However, norms have changed. We consider a recent paper on the use of interaction terms in regression models. The authors requested the raw data from more than 20 scholars and all but two authors complied (Hainmueller, Mummolo, and Xu 2016). This is genuine progress; however, even with this progress, we cannot rely only on replication to identify flawed findings and errors, if only because self-correction works very slowly.

More recently, the Berkeley Initiative for Transparency in the Social Sciences (BITSS) was founded under the leadership of Edward Miguel (University of California, Berkeley) and Kevin Esterling (University of California, Riverside). With strong support from philanthropy, BITSS is now active in a range of transparency and replication initiatives. Most prominently, it is promoting preanalysis plan registration that *inter alia* would deter common practices of ex-post subgroup analyses that inappropriately permit researchers to attain standard significance levels for their statistical models—or, in more common parlance, to "go fishing" for a compelling result.[3]

These self-monitoring efforts have two main effects. First, the emphasis on replication and preanalysis plan registration increases the likelihood that *p*-hacking will be identified, thereby deterring fishing for significant results. Second, these efforts correct well-intentioned but flawed research findings. On this latter dimension, it is perhaps more appropriate to say that replication and preanalysis plans are useful but limited mechanisms for the practice of good science.

We think more should be done. Political science should promote further institutional mechanisms to ensure ethical practices that are incentive-compatible with encouragement to sustain scientific creativity. Applying the evocative metaphor of McCubbins and Schwartz (1984), we can assume that scientific police patrols will stifle creativity. In their stead, we need to cultivate effective "fire alarms" (that actually activate responders and need not assume scientific malpractice) to serve as warnings of error or violations of scientific norms. In addition, we must cultivate—early in graduate training—clear norms of good scientific practice that will diminish the likelihood of "fire" in the first place. We need the equivalent of flame retardants in our labs.

This is no easy task. We are humbled by the evaluation of Neuroskeptic (2012), which showed "circles of hell" housing manifold levels of bad science. Moreover, it is no mean task to separate ethical misconduct from intradisciplinary debates about what constitutes best practices. We do not want to provide ammunition to scholars embroiled in disciplinary debates that allows them to discredit work from other research traditions through accusations of unethical practices. However, fears that we might move from disuse to overuse of ethical fire alarms should not deter us from a search for better practices. A few suggestions follow, more to sustain debates in our discipline than to lay out a full-scale plan.

GRADUATE TRAINING

Political science departments do not have required courses in research ethics. Professional-ethics courses are standard in business and law curricula, and it seems a mistake to assume that our students learn our ethical aspirations through osmosis or tangentially through methods courses. In other words, we think the first steps in addressing our collective concerns ought not to be through policing or fire alarms; rather, there is need to instruct and encourage good behavior. Disciplinary-ethics courses should build on the increasing attention in the APSA to set new standards of ethical conduct.[4] To be sure, an ethics course alone will never inoculate against dishonesty, but the requirement to take such a course sends a clear signal to new and aspiring scientists: it conveys a disciplinary consensus about best practices—and all the better if it is instructed by tenure-line faculty.

Dissertation advisers should institutionalize practices that require students to regularly archive or version-control their code and data (including raw data—and not only final cleansed "analysis" data) such that they are readily available to committee members and future peer reviewers when solicited. Currently, few advisers require their students to present batch files with submitted work. Furthermore, our PhD programs rarely provide in methods courses the detailed instructions on how to prepare files so as to be transparent not only to their advisers but also to future reviewers.[5] This failure leads to long lag times in delivering data to potential replicators, thereby slowing down scientific progress. It also is an invitation for uncorrected mistakes to persist throughout a research project.

JOURNAL PRACTICES

Rigid criteria for assessing the value of a scientific contribution inadvertently, although predictably, can encourage misrepresentation. For example, journals that refuse to consider claims to significance in submissions that do not meet stipulated *p*-values, as has been the practice of the *American Journal of Political Science (AJPS)*, are not only arbitrary; they also encourage researchers to "*p*-hack"—that is, to present only those functional forms in which the required *p*-values are met.[6] The more that we bureaucratically set standards for scientific validity, the more that scholars will attempt to re-present their results to meet those standards.

Publication Outlets for Replications and Null Findings

On the other side of the *p*-hacking coin, if the discipline supported the publication of replications and null findings, through either reserved space in a leading journal (e.g., *American Political Science Review* [*APSR*]) or in a new journal, there would be less incentive to put failed experiments in the file drawer or to use strategic subsamples of the data to achieve conventional significance levels.[7] The systematic publication of replications has the added advantage of lowering confidence in previous findings that were held to be solid; the systematic publication of null results has the added advantage of saving the time of future scholars who otherwise might follow that same dead-end path. Of course, a paper reporting on failed replications must do far more than show how a result can disappear (or reduce its coefficient) if the "kitchen sink" were put on the right-hand side of a statistical equation. Furthermore, a paper on null findings would need to convince peer reviewers that there were strong theoretical or empirical reasons to have expected a significant result. In this case, the null finding would be a (partial) discovery.

Enable Replication of "Un-Deidentified" Datasets

In some instances, information such as geolocation may be central to analysis, rendering a deidentified dataset useless for the purposes of replication and reinvestigation. A "Trusted Human Subjects Data Intermediary," an adjunct to the Institutional Review Board, is a possible remedy. This intermediary could answer questions about the data without compromising the identity of human subjects. Similar to the proposal for "information escrows" (discussed later in this article), the intermediary role also would allow those outside of a project who have questions or suspicions to have their concerns addressed.

Sharing of (Raw) Data during the Peer-Review Process[8]

Double-blind peer-reviewing is at the heart of merit-based evaluation. However, the procedures have not kept up with standard practices involving large data files. Without the raw data and accompanying ".do" files, peer reviewers cannot thoroughly evaluate the quality of the submission. Moreover, without being required to send the raw data, authors can submit work that remains unfinished—or at least insufficiently cleansed. One concern is that with the raw data, reviewers might have incentives to recommend rejection and then to run the data for their own research purposes. This would not only be unethical, it also would remove the initial researcher's advantage in publishing results from hard-earned data collection. One solution is to require reviewers to sign a nondisclosure agreement with a journal, promising not to rely in their own work on the data of a reviewed paper until it reaches published form within a time limit. How best to reward reviewers with this additional task—that is, reviewing raw data and accompanying files—remains an issue. Given what we know about intrinsic motivation, recommending that reviewers be paid (as in economics) may not increase their willingness to review in a timely manner.

DISCIPLINARY PRACTICES

Technical support for disciplinary journals can assure readers that all published work in our discipline has been vetted for reproducibility of results. When a final revised manuscript reaches editors at journals such as *Proceedings of the National Academy of Sciences*, paid technical personnel rerun the models from the accompanying data to assure editors that all results are reproducible. For this to be implemented in political science, the APSA (or perhaps philanthropists) would need to fund employment for technical experts capable of assuring all readers that basic reproducibility has been achieved.[9] Whereas some might argue that these experts should be assigned more complex tasks (e.g., assuring readers that best scientific practices were met in the analyses), we believe that we can achieve disciplinary consensus on the reproducibility criterion.

In other words, we think the first steps in addressing our collective concerns ought not to be through policing or fire alarms; rather, there is need to instruct and encourage good behavior.

Information Escrows to Provide Incentives to Firefighters

According to Ayres and Unkovic 2012, research transparency is insufficient to monitor scientific malpractice. There are powerful disincentives for researchers to search for and detect mistakes and misconduct because this requires considerable work and invites personal reprisal with potential negative career implications. The resulting professional "code of silence" not only prevents rapid disclosure of error, it also allows the spread of unfounded rumors that can harm the reputation of innocent authors. A web platform in which an accuser has revocable anonymity (to avoid reprisal) and authors have a parallel revocable confidentiality (allowing them to publicly respond to critics before the unjust rumor spreads uncontrollably) can play a part in aligning incentives for critics to state their concerns and authors to defend their scientific practices. Thus, the web would serve as an "information escrow" for the critic–author communication, with the rules of public disclosure essentially allowing both sides at low cost to dispel concerns without either side losing face. Many details of this escrow are yet to be worked out, especially in thinking about ways to prevent "trolling"—that is, activities by critics to bombard authors with unending queries that eventually overwhelm their capacity in the confidential-exchange period. (This could be reduced if there were a nominal fee to submit a query or an independent editor to vet submissions.) Legal issues involving possible defamation suits would need to be addressed. The aim is that an escrow of this type would play a role similar to that of "Retraction Watch," which raises the professional costs of deceit and, more important, quickly identifies inadvertent errors.[10]

These suggestions comprise only a small step in thinking about the larger issues of ethical practices in our profession. Many questions remain unanswered. First, despite the fanfare in the wake of a few egregious ethical missteps, we can only speculate on the scope of the problem. Should we work more to optimize good practices (e.g., transparency and replicability) or to deter the worst abuses?

Second, we have not yet addressed the political-economy question of who should bear the burden of educating and monitoring. Some of our proposals rely on senior researchers to set new standards—for example, special sections of journals for replications, or advisers and senior collaborators as models of proper behavior, or new graduate courses in professional ethics. Others rely on junior researchers to hold senior researchers more accountable—for example, graduate seminars in which students aim to replicate recent journal articles. There are career risks in serving as firefighters and opportunity costs for teaching ethics. We need to address how those costs and risks should be allocated.

There are powerful disincentives for researchers to search for and detect mistakes and misconduct because this requires considerable work and invites personal reprisal with potential negative career implications.

Third, any proposals that suggest new institutions or courses need to better address who would establish and support them. If information escrows require monitors to contain cheating, from where would they be recruited—and what makes us think they would not be subject to malpractice themselves? Concerning nondisclosure agreements for peer reviewers, who would be assigned the role of ensuring compliance?

Fourth is the issue of implementation. Is it better to pursue efforts to promote *decentralized* norm adoption and diffusion or to promote *centralized* rules to govern research practices? Although the APSA is currently active in promulgating new rules to promote transparency, the question of enforcement remains unanswered. Should fire alarms be sounded in APSA committees appointed by the APSA Council, or would be it better to leave the monitoring function to individual journals, departments, and advisers? The answer to this question turns on the issue of whether leading researchers—given the perverse incentives we all face—are willing to respond to fire alarms. Political science must confront the situation as it is presented: researchers, like humans more generally, are not moral saints and they do respond to perverse incentive structures.

Although there remain many unanswered questions on the issues of reform and implementation, we have no doubt that addressing the assortment of concerns about the practice of good science is of great importance—and not only to improve science. Indeed, to earn the public trust that is the foundation of scientific funding and influence over policy, science also must be held to the highest ethical standards. It must do right and be seen to be doing right. Without monitoring ourselves, we could easily (and further) erode that trust, sacrificing the real gains that have been made in the maturation of our discipline as a science and as a resource for policy improvement. ■

NOTES

1. This move was foreseen and spurred by Green and Gerber (2002).
2. The McCoy Family Center for Ethics in Society at Stanford University sponsored a discussion of these issues on October 26, 2015. We thank Joan Berry and Anne Newman for their support. Chatham House Rules applied ensuring nonattribution, and we therefore do not acknowledge particular contributions to our discussions. In any event, we take full responsibility for the proposals offered herein.
3. For the mission statement, see Miguel (2014). The larger mission of BITTS is available at http://bitss.org. See Laitin (2013) for a discussion on how preanalysis plans can stifle creativity.
4. The APSA-supported initiative for implementation of transparency standards is available at www.dartstatement.org.
5. Tools are now available making batch files and one-click reproducible workflows (i.e., version control with Git and GitHub and dynamic documents using R Markdown and knitr) relatively easy.
6. On *p*-hacking, see Ioannidis (2005). On the ubiquity of *p*-hacking in psychology, see Simonsohn, Nelson, and Simmons (2014). On the criteria for scientific claims in the *AJPS*, see its Guidelines for Manuscripts, available at https://ajps.org/guidelines-for-manuscripts.
7. Laitin (2013) proposed the annual *APSR* issue for replication and null findings. The new *Journal of Experimental Political Science* has promised its readers that it will consider papers reporting null results.
8. Rose McDermott has advocated this idea in the APSA Council, and we borrow it from her.
9. *Political Analysis* is the exception in our discipline. It employs a graduate student who, for provisionally accepted papers, reruns all of the models and checks whether the results replicate. The manuscript is not published until they do.
10. See http://retractionwatch.com, which recently received a MacArthur Foundation grant to support its policing activities in the life sciences.

REFERENCES

Ayres, Ian and Cait Unkovic. 2012. "Information Escrows." *Michigan Law Review* 111 (145): 145.

Green, Donald and Alan S. Gerber. 2002. "Reclaiming the Experimental Tradition in Political Science." In *Political Science: State of the Discipline*, ed. Ira Katznelson and Helen V. Milner, 805–32. New York: W.W. Norton.

Hainmueller, Jens, Jonathan Mummolo, and Yiqing Xu. 2016. "How Much Should We Trust Estimates from Multiplicative Interaction Models? Simple Tools to Improve Empirical Practice." *Social Science Research Network*. Available at http://ssrn.com/abstract=2739221.

Ioannidis, John. 2005. "Why Most Published Research Findings Are False." *PLoS Medicine* 2 (8): 124.

King, Gary. 1995. "Replication, Replication." *PS: Political Science and Politics* 28: 443–99.

Laitin, David D. 2013. "Fisheries Management." *Political Analysis* 21 (1): 42–7.

Malesky, Edmund, Paul Schuler, and Anh Tran. 2012. "The Adverse Effects of Sunshine: A Field Experiment on Legislative Transparency in an Authoritarian Assembly." *American Political Science Review* 106: 762–86.

McCubbins, Mathew D. and Thomas Schwartz. 1984. "Congressional Oversight Overlooked: Police Patrols versus Fire Alarms." *American Journal of Political Science* 28 (1): 165–79.

Miguel, Edward. 2014. "Promoting Transparency in Social Science Research." *Science* 343 (6166): 30–31.

Moore, Barrington. 1966. *Social Origins of Dictatorship and Democracy: Lord and Peasant in the Making of the Modern World*. Boston: Beacon Press.

Neuroskeptic. 2012. "The Nine Circles of Scientific Hell." *Perspectives on Psychological Science* 7 (6): 643–4.

Simonsohn, Uri, Leif D. Nelson, and Joseph P. Simmons. 2014. "*p*-Curve: A Key to the File Drawer." *Journal of Experimental Psychology: General* 143 (2): 534–47.

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