Article

Comparative Political Studies
1–37

DOI: 10.1177/0010414016655539
cps.sagepub.com
SAGE

Can Results-Free Review Reduce Publication Bias? The Results and Implications of a Pilot Study

Michael G. Findley[1], Nathan M. Jensen[1], Edmund J. Malesky[2], and Thomas B. Pepinsky[3]

Abstract

In 2015, *Comparative Political Studies* embarked on a landmark pilot study in research transparency in the social sciences. The editors issued an open call for submissions of manuscripts that contained no mention of their actual results, incentivizing reviewers to evaluate manuscripts based on their theoretical contributions, research designs, and analysis plans. The three papers in this special issue are the result of this process that began with 19 submissions. In this article, we describe the rationale for this pilot, expressly articulating the practices of preregistration and results-free review. We document the process of carrying out the special issue with a discussion of the three accepted papers, and critically evaluate the role of both preregistration and results-free review. Our main conclusions are that results-free review encourages much greater attention to theory and research design, but that it raises thorny problems about how to anticipate and interpret null findings. We also observe that as currently practiced, results-free review has a particular affinity with experimental and cross-case

[1]University of Texas at Austin, TX, USA
[2]Duke University, Durham, NC, USA
[3]Cornell University, Ithaca, NY, USA

Corresponding Author:
Michael G. Findley, Department of Government, University of Texas at Austin, 3.108 Batts, Department of Gov, Austin, TX 78712, USA.
Email: mikefindley@utexas.edu

methodologies. Our lack of submissions from scholars using qualitative or interpretivist research suggests limitations to the widespread use of results-free review.

Keywords
experimental research, quantitative methods, qualitative methods, results-free review, transparency, preregistration

Introduction

In the past decade, political science has witnessed a growing movement for greater transparency in research. Prominent examples include efforts by the Evidence in Governance and Politics Network (EGAP; www.egap.org), the Berkeley Initiative for Transparency in the Social Sciences (BITSS; www.bitss.org), a recent symposium on transparency in qualitative methods (Moravcsik, 2014), and the recent Data Access and Research Transparency (DART) statement signed by the editors of 27 leading journals (http://www.dartstatement.org/).

Although there are varied objectives driving the shift toward greater transparency, one of the key motivations is to avoid *publication bias*, which can emerge as a result of a peer-review process that privileges the significance of results over their theoretical contribution, research design, quality of the data and analysis, and even the importance of the motivating research question. So long as the significance of results is the overriding concern among editors and reviewers, authors will have few incentives to report all of the empirical tests they conduct. Publication bias can manifest itself through bias in individual studies, but aggregated across studies an overall bias manifests itself in the scholarly record in a given area. Moreover, it can lead to serious questions about the overall quality of research in the field, as evidenced by the recent crisis in psychology (Open Science Collaboration, 2015; Yong, 2012).

A potentially simple and yet powerful way to mitigate publication bias is for journals to commit to publish manuscripts without any knowledge of the actual findings. Authors might submit sophisticated research designs that serve as a registration of what they intend to do.[1] Or they might submit already completed studies for which any mention of results is expunged from the submitted manuscript. Reviewers would carefully analyze the theory and research design of the article. If they found that the theoretical contribution was justifiably large and the design an appropriate test of the theoretical logic, then reviewers could recommend publication regardless of the final outcome of the research. In theory, this could mitigate publication bias (see Nyhan, 2014).

Implementing such a system is challenging, and full of uncertainty. This Special Issue of *Comparative Political Studies* (*CPS*) helps to assess the potential benefits and costs associated with new models of the publication process by studying how this particular model works in practice. In so doing, we shed new light on the transparency debate in the social sciences. We consider it to be self-evidently true that transparency should be one central objective in contemporary social science, but what are the costs and benefits of different transparency approaches and in what ways would current publication practices have to change to accommodate a results-free review model? Some critics of results-free review, for example, may worry that it will inevitably lead to journals full of null results, and to projects that are less theoretically innovative and path breaking than would otherwise be possible. In other words, journals would receive and publish "boring" work. Does results-free review commit scholars to carry out projects that are unfeasible, or dissuade creative dialogue between theory and data? How will manuscript referees respond to manuscripts without results or conclusions? These questions cannot be settled in the abstract.

We investigated these questions through a special issue on research transparency. The goal of this special issue was to consider papers that fit within the mandate of *CPS*, but were submitted as standalone designs or completed papers without any of the results reported. Thus, our special issue is not on a substantive theme or topic, but is rather defined by the process whereby authors submitted manuscripts and referees reviewed them. As special issue editors, we were involved in the entire process, which meant that we observed the submissions through to acceptance. Like the peer reviewers, though, we too never once saw the results of any of the papers until the final stage, well after publication decisions were finalized.

We created the call for papers (CFP) and advertised broadly, we observed what kinds of submissions we received, added our own evaluations of the manuscripts, decided which pieces to desk reject or send out for review, selected reviewers, received reviewer comments, and made final recommendations to the *CPS* standing editors. Of course, we worked closely with the standing editors throughout the entire process. Our close involvement gave us helpful insights into how results-free review works in practice, which we share here.

Based on this experience, in this introductory essay, we make three key observations about results-free peer review in practice. First, contrary to fears that greater emphasis on transparency creates more incentives for clever research designs and methodological perfection, reviewers placed an overwhelming emphasis on theoretical consistency and substantive importance. In this regard, results-free review worked better than we could have

hoped in incentivizing theory and research design over narrow concerns about novelty of methodology or empirical causal identification. Of course, our pilot was not a direct comparison between results-free review and the same exact papers undergoing standard reviews, so we lack the counterfactual to make definitive conclusions about the benefits of the process. We can say with confidence, however, that reviewers in our pilot were explicitly concerned about well-articulated theories and null results that moved literatures forward scientifically, and were not tolerant of long lists of ambiguous hypotheses to be tested (what we refer to as hypothesis trolling). Atheoretical and "boring" work stood very little chance of publication in this pilot. Relatedly, we hasten to add that the overall quality of the reviews for the special issue were quite strong. Indeed, it appears that by needing to engage the theory and hypotheses, reviewers could not simply nitpick over the credibility of statistical results. Improved engagement by reviewers is one argument that editors may consider when allowing results-free review as a submission option.

Second, it was nevertheless immensely challenging for reviewers and authors alike to argue coherently about the proper role of null findings, often referred to tellingly as "non-results." The challenges for current practices in this regard are steeper than we had anticipated, and speak to general debates about null-significance hypothesis testing and the relationship between theory, data, and models (Clarke & Primo, 2012).

And, third, results-free review has a particular affinity for certain methodologies, reflected in the types of submissions we received for this special issue. In particular, our submissions were almost exclusively experiments and observational studies that were testing general propositions using cross-case inferential techniques. Each of these three observations, we argue, has substantial implications for social science in general, for comparative politics in particular, and also for contemporary debates about transparency.

The rest of this essay proceeds as follows. In the next section, we provide an overview of publication bias and then consider the premise that results-free peer review could be a potential solution to the problem. We then briefly outline the procedures that we followed in our special issue. The subsequent section discusses what we learned—the importance of theory, null findings, and methodological affinities. Next, we summarize the findings in the three articles that successfully completed the peer-review process and which appear in the special issue, noting ways in which the process behind each reflects these general concerns. A final section provides some practical considerations about how to manage a results-free process in a top-flight journal.

The Problem of Publication Bias

Before diving into the details of our special issue, we take a step back and review the general challenge of publication bias. How do we know such bias exists in political science journals? What factors have driven it? What steps have been taken thus far to address it? And how successful have those steps been? In this section, we answer these questions before discussing how results-free review might assist in the battle against publication bias.

What Is Publication Bias?

In its most basic form, publication bias exists when a set of published studies is not representative of all available or possible studies. There are myriad reasons for a non-representative set of available studies. In much scientific work, publication bias is most pronounced when publication decisions are based on the realized outcomes of a study—typically statistical significance of a result—rather than the merits of the approach and design (Dickersin, 1990; Humphreys, de la Sierra, & van der Windt, 2013; Sterling, 1959).

The problem of publication bias is not complicated, but it is rampant and consequential. One goal in political science (and social science more generally) is the correct measurement of causal effects. If a study is carried out correctly, then the results should matter regardless of whether they confirm preexisting hypotheses about those causal effects. Indeed, null results from a well-designed study are just as meaningful as strong positive or negative effects. To take one prominent example from comparative politics, what if there is no causal relationship between political culture and democratic rule? If this is so, then this guides our understanding as a discipline of the origins of democratic regimes, and may also serve to guide policy makers who wish to promote democracy. But what if results that show no link between values and political regimes are less likely to be published than are results that support the existence of such a relationship? If so, then even for a question in which there is vigorous debate between findings and null findings (see, recently, Welzel & Inglehart, 2009), the *published* evidence will imply that the claim has more empirical support than it does.

Unfortunately, existing publication practices in social science and other disciplines privilege strong (i.e., statistically significant and substantively large) positive or negative findings, thus making null effects less likely to emerge. As such, scholars have compelling incentives to engage in data fishing (Humphreys et al., 2013) to obtain results that will be publishable. What is especially problematic is that even if individual scholars find ways to precommit to not engaging in data fishing, the publication process could lead to

bias. Studies with null results may not survive the publication process due to reviewer and editorial decisions, as other studies with large (and perhaps more counter-intuitive results) are more likely to be published. And if those published results are not representative of the actual distribution of causal effects (or lack thereof) in the real world, then publication bias exists and skews our knowledge base as well as any public policy that results from a given corpus of studies. Even in a world of angels writing research papers, the devil may still hide in the peer-review process.

How prevalent is publication bias? In a recent examination of the *American Political Science Review* and the *American Journal of Political Science*, Gerber and Malhotra (2008) conducted an extensive survey and test the hypothesis of whether publication bias exists. In their own staggering words, "we can reject the hypothesis of no publication bias at the 1 in 32 billion level" (p. 313). In a more recent study, Franco, Malhotra, and Simonovits (2014) document publication bias across the known population of studies utilizing the Time-Sharing Experiments in the Social Sciences (TESS) program and demonstrate that fielded survey experiments with null findings are substantially less likely to be published than those with significant effects, and the principal investigators themselves admit to abandoning such "unsuccessful" projects. Sadly, these results confirm what others have found across the social sciences (see, for example, Dickersin, 1990; Gerber, Malhotra, Dowling, & Doherty, 2010; Glewwe & Kremer, 2006; Ioannidis, 1998).

The implications of publication bias for the social sciences may be more consequential than distorting the findings in academic journals. The consequences of incorrect findings, championed as scientific evidence, are obvious in medicine, and can lead to incorrect diagnosis or treatment, such as decades of demonizing saturated fats despite clear, unpublished experimental evidence to the contrary (O'Connor, 2016). Similar problems may confront the social sciences.[2] Returning to the democracy example from above, academic research by Finkel, Pérez-Liñán, and Seligson (2007) has informed democracy promotion decisions by the U.S. Agency for International Development (USAID), and yet the study's few challengers find little to no evidence that the statistically significant results in Finkel et al. (2007) hold. Although far from a settled debate, these studies illustrate that research has the potential to influence the decisions of well-meaning practitioners and policy makers in government, even before scientific consensus is reached. And, needless to say, public policy decisions in the social sciences can have large and lasting impacts on peoples' lives.

One prominent example from political science serves to show the most extreme case. Most political scientists are familiar with the controversy surrounding a paper published by LaCour and Green (2014) on the impact of

canvasing on support for gay marriage. This article is notable for two reasons. First, it was retracted due to alleged fraud in essentially every aspect of the project including false statements about grant funding, compensation of subjects, preregistration of hypotheses, Human Subjects approval, and even the very data collection itself.[3] Second, what may have made this article especially interesting in the first place was the massive size and persistence of the impact of gay canvasing on support for gay marriage that stood in sharp contrast to nearly all existing studies on persuasion.[4]

Our special issue has little to say about outright fraud in research. But the counterfactual we would like *CPS* readers to consider is as follows: What would have happened to this paper if it had been reviewed without results? Would this have led to additional scrutiny of the paper that would have uncovered the fraud? Perhaps. More important for our exercise, the merits of publishing this article would not have been based on its splashy results. Scrutiny of how this research design relates to existing work in the field, data collection efforts, and an analysis plan would have been central to the success or failure of the paper.

This is clearly an extreme case of fraud, and our counterfactual is speculative. But it is helpful to recognize that there are two distinct drivers of publication bias. The first source is the career-oriented motivations of individual authors. Job placement and tenure decisions can generate incentives for (a) prioritizing work in the pipeline that has "significance stars" next to key coefficients, which was the key lesson of Franco et al. (2014); (b) choking unresponsive data with a barrage of specification choices and subsample analyses until the data confess (Nuzzo, 2014);[5] and, rarely but it happens, (c) outright manipulation or fraud.[6]

The second source of bias occurs when reviewers and editors evaluate the publication merits of null results. The burden of proof appears to be higher for null rather than significant results, because reviewers are forced to decide whether incorrect theory or a problematic research design generated the insignificant result. Recently, one of the editors of this special issue received a revise and resubmit decision from a prominent journal with encouragement to abandon null results. The reviewers cited theoretical deficiencies, leaving a difficult decision of whether to push back and keep the null results or drop them at the behest of the reviewers. Academic incentives for junior faculty or grad students likely result in following the reviewers in cases like this.

First and foremost, null results often call into question the larger theoretical enterprise of the paper. Skeptical reviewers might give the benefit of the doubt to an implausible theory that, despite reviewer misgivings, yielded observable implications that were tested and identified with significant findings. The same generosity would not be provided to null findings. If a

reviewer never believed that sunspots influenced social movements to begin with, why would an empirical test that finds no evidence of such a relationship be worth publishing? Gelman and Carlin (2014) describe several examples of theoretically implausible but highly provocative findings that were indeed published because of their statistical significance, and show that reasonable calculations of the likely effect sizes in the studies in question imply that the reported effects are massive overestimates—and very possibly have the wrong sign.

The second issue is empirical, and the frequentist language for hypothesis testing is helpful for elucidating the problem. With a null finding, we "fail to reject" a null hypothesis, we do not "disprove" the alternative. Why does this matter? As a thought experiment, imagine an accounting ledger of research design flaws that might bias in favor of rejecting the null hypothesis when it is correct (Type I error) or failing to reject the null hypothesis when the alternative is, in fact, correct (Type II error). A number of mistakes can lead to biased coefficients and Type I errors, including simultaneity, biased selection into treatment, systematic measurement error in the independent variable, omitted variable bias, and unobserved heterogeneity. The list for Type II errors, however, includes all of those issues and a few more that either bias coefficients to zero or increase inefficiency, including insufficient power, stochastic measurement error in the independent variable (leading to white noise), and stochastic measurement error in the dependent variable (leading to attenuation bias). Thus, on a simple accounting basis, papers with null findings have to overcome a greater set of inferential obstacles than those with significant coefficients.

The problem actually goes deeper, however, as even when a flaw is apparent, the burden of demonstrating the robustness of findings to a potential correction is easier for Type I errors. First of all, the most common stratagem for side-stepping a research design flaw that would bias against a significant coefficient is not available to scholars with null findings. Authors aware of measurement error that creates noise and increases their standard errors, for example, will often claim that they obtained significance despite the stochastic measurement error. Similarly, authors aware of omitted variable bias can argue that the excluded variable would have likely biased the coefficient on their key causal variable toward zero. The fact that they still found a significant effect despite the bias indicates that their main effects would even be stronger if they had a properly specified regression.

The appeal to the persistence of stars in the presence of bias is not available for null findings. Similarly, scholars with significant findings can demonstrate the insensitivity of their significant coefficient to multiple measures, introduction of confounders, and specification choices, "Despite multiple

attempts, I could not make those stars disappear." This approach, however, rings hollow when the stars were never there. Adding more models with null findings only appears to reinforce that the alternative specifications have not addressed the underlying design flaw.

All that said, there are techniques to mitigate the threat of null findings, distinguishing between design problems and deeper concerns. The most common of these is assessment of a study's power. By ensuring adequate statistical power, a study attempts to avoid Type II errors. As commonly defined, statistical power is the probability of correctly rejecting the null hypothesis, or, alternatively, 1 minus the probability of committing a Type II error. Put this way, if a study is sufficiently powered, then authors can argue that they have solved at least some design issues that could have yielded the null result and therefore offer greater confidence that the null result is meaningful.[7]

The specific point on statistical power is indicative of a more general point about the quality of the research question, theory, and design. It may be that few null results get published precisely because the quality of the research that produced those results was so low. Given the set of possible statistical relationships that could be explored, scholars typically begin by theorizing about those that ought to be related, thereby leaving aside the investigation of true null relationships. Thus, a disproportionate share of significant findings could reflect on scholars carefully choosing questions and designing research appropriately, whereas null results could reflect on scholars' attempts to understand what should have been non-null relationships, but with low-quality research approaches.

Replication as a Solution to Publication Bias?

Any prescription for overcoming publication bias must first begin from the premise that null findings face a greater uphill battle for publication. This problem affects reviewers and is also clearly understood and appreciated by authors, which is why we suspect that researchers may consider efforts to publish null findings a fool's errand, and therefore do not even attempt to publish them, as Franco et al. (2014) showed.

Thus far, most efforts to address publication bias have focused on the first driver, the intentional actions of authors to produce work with p values below 0.05. To this end, efforts at producing greater transparency in research have thus far emphasized better replication practices. Most notably, the practice of making replication data available is increasingly common. The *Quarterly Journal of Political Science* (*QJPS*) has a staff member replicate all reported findings before publication. Other journals are beginning to follow suit, including the *American Journal of Political Science* and the *Journal of*

Experimental Political Science. Other journals require posting replication files (e.g., *CPS, Journal of Politics, Journal of Conflict Resolution, Journal of Peace Research*, among others), and full platforms have made cataloging data both uniform and accessible (e.g., *Dataverse*: http://thedata.org/). These decentralized efforts on replication have culminated in a recent, but currently embattled, initiative by the American Political Science Association (APSA) to establish the DART standards, to which about 27 journals—including *CPS*—have agreed.

Perhaps due to the increased scrutiny, replication exercises have uncovered high-profile cases of academic fraud. As noted above, a study by Michael Lacour and Donald Green (2014; now retracted) was alleged to have used fabricated data through the process of replication (see Broockman, Kalla, & Aronow, 2014). One of the highest profile cases to hit political science, this scandal not only suggests the value of stronger transparency standards but also provides some initial validation that the system may at some level work in detecting unscrupulous behavior. Equally notable, but in psychology, is a fraud case by a Dutch researcher who falsified data and made up entire experiments (Carey, 2011). Although not academic fraud, not so long ago, a high-profile working paper by Reinhart and Rogoff (2010) found that government debt in excess of 80% of gross domestic product (GDP) had devastating consequences. Numerous politicians heralded this research as justification for fiscal policy reforms. And yet the authors provided little information on their coding rules and procedures for constructing their sample.[8] Eventually, it was uncovered that their results were driven by a questionable decision to drop some countries and a major Excel coding error (see Herndon, Ash, & Pollin, 2014). These are just a few examples, whereas organizations such as *Retraction Watch* document retractions (or discuss proposed retractions) of any kind and include a leader board of authors with the most retractions.[9]

Although data replication addresses some of the challenges of transparency, unfortunately it cannot set formidable standards against data fishing. For example, with few exceptions (see, for example, Nielsen, Findley, Candland, Davis, & Nielson, 2011), most replication studies provide only the data for the final set of results in a manuscript, thus leaving unknown the full set of data preparation operations conducted.

Preregistration as a Solution to Publication Bias

One possible way to increase transparency is through preregistration, which specifically requires that, prior to carrying out a study, scholars provide details about the research design of the study, how the study's data will be analyzed, as well as any potential conflict of interests with funders.

These registries record all studies, including those that have been conducted and have not been submitted to journals. In this piece, we focus on the role of preregistration for submitted studies.

The field of medicine was the first to set up preregistration standards (De Angelis et al., 2004). The process of establishing mandatory preregistration in medicine was not easy; indeed, significant opposition contributed to a number of false starts, which delayed the adoption of preregistration (Dickersin & Rennie, 2003). But it is now possible to trace most research from inception through to completion, and it is clearer which research is funded by private donors who may have an interest in the outcomes of the research.

Drawing on the example of medicine, there is a broader movement toward preregistration of research that is just now entering the social sciences, including psychology, economics, and political science. Indeed, there is much optimism that adopting more stringent transparency standards should improve social science research (Humphreys et al., 2013; Miguel et al., 2014; Monogan, 2015). Registries for research designs have been established by the Evidence in Governance and Politics Network (EGAP), the BITSS, the American Economic Association's Randomized Control Trial (RCT) registry, the Registry for International Development Impact Evaluations (RIDIE) registry, and the Center for Open Science's Open Science Framework, among others. The expectation of these registries is as follows: preregistration creates the proper incentives to report (and publish) based on research design rather than the results, which makes it more likely that accurate causal effects—be they directional or null—come to light for the scientific community to be aware.

Preregistration admits a wide variety of possible designs. They could range from providing basic information about hypotheses and expected tests as with the basic EGAP preregistration option (http://egap.org/content/registration) to extraordinarily detailed analysis plans and mock reports as in many of the designs posted to EGAP (http://egap.org/design-registrations).

Although there is substantial theorizing about research transparency, there is very little empirical evaluation of actual preregistration practices. Some scholars have begun to register their designs, but very few of those designs have been published in political science (see Findley, Nielson, & Sharman, 2013, 2014, 2015; Gottlieb, 2016; Monogan, 2015 for additional examples). There are a few published examples in Economics and Psychology (e.g., Casey, Glennerster, & Miguel, 2012) in *Quarterly Journal of Economics* and the landmark preregistered observational study by Neumark (2001), but published results are not yet keeping pace with the growing interest in preregistration.

Preregistration complements replication policies by setting standards for earlier phases of the research, including the full set of data preparation and analysis operations to be conducted, which should reduce data fishing and in turn reduce publication bias. Although these incidents provide lessons about replication and data sharing, they are primarily aimed at catching publication bias caused by the researchers. They do little to address the role that the review process indirectly plays in the censorship of null findings

Our special issue received nine (out of 19) submissions for preregistered work that was yet to be fielded. As special issue editors, we were especially excited about submissions in which scholars submitted their plans prior to implementing their research. This allowed greater transparency in the research process and also enabled detailed feedback to the researcher before they went into the field.

Nevertheless, it is not clear that preregistration alone is a sufficient solution for publication bias. First, it is possible for authors to engage in hypothesis trolling, preregistering multiple measurements, hypotheses, and subsample analyses, thereby allowing themselves ample room to *p*-fish within their stated research plans. As multiple comparison corrections become increasingly common, these problems may not be as acute, because there is a penalty for each additional test conducted. In the most conservative case, for example, the Bonferroni correction requires a revised significant level cutoff at α/n where α is the standard significance level set by the researcher (typically .05) and n is the number of tests considered. Thus, as more outcome measures are considered, the significance level required goes to zero very quickly.

Second, as Simmons, Nelson, and Simonsohn (2011) point out, reviewers must be willing to take on the extra burden of downloading preanalysis plans and carefully determining whether the authors were faithful to their proposed design in addition to their other duties. If they do not, preregistration will not provide a reliable check on false positives. Even in the field of neurology, which has a much longer history of preregistration than the social sciences, follow-up studies have shown that 74% of work was never preregistered and the even work with preanalysis plans diverges considerably from the hypotheses and specifications of the preanalysis plan (Rayhill, Sharon, Burch, & Loder, 2015).

How Can Results-Free Peer Review Help?

In addition to considering a set of research designs, we explored an additional and complementary mechanism to address publication bias: results-free review. The special issue authors are in agreement about the merits of

replication, and three out of the four editors have preregistered their own research designs. We believe that results-free review has the ability to complement existing strategies for mitigating publication bias and increasing the transparency of the research process. What our special issue accomplished was reviewing all submissions—designs or completed studies—"results free."

Results-free review consists of authors submitting their manuscripts to the journal devoid of empirical results and then journals reviewing the manuscripts through to an accept–reject decision without ever seeing the results. One format is for a complete paper with all of the details of a normal submission with the exception of the empirical analysis and no mention of the actual results anywhere else in the manuscript. In another format, a manuscript is submitted that is close to a preanalysis plan, describing a study that has yet to be conducted. Both of these formats are "results free" in the sense that the results of the analysis are unknown to the reviewers and editor(s), but in the latter, the results are also unknown to the author(s).

How does results-free review help to solve the problem of publication bias? In short, reviewers assessed whether a theory was innovative, whether empirical tests were appropriate, and whether there were any obvious flaws in the design. If a research plan overcame all of these hurdles, it would be preaccepted for publication. As long as the researchers adhered to their plan, their work would be published regardless of the p values on their key causal variables. The idea behind this process is to encourage researchers to be more open and precise about their design on the front end by liberating them to be as open as possible about the fruits of their work on the back end.

This process can influence the decisions of both authors and reviewers. Authors had full knowledge that their work would be reviewed results free. As our focus was not on author incentives, we leave it up to the reader to provide conjectures on how this type of review shapes initial author decisions. Our focus in this special issue was to examine how reviewers evaluate papers without knowledge of the empirical results. This removes the bias of reviewers wanting to see work that is statistically significant or perhaps even counter-intuitive. In the next section, we discuss our exact process and then provide an evaluation of this process based on interpretation of the reviews.

Before doing so, it is important to emphasize that results-free review addresses only one set of problems that can lead to publication bias: professional incentives to produce significant findings that affect how authors analyze their data. We are skeptical that there is any kind of institutional design that will eliminate manipulation or fraud, and emphasize that norms of scholarly inquiry such as honesty, trust, and acceptance of fallibility are foundational to knowledge accumulation. However, results-free peer review can be helpful even in such a norm-based community, especially

when researchers' and reviewers' incentives lead them to privilege certain kinds of results over others.

Our Process

Our process began with a CFP in which we encouraged two types of submissions for results-free review. (See the appendix for the CFP.) This call was published on the *CPS* website, circulated through numerous email lists, and we wrote a short CFP for the *Washington Post's Monkey Cage*. We attempted to solicit manuscripts on all types of research, substantively and methodologically, that fit within the mandate of *CPS*. The *CPS* editors and special issue editors agreed that decisions on the manuscript would be made before seeing the final results, and the special issue editors would only check manuscripts to ensure they faithfully implemented the tests and analysis that were part of the final results-free submission.

In the first type of submission, we asked for a submission that approximated a preanalysis plan, instructing prospective authors that submissions for this special issue should provide designs that enable a reviewer to assess as fully as possible the theory, main hypotheses, design, feasibility, and potential contributions of the results. In the second type of submission, we invited submissions of otherwise complete manuscripts in which the results and discussion had been removed. For these submissions, the author(s) needed to provide a similar level of detail on the theory, design, *and* credible documentation that the results of the study were not posted or circulated in any way such that a peer reviewer could find and view the results and make a judgment on the paper with conclusions in mind. Preference was given to submissions that had not been previously reviewed at another journal. What united both types of submissions was that *reviewers could not use the results of the analysis to judge the value of the contribution*. The key difference between the two submissions was that the first type had not actually been carried out, whereas the second type had. In the end, this special issue features two articles where the data were only collected and analyzed after peer review (Bush, Erlich, Prather, & Zeira, 2016; Huff & Kruszewska, 2016), and one where the data were collected but the results unknown to the authors and reviewers (Hidalgo, Lima-de-Oliveira, & Canello, 2016).

As special issue editors, we were active in evaluating all manuscripts. Some manuscripts were judged not to fit the special issue, although we were open to any topic relevant to comparative politics. In most cases, these were relatively easy choices, but the harder decisions were about which manuscripts were of sufficient quality and provided enough detail to merit peer review. Given the novelty of this special issue, there seemed to be some

confusion about what constituted a results-free submission. Some submissions were very speculative and provided even less detail on the data collection and analysis plans than a research design section in a regular journal submission. Other submissions were on the topic of research transparency and did not conduct original research that fits within the mandate of *CPS*. The special issue editors read manuscripts and consulted with the standing *CPS* editors in a number of cases. In all, one manuscript was withdrawn by the author, eight manuscripts were desk rejected, and a final 10 manuscripts were sent out for peer review.

The special issue editors also helped with the selection of manuscript reviewers. Reviewers were largely chosen based on the substantive topic of the submitted manuscript, although some reviewers were selected based on their methodological expertise. Again, most of these decisions were easy, and for manuscripts, we had a long list of potential reviewers. Of the total of 43 reviewer requests we sent, 16 declined to review the manuscripts (37% turndown rate). This turndown rate is lower than the average *CPS* turndown rate of 47%.[10]

Reviewers submitted their comments through the regular *CPS* editorial mechanism, and the reviews were then sent to the special issue editors by the *CPS* standing editors. Both sets of editors jointly made the final decisions. Three of the 10 papers sent out for review were offered revise and resubmits. After revisions, all three papers were sent back to the original reviewers who all commented relatively positively, and the papers were then accepted for the special issue.

Once the decision to accept the manuscript had been made, that decision was the near-final decision on the manuscript, subject only to the constraint that the research was executed as planned. We instructed authors that deviations from the accepted research designs were acceptable, but had to be documented rigorously and discussed thoroughly. By asking that authors delineate the alterations made as a result of reviewer suggestions in the final article to clearly and publicly differentiate them from analyses that were preregistered, we gained novel insights into how the peer-review process shapes knowledge production and accumulation in comparative politics.

The Huff and Kruszewska piece is particularly enlightening in this regard. In the published article that follows, they present and interpret their results in line with their preanalysis plan. In addition, however, throughout the manuscript, they document how slightly different specifications from their preanalysis plans (i.e., specifying significance tests at 0.1 rather than 0.05 level or employing different baselines) would have altered their results. In discussing this presentational choice in their comments to the editors, the authors wrote:

> Presenting the results in this way is consistent with the goal of the special issue in promoting research transparency as it allows the maximum opportunity for the reader to draw their own conclusions from our results. Moreover, we think that doing so helps emphasize the importance of results-blind peer review in that it removes the incentive for authors to only present findings that are statistically significant while omitting models that are sensitive to design choices and outcome variable specifications.[11]

Potential Pitfalls and What We Learned

We began the pilot with optimism, yet we knew at the outset there were potential problems with the process. First, we were concerned that reviewers would be unwilling to review manuscripts without results, or that they might provide only cursory reviews not of the same quality as regular reviews. This was clearly not the case, where already burdened reviewers were willing to evaluate these manuscripts, and we were especially impressed with the quality of the reviews. The standing *CPS* editors agreed that these reviews were of higher quality than the average review.

A second concern is that these sorts of new forms of review can have implications for the types of authors willing to submit their work. For example, in a blog about preregistration, Joshua Tucker notes that untenured scholars may feel the most pressure to adhere to stronger norms of research transparency.[12] This could lead to imposing higher costs on more junior researchers, although we note that all of the authors in this special issue are junior scholars. Alternatively, we could observe faculty with tenure willing to embark on more "risky" forms of publication. In the case of this special issue, though, the review process generated submissions from all levels, ranging from graduate students to tenured faculty.

Third, results-free review, and especially preregistered designs, could actually lead scholars to invest *less* in theory development and select research questions that allow for hypotheses in different directions. In plain language, we worried that researchers would focus more on research projects where any empirical tests—positive, negative, or null—are interesting to readers. At the worst, this could lead to a type of hypothesis trolling where researchers propose a laundry list of hypothesis in a preregistration document, assuring themselves that there will be at least some significant results. We could have moved the discipline from data mining to hypothesis trolling.

Ironically, we are limited by research ethics and journal policy on how much of the insights we gained from the review process can be formally documented in this special issue on research transparency. Ideally, we could create an online archive of manuscript submissions, all reviews for the manuscripts, and include direct quotes from these reviews in this introduction.

Without going into detail, it is obvious that this leads to a number of ethical issues in that individual reviewers graciously provided reviews without the knowledge that these reviews would be quoted in this special issue to defend the claims of the special issue editors. As a compromise, we simply summarize reviewer comments in this special issue and provided detailed quotes from which we draw these summaries to the *CPS* editors. Thus, the full manuscripts, reviews, and the passages we are drawing upon have been verified by the *CPS* editors.

Our major concern turned out to have been largely unfounded; hypothesis trolling was specifically targeted and rejected by reviewers. Again, reviewer anonymity and author confidentiality prevent us from revealing specific comments, but reviewers noticed when, for example, manuscripts focused primarily on empirical data and proposed a wide range of theories and hypotheses to anticipate any and all findings. Such observations suggest that hypothesis trolling might be more common than we know in the work that is currently published. One reviewer was moved to comment to us that perhaps most manuscripts begin this way, with the theory being constructed post hoc and only then "sold" to the reader based on the results themselves. One advantage of results-free review over preregistration alone is that it nips this problem in the bud before the authors hit the jackpot on one of many hypotheses and rewrite the paper highlighting only the successful expectation and conclusion.

Our main findings from this exercise are in retrospect intuitive, but they were largely unanticipated. First, we found that reviewers placed a much *greater* focus on theory, the importance of the question, and most notably the relationship between theory and research design. This last point is worth emphasizing as some of our submissions had important theoretical contributions and rigorous research designs, but reviewers consistently commented on weak links between theory and analysis.

Relatedly, reviewers in our pilot insisted on a great deal of country context and knowledge to understand the design choices, adjudicate their importance, and think about external validity. The combination of designs focusing on causal inference and results-free review appeared to emphasize the importance of area-specific knowledge.

Third, reviewers (and the special issue editors) struggled to identify the criteria for which studies would be publishable even with null findings. Which null results are valuable and which can be dismissed due to research design issues? Although null findings have given considerable discomfort to scholars in the social sciences, relatively little discussion exists on how null findings should be treated in the review and publication process.

Finally, we did not receive a single qualitative submission. We attempted to reach out to qualitative researchers through explicitly qualitative research

channels, and were hopeful that we would receive at least some non-quantitative papers to review. Interestingly, our first finding that authors and reviewers valued substantive importance and theory could very well have privileged qualitative work. Alas, we had no submissions of this type and we speculate below as to the causes of this bias.

We flesh out the discussion of each of these issues below

Theory and Substance

The independent evaluations of the four special issue editors were in complete agreement regarding the rigor and focus of the reviews. All four of us were struck by the reviewers' extensive focus on each manuscript's theory and substance. The reviews were in comparable length to a regular journal review but did not have the same focus on the interpretation of results. Reviewers obviously made comments on the methodology, control variables, and issues with the empirical research design. But we judged these reviews as focusing much more on the "substance" of the manuscript and the relationship between the question, the theory, research design, and the potential contribution.

We believe that this outcome could very well be the greatest success of the special issue. Experimentalists, who focus intently on the identification of causal effects, have been a key group pressing for greater transparency, including preregistration and results-free review. And yet scholars point out that theory may be left behind in the race toward better and better identification. John Huber (2013) lamented that a laser like focus on causal identification in research designs might lead scholars to eschew difficult social science questions in favor of queries that allowed for designs more closely approximating randomization. Huber was making a nuanced point, but the article triggered broader water cooler discussions about whether well-identified work was also theoretically grounded. And David Laitin (2013) articulated a related concern that preregistration might undermine the productive feedback loop between empirical research and theoretical exploration. These concerns may be warranted, but the results of this exercise demonstrate that theory need not be lost; indeed, given the right peer-review incentives, theory and substance may carry the day.

Of course, political scientists may disagree on what should be emphasized during the review process, but our special issue clearly demonstrated that this process shifted the focus of these manuscripts toward the substance. By far, the most common concern from our pool of reviewers was inattention to theory. Of course, we observed the common gripes about lack of acknowledgment of the extant literature and limited engagement with major

contributions. More poignantly, however, what became clear is that it was impossible for a research design to be atheoretical and survive results-free review. Every stage of the enterprise from choice of location to operationalization to specification to analysis of heterogeneous effects depends on well-defined theory as the guiding light. Again, anonymity and confidentiality prevent us from providing examples, but time and again, imprecise theory made it impossible for reviewers to determine whether the research designs could help answer the author's ultimate question.

Reviewers were also quick to note when the theory section seemed off the mark, responding to the wrong literature and missing critical antecedents that would make it more broadly appealing. Whether or not journals should implement this process, either for all reviewers or a subset of reviewers for each manuscript, is a question that we cannot answer. But, in our experience as authors and reviewers, the real effort to unpack the theoretical questions as a way of understanding the research design was quite different from what we have seen in our experience writing and reviewing otherwise. In our experience, when results are available, the discussion between authors and reviewers becomes one of "what theory are these results consistent with?" When results are not available, then the theory has to stand on its own. We now conjecture that results-free review could reinforce a more productive interplay between theory and empirics.

The Return of Area Expertise

For decades, there has been a deep tension between students of comparative politics and regional specialists (see, for example, Bates, 1996). Traditional area specialists have criticized mainstream comparative politics, especially large-N cross-national work, as devoid of local context. Many cross-national comparative politics scholars have flipped the criticism on its head, claiming that country experts' work is of little utility beyond their very small and tightly knit community (Pepinsky, 2015).

As with the critical importance of theory, one key lesson of our pilot is that results-free review of field research rewarded greater emphasis on areas studies knowledge as essential for building more compelling research designs. In numerous reviews, referees demanded greater local specificity to understand the implications, internal validity, and generalizability of the design and predicted results. In 100% of the submissions that had a field-based component, the question of whether the authors had the adequate area expertise to carry out and make sense of the research results came up.

This happened in a number of different ways. Some reviewers wondered whether treatments would be effective in a particular country context given

reviewers' specific knowledge of how institutions work there. Others asked about the meaning of key variables in particular national contexts. Others focused on external validity and the broader theoretical impact of findings from a particular country, given the unique climate for an experiment there. These questions even came up in two of the successful manuscripts, forcing the authors to better defend their choices and offer more thorough descriptions of context.

As with theory, it appears to us that reviewers were liberated to challenge researchers on these fundamental questions, because they did not have to deal with the distraction of the empirical results. In the three successful submissions, such focused attention on context and local knowledge led to what we perceive as major improvements in the authors' research designs. Authors were forced to address local nuances that might affect interpretation and choose designs that would best help readers think about generalizing to other contexts. For proponents of a new synthesis of area studies and comparative politics, one that eschews the battles between local context and general social science (see Malesky, 2008; Pepinsky, 2015), this is an encouraging result.

Null Results

The third result of our pilot is so provocative it divides this special issue team. As noted above, numerous reviewers expressed frustration in reviewing work without results, in some cases admitting their own biases, and in other cases making clear that the direction and size of the results are a core part of the intellectual contribution. There are two interrelated problems that the subject of null findings poses for review. The first has to do with acclimating to a new way of thinking about null findings—that they may be meaningful theoretically. The second is the question of what types of null findings are worthy of publication.

It seems especially difficult for referees and authors alike to accept that null findings might mean that a theory has been proved to be unhelpful for explaining some phenomenon,[13] as opposed to being the result of mechanical problems with how the hypothesis was tested (low power, poor measures, etc.). Making this distinction, of course, is exactly the main benefit of results-free peer review. Perhaps the single most compelling argument in favor of results-free peer review is that it allows for findings of non-relationships. Yet, our reviewers pushed back against making such calls. They appeared reluctant to endorse manuscripts in which null findings were possible, or if so, to interpret those null results as evidence against the existence of a hypothesized relationship. For some reviewers, this was a source of some consternation: Reviewing manuscripts without results made them aware of how they were

making decisions based on the strength of findings, and also how much easier it was to feel "excited" by strong findings

This question even led to debate among the special issue editors on what are the standards for publishing a null finding? For example, let us return to the LaCour and Green (2014) paper once again. Imagine that this research was faithfully conducted and submitted without results. Would this paper merit publication in a prominent journal? If our expectation was a null or small impact based on substantial prior research indicating just that, would the study be worth publishing? If we knew there was a large finding, would that change our evaluation of this paper? Again and again, reviewers posed some version of the question: If the tested hypotheses proved insignificant, would that move debates in this subliterature forward in any way? In many of the rejected papers and even one of the accepted papers, the answer was no.

There were three reasons that reviewers reached this conclusion. First, a null finding would not be interesting because the reviewer found the theory to be implausible in the first place. Proving that the implausible was in fact implausible is not a recipe for scintillating scholarship.

The second was a variant of Occam's razor. Reviewers did not believe that the author had adequately accounted for the simpler, alternative theory to explain the underlying puzzle that motivated their research. In this instance, a null result would only reinforce the notion that the more parsimonious theory was superior, or that a natural experiment was confounded by unobservable selection.

Third, there was too much distance between the articulated theory and the abstract field, lab-in-field, or survey experiment articulated in the paper. The theory invoked a compelling concept, but the proposed research design failed to adequately capture it or stretched the meaning of the concept to the point of unrecognizability. In this case, a null result would only prove the empirical test was inadequate for the bigger question. This was a common criticism of experimental research.

None of these dismissals of proposed research plans are new problems or unique to results-free review. They are a standard part of the way scholars evaluate research. The interesting implications for results-free review manifest themselves in how strategic authors may alter their research agenda to survive the review process. Knowing that they have to convince a skeptical reviewer that a null finding is interesting, they may choose to abjure big questions and paradigmatic shifting scholarship for incremental research designs. Remember also that a laundry list of hypotheses and potential heterogeneous effects will not suffice either. Our reviewers were quick to spot and reject this type of hypothesis trolling.

Three author strategies would seem most plausible. First, authors place themselves between two competing theories with contrasting observable implications, posing their research design as the distinguishing test. For example, does fiscal decentralization decrease or increase corruption? Here, a null finding might rule out one of the competing hypotheses.

Second, authors may offer their research design as the first or a better test of prevailing theory or logic that has been inadequately tested in the literature. The theory of deliberative democracy, for instance, offers a number of very clear implications about how deliberation should affect the thinking and behavior of citizens, yet, most of these have been subjected to only limited empirical testing. If designed properly, this would be interesting purely because the potential target would be well known. Again, reviewers reacted quite negatively to this type of approach. Most referees wanted authors to build on the existing literature in important ways or to thoroughly explain why the observational work of previous generations was flawed.

Finally, authors might offer a test of a hypothesis that is the next logical step within a prevailing and well-traveled research paradigm. In the American politics literature, theories regarding voter mobilization efforts and turnout are the closest to the type of incremental progress we have in mind.

All of these strategies would likely fare better in results-free review than a brand new theory, built directly from first principles, or paradigm-shifting theory that challenges the prevailing wisdom in the literature. However, all three approaches are predominantly empirical, building upon existing theory, rather than creating it. In Thomas Kuhn's terminology, results-free review would engender a lot more normal science.

And here is where the disagreement among the co-editors is most severe. Some of the special issue editors applaud this potential trajectory, arguing that it is time that political science de-emphasized grand theorizing, focusing on a gradual accumulation of knowledge that specifically includes a large catalog of theories that have not proved useful. These editors argue that there will always be outlets for big think pieces, but there is still not enough room for the hard, plodding empirical confirmation of the discipline's theorists.

The other co-editors worry about the damaging result this trajectory would have on creative scholarship. They worry that there are still big questions out there to be asked. In fact, as recent events have shown, on some of the most vital questions to mankind such as economic inequality, international immigration patterns, the role of aid in disaster relief, and the resilience of state institutions to global pandemics, the depth of political science scholarship has proved wholly inadequate to society's needs. This is not the time, they argue, to discourage big theory and narrow the lens of the field's most ambitious scholars.

There is one alternative that we have not discussed that may provide a way around the problem of what to do with null results. That is for authors and reviewers alike to abandon null significance hypothesis testing altogether. The conceptual problems with null significance hypothesis testing should be well known to political scientists (e.g., Gill, 1999), but periodic calls for a Bayesian alternative have yet to unsettle long-established practice. In a blog post reflecting on problems of *p*-fishing and experiments, Simon Jackman commented, "From the Bayesian perspective, all this stuff is kind of ridiculously overblown, a consequence of an unthinking acceptance of $p < .05$ as a model for scientific decision making, point null hypothesis testing, the whole box and dice."[14] But the problems are deeper. Jackman invokes Berger and Sellke (1987), who demonstrate that small *p* values do not (necessarily) correspond to strong evidence against a null hypothesis that a parameter is zero. And as is well known, even in the standard frequentist setting, large *p* values are not evidence that a parameter *is* zero.

If reviewers and authors did not attribute substantive meaning to tests of statistical significance then there would be no statistical significance filter. What would replace null significance hypothesis testing remains unknown. But we emphasize that authors and reviewers used statistical significance as a shorthand for adjudicating whether effects exist or not. This indicates to us that getting away from the very premise that there is such as thing as "null" results (to say nothing of "non-results") will require a significant departure from current practice. Perhaps one result of our pilot study is to highlight not just the practical difficulties that reviewers face with null results, but the conceptual and theoretical problems with null results that extend to the vast majority of published research in political science.

Method

A final observation is that our special issue generated a very particular type of submission. The vast majority of submissions that we received were for survey or field experiments, and the remainder involved the statistical analysis of quantitative data. We received no submissions of qualitative case studies, historical comparisons, or ethnographic research.[15]

Why would this be? It is not possible to answer this question definitively based on our own experiences, but there are at least four possibilities. One is that our CFP happened to have been read primarily by people working in the new experimental tradition in political science. If so, and despite our efforts to the contrary, we simply failed to reach out broadly enough to include a representative sample of research in contemporary political science.

Another is that authors using different kinds of methodologies saw our announcement, but believed that we were looking primarily for experimental, or at least statistical, research. We did not intend to elicit only statistical or experimental submissions, but we also did not take extra steps to encourage specifically qualitative methods in our submissions. Although we did attempt to reach out to a group that coalesces around the study and practice of qualitative methods, the effort appears not to have been sufficient. This would be our own failure and not a limitation of results-free review.

A third reason why we did not receive qualitative submissions may stem from the reputation of *CPS* as a quantitative journal. The journal maintains no explicit policy about methodology, and is working to change the reputation, but it nonetheless is still largely known as the central quantitative journal in comparative politics. As qualitative researchers are well represented within comparative politics, a large number of possible submitters may have been influenced by the journal's reputation.

Still another is that qualitative case studies, comparative historical analyses, and other similar types of research *cannot be preregistered* and that *results cannot be removed from case studies*. These are types of research in which scholars generally accept that theories and arguments are informed by the interaction between a researcher's initial hypotheses—in some cases little more than hunches—and the specifics of a case. We believe this type of work is a valuable contribution to political science scholarship, but we can imagine the complexity of submitting this work preregistered and/or results free.

The core principle of preregistration, that hypotheses must be specified before the researcher collects and analyzes the data, is simply incompatible with approaches that prioritize the reciprocal engagement between theory and evidence. This seems especially difficult for qualitative and historical types of research. More fundamental to our special issue, where half of the manuscripts were not preregistered, all of the papers were submitted results free. The premise of results-free peer review—that it is possible to describe a research enterprise without reference to the data it produces—is inconsistent with the way that we actually conduct qualitative comparative analyses and in-depth case studies (see, for example, Yom, 2015). Importantly, our argument is not that such research is unscientific—it certainly can be, and in fact, such research can fit well within a positivist epistemology. The point is simply that inductive research, and various kinds of mechanismic-centered qualitative and historical research, cannot be described without reference to the data.

The special issue editors all have different kinds of methodological expertise, but for those of us who have worked with historical and archival materials, and who have paired these with in-person interviews with important policy makers, the tensions between preregistration and results-free peer

review, on one hand, and careful qualitative research, on the other, seem insurmountable. Although we have no reason to conclude that results-free peer review must prevent such research, such that shared standards can never be developed, results-free peer review would likely have serious implications for current practice.

Specifically, we suggest that results-free peer review has an affinity for a normal science view of social scientific research. Results-free peer review is most feasible when authors are working within established research traditions that work with clear and long-established hypotheses. It also has an affinity for research that uses formal analytical tools to generate deductive hypotheses. In both of these cases, a positivist epistemology undergirds the research enterprise. It is for this reason, we suggest, that experimental methods were particularly attractive to authors who submitted manuscripts to *CPS*; they themselves have a natural affinity for estimating causal effects from designs drawn from well-established theory.

By contrast, it is difficult to see how interpretivist and other post- or non-positivist epistemologies would work with results-free review. A core feature of interpretivism is the rejection of any strict distinction between theory and data, so that the struggle for many interpretivists is to leave aside theories and assumptions about the social world. Especially in ethnographic or hermeneutic research, the research enterprise seeks to uncover how meaning is made, or to come to understand the lived experiences of interlocutors or the texts that they have produced. Only through research itself do these meanings become clear. It is certainly possible to plan an ethnographic project, or to list a series of texts or archives one plans to consult, but it makes little sense to list hypotheses and the data to be collected to test them, because neither the hypotheses nor the data can be known in advance.

Case study research in the standard positivist mode lies in between these two extremes. We find it useful to draw on Lieberman's (2005) distinction between "model-building" and "model-testing" small-n analyses. The former is characterized by much less certainty about the theory or the data—in Lieberman's words, "the scholar engaged in [model-building small n analysis] does *not* proceed with the notion that a fully specified model is available and must develop explanations for the puzzle of varied outcomes" (pp. 443). This type of inductive, exploratory research fits less obviously with a results-free model of peer review—even when the research is complete and has been written up—because describing the qualitative data being used may itself be part of the process that generates the new theoretical insight. By contrast, model-testing small-n analysis that draws on cross-case statistical findings to justify intensive study of whether particular cases are consistent with those findings fits more naturally in the experimental or observational templates

described above. Although we did not encounter any such submissions, we find it easier to conceive of results-free description of such a case study or historical analysis.

Future research by qualitative methodologists into the possibility of preregistering historical, ethnographic, or otherwise exploratory research may help to tell us whether the very nature of our special issue itself discouraged qualitative submissions. But we see here a parallel with the tensions that we identified in the previous section on null results. Some political scientists may welcome preregistration precisely because it places greater emphasis on political science as a normal science. Others will see that as a substantial drawback.

It is hard to escape the conclusion, though, that any requirement that research manuscripts have been preregistered will almost certainly affect the types of submissions that a journal receives. One possible consequence is a bifurcation of publication outlets, and as a result, of researchers. One set of researchers adheres strictly to a normal science template to produce manuscripts that are eligible for journals that insist on results-free review, while others adhere to and are assessed on a very different set of standards in a different set of journals. For the discipline as a whole, this would almost certainly generate divisions and inequalities.

An Overview of the Articles

Of the 19 articles originally submitted, and then the 10 sent out for review, three submissions were granted revise and resubmit status. The authors of the three papers made revisions to the designs or results-free submissions and then resubmitted, and each was subsequently conditionally accepted for publication. The papers were accepted conditional only on finalizing the papers without any deviations so large as to be out of the spirit of what was reviewed by the referees. Otherwise, regardless of whether results came back weak or strong substantively, be they null, positive, or negative, the papers would still be published. We emphasize that the submissions were conditionally accepted prior to any results being available to the standing CPS editors, to the special issue editors, and to all of the anonymous reviewers. The three accepted papers appearing in this volume are Bush et al. (this issue) "The Effects of Authoritarian Iconography: An Experimental Test," Hidalgo et al. (this issue) "Can Politicians Police Themselves? Natural Experimental Evidence from Brazil's Audit Courts," and Huff and Kruszewska (this issue) "Banners, Barricades, and Bombs: The Tactical Choices of Social Movements and Public Opinion."

In their paper, Bush et al. experimentally examine the effects of authoritarian images, or iconography, on citizen behavioral compliance and regime

support. Tying into the broader literature on authoritarian survival and cults of personality, the authors contend that the use of iconography—posters, sculptures, seals, or insignias bearing images of authoritarian leaders—should work through mechanisms of legitimacy, self-interest, and coercion to generate greater compliance and regime support among citizens. Bush et al. carried out their laboratory experiment in the United Arab Emirates, a rising regional, authoritarian power. The authors submitted a research design for the special issue and only carried out the actual experiment once the review process was complete and they received a conditional acceptance. They find no evidence that iconography affects support for the Emirati regime in their experimental analysis. The authors probe the null result with additional tests, finding that standard explanations such as insufficient power or measurement error are unlikely explanations for the lack of significance. As a result, this article is an excellent example of a research project demonstrating that existing theoretical expectations may have been overblown.

Hidalgo, Lima, and Canello examine horizontal accountability mechanisms for government political institutions. They explore how differential assignment to public audit courts affects punishment of lawmakers explicitly leveraging a natural experiment whereby government agencies and subnational governments are assigned by lottery to municipality-level councilors. This paper was submitted after data had been collected but before it had been analyzed. In lieu of results, the authors provided a detailed preanalysis plan identifying how each hypothesis would be tested so that the reviewers and editors would have a comprehensive picture of all procedures. Once conditionally accepted, the authors then added the results, showing that when institutions shield auditors from political interference, they punish lawbreaking politicians more than do auditors who do not enjoy such insulation.

The Huff and Kruszewska study engages the broad contentious politics literature with an examination of how the degree of extremeness of tactical choice influences public opinion about a movement. Although this central question has long been a concern among scholars, identifying the effects of tactical choice is an exercise laden with pitfalls. The authors use a novel survey experiment design in Poland that systematically varies tactical choice and then considers the extent to which subjects believe that the government should negotiate with a movement as well as how many concessions should be offered. The authors' submission was a research design that had yet to be implemented and the authors updated the design in response to the reviewers' and editors' feedback. The authors set up a theoretical horse race among three approaches: the benefits of extremism, the benefits of moderation, and the no-concessions policy. Supplementing standard statistical techniques with a

structural topic model that allowed for text analysis, the results are broadly consistent with the no-concessions theoretical approach.

Final Thoughts

In many ways, this special issue exceeded our expectations. First, *CPS* will publish three excellent studies that we believe will have an impact on the discipline. This is obviously first and foremost due to the authors' hard work, but we also credit the reviewers for their contributions. Our subjective evaluation was that the reviewers provided extremely high-quality reviews and most likely set the bar for these papers higher than for the normal submission. Whether or not raising this bar is advisable is debatable, but we can clearly state that this form of review lead to papers that were of the highest quality. We would love to see a top journal adopt results-free review as a policy, at very least allowing results-free review as one among several standard submission options. In thinking about whether the pilot could be applied more generally, and reflecting on some of the logistical issues that we faced in this process, three practical considerations arose.

First, this form of review could lead to new incentive problems. The least serious of which is that journals that accept results-free papers might be flooded with null results as academics open up their file drawers. This could be a problem for an individual journal where this form of review leads it to specialize in null results, and that reviewers, knowing it is results-free, have priors that the submitted paper probably has weak or null results. Although this may lead to some immediate problems as journal pipelines get flooded, overall it should correct for the biased distribution in p values and open up outlets for high-quality papers with null findings. We note that for the three papers published here, that one of them had null results as the main finding, although this result was not known to the authors at the time of submission.

Second, as a discipline, we might want to rethink when it is appropriate to present and circulate work. The requirement that papers are not fielded or have not been presented put the work submitted to the special issue at an enormous disadvantage. Polished working papers frequently have benefited from insights of discussants and participants, are often sent to peer-reviewed journals and are rejected by the first journal, (hopefully) revised based on the reviewer criticisms, and then submitted to another journal. The submissions to this special issue did not benefit from this same cycle of presentation, feedback, and revision, leading at least some initial submissions to be more "green" than what we imagine is normally submitted to *CPS*. The trade-off of an author hiding results from reviewers is that they may also hide the work from the research community, limiting the amount of feedback they receive,

and possibly affecting the amount of scholars who know about this work even after publication. Currently, many papers become well known even before they are published, which may or may not be desirable, but that might not continue under a largely results-free review standard. This concern is not as serious for preregistered designs that would be reviewed results free—in those cases, one could imagine scholars circulating the designs widely with no fear of results getting out. To the extent that designs were to be accepted ahead of conducting the research, this could mitigate the problem of publication bias as well as under developing papers. However, for other types of research, there is a clear trade-off, and results-free review will not work if authors are forced to strip out results from papers that have already been circulated.

Third, for preregistered studies, a results-free review mechanism could add yet another bottleneck to an already long research process. Field researchers often have to develop a project idea, conduct a pilot study, secure funding, go to the field and implement the research, and write up the results. And this sentence even sugarcoats the difficult road all researchers travel in the timing of research projects, where formal deadlines (grant applications, tenure clocks) all to some extent dictate the timing of projects, the academic calendar and teaching obligations shape when time is available, and personal factors, such as the timing of breaks in child care, require a Tetris style mastery of scheduling to pull off successfully.

Field researchers submitting preregistered designs, if they followed our framework, must now wait for peer reviews and editorial decisions before they can move onto the next step. For researchers who have not fielded a study, they must wait for suggestions that may completely reshape their proposed study. For researchers who have already fielded the study, they may be waiting for reviewers before they even begin a preliminary analysis of the data. If the author receives a rejection from the journal, she or he may consider sending their proposal to another journal, further pushing back the fielding of their project. This process could add years to the time between idea generation and the beginning of the field research. So it is important to note that what we perceived as the gold standard of transparency (submission of the design prior to fielding the survey) imposed clear costs on the researcher.

On the flip side, there may be some benefits only possible with this style of review. For one, currently when a scholar carries out a research project, there is often a very long delay before those findings get published. Some may argue, drawing on ethical principles, that policy-relevant research should be made available as soon as possible. If the review process is fully complete before the research is executed, then the study can be put into the publication pipeline almost immediately following the time that researchers carry out the

study, substantially increasing the timeliness of research for the political problems under study. Moreover, if a scholar's design is already accepted for publication, they may find it easier to secure funding if they have not already done so. And the difficulty of carrying out the research may be lessened if organizational partners and other stakeholders are aware that the research will not only eventually see the light of day but will do so in the near future.

These are practical considerations. But moving forward, our reflections in this essay do raise some consistent themes.

1. Is a full recording of all steps of a research project, from conceptualization to empirical testing, actually possible? What is the limit to the information that we wish to record, and how much of the intellectual architecture should peer reviewers have access to when evaluating a manuscript?
2. How should we interpret null results? Why are authors and reviewers alike so willing to accept the null hypothesis significance testing paradigm, yet reluctant to conclude that insignificant results are evidence against particular hypotheses? Might a Bayesian framework provide an alternative foundation for hypothesis testing, one that puts more structure on hypothesis testing as a decision problem rather than declaring results to be significant or not?
3. To what extent is the affinity of results-free peer review with a normal science view of comparative politics inevitable? Can inductive and exploratory research fit within such a research paradigm?
4. Is political science mature enough of a discipline that researchers should only ask questions in which any results, null or otherwise, are interesting? Does this vary by subfield, topic area, or research question?

Both supporters and critics of results-free peer review will benefit from keeping these in mind.

Appendix

Research Transparency in the Social Sciences

Issue editors: Michael G. Findley, Nathan M. Jensen, Edmund J. Malesky, and Thomas B. Pepinsky.

We invite proposals for a special issue of *Comparative Political Studies* (*CPS*) on research transparency in the social sciences. Proposals for original research papers using quantitative or qualitative approaches, and collecting

quantitative or qualitative data are all encouraged. The deadline for submitted proposals is October 15, 2014, for a Special *CPS* issue scheduled to appear in 2015-2016 academic year.

There is growing momentum in the natural and social sciences for greater transparency in research. For example, see the Evidence in Governance and Politics Network (EGAP; www.egap.org) and the Berkeley Initiative for Transparency in the Social Sciences (BITSS; www.bitss.org). Although there are varied objectives driving the shift toward greater transparency, one of the key motivations is to avoid publication bias, the result of peer-review processes that privilege the significance of results over the theoretical contribution or integrity of the research design. On the contrary, critics of preregistration argue that it can handcuff authors, leading to journals filled with projects that are less theoretically innovative and path breaking than would otherwise be possible.

This Special Issue of *CPS* will help to assess the potential benefits and costs associated with new models of the publication process by studying how new models can work in practice. Transparency should obviously be a central objective in contemporary social science, but what are the costs? Do strict preregistration protocols commit scholars to carry out projects that are unfeasible, or dissuade creative dialogue between theory and data? Is it possible for a full recording of all steps of a research project, from conceptualization to empirical testing? How will manuscript referees respond to manuscripts without results or conclusions? These questions cannot be settled in the abstract.

The Special Issue aims to study the role of full transparency in research in two ways: (a) accepting work based on prospective research designs, and (b) opening up field notes and last-minute alterations in the research design through online archiving (as with replication data). Articles in the Special Issue will be bookended by two articles by the editors, which introduce the goals of the project and critically evaluate the pros and cons of preregistration and research transparency in political science.

To this end, we invite one of two types of submissions:

1. Full research designs for *prospective* research projects that have not yet been conducted.
2. Full research designs for projects that have already been conducted, and for which any discussion of results has been stripped out of the manuscript.

If the first type of submission, the design needs to be a thorough project prospectus, sometimes referred to as a preanalysis plan. Although there are

multiple ways to construct a preanalysis plan, submissions for this special issue should provide designs that enable a reviewer to assess as fully as possible the theory, main hypotheses, design, feasibility, and potential contributions of the results. This information should be sufficient to allow reviewers to reach a firm conclusion on the project, and ultimately accept or reject the project for publication in the special issue.

If the second type of submission, the author(s) need to provide a similar level of detail on the theory, design, *and* credible documentation that the results of the study are not posted or circulated in any way such that a peer reviewer could view the results and make a judgment on the paper with conclusions in mind. Preference will be given to submissions that have not been previously reviewed at another journal.

Once the designs have been submitted, they will be sent out for full peer review. Designs will be accepted, rejected, or invited to make revisions with resubmission. Once a determination has been made on the design, that decision will be the near-final decision on the manuscript, subject only to the constraint that the research is executed. Deviations from the accepted research designs are acceptable, but need to be documented rigorously and discussed thoroughly. In fact, it is expected that authors of projects that have already been conducted will be asked by reviewers to perform analyses outside of their initial research protocol. This is a normal part of the peer-review process: We ask that authors delineate the alterations made as a result of reviewer suggestions in the final article to clearly and publicly differentiate them from analyses that were preregistered. This will provide the editors with unique insight into how the peer-review process shapes scientific knowledge and accumulation.

Authors of research papers that are invited to move forward with publication will need to make available all background documents including coding notes, full replication files, and so on. To facilitate this process, authors will be eligible for a US$5,000 grant provided through the University of Texas at Austin to offset the costs of gathering and making available the required documents, notes, and so on.

As with regular submissions, the *CPS* permanent editors will make a definitive acceptance or rejection based on how authors address the reviewers' comments, but will not make an independent evaluation of the paper based on the final results.

Proposals should follow the standard *CPS* submission requirements for normal articles, but should be submitted to directly to the special issue editors at transparency@ipdutexas.org. Please indicate "*CPS* Special Issue Submission" in the subject line. We encourage you to contact the special issue editors if you have any questions at the above email.

Acknowledgments

We thank the standing *Comparative Political Studies* (*CPS*) editors, David Samuels and Ben Ansell, for supporting this new and unique special issue on preregistration and results-free review. As this had never been tried in political science, we recognize they put in a great deal of effort accommodating the new challenges of this review process.

Declaration of Conflicting Interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) received no financial support for the research, authorship, and/or publication of this article.

Notes

1. Study registries are also valuable to help document research projects that did not result in publication. These design registries can be valuable in documenting which studies do not ultimately get written up as part of a research project. In this introduction, we do not focus on this important question and rather address how registration or results-free review affects the evaluation of manuscripts.
2. See a recent discussion in the *Economist* for some examples: "Trouble at the Lab" (2013).
3. We discuss this in more detail below.
4. A follow-up study by Broockman and Kalla (2016) finds large canvasing effects but this is unrelated to the gender identity of the canvaser.
5. Six common *p*-hacking tactics are as follows: (a) stopping data collection once $p < .05$; (b) analyzing many measures but report only those with $p < .05$; (c) using many specification but only report those with $p < .05$; (d) using covariates to get $p < .05$; (e) excluding participants to get to $p < .05$; and (f) transforming the data to get $p < .05$ (Simonsohn, Nelson, & Simmons, 2014).
6. There have been research fraud scandals in several disciplines, but one of the most public involving political science is LaCour and Green (2014); see above for discussion.
7. The issue of significance versus insignificance should also be separated from the issue of whether a result is close to or far from zero. On this point, see Hartman and Hidalgo (2015).
8. For an excellent overview of this controversy, see Cassidy (2013).
9. At the time of writing this paper, the "first place" author had 183 retractions. See http://retractionwatch.com/the-retraction-watch-leaderboard/
10. We thank the *Comparative Political Studies* (*CPS*) editors for providing this data.

11. Correspondence with editors (July 28, 2015).
12. See http://blog.oup.com/2014/09/pro-con-research-preregistration/
13. We do not use the language here of a theory being "wrong" or "false," because theories are neither true nor false (Clarke & Primo, 2012).
14. The post is available here: http://webcache.googleusercontent.com/search?q=cache:DVdwW5x20DcJ:jackman.stanford.edu/blog/%3Fp%3D2708&hl=en&gl=us&strip=1&vwsrc=0
15. We did receive a few proposals for theoretical analyses of important concepts in political science. These were uniformly of a lower quality than our other submissions, and were from scholars who had not yet received a PhD, so we hesitate to conclude much about them.

References

Bates, R. H. (1996). Letter from the president: Area studies and the discipline. *American Political Science Association–Comparative Politics*, *7*(1), 1-2.

Berger, J. O., & Sellke, T. (1987). Testing a point null hypothesis: The irreconcilability of p values and evidence. *Journal of the American Statistical Association*, *82*, 112-122.

Broockman, D. E., & Kalla, J. L. (2016). Durably reducing transphobia: A field experiment on door-to-door canvassing. *Science*, *352*, 220-224.

Broockman, D. E., Kalla, J. L., & Aronow, P. (2014). *Irregularities in LaCour (2014)*. Unpublished manuscript, University of California, Berkeley.

Bush, S., Erlich, A., Prather, L., & Zeira, Y. (this issue). The effects of authoritarian iconography: An experimental test. *Comparative Political Studies*.

Carey, B. (2011, November). Fraud case seen as a red flag for psychology research. *The New York Times*. Retrieved from http://www.nytimes.com/2011/11/03/health/research/noted-dutch-psychologist-stapel-accused-of-research-fraud.html?_r=1

Casey, K., Glennerster, R., & Miguel, E. (2012). Reshaping institutions: Evidence on aid impacts using a preanalysis plan. *Quarterly Journal of Economics*, *127*, 1755-1812.

Cassidy, J. (2013, April). The Reinhart and Rogoff controversy: A summing up. *The New Yorker*. Retrieved from http://www.newyorker.com/rational-irrationality/the-reinhart-and-rogoff-controversy-a-summing-up

Clarke, K., & Primo, D. (2012). *A model discipline: Political science and the logic of representations*. Oxford, UK: Oxford University Press.

De Angelis, C., Drazen, J. M., Frizelle, F. A., Haug, C., Hoey, J., Horton, R., …Van Der Weyden, M. B. (2004). Clinical trial registration: A statement from the international committee of medical journal editors. *The New England Journal of Medicine*, 351, 1250-1251.

Dickersin, K. (1990). The existence of publication bias and risk factors for its occurrence. *Journal of the American Medical Association*, *263*, 1385-1389.

Dickersin, K., & Rennie, D. (2003). Registering clinical trials. *Journal of the American Medical Association*, *290*, 516-523.

Findley, M. G., Nielson, D. L., & Sharman, J. C. (2013). Using field experiments in international relations: A randomized study of anonymous incorporation. *International Organization*, *67*, 657-693.

Findley, M. G., Nielson, D. L., & Sharman, J. C. (2014). *Global shell games: Experiments in transnational relations, crime, and terrorism*. Cambridge, UK: Cambridge University Press.

Findley, M. G., Nielson, D. L., & Sharman, J. C. (2015). Causes of non-compliance with international law: Evidence from a field experiment on financial transparency. *American Journal of Political Science*, *59*, 146-161.

Finkel, S. E., Pérez-Liñán, A., & Seligson, M. A. (2007). The effects of U.S. foreign assistance on democracy building, 1990-2003. *World Politics*, *59*, 404-439.

Franco, A., Malhotra, N., & Simonovits, G. (2014). Publication bias in the social sciences: Unlocking the file drawer. *Science*, *345*, 1502-1505.

Gelman, A., & Carlin, J. (2014). Beyond power calculations: Assessing type S (sign) and type M (magnitude) errors. *Perspectives on Psychological Science*, *9*, 641-651.

Gerber, A., & Malhotra, N. (2008). Do statistical reporting standards affect what is published: Publication bias in two leading political science journals. *Quarterly Journal of Political Science*, *3*, 313-326.

Gerber, A., Malhotra, N., Dowling, C., & Doherty, D. (2010). Publication bias in two political behavior literatures. *American Politics Research*, *38*, 591-613.

Gill, J. (1999). The insignificance of null hypothesis significance testing. *Political Research Quarterly*, *52*, 647-674.

Glewwe, P., & Kremer, M. (2006). Schools, teachers, and education outcomes in developing countries. In E. Hanushek & F. Welch (Eds.), *Handbook of the economics of education* (Vol. 2, pp. 945-1017). New York: Elsevier.

Gottlieb, J. (2016). Greater expectations? A field experiment to improve accountability in Mali. *American Journal of Political Science*, *60*, 143-157.

Hartman, E., & Hidalgo, F. D. (2015). *What's the alternative? An equivalence approach to balance and placebo tests*. Unpublished manuscript, Princeton University, NJ.

Herndon, T., Ash, M., & Pollin, R. (2014). Does high public debt consistently stifle economic growth? A critique of Reinhart and Rogoff. *Cambridge Journal of Economics*, *38*, 257-279.

Hidalgo, F.D., Lima-de-Oliveira, R., & Canello, J. (this issue). Can politicians police themselves? Natural experimental evidence from Brazil's audit courts. *Comparative Political Studies*.

Huber, J. (2013, June). Is theory getting lost in the "identification revolution"? *The Political Economist*, pp. 1-3.

Huff, C., & Kruszewska, D. (this issue). Banners, barricades, and bombs: The tactical choices of social movements and public opinion. *Comparative Political Studies*.

Humphreys, M., de, la, Sierra, R. S., & van der Windt, P. (2013). Fishing, commitment, and communication: A proposal for comprehensive nonbinding research registration. *Political Analysis*, *21*, 1-20.

Ioannidis, J. (1998). Effect of the statistical significance of results on the time to completion and publication of randomized efficacy trials. *Journal of the American Medical Association*, *279*, 281-286.

LaCour, M., & Green, D. (2014). When contact changes minds: An experiment on transmission of support for gay equality. *Science*, *346*, 1366-1369.

Laitin, D. (2013). Fisheries management. *Political Analysis*, *21*, 42-47.

Lieberman, E. S. (2005). Nested analysis as a mixed-method strategy for comparative research. *American Political Science Review*, *99*, 435-452.

Malesky, E. J. (2008). Battling onward: The debate over field research in developmental economics and its implications for comparative politics. *Qualitative Methods*, *6*(2), 17-21.

Miguel, E., Camerer, C., Casey, K., Cohen, J., Esterling, K. M., Gerber, A., … Van der Laan, M. (2014). Promoting transparency in social science research. *Science*, *343*, 30-31.

Monogan, J. E. (2015). Research preregistration in political science: The case, counterarguments, and a response to critiques. *Political Science & Politics*, *48*, 425-429.

Moravcsik, A. (2014). Transparency: The revolution in qualitative research. *Political Science in Politics*, *47*, 48-53.

Neumark, D. (2001). The employment effects of minimum wages: Evidence from a prespecified research design the employment effects of minimum wages. *Industrial Relation*, *40*, 121-144.

Nielsen, R., Findley, M. G., Candland, T., Davis, Z., & Nielson, D. L. (2011). Foreign aid shocks as a cause of violent armed conflict. *American Journal of Political Science*, *55*, 219-232.

Nuzzo, R. (2014). Scientific method: Statistical errors. *Nature*. Retrieved from http://www.nature.com/news/scientific-method-statistical-errors-1.14700

Nyhan, B. (2015). Increasing the credibility of political science research: A proposal for journal reforms. *PS: Political Science and Politics 48*(S1): 78–83.

O'Connor, A. (2016, April 13). A decades-old study, rediscovered, challenges advice on saturated fat. *The New York Times Well*. Retrieved from http://nyti.ms/1SynN1M

Open Science Collaboration. (2015). Estimating the reproducibility of psychological science. *Science, 349*. doi:10.1126/science.aac4716

Pepinsky, T. B. (2015). Context and method in southeast Asian politics. *Pacific Affairs*, *87*, 441-461.

Rayhill, M., Sharon, R., Burch, R., & Loder, E. (2015). Registration status and outcome reporting of trials published in core headache medicine journals, *Neurology*, *85*, 1789-1794. doi:10.1212/WNL.0000000000002127

Reinhart, C. M., & Rogoff, K. S. (2010). Growth in a time of debt. *American Economic Review: Papers and Proceedings*, *100*, 573-578.

Simmons, J. P., Nelson, L. D., & Simonsohn, U. (2011). False-positive psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychological Science*, *22*, 1359-1366.

Simonsohn, U., Nelson, L. D., & Simmons, J. P. (2014). P-curve: A key to the file-drawer. *Journal of Experimental Psychology: General*, *143*, 534-547.

Sterling, T. (1959). Publication decisions and their possible effects on inferences drawn from statistical tests—Or vice versa. *Journal of the American Statistical Association*, *54*, 30-34.

Trouble at the lab. (2013, October 19). *The Economist*. Retrieved from http://www.economist.com/news/briefing/21588057-scientists-think-science-self-correcting-alarming-degree-it-not-trouble

Welzel, C., & Inglehart, R. (2009). Mass beliefs and democratic institutions. In C. Boix & S. C. Stokes (Eds.), *The Oxford handbook of comparative politics* (pp. 297-316). New York, NY: Oxford University Press.

Yom, S. (2015). From methodology to practice: Inductive iteration in comparative research. *Comparative Political Studies*, *48*, 616-644.

Yong, E. (2012). Nobel laureate challenges psychologists to clean up their act. *Nature*. Retrieved from http://www.nature.com/news/nobel-laureate-challenges-psychologists-to-clean-up-their-act-1.11535

Author Biographies

Michael G. Findley is associate professor in the Department of Government at University of Texas at Austin. His research addresses political violence, international development, and international law, and has appeared in *Cambridge University Press, American Journal of Political Science, International Organization*, among others.

Nathan M. Jensen is professor in the Department of Government at University of Texas at Austin. His research interests include multinational enterprises and political risk, the relationship between foreign direct investment and corruption, and tax competition for investment.

Edmund J. Malesky is associate professor in the Department of Political Science at Duke University. His research interests include the political development in Vietnam and China, comparative political economy in Southeast Asia, as well as economic transitions in developing economies. His work appears in the *American Political Science Review, Journal of Politics*, and other venues.

Thomas B. Pepinsky is associate professor in the Department of Government at Cornell University. Currently, he is working on issues relating to Islam, politics, and political economy in Southeast Asia and beyond. His work has appeared in *Cambridge University Press, American Journal of Political Science, World Politics*, and other venues.

Original Article

The Effects of Authoritarian Iconography: An Experimental Test

Comparative Political Studies
1–35

DOI: 10.1177/0010414016633228
cps.sagepub.com
SAGE

Sarah Sunn Bush[1], Aaron Erlich[2], Lauren Prather[3], and Yael Zeira[4]

Abstract

Do public images of state leaders affect individuals' political attitudes and behaviors? If so, why do they have that effect and among whom? Authoritarian iconography could increase compliance with and support for the state via three causal mechanisms: legitimacy, self-interest, and coercion. This article uses a laboratory experiment in the United Arab Emirates to evaluate the effect of public images of state leaders on individuals' compliance with and support for an authoritarian regime. Using a pre-registered research design, it finds no meaningful evidence that authoritarian iconography increases political compliance or support for the Emirati regime. Although these null results may be due to a number of factors, the findings have important implications for the future research agenda on how and why authoritarian leaders use political culture to maintain power.

Keywords

non-democratic regimes, experimental research

[1]Temple University, Philadelphia, PA, USA
[2]University of Washington, Seattle, USA
[3]University of California, San Diego, USA
[4]University of Mississippi, Oxford, USA

Corresponding Author:
Sarah Sunn Bush, 1115 Polett Walk, Gladfelter Hall 455, Philadelphia, PA 19122, USA.
Email: sarah.bush@temple.edu

Introduction

Does authoritarian iconography encourage the public to obey and support the state? Many observers remark that leaders' images, which are often omnipresent in authoritarian states, must clearly perform important functions. As Daniel Kahneman (2011) puts it,

> Some cultures provide frequent reminders of respect, others constantly remind their members of God, and some societies prime obedience by large images of the Dear Leader. Can there be any doubt that the ubiquitous portraits of the national leader in dictatorial societies not only convey the feeling that "Big Brother Is Watching" but also lead to an actual reduction in spontaneous thought and independent action? (p. 56)

Inspired by the question about the role authoritarian iconography plays, this article uses a laboratory experiment in the United Arab Emirates (UAE) to evaluate the effects of public images of state leaders on individuals' compliance with and support for the state. Although there is a long tradition in political science of studying political culture, personality cults, and authoritarian iconography as sources of authoritarian rule, our study is the first to use experimental methods to examine the effects of authoritarian iconography, and one of the only studies that uses experimental methods to evaluate the effect of any authoritarian strategies on individuals' behavior (see also Lawrence, 2013). The use of experiments to address this research question is important because, although prior research treats symbolic displays of power as "independent variables" (Wedeen, 2010, p. 261), these displays are not randomly assigned to individuals. Instead, they are the deliberate results of political action on the part of both the state and individuals. Because leaders' symbols are displayed as a consequence of deliberate political actions, inferences about their effects may be biased if the researcher does not control for variables that could confound the relationship between images and political compliance and support. Experimental methods can help overcome this problem.

Our study randomly assigns some participants in the laboratory experiment to be exposed to images of Sheikh Khalifa bin Zayed Al Nahyan (Sheikh Khalifa), the ruler of the UAE, on their computer screens. This process allows us to test two hypotheses: participants will behave more compliantly when exposed to the leader's image, and participants will support the regime more when exposed to the leader's image. As we explain below, compliance and support often, but not always, go together; although an individual's support for the rules and institutions of the state typically implies her compliance with its rules and institutions, an individual's compliance does not necessarily imply her support. Finally, we also differentiate the possible impact of the leader's

image in particular from that of surveillance in general through a second treatment in which subjects are exposed to an image of stylized eyes in addition to the leader's image.

We measure compliance and support using a combination of behavioral and attitudinal indicators. In one prominent previous study of authoritarian iconography, Wedeen (1999) measures political compliance using evidence from individuals' "hidden transcripts" (Scott, 1990)—their private stories, ironical jokes, cartoons, and speculations. Yet these obeisant semiotic practices, which are often fictitious, are challenging to replicate. Moreover, they represent just some of the many potential outcomes that may be affected by political symbols. Therefore, we propose four new measures. First, we measure compliance using subjects' behaviors in a compliance game borrowed from behavioral economics in which participants are given a sum of money and asked to report honestly how much they have received so that their income can be taxed. Second, we measure compliance using subjects' willingness to donate some of their winnings to a charity to which they are (truthfully) told Sheikh Khalifa has directed people to donate. Third, we measure compliance using political attitudes questions. Finally, although some of these measures of compliance capture aspects of support, we specifically measure regime support using subjects' attitudes toward policies that they are (truthfully) told Sheikh Khalifa has endorsed.

Contrary to the existing literature and our own expectations, we do not find evidence that images of the UAE's leader affect political compliance or support for the regime among UAE residents. After correcting for multiple comparisons, we do not find any statistically significant differences between subjects exposed to images of the leader and subjects in the control group on any of our measures of political compliance or support.[1] We do find some statistically significant differences in support for the regime between subjects exposed to the second treatment, which combines images of the leader with those of stylized eyes, and subjects in the control group. However, these differences are in the opposite of the expected direction: Subjects who were randomly exposed to the combined leader and eyes image were significantly *less* likely to express support for regime-endorsed policies. The results of a power analysis indicate that these findings are not due to a mere shortage in sample size although they may be due to other features of our research design such as its laboratory setting.

In spite of these null findings, the study makes several contributions. First, we draw on the literature to develop a theory that links authoritarian iconography to political attitudes and behavior. Second, we study the UAE, a case that is under-studied in political science and yet "intrinsically important" (Jones, 2015, p. 25) given its regional power and rapid liberalization.

Authoritarian iconography is hypothesized to be an important part of the Emirati regime's survival strategy (Davidson, 2013; Jones, 2015), making it a good "test case" for assessing the impact of iconography. Third, we test the effect of authoritarian iconography experimentally to overcome some of the limitations of previous observational approaches. Although we do not find any effect of authoritarian iconography on political compliance or support for the regime, the study establishes a baseline for future experimental research that can adapt our research design. Finally, although a single experimental study cannot on its own invalidate a theory, the study's findings suggest that authoritarian iconography could have null or different effects than previously believed, opening up new lines of theoretical inquiry.

The rest of the article proceeds as follows. We begin by reviewing the literature on authoritarian iconography and survival. We then develop a theory about how leaders' images may affect individuals' political attitudes and behaviors. We next describe our experimental study, which is designed to test the main predictions of the theory. Finally, we present our findings and discuss their interpretation.

Literature Review

Authoritarian leaders use a number of tools to stay in power. Coercion is perhaps the most obvious way in which authoritarian leaders survive. Through their monopoly on coercive power, autocrats deter transgressions and secure compliance from potential opponents and the masses. Compliance under coercion is involuntary, as it is achieved under the threat of harm. Scholars have pointed to the importance of coercion in sustaining autocratic rule, arguing that the coercive power of the state is a stronger predictor of autocratic survival in some countries than the strength of the opposition (Way & Levitsky, 2006). Research has also pointed to autocrats' coercive potential as an important explanation for the failure of democratization (Bellin, 2004; Diamond, 2010).

Autocratic leaders do not, however, rely on coercion alone. Rather, they use a variety of strategies of accommodation (Diamond, 2010). In neopatrimonial regimes, leaders maintain authority through personal patronage (Roth, 1968; see also Weber, 1978, vol. 1 on patrimonialism). The fundamental feature of such regimes is the awarding by public officials of favors (e.g., jobs or licenses) in return for loyalty. Other strategies of accommodation, which can work in concert with the use of patronage, include mechanisms of representation and consultation such as limited, pluralistic elections or the passage of laws designed to include women in politics (Bush & Jamal, 2015). Paradoxically, elections can co-opt opposition and further entrench authoritarian regimes in

power, including by offering regimes a way to dispense patronage to key supporters (see, for example, Blaydes, 2011; Gandhi & Przeworski, 2007; Lust-Okar, 2006; Magaloni, 2006).

Studies of authoritarian survival have paid relatively less attention to political culture and authoritarian iconography of late. Classic studies, however, examined the cultural and iconographic elements of "personalistic" or "sultanistic" regimes in which the leader retains personal control over policy decisions and selects regime personnel (Chehabi & Linz, 1998).[2] In such regimes, iconography may generate support by enhancing legitimacy.[3] Cults of personality, for example, often exalt the nation's history and heritage and draw on "invented traditions" that differentiate the nation ethnically or culturally (Chehabi & Linz, 1998). This legitimacy-enhancing role of iconography and personality cults can be combined with other functions. In the Gulf monarchies, for example, "gentle" portraits featuring current rulers in a soft or flattering manner are displayed near portraits that display the same rulers as "hard men." The goal, according to Davidson, is to demonstrate to the population that the rulers should be "both loved and feared, and certainly never crossed" (Davidson, 2013, pp. 66-67).

Thus, previous studies have generated important insights into the sources and consequences of personality cults. Recent studies on this topic have relied on formal, theoretical models (Márquez, 2013) and interpretive methods (Wedeen, 1999) to further uncover the logic and meanings of iconography. Yet such studies are challenging for other researchers to independently verify and replicate. Moreover, observational studies may suffer from selection bias. As we describe below, authoritarian iconography is not deployed randomly but used in particular times and locations to achieve political goals. Thus, the impact of iconography may derive from the conditions under which it is displayed just as much as from the iconography itself. The experiment we described below overcomes some of these limitations, though it is not without limitations itself. Before describing it, however, we first synthesize the existing literature and distill its arguments into a general theory that links iconography to political behavior in stable authoritarian regimes with at least some element of "personalism."[4]

Conceptualizing Authoritarian Iconography

The Oxford Dictionary of English defines iconography (2010) as "the visual images, symbols, or modes of representation collectively associated with a person, cult, or movement." As such, authoritarian iconography includes posters, sculptures, seals, insignia, public spectacles, and other visual representations that are associated with authoritarian regimes. The representations

may be particular to a leader, a political party, the military, or the nation—or they may be affiliated with multiple referents. We focus on authoritarian iconography that is associated with individual rulers, although we suspect our argument applies to other iconographic forms.

Broadly speaking, two types of actors deploy authoritarian iconography: states and citizens. Although it is possible that the state deploys these representations for purely expressive or normative purposes, many scholars argue that they serve a strategic function. Iconography may build and reinforce the legitimacy of the regime; for example, the leader may be pictured with previous rulers, which suggests continuity, or with other symbols of the nation or culture, which provide external sources of legitimacy. Iconography may also be used to remind the public about the regime's presence and coercive power; for example, the enormity or ubiquity of visual representations may convey the power and omnipresence of the regime. The public also deploys authoritarian iconography. People may do so to express their support of the government or because they want to signal their support—whether sincere or insincere—to state officials and other people. They may also do so because of cultural norms.

In many cases, it is clear who is deploying authoritarian iconography and why. In Amman, for example, the state is likely responsible for erecting a highway billboard featuring King Abdullah of Jordan, whereas the shopkeeper is likely responsible for hanging up a framed photo of the king on her wall.[5] In other cases—the large poster of the leader in a shopping mall, perhaps—the origins of iconography is less readily discernable. Yet even if the audience knows who has displayed the images, the audience rarely knows *why* with confidence. Both the state and the public have multiple potential motives for producing and displaying authoritarian imagery, and it is often difficult to know which motives are in operation. After all, the state rarely explains why it uses such images, and ordinary people have reasons to dissemble. Thus, we do not argue that authoritarian iconography requires the audience to recognize a particular source or ascribe a particular set of motives to that source for it to influence political behavior.

Individuals also process authoritarian iconography in diverse ways. Iconography can be processed subconsciously, which is to say that the images register in someone's mind without her awareness. Many kinds of political iconography have subconscious effects on individual behavior, including judicial symbols (Gibson, Lodge, Taber, & Woodson, 2010), national flags (Hassin, Ferguson, Shidlovski, & Gross, 2007; Robinson, in press), and affective cartoons (Erisen, Lodge, & Taber, 2014). As authoritarian iconography can be pervasive in countries where it is found, it may be so ordinary that the intended audience does not consciously notice it, yet it still subconsciously

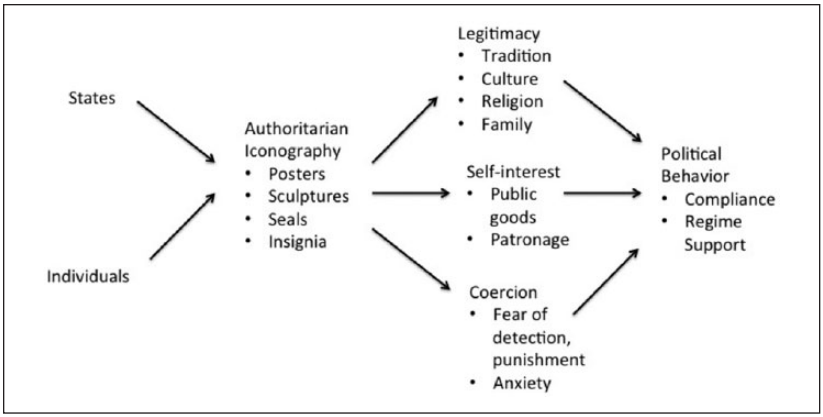

Figure 1. Causal pathways from iconography to behavior.

affects individuals' behavior. It is also possible that authoritarian iconography is processed consciously, which is to say that the representations of the leader affect someone's mind with her awareness. Even in environments where authoritarian iconography is pervasive, non-routine representations of the leader—for example, new or unusually large images—will be noticeable. In this case, the effects of authoritarian iconography on behavior may be conscious.

How Authoritarian Iconography Shapes Political Behavior

How might authoritarian iconography affect political behavior? Regardless of whether the state or individuals display images and whether they do so in ways that are likely to generate conscious or subconscious effects, we argue that iconography can affect people's political behavior through at least three causal pathways. Figure 1 summarizes our argument.

Iconography can promote both individuals' compliance with the state and their support for the state.[6] *Political compliance* refers to the extent to which people obey the state's rules and institutions. *Regime support* refers to the extent to which people actually agree with the state's rules and institutions. Although compliance and support often result in behaviors that are observationally equivalent, they are not the same outcomes conceptually. People may publicly comply at the same time as they privately make fun of their ruler and resist his authority (Scott, 1990; Wedeen, 1999). In other words, although an individual's support for the state typically implies her compliance, an individual's compliance does not necessarily imply her support. For these

reasons, we believe that iconography can affect both individuals' compliance with and support for the state or that it can affect only individuals' compliance. The likelihood of iconography affecting both compliance and support instead of merely compliance may depend on how images are used, the audience, and the regime in question.

We propose that authoritarian iconography affects compliance and support via three mechanisms. We follow scholars such as Hurd (1999) in labeling these mechanisms legitimacy, self-interest, and coercion. First, the individual complies because she believes a legitimate authority has made the rule. Second, it is in her material self-interest to comply. Third, she fears reprisal for failing to comply. It follows, then, that by affecting perceptions of legitimacy, self-interest, or coercion, iconography could increase compliance and support.

The mechanism through which iconography affects compliance and support depends on the content and context of the image as well as individual factors. To make this argument more precise, it is helpful to take each mechanism in turn. Beginning with legitimacy, one of the primary ways states can augment their legitimacy is by calling on an external source (Schaar, 1981). In democracies, people are the external source, who validate popular acceptance of authority via elections. For autocrats, the external source is often tradition, history, or religion. Thus, to increase beliefs in the autocrat's legitimacy, iconography may depict the leader in traditional dress or include historic or religious symbols with the image. Moreover, the leader's image may appear alongside past leaders of the country or familial figures (Davidson, 2013; Wedeen, 1999). These associations suggest that the leader is the rightful authority and thereby may increase compliance with and support for his rules.

Turning to self-interest, it is well known that regimes can increase compliance by offering private benefits to individuals. In the UAE and other wealthy Gulf monarchies, states provide an extensive list of goods and services that includes free land, health care, and education (see, for example, Gause, 1994; Kamrava, 2009). In addition, regime survival may be linked to the ability of the leader to competently provide public goods (Bueno de Mesquita, 2003). Thus, iconography can remind individuals of the leader's role as provider by its placement near public works, such as bridges, ports, medical clinics, and schools. For example, the image of Sheikh Khalifa and Sheikh Mohammed of Dubai in Figure 2 is featured next to the Dubai World Trade Center. This image reminds individuals of the leader's role as provider. The online appendix contains other examples of iconography in the UAE, many of which fall into the self-interest category.

The last mechanism available to autocrats is coercion. Coercion may lead to compliance as individuals are reminded of the possibility of being caught

Figure 2. The Dubai World Trade Center, United Arab Emirates.
Photo taken May 27, 2011. ©Typhoonski | Dreamstime.com—Dubai World Trade Center Photo.

and punished for non-compliance. Although coercion may also lead to support, that will not always be the case. That is to say, coercion may motivate individuals to follow rules, but may not necessarily lead individuals to agree with them. Leaders may employ their image to remind individuals of their coercive power in several ways.[7] First, they may deploy large images and large numbers of images. The omnipresence of the images suggests that the state has the resources to dominate individuals. Moreover, it conveys a sense

of being watched, which suggests to individuals that the state has the ability to detect their non-compliance. Second, the leader may appear in military dress or with other symbols of the coercive apparatus of the state.

Each of these mechanisms—legitimacy, self-interest, and coercion—yield the same prediction: Exposure to authoritarian iconography will increase compliance with the state's rules and institutions and may also increase support for those rules and institutions. Yet it is important to acknowledge that certain mechanisms may be privileged across countries and individuals. Although we expect iconography to lead to compliance and possibly support in most stable autocracies, the predominant mechanism is likely to vary across different contexts. For example, the legitimacy mechanism may be more likely in a country such as Morocco, which is a monarchy that calls on religious motifs, whereas the coercion mechanism may be more likely in a country such as Saddam's Iraq. Moreover, the predominant mechanisms may also vary within countries. Consider again the example of Saddam's Iraq. Although the dominant mechanism linking Saddam's image to compliance in Iraq may have been coercion, the particular pathway for any individual could have varied. Individuals with contentious relationships with the regime may have complied when presented his image because they were reminded of coercion, whereas individuals who benefited from regime patronage may have complied out of self-interest. Again, however, and regardless of the mechanism, the leader's image points toward greater compliance.

Below, we specify a research design for examining the relationship between authoritarian iconography and compliance and support. We focus on testing the main effect of authoritarian iconography on compliance and support because a causal effect has not yet been established. Once the main effect of authoritarian iconography has been identified, future research can investigate the underlying causal mechanisms.

A Lab Experiment in Abu Dhabi

Our study examines the effects of authoritarian iconography on compliance and support using a laboratory experiment in the UAE. As a firmly authoritarian country where authoritarian iconography is common, the UAE is an appropriate first location for an experimental investigation of the effects of iconography.[8]

Using a laboratory—rather than field—setting also confers several advantages. First, it allows us to conduct an experiment, which ensures that the main treatment—an image of the country's leader—is randomly assigned to participants. Although the logic according to which regimes and individuals display images of the country's leader is not fully understood, leaders' images

are not placed in locations at random. Thus, a real-world study that compared individuals' actions in the presence and absence of a leader's images could make an incorrect inference about the effect of those images due to omitted variable and selection bias.

Second, the laboratory allows us to employ more treatments than is likely possible in the field. Studies (e.g., Panagopoulos, 2014) have shown that people respond strongly to the sense that they are being observed in general, not just by an authority figure. Thus, it is important to distinguish the effects of exposure to an image of the leader from those of being watched more generally. This research design is more feasible in a controlled environment.

Finally, the laboratory allows us to conduct an experiment relatively freely. Conducting a study of iconography in a public place would be problematic in an autocracy from a number of perspectives: the difficulty of obtaining permission; the possibility of harassment or arrest of the researchers, subjects, or both; and the fear that subjects may have in terms of participation. Conducting the study in a laboratory—especially a lab with some of the freedoms described below—obviates most of those problems. Before elaborating the specific advantages of the UAE laboratory that we use, we provide some general background on the UAE.

Background on the UAE

The UAE emerged as a sovereign nation during British decolonization. Originally known as Trucial States because of the truces signed between Britain and various tribal leaders, the UAE unified seven Trucial States into one country. To this day, the UAE remains a federation; seven emirs govern their own territories, or emirates. Between the emirates, power is unequal, in large part due to differences in oil resources and land. The emirate of Abu Dhabi, which includes more than three quarters of the UAE's landmass, and the emirate of Dubai, whose royal family is related to the Abu Dhabi family, are most powerful. Hence, in some ways the different Emirati lines represent a larger "royal grouping," even if they are not related by blood. The UAE's federalist structure is similar to other Gulf monarchies, which often have this feature within their families.

The UAE is also typical of other monarchies in that it has experienced one monarchical succession of power during independence. The succession occurred within the Bani Yas tribe of the ruling Al Nahyan family in 2004. At that time, Sheikh Khalifa succeeded his father Sheikh Zayed ("the Great") as the president of the UAE and ruler of Abu Dhabi. As in other monarchies, Sheikh Zayed is portrayed as the founding father of an independent UAE. It is possible that this hereditary succession will continue, though it

was previously uncommon in tribal settings for a son to succeed his father (Heard-Bey, 2005). One can continue to see images of Sheikh Zayed in many of the emirates as well as images of Sheikh Khalifa and other emirs.

Turning to representation, the UAE is again broadly similar to other Arab monarchies, although it allows less popular representation than some other members of the Gulf Cooperation Council, such as Kuwait. There is a unique form of competition between the federal units, however, which is firmly entrenched in the UAE's constitution and may partially substitute for competition on the legislative council. The most important legislative body in the country is the Supreme Council of the Union, which is made up of 25 ministers and has a cabinet with the seven emirs. Although the president is technically renewed every 5 years, in practice, the presidency is reserved for the ruler of Abu Dhabi. Also reflecting the disparate power relations within the emirates, only Dubai and Abu Dhabi have vetoes (Davidson, 2013).

Also similar to other Arab monarchies, freedom of speech in the UAE is restricted, particularly with regard to criticism of the royal family and other royal families in the region. Although criticism of government services and policies is generally permitted, rarely, if ever, do residents of the UAE criticize government officials by name. Indeed, even activists typically see criticism as unproductive or unnecessary (Krause, 2008). Criticism of the monarchy is further suppressed because it can result in jail and deportation for non-citizens and jail time for citizens.

Finally, similar to other Gulf nations and many other Arab countries, the UAE has a large expatriate community. Most long-term expatriate residents hail from the Indian sub-continent, with other smaller groups coming from the greater Arab world. These immigrants cannot currently obtain citizenship (Vora, 2013). Willingness to openly criticize the government may be lower among non-citizens for fear of loss of business licenses (through the *kafala* system, in which a non-citizen has to have a UAE citizen business partner) or deportation. Indeed, despite discrimination and their anger at not having access to benefits available to citizens, expatriate residents often praise the government for allowing them to accumulate wealth (Vora, 2008). As documented in anthropological work in Dubai, higher skilled expatriate workers are also in charge of monitoring and regulating the lower skilled workers they bring to the country and, hence, also become tools in maintaining authoritarian control (Vora, 2011).

Perhaps most pertinently for our study, the UAE is a typical case in the region in its use of iconography. In his recent book, Davidson (2013) analyzes political iconography in the Gulf, where images frequently occur in diptychs or triptychs that link together multiple generations or types of rulers. For example, Sheikh Zayed is often pictured next to Sheikh Khalifa in a diptych or the Vice President and Premier are present together in a diptych. The

online appendix contains several examples. This practice is also typical outside of the Gulf. For example, the leader of Azerbaijan, Ilham Aliyev, is often pictured with his father, Heydar Aliyev, in billboards. In Jordan, pictures of the current King Abdullah are often near pictures of his father, Hussein. These pictures, similar to the ones in the UAE, appear to stress the continuity of the regime.

Jones (2015) adopts the term "social engineering" to discuss the methods, including iconography, that the UAE regime uses to perpetuate its rule. Jones notes that the form of this engineering has changed over time to embrace a more liberal veneer than in cases such as the Soviet Union, Nazi Germany, or modern North Korea. Emirati spectacles and symbolism, while still seeking to promote regime stability, attempt to do so in a less coercive manner. Instead of holding a political rally with forced attendance, the UAE and similar regimes hold festivals and events, such as a Festival of Thinkers (Jones, 2015). This change in symbolism is particularly prevalent in many monarchies in the Gulf but has also occurred in China, Singapore, and Kazakhstan (Jones, 2015).

Background on the Laboratory in Abu Dhabi

The above discussion suggests that the UAE is a plausible case for testing our theory, but our selection of it is not random. The key reason we choose the UAE is that it is the first country in the region to open an experimental social science laboratory. We employ the Social Science Experimental Laboratory (SSEL) at New York University (NYU), Abu Dhabi. Because it is part of NYU Abu Dhabi, the SSEL provides an environment of relative academic freedom in which we can study political dynamics that might otherwise be dangerous, prohibited, or both. While the SSEL lab is the first in the region, the UAE is no exception in its goal to internationalize education and increase critical thinking, even among authoritarian regimes. Most of the Gulf monarchies and other similar countries are currently attempting to expand the presence of international universities in their midst (Altbach & Knight, 2007).

Generalizability

As with any single country study, readers may wonder whether it is possible to draw broader conclusions. We use residual analysis (Lieberman, 2005; Seawright & Gerring, 2008) to show that the UAE is a typical case in terms of regime stability, an outcome associated with political compliance and support, our main dependent variables. If the effects of iconography across authoritarian regimes are the same—to increase compliance with and support for the regime—then we want to know whether, controlling for observable

factors, the UAE is a typical case in the relationship between iconography and political behavior. In a perfect world, we would measure the usage of iconography by authoritarian leaders to examine whether the frequency and type of usage in the UAE is typical in relationship to regime stability. Unfortunately, no measures of iconography across countries exist. In other words, it is a missing variable in most models of regime stability. Hence, the best we can do is to examine the typicality of the case when predicting regime stability using other known factors.

To establish that the UAE is well-predicted by recent models of authoritarian survival, we turn to a recent, prominent study: Menaldo (2012). Menaldo examined the relationships between a variety of factors and regime stability in the Middle East. Given that Menaldo uses a time series, cross-sectional model, we pay particular attention to his prediction of stability in terms of the more recent residuals, rather than the historic residuals. As shown in the online appendix, the UAE fits a fairly typical model of an authoritarian monarchy. We believe this analysis suggests we can make comparisons between the UAE and other Arab monarchies with a fair level of confidence and can also comment on the relationship between iconography and political behavior more generally in other modern authoritarian regimes.

The Experimental Design

Subjects

The experiment was conducted on a subject pool consisting of 123 adult residents of the UAE. Although our pre-analysis plan (PAP) called for at least 150 subjects, we were unable to reach that number due to several challenges in subject recruitment. Nevertheless, the results of a power analysis do not indicate that we would have found substantially different results with 150 subjects. The results of the power analysis and a discussion of the specific challenges to recruitment are in the online appendix.

The subjects are part of a subject pool maintained by the SSEL that includes students from local universities. Some of the subjects are Emirati, but most are long-term, non-citizen residents, who were typically born in the UAE or immigrated with their families at a young age. Nearly all are Muslim. Their parents' (or, in some cases, their) countries of origin include states with large emigrant populations in the Middle East and South Asia, including India, Lebanon, Pakistan, Palestine, and Syria. See the balance table in the online appendix for descriptive statistics on the demographic characteristics of the sample. The online appendix also compares the subject pool with the Emirati population, showing that our subjects were more likely to be female, from the emirate of Abu Dhabi, and non-Emirati nationals than the population as a whole.

The composition of the sample may privilege some causal mechanisms over others. In particular, non-citizen residents are not only more vulnerable to state authority than Emirati citizens but also lack their networks of clientelism and patronage. Thus, any impact of authoritarian iconography that we find among them is more likely to be driven by coercion and is less likely to be driven by self-interest and legitimacy. Because our interest in this study is establishing the main effect of authoritarian iconography, not testing causal mechanisms, we do not expect this dynamic to bias our findings.

Treatments

To test our argument about how authoritarian iconography affects political behavior, we expose randomly selected individuals in our study to images of the UAE's ruler—Sheikh Khalifa—via subliminal priming. In subliminal priming, a stimulus—in this case, an image—is displayed to the subject for a short amount of time such that it lies outside of the subject's conscious awareness. In our experiment, as in many others, the stimulus "flashes" repeatedly on the computer screen at which the subject is seated.[9] Computer stations have barriers that prevent contamination between subjects.

Social scientists have widely studied subconscious political stimuli and have often found significant effects.[10] Although authoritarian iconography can have both conscious and subconscious effects, subconscious stimuli allow for a more precise administration of the treatment and thereby minimize non-compliance. In subliminal priming, subjects are told to stare at the screen while the prime "hits" them between the eyes. Thus, to the extent that they follow the instructions, all subjects are affected by the treatment. We therefore initiate the experimental research agenda on authoritarian iconography by exploring the theorized subconscious effects.

Using this subliminal approach, we randomly assign subjects to one of three experimental conditions. Subjects assigned to the main treatment (the Leader Treatment) are exposed to images of Sheikh Khalifa. The images we use (e.g., Figure 3) are reproductions of some of the many official photographs of Sheikh Khalifa that are displayed in the UAE. We selected images that do not include anything beyond the leader (e.g., they do not contain flags or other national symbols), which therefore should not privilege any mechanism. For each priming sequence, we randomly expose subjects to one of four images of the Sheikh facing the camera; we use multiple images of the Sheikh to reduce the likelihood that subjects recognize the image and it becomes supraliminal. Image selection is randomized.

One potential critique of our main treatment is that its effects may have nothing to do with the authoritarian leader, but rather may be due to a generic

Figure 3. Example of Sheikh Khalifa treatment image.

monitoring effect. Scholars have shown that images that signal being observed, such as stylized eyes, can affect behavior (Haley & Fessler, 2005; Panagopoulos, 2014). To address this concern, we also include a second experimental condition (the Leader and Eyes Treatment), in which subjects are exposed to an image of stylized eyes in addition to Sheikh Khalifa.[11] As in the main experimental condition, the treatment occurs when subjects are twice exposed subliminally to Sheikh Khalifa's image and twice exposed to the stylized eyes. The order of exposure to the images is randomized. If the additional exposure to the eyes has no effect on compliance and support relative to the effect of the image of Sheikh Khalifa on its own, this would suggest that no general monitoring effect is in operation. If it does have an effect, this would suggest a monitoring effect.[12]

Finally, subjects assigned to the control condition are exposed to the forward and backward masks. These masks can be found in the online appendix and are simply disarticulated versions of the images of the Sheikh. We follow this procedure to make sure that all subjects experience the experiment in the same way.

As summarized in Table 1, we administered the complete priming sequence to subjects before each outcome measure. Repeating the prime means that the "dosages" of our treatments could increase. But because the treatments are subliminal, our inferences should not be affected; indeed,

Table 1. Sequencing of Lab Instrument.

Sequence	Category	Measure
1	Demographic and mood questions	Potential covariates
2	Priming	
3	Compliance game	Compliance outcome 1
4	Priming	
5	Charity task	Compliance outcome 2
6	Priming	
7	Policy support questions	Support outcome
8	Priming	
9	Compliance attitudinal measures	Compliance outcome 3
10	Manipulation check	

repeated exposure to treatments over an experimental session is common in studies using subconscious primes (Erisen et al., 2014; Kam & Zechmeister, 2013). Moreover, repeatedly priming subjects ensures that the treatment is fresh for each outcome and mimics real life, as in walking past images repeatedly in the street. As is standard in subliminal priming studies, we ask participants at the end of the study if they saw any images during the experiment and, if so, what the images were. These results are reported in the online appendix.

Randomization Strategy

Several strategies of assigning the treatments to subjects are possible. The simplest strategy would involve assigning the subjects to experimental groups using a random number generator upon arrival. In small- to medium-sized lab experiments, however, precision can be improved through blocked designs. Blocking should focus on variables likely to affect the outcomes, as well as the variables that define any subgroups to be used in subgroup analyses (Moore, 2012). Of course, blocking requires obtaining data on subjects before they arrive in the lab. The SSEL collects data in advance on a limited number of demographic factors including gender. As gender could be related to individuals' compliance (e.g., Blass, 1999; Hasseldine & Hite, 2003), we block on it.

Outcome Measures

We measure political compliance and support using a mix of behavioral and attitudinal measures that are portable to other contexts. Below, we outline

each test and specify how we will analyze and interpret our results across the different tests. We conclude with a discussion of plausible alternative tests and measures.

Political Compliance. We first measure compliance behaviorally. Many psychological experiments have studied individuals' compliance with authority, the most famous of which is Milgram (1963). For our study, however, it is important to measure compliance with rules and institutions of the state. Subjects therefore play one round of a game developed to measure tax compliance. In the basic game, participants are given a sum of money. They are told that a percentage of their money must be given back to the researchers and that the researchers only know with some probability how much they received based on what they report. Finally, the participants are told the penalty they incur if they do not report the true amount they received. Compliance with the rules of this game is expected to reveal the motivations to comply when faced with a risky task (Andreoni, Erard, & Feinstein, 1998). We follow the procedures used in the widely cited study that developed this game (Alm, Jackson, & McKee, 1992): The proportion of the declared money that subjects are asked to return to the researchers is 40%, the probability of their true amount being discovered is 1%, and the penalty that subjects pay is an additional 30% of their earned income. To ensure that subjects understand the protocol, we conduct two practice rounds of the game. Because compliance in the game was strongly bimodal, we measure compliance using a binary measure indicating whether individuals did not report any of their income (e.g., Cadsby, Maynes, & Trivedi, 2006).[13]

Our second measure of compliance relies on the subject's willingness to donate to a regime-endorsed charity. In January 2015, Sheikh Khalifa issued a public directive to donate to charity as a part of a campaign called "UAE Compassion." At the end of the compliance game, we inform participants of the directive and ask them if they would like to donate some of their earnings to the Red Crescent Society of the UAE, which was one of the main partner charities that participated in the UAE Compassion campaign. Because charitable giving was also bimodal, we measure compliance using a binary measure indicating whether individuals donated at all and report results using an alternate, continuous measure in the online appendix (Karlan, List, & Shafir, 2011; Landry et al., 2010).

Finally, we measure compliance using survey questions that permit us to examine how leaders' images affect reported attitudes about political authority. Openly expressing opposition to the regime is risky in the UAE and asking about compliance with or support for the government is unlikely to elicit

truthful responses. Thus, we examine attitudes toward compliance with the state *in general* using survey measures adapted from questions that were previously posed to respondents in the Gulf region and other authoritarian states via the Arab and Afrobarometer surveys. We ask respondents five paired survey questions intended to measure our underlying concept of political compliance, as it relates to individuals' respect for authority, relationships with the government, freedom to organize, freedom of speech, and freedom of expression. Contrary to our expectations, responses to these five survey questions did not load onto a single factor, suggesting that they do not reflect a single, underlying concept of compliance. As Cronbach's alpha fell well below the threshold of at least .70 specified in the PAP, we analyze responses to each survey question separately rather than using an index or other composite measure.[14]

We view the behavioral and attitudinal measures of compliance as complementary. That is, we view each measure as capturing the same, underlying concept: political compliance. For this reason, we interpret results across the three measures of compliance using a strict test and false discovery rate corrections across our measures, including the five survey measures as they do not load onto one factor. In both cases, we adjust the p values to control for the false discovery rate using the Benjamini–Hochberg correction (Benjamini & Hochberg, 1995). In the results that follow, we report both the uncorrected p values and the Benjamini–Hochberg corrected p values.

Political Support. We measure political support using three survey questions. As in the charity donation task, we give respondents information about policies that Sheikh Khalifa has endorsed. We then ask respondents the extent to which they agree or disagree with these policies. Two of the policies are foreign policies—international climate change and nuclear agreements that Sheikh Khalifa has supported—and one is a domestic policy—the creation of a space program by the UAE government that, as a UAE government policy, has the implicit endorsement of Sheikh Khalifa. We selected these policies as they are not sensitive policy issues in the UAE, allowing for honest reporting by subjects. The order of the questions is randomly assigned.

As with the compliance survey, we are unable to aggregate these survey questions into a single index because the three regime-endorsed policies do not correlate highly with one another. Given a Cronbach's alpha that falls well below the conventional threshold of .70 specified in our PAP, we analyze the three regime-endorsed policies separately. As such, we again adjust for

the false discovery rate across the three measures of support using the Benjamini–Hochberg correction and report the unadjusted and adjusted p values.

Measuring Compliance and Support Using Other Strategies. Other measurement strategies are possible. One would be to ask participants directly about compliance and support using a list experiment (see, for example, Corstange, 2013; Kuklinski, Cobb, & Gilens, 1997). However, because a list experiment exposes subjects in each condition to separate lists containing different items, it effectively doubles the number of conditions. Given that our study already includes three groups and multiple outcome measures, using a list experiment would unacceptably diminish statistical power.

An implicit association test could also be used to understand how positively subjects feel toward the Emirati government. We could ask subjects to categorize regime-related and regime-unrelated stimuli into categories with either positive or negative associations. Although this type of procedure might accurately measure support for the Emirati regime, it would confound the causal inferences that our study is designed to make. Exposing participants in the control group to the regime-related stimuli would essentially expose them to the leader's image.

Results

Contrary to the existing literature and our own expectations, the results of the study do not indicate that images of the UAE's leader affect compliance with or support for the ruling regime among UAE residents. Although the direction of our findings is generally in keeping with our hypotheses, the overall finding of our study is null. As we elaborate later in the article, this null finding is probably not due to insufficient sample size although it could be due to other features of the research design. As we do not find that images of the authoritarian leader affect either political compliance or support, we focus our presentation below on comparing each of the two treatment groups with the control group.[15]

Before beginning our analysis, we assess balance across the three experimental conditions on 13, key pre-treatment covariates.[16] The three experimental conditions are generally balanced across the covariates. However, as may be expected given our relatively small sample size, we have some imbalance across groups. Specifically, we find that subjects in the Leader Treatment were somewhat older and came from wealthier families than subjects in the control group ($p < .10$). These subjects also reported themselves to be more

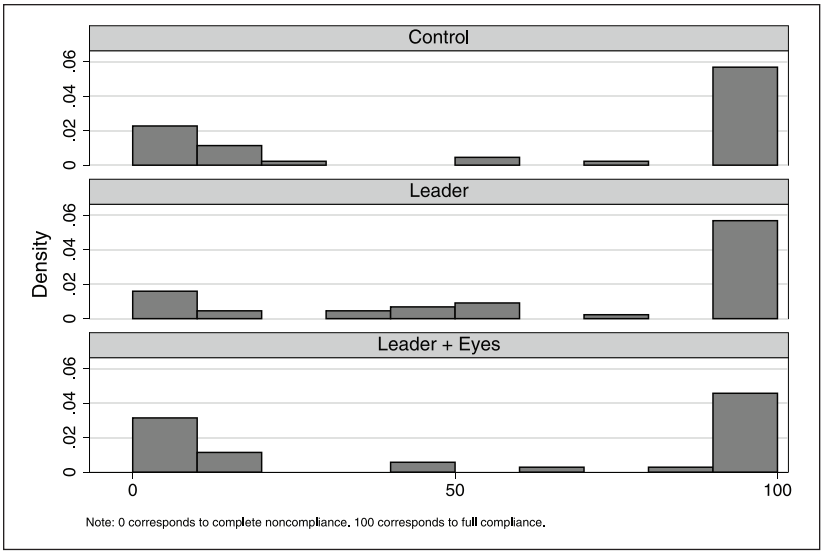

Figure 4. Percent of income reported in tax compliance game.

religious than subjects in the Leader and Eyes Treatment group ($p < .10$). These results are likely due to chance, as a model with all of the covariates combined does not predict assignment to treatment better than a null model. Nonetheless, in the following analyses, we report the results of unadjusted comparisons between groups as well as of regression analysis controlling for all three imbalanced covariates.

Compliance Results

We first investigate the impact of the treatments in the compliance game. We measure compliance as a dichotomous variable, which is coded "1" for subjects who reported all their earned income (fully complied) and "0" for subjects who reported less. Following our PAP, we use this dichotomous measure because the proportion of income reported in the game is strongly bimodal (see Figure 4). We also report the results of the following analyses using the continuous measure of compliance in the online appendix.

Given this dichotomous dependent variable, we report the difference in the proportion of "compliers" across experimental conditions, as well as analyze the results using logistic regression models that also adjust for imbalanced covariates and the method of randomization. Following Bruhn and

Mackenzie (2008), we adjust for the method of randomization by including gender in all regression models as we blocked on gender in assigning subjects to experimental conditions.

As seen in Table 2, the proportion of subjects who complied (i.e., reported their entire income) was indeed higher in the Leader Treatment than in the control group. Whereas 52% of subjects in the Leader Treatment group complied, only 41% of control group subjects did so. This difference is not, however, statistically significant at conventional levels ($p = .29/p = .57$).[17] We also do not find any substantively meaningful or statistically significant differences between the Leader and Eyes Treatment and control groups. We find similar results using regression analysis that also adjusts for imbalanced covariates and the method of randomization, as shown in Table 3.

We next examine the impact of the treatments on the charity task. Because this variable is also highly skewed (see Figure 5), we use a dichotomous variable coded "1" if subjects donate any income (i.e., complied) and "0" otherwise. Given this dichotomous dependent variable, we again present the results of a difference in proportions test as well as of a logistic regression model below. The results of a difference in means test and ordinary least squares (OLS) regression using the original, continuous measure of donation are similar (see Table 10 in the online appendix for results using the continuous measure).

Contrary to our hypotheses, we do not find any statistically significant effect of either the Leader Treatment or the Leader and Eyes Treatment on donation to a regime-endorsed charity after controlling for multiple comparisons. The proportion of individuals in the Leader Treatment group who complied (i.e., donated to a regime-endorsed charity) is higher than the proportion of individuals in the control group who did so, but the difference is fairly small (see Table 2) and far from statistically significant (5 percentage points). Table 3 shows a similar pattern in the regression analysis ($p = .93/.94$).

We do find some suggestive evidence that the Leader and Eyes Treatment may have increased compliance with the donations directive. 57% of subjects who were randomly assigned to this condition complied, as compared with 32% of subjects in the control group. However, this effect is only significant when we do not adjust for multiple comparisons ($p = .02/.33$). These results hold when moving to logistic regression analysis that also adjusts for imbalanced covariates and the method of randomization, as Table 3 shows.[18]

We now turn to our final, attitudinal measures of compliance. As seen in Tables 2 and 3, we do not find any consistent effect of either treatment. The treatments are positively associated with some measures of compliance and negatively associated with others; none of the effects are statistically significant. This pattern is perhaps not surprising, given that our factor analysis suggests that the questions fail to reflect a single, underlying concept of compliance.

Table 2. Differences in Proportions and Means by Treatment Status for Compliance Outcomes.

	Mean			Difference		*p* value (*t* or *z*)		*p* value (rank)	
	Control (1)	Leader (2)	Leader + eyes (3)	(2) − (1)	(3) − (1)	(2) vs. (1)	(3) vs. (1)	(2) vs. (1)	(3) vs. (1)
Complier (game)	0.41	0.52	0.40	0.11	−0.01	.29/.57	.93/.93		
Complier (donate)	0.32	0.36	0.57	0.05	0.25	.65/.93	.02/.33*		
Don't question leaders	0.14	0.05	0.26	−0.09	0.12	.74/.93	.69/.93	.73	.75
Treat as child	0.98	1.02	0.60	0.05	−0.38	.85/.93	.17/.57	.66	.20
Ban organizations	0.57	1.00	0.83	0.43	0.26	.10/.49*	.37/.65	.07	.30
Close newspapers	−0.70	−0.66	−0.34	0.05	0.36	.86/.93	.21/.57	.79	.28
Don't express views	−0.86	−0.57	−0.40	0.30	0.46	.24/.57	.08/.49*	.35	.11

All rows show estimates from either difference-in-means or difference-in-proportions tests and Wilcoxon rank sum tests with *p* values, both with and without multiple comparison corrections. Rank sum *p* values not shown for complier measures as these variables are dichotomous.
*$p < .10$ without multiple comparison. **$p < .10$ with Benjamini–Hochberg multiple comparison adjustment.

Table 3. Treatment Effects on Compliance Outcomes From Regression Analysis With Covariate Adjustment.

	Complier (game) logit	Complier (donate) logit	Don't question leaders	Treat as child	Ban organizations	Close newspapers	Don't express views
Leader	0.59 (.21/.49)	0.04 (.93/.94)	−0.34 (.21/.49)	0.05 (.83/.94)	0.32 (.24/.49)	−0.10 (.72/.94)	0.04 (.87/.94)
Leader + eyes	0.10 (.84/.94)	0.99* (.04/.49)	−0.02 (.94/.94)	−0.34 (.22/.49)	0.20 (.48/.85)	0.34 (.24/.49)	0.38 (.17/.49)
Control variables	Yes	Yes	Yes	Yes	Yes	Yes	Yes
n	123	121	123	123	123	123	123

All regressions show coefficients and *p* values in parentheses, both with and without multiple comparison corrections.
*$p < .10$, without multiple comparison. **$p < .10$, with Benjamini–Hochberg multiple comparison adjustment.

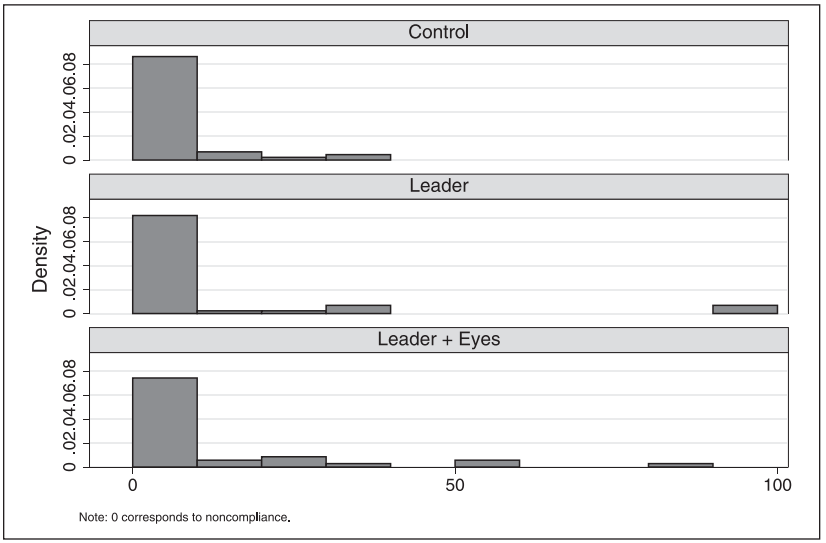

Figure 5. Percent of income earned donated in charity task.

Support Results

Finally, we investigate the impact of the experimental treatments on support for the regime, as measured by support for three policies endorsed by the regime leader. Contrary to our expectations, exposure to both treatments is

Table 4. Differences in Means by Treatment Status for Support Outcomes.

	Mean			Difference		*p* value (*t* test)		*p* value (rank)	
	Control (1)	Leader (2)	Leader + eyes (3)	(2) − (1)	(3) − (1)	(2) vs. (1)	(3) vs. (1)	(2) vs. (1)	(3) vs. (1)
Climate policy	83.55	78.20	71.97	−5.34	−11.57	.28/.49	.03/.09**	.41/.61	.03/.08**
Space program	79.39	78.86	63.57	−0.52	−15.81	.93/.93	.03/.09	.60/.72	.01/.07**
Nuclear deal	65.07	61.43	57.66	−3.64	−7.41	.62/.75	.33/.49	.79/.79	.31/.61

All rows show estimates from both *t* tests and Wilcoxon rank sum tests with *p* values, both with and without multiple comparison corrections.
*$p < .10$, without multiple comparison. **$p < .10$, with Benjamini–Hochberg multiple comparison adjustment.

generally negatively associated with support. Assignment to the Leader Treatment results in more negative appraisals of regime policies related to the environment, space exploration, and nuclear proliferation, though none of these effects are statistically significant. Assignment to the Leader and Eyes Treatment reduces support for the regime's climate policy by approximately 12 points on the 101-point feeling thermometer ($p = .03/p = .08$) and for the regime's space exploration policy by almost 16 points ($p = .01/p = .07$),[19] even after adjusting for multiple comparisons (see Table 4). In contrast, assignment to the Leader and Eyes Treatment does not have a statistically significant effect on support for the regime's nuclear proliferation policy ($p = .31/p = .61$), though the effect again points in a negative direction. As seen in Table 5, these results continue to hold after adjusting for imbalanced covariates and the method of randomization.

Table 5. Treatment Effects on Support Outcomes From Regression Analysis With Covariate Adjustment.

	Climate policy	Space program	Nuclear deal
Leader	−6.12 (.23/.47)	−1.00 (.88/.99)	0.05 (.99/.99)
Leader + eyes	−12.04** (.02/.07)	−17.50** (.02/.07)	−3.96 (.58/.87)
Control variables	Yes	Yes	Yes
N	123	123	123

All regressions show coefficients and *p* values in parentheses, both with and without multiple comparison corrections.
*$p < .10$, without multiple comparison. **$p < .10$, with Benjamini–Hochberg multiple comparison adjustment.

Interpretation of Results

In summary, our results do not indicate that images of the authoritarian leader affect political compliance or regime support among UAE residents. The most obvious reason for this null finding is insufficient statistical power; however, the results of a power analysis presented in the online appendix suggest that it is likely not the culprit. Although our sample size is small, expanding the sample to any sample size typical for a laboratory experiment is unlikely to change the key results of the study.

The specific design of our experiment is a likelier culprit. One issue relates to our use of a subliminal prime. Although subliminal priming allows for a more precise administration of the treatment and thereby minimizes non-compliance, it does not offer a way to independently verify that subjects received the treatment and it worked as intended. Therefore, it is in principle possible that subjects did not receive the treatment or that it worked differently than intended. Moreover, authoritarian iconography may in fact work primarily through a supraliminal mechanism. Future research could investigate this possibility by replicating our experiment using a supraliminal prime, such as through longer exposure to the images we used or by placing physical images outside the computer stations (e.g., on a laboratory wall).

Another issue relates to our use of primarily non-citizen subjects who differ from Emirati citizens in a number of ways, particularly their access to state networks of patronage. If it is the case that authoritarian iconography's effects inhere largely through the legitimacy or self-interest (i.e., patronage) mechanisms, then our sample may be the source of our null results. Future research could replicate our experiment using citizen subjects, though as we document in the online appendix, doing so will be challenging.

The use of a lab in the first place may also explain the null findings. Moving away from the real-world setting described in our theory removes the leader's image from its context. Before conducting the experiment, we hypothesized that the contextual associations of the image should remain with people, likely by leaders' design. Nonetheless, it may be that, absent the context that associates the leader with coercion, self-interest, or legitimacy, an image has no effect on compliance or support for the regime. Future research could investigate this possibility by modifying the treatment to pair the leaders' image with contextual factors. Similarly, because the authoritarian imagery in our study is ubiquitous in everyday life, it is possible that the main "treatment effect" of this imagery has already occurred and additional exposure to it through our experiment has no additional effect. Again, future research could investigate this possibility by priming individuals with images of leaders in contexts where they are not ubiquitous.

If this future research has null findings similar to our own, then this would indicate that our study was the first to establish empirically that authoritarian iconography does not affect individuals' political compliance and support. Given that iconography is widespread in authoritarian regimes, scholars may then need to develop new theory to account for its use. While it may be that authoritarian leaders are simply unaware that iconography has no effect on compliance and support, it may also be that its use by leaders fulfills a different purpose. The audience may not be domestic or the erecting of images may be due to cultural norms. While there are a number of steps in the research agenda before this conclusion can be made, we nevertheless note that authoritarian iconography may not affect compliance and support in the way scholars have traditionally theorized.

Conclusion

This article evaluated the effect of public images of state leaders on individuals' political compliance and support in the UAE. Large public portraits of state leaders are ubiquitous in the UAE and other authoritarian states and were hypothesized to increase peoples' compliance with and support for the state. In our experiment, we randomly assigned some subjects to be exposed to an image of the state leader. The treatment—the portrait of the state leader—was naturalistic and closely mimicked the concept it was supposed to represent. Thus, in addition to the strong internal validity typically enjoyed by experiments, our experiment also had strong ecological validity. Contrary to our expectations, we did not identify any meaningful effects of the portrait of Sheikh Khalifa on subjects' compliance with or support for the regime.

In conducting our study—particularly for the special issue on transparency in the social sciences—we learned that many aspects of published social scientific laboratory experiments are not fully documented. Though these details—how long to prime for given a particular computer monitor's refresh rate, or how to effectively "teach" participants how to play the tax compliance game—may seem mundane, they are essential to making accurate descriptive and causal inferences. An important contribution of the study is therefore its transparent research design including online materials, which provide information about the details of our study that others researchers can use, refine, and adapt. We also provide the full script to run the study using the Inquisit Millisecond software. We encourage other researchers to replicate our findings, as the results of our experiment may be a statistical outlier. As readers who examine the online appendix will note, it was costly to try to record all steps in the research process. Time will tell whether the benefits to the research community of such "extreme transparency" outweigh the costs to the individual scholars of providing it.

In light of what we have learned, what should future researchers who want to study the effects of authoritarian iconography do beyond replicating our laboratory experiment? We offer the following suggestions for other researchers interested in adopting an experimental approach.

First, if it is possible to safely study iconography's effects in a non-laboratory setting, try it. Even in contexts that may be less challenging than the UAE in terms of recruiting subjects, researchers' budgets and laboratories' capacities make it difficult to conduct studies with sample sizes of more than a few hundred subjects. A field experiment that, for example, randomly exposed thousands of people to authoritarian iconography may improve upon our study in terms of verisimilitude as well as sample size—something which is likely to be necessary based on our revised power analyses.

Second, if it is not possible to safely study iconography's effects outside of the laboratory, try a lab experiment with a supraliminal treatment. This supraliminal treatment could be displayed on the computer or in the laboratory itself. What our study gained in terms of precision in treatment administration by using a subliminal treatment may have been lost in terms of treatment subtlety. Indeed, were we not committed to executing our study as outlined in the PAP, we might have switched to a supraliminal prime ourselves after our null pilot results.

This article therefore illustrates one of the core trade-offs involved with accepting work based on prospective research designs. On one hand, we are making public the null results of what we thought was a well-designed plan; and three reviewers, two editors, and four guest editors for *Comparative Political Studies* agreed sufficiently to give the manuscript a conditional acceptance based on our prospective research design. Yet under other circumstances, we might have struggled to publish the null results or left them in the proverbial "file drawer," thus contributing to publication bias and inhibiting the aggregation of scientific knowledge. On the other hand, we tied our hands, choosing not to adapt our research plan once the pilot results suggested that it was unlikely to generate significant findings. This choice, which was encouraged by our understanding of the mission of the special issue, may have prevented us from uncovering suggestive evidence of a different way that authoritarian iconography matters. Given limited subjects, lab time, and money, it was not possible to fully execute both our planned study and an adapted study that tried other types of treatments and might have more clearly pointed to the best avenue for future studies.

On balance, we think the trade-off associated with this publication model is worthwhile because it is possible that our study's findings reflect a "true null"; that is, that authoritarian iconography does not affect citizens' compliance with and support for regimes at all. Though previous studies argue such iconography is impactful, it is difficult to know whether the results generalize and also

whether they are truly due to iconography or to some other aspect of the state's survival strategy. Thus, it is important to continue to study this topic using a variety of research methods, including (but certainly not limited to) experimental ones. The need for future studies is underscored by alternative, plausible explanations for why leaders use iconography, such as that leaders use these images to express themselves and their authority or to mimic other authoritarian rulers. If those accounts are true, then authoritarian iconography may not cause citizens and residents to be more obeisant. Thus, in addition to replicating and extending this study, we encourage future researchers to theorize further about the conditions under which authoritarian leaders elect to display iconography and its possible, alternative effects. Ideally, researchers will jointly study the causes and consequences of authoritarian iconography.

Finally, the research program on political iconography should be extended to compare the use of images in democratic and authoritarian regimes. Many topics in comparative and international politics have been profitably studied in recent years by examining how phenomena that were once thought of as "democratic" work in authoritarian regimes, and vice versa. As any visitor to a U.S. post office, courtroom, or other public building will attest, images of the president are displayed prominently. If public images of state leaders genuinely convey meaning, then they may have effects even in democracies. Alternatively, the effects of leader images on citizens' compliance may be moderated by the institutional context in which they are located. Thus, we hope this article represents the first case in a new cross-national research agenda on the effect of leaders' images on individual compliance and support.

Acknowledgments

For feedback and advice on this project, we thank the special issue editors, the *Comparative Political Studies (CPS)* editors, three anonymous reviewers, Christopher Adolph, Lisa Blaydes, Matt Buehler, P. J. Henry, James Hollyer, Calvert Jones, Victor Menaldo, Jon Rogers, Julie Wronski, and Sean Yom, and participants at University of Mississippi's Institutional Relations (IR) Working Group and the Midwest Political Science Association Annual Meeting. For graphic design help, we thank Erika Steiskal. For assistance at the Social Science Experimental Laboratory (SSEL), we thank Alizeh Bahtra and Arusyak Hakhnazaryan. Finally, we are deeply indebted to Becky Morton for assistance throughout the project; without Becky, this undertaking would have not been possible.

Authors' Note

The authors are listed in alphabetical order and contributed equally. Any remaining errors are their own. Replication files and appendices are available at https://dataverse.harvard.edu/dataset.xhtml?persistentId=doi:10.7910/DVN/MITC3W

Declaration of Conflicting Interests

The authors declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The authors disclosed receipt of the following financial support for the research, authorship, and/or publication of the article: The authors thank Michael Findley and University of Texas at Austin (all), University of Mississippi Office of Research and Sponsored Programs Investment Grant (Zeira), Temple University Office of the Vice President for Research Administration (Bush), and the Program on Middle East Political Science (Erlich and Prather) for generously contributing funding to this project.

Notes

1. The uncorrected (for multiple comparisons) results are largely similar.
2. Personalism often manifests itself in personality cults. Personality cults are frequently characteristic of totalitarian regimes but are not sufficient for them to exist (Linz, 2000) and are sometimes seen as an extreme form of neopatrimonial rule (see, for example, Chehabi & Linz, 1998).
3. Another function of personality cults is to reveal information about the regime's base of support. By forcing citizens to publicly accept fabulous claims (and seeing when citizens stop being willing to do so), they allow leaders to gauge their true level of support in a context of "preference falsification" (Kuran, 1991; Márquez, 2013; Wedeen, 1999). In contrast, our article examines *individual* rather than *societal* outcomes.
4. That is to say, our theory's scope is authoritarian regimes in which power is concentrated in the personal hands of the leader, whether he is supported by a party, military, or family. Of course, some types of iconography will be more likely in certain contexts than others.
5. Though in some countries, she is required by the state to do so.
6. It is also possible that a leader's image could lead to less compliance. For example, a leader's image could lead to everyday forms of resistance such as rumors, jokes, or other "weapons of the weak" (Scott, 1990). Leaders' images could also lead to less compliance when the leader is in the process of being deposed and might vary across individuals. Theorizing about the effect of images in regime transitions would be a fruitful path for future research.
7. *Individuals*, too, may display authoritarian iconography in ways that will generate compliance via coercion: When people see others displaying images, they are reminded that people support the state and can inform on them.
8. An alternative research design would remove individuals from the authoritarian context and examine the effects of exposing them to the images—perhaps, following Miguel, Saiegh, and Satyanath (2011), using a sample of immigrants in Europe. This approach would isolate the effects of the images from the country context.

9. The sequence occurs as follows: subjects are asked to stare at their computer, where they observe a fixation point (a black dot in the center of the white screen) for ~1,000 ms, a disarticulated version of a treatment image for 29 frames (the "forward mask"; ~484 ms on a 60 Hz monitor), a treatment image for one frame (~16.7 ms on a 60 Hz monitor), a disarticulated version of a treatment image for ~100 ms, a fixation point for ~1,000 ms, a disarticulated version of a treatment image for 29 frames (the "backward mask"), a treatment image for one frame, and a disarticulated version of a treatment image for ~100 ms. The one-frame exposure and two-time prime is consistent with previous research on priming (Weinberger & Western, 2008). Monitors in our study report refresh rates of 59 to 60 Hz.
10. Some recent examples include Erisen, Lodge, and Taber (2014), Kam and Zechmeister (2013), Olivola and Todorov (2010), and Weinberger and Western (2008). Although social scientists widely use priming, it has recently come under criticism because several prominent priming studies have been difficult to replicate and, in some cases, found to be fraudulent. To restore credibility to the social science priming research agenda, Kahneman (2012) recommends that researchers publicly commit to their planned study and "[p]re-commit to publish the results, letting the chips fall where they may, and make all data available for analysis by others" (p. 2). We follow this strategy.
11. We use the stylized eyes used in Haley and Fessler (2005) although, as far as we know, there is no experiment that indicates that the specific type of eyes matters. The stylized eyes are displayed on a black background to ensure the comparability of the treatments.
12. Ideally, we would also have had a condition with only the eyes to identify how much of the effect of exposure to the leader's image is due to the general monitoring treatment. Because of the difficulties associated with assembling a large sample, we could not pursue this avenue.
13. See the online appendix for results using a continuous measure of the money paid as a proportion of the money owed.
14. See the online appendix for results of the factor analysis.
15. That is, because we do not find that images of the authoritarian leader have an effect on political compliance or support, we do not attempt to differentiate the effect of the leader from a generic monitoring effect. Thus, we also do not include the comparison between the two treatment groups in our corrections for the multiple comparison problem.
16. See the online appendix for the balance table, as well as a discussion of how we treat missing values of the covariates.
17. The first p value reported is the unadjusted p value, and the second is the adjusted p value; we use this format throughout the rest of the article.
18. In this specification, we omit the two respondents who were audited in the first round. Respondents who are audited should be very unlikely to donate money; due to the small number of these respondents ($n = 2$), we simply omit them. Furthermore, we are reporting the rank p values but the interpretation is the same of the t test p values.

19. We report *p* values here from a rank sum test, which does not assume normality, but the results are the same either way.

References

Alm, J., Jackson, B. R., & McKee, M. (1992). Estimating the determinants of taxpayer compliance with experimental data. *National Tax Journal*, *45*, 107-114.

Altbach, P. G., & Knight, J. (2007). The internationalization of higher education: Motivations and realities. *Journal of Studies in International Education*, *11*, 290-305.

Andreoni, J., Erard, B., & Feinstein, J. (1998). Tax compliance. *Journal of Economic Literature*, *36*, 818-860.

Bellin, E. (2004). The robustness of authoritarianism in the Middle East: Exceptionalism in comparative perspective. *Comparative Politics*, *36*, 139-157.

Benjamini, Y., & Hochberg, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society. Series B (Methodological)*, *57*, 289-300.

Blass, T. (1999). The Milgram paradigm after 35 years: Some things we now know about obedience to authority. *Journal of Applied Social Psychology*, *29*, 955-978.

Blaydes, L. (2011). *Elections and distributive politics in Mubarak's Egypt*. New York, NY: Cambridge University Press.

Bruhn, M., & McKenzie, D. J. (2008). *In pursuit of balance: Randomization in practice in development field experiments* [Policy research working paper]. World Bank. Retrieved from http://elibrary.worldbank.org/doi/abs/10.1596/1813-9450-4752

Bueno de Mesquita, B., Smith, A., Siverson, R. M., & Morrow, J. D. (2003). *The logic of political survival*. Cambridge, MA: MIT Press.

Bush, S. S., & Jamal, A. (2015). Anti-Americanism, authoritarian politics, and attitudes about women's representation: Evidence from a survey experiment in Jordan. *International Studies Quarterly*, *59*, 34-45.

Cadsby, C., Maynes, E. B., & Trivedi, V. U. (2006). Tax compliance and obedience to authority at home and in the lab: A new experimental approach. *Experimental Economics*, *9*, 343-359.

Chehabi, H. E., & Linz, J. J. (1998). A theory of sultanism 1: A type of non-democratic rule. In H. E. Chehabi & J. J. Linz (Eds.), *Sultanistic regimes* (pp. 3-25). Baltimore, MD: Johns Hopkins University Press.

Corstange, D. (2013). Ethnicity on the sleeve and class in the heart. *British Journal of Political Science*, *43*, 889-914.

Davidson, C. M. (2013). *After the sheikhs: The coming collapse of the Gulf monarchies*. New York, NY: Oxford University Press.

Diamond, L. (2010). Why are there no Arab democracies? *Journal of Democracy*, *21*, 93-112.

Erisen, C., Lodge, M., & Taber, C. S. (2014). Affective contagion in effortful political thinking. *Political Psychology*, *35*, 187-206.

Gandhi, J., & Przeworski, A. (2007). Authoritarian institutions and the survival of autocrats. *Comparative Political Studies*, *40*, 1279-1301.

Gause, F. G. (1994). *Oil monarchies: Domestic and security challenges in the Arab Gulf states*. New York, NY: Council on Foreign Relations Press.

Gibson, J. L., Lodge, M., Taber, C., & Woodson, B. (2010). *Can judicial symbols produce persuasion and acquiescence? Testing a micro-level model of the effects of court legitimacy*. Retrieved from https://www.researchgate.net/publication/229049353_Can_Judicial_Symbols_Produce_Persuasion_and_Acquiescence_Testing_a_Micro-Level_Model_of_the_Effects_of_Court_Legitimacy

Haley, K. J., & Fessler, D. M. T. (2005). Nobody's watching? Subtle cues affect generosity in an anonymous economic game. *Evolution and Human Behavior*, *26*, 245-256.

Hasseldine, J., & Hite, P. A. (2003). Framing, gender, and tax compliance. *Journal of Economic Psychology*, *24*, 517-533.

Hassin, R. R., Ferguson, M. J., Shidlovski, D., & Gross, T. (2007). Subliminal exposure to national flags affects political thought and behavior. *Proceedings of the National Academy of Sciences*, *104*, 19757-19761.

Heard-Bey, F. (2005). The United Arab Emirates: Statehood and nation-building in a traditional society. *Middle East Journal*, *59*, 357-375.

Hurd, I. (1999). Legitimacy and authority in international politics. *International Organization*, *53*, 379-408.

Iconography. (2010). In *The Oxford dictionary of English* (3rd ed.). Oxford, UK: Oxford University Press. Available from www.OxfordDictionaries.com

Jones, C. W. (2015). Seeing like an autocrat: Liberal social engineering in an illiberal state. *Perspectives on Politics*, *13*, 24-41.

Kahneman, D. (2011). *Thinking, fast and slow*. New York, NY: Farrar, Straus and Giroux.

Kahneman, D. (2012). *Open letter*. Retrieved from http://www.nature.com/polopoly_fs/7.6716.1349271308!/suppinfoFile/Kahneman%20Letter.pdf

Kam, C. D., & Zechmeister, E. J. (2013). Name recognition and candidate support. *American Journal of Political Science*, *57*, 971-986.

Kamrava, M. (2009). Royal factionalism and political liberalization in Qatar. *The Middle East Journal*, *63*, 401-420.

Karlan, D., List, J. A., & Shafir, E. (2011). Small matches and charitable giving: Evidence from a natural field experiment. *Journal of Public Economics*, *95*, 344-350.

Krause, W. (2008). *Women in civil society: The state, Islamism, and networks in the UAE*. New York, NY: Palgrave Macmillan.

Kuklinski, J. H., Cobb, M. D., & Gilens, M. (1997). Racial attitudes and the "New South." *The Journal of Politics*, *59*, 323-349.

Kuran, T. (1991). Now out of never: The element of surprise in the East European Revolution of 1989. *World Politics*, *44*, 7-48.

Landry, C. E., Lange, A., List, J. A., Price, M. K., & Rupp, N. G. (2010). Is a donor in hand better than two in the bush? Evidence from a natural field experiment. *American Economic Review*, *100*, 958-983.

Lawrence, A. (2013, August). *Repression and activism among the Arab Spring's first movers: Morocco's (almost) revolutionaries*. Presented at the 2013 Annual Meeting of the American Political Science Association, Chicago, IL.

Lieberman, E. S. (2005). Nested analysis as a mixed-method strategy for comparative research. *American Political Science Review*, *99*, 435-452.

Linz, J. J. (2000). *Totalitarian and authoritarian regimes*. Boulder, CO: Lynne Rienner.

Lust-Okar, E. (2006). Elections under authoritarianism: Preliminary lessons from Jordan. *Democratization*, *13*, 456-471.

Magaloni, B. (2006). *Voting for autocracy: Hegemonic party survival and its demise in Mexico*. New York, NY: Cambridge University Press.

Márquez, X. (2013, August). *A model of cults of personality*. Presented at the 2013 Annual Meeting of the American Political Science Association, Chicago, IL.

Menaldo, V.A. (2012). The Middle East and North Africa's resilient monarchs. *Journal of Politics*, *74*, 707-722.

Miguel, E., Saiegh, S. M., & Satyanath, S. (2011). Civil war exposure and violence. *Economics & Politics*, *23*, 59-73.

Milgram, S. (1963). Behavioral study of obedience. *The Journal of Abnormal and Social Psychology*, *67*, 371-378.

Moore, R. T. (2012). Multivariate continuous blocking to improve political science experiments. *Political Analysis*, *20*, 460-479.

Olivola, C. Y., & Todorov, A. (2010). Elected in 100 milliseconds: Appearance-based trait inferences and voting. *Journal of Nonverbal Behavior*, *34*, 83-110.

Panagopoulos, C. (2014). Watchful eyes: Implicit observability cues and voting. *Evolution and Human Behavior*, *35*, 279-284.

Robinson, A. L. (in press). National identification and inter-ethnic trust: Evidence from an African border region. *Comparative Political Studies*.

Roth, G. (1968). Personal rulership, patrimonialism, and empire-building in the new states. *World Politics*, *20*, 194-206.

Schaar, J. H. (1981). *Legitimacy in the modern state*. New Brunswick, NJ: Transaction Books.

Scott, J. C. (1990). *Domination and the arts of resistance: Hidden transcripts*. New Haven, CT: Yale University Press.

Seawright, J., & Gerring, J. (2008). Case selection techniques in case study research: A menu of qualitative and quantitative options. *Political Research Quarterly*, *61*, 294-308.

Vora, N. (2008). Producing diasporas and globalization: Indian middle-class migrants in Dubai. *Anthropological Quarterly*, *81*, 377-406.

Vora, N. (2011). From golden frontier to global city: Shifting forms of belonging, "freedom," and governance among Indian businessmen in Dubai. *American Anthropologist*, *113*, 306-318.

Vora, N. (2013). *Impossible citizens: Dubai's Indian diaspora*. Durham, NC: Duke University Press.

Way, L. A., & Levitsky, S. (2006). The dynamics of autocratic coercion after the Cold War. *Communist and Post-Communist Studies*, *39*, 387-410.

Weber, M. (1978). *Economy and society: An outline of interpretive sociology*. Berkeley: University of California Press.
Wedeen, L. (1999). *Ambiguities of domination: Politics, rhetoric, and symbols in contemporary Syria*. Chicago, IL: University of Chicago Press.
Wedeen, L. (2010). Reflections on ethnographic work in political science. *Annual Review of Political Science*, *13*, 255-272.
Weinberger, J., & Westen, D. (2008). RATS, we should have used Clinton: Subliminal Priming in Political Campaigns. *Political Psychology*, *29*, 631-651.

Author Biographies

Sarah Sunn Bush is an assistant professor of political science at Temple University. She is the author of *The Taming of Democracy Assistance: Why Democracy Promotion Does Not Confront Dictators* (Cambridge University Press, 2015). Her other research on international politics and democracy has been published or is forthcoming in *International Organization, International Studies Quarterly, The Review of International Organizations*, and *Comparative Politics*.

Aaron Erlich is a PhD candidate at University of Washington, Seattle, and from August 2016 an assistant professor at McGill University. His current research interests include the impact of information in developing countries, measuring uncertainty, democratization, and experimental design. Previous work has appeared in *American Political Science Review*.

Lauren Prather is an assistant professor of political science in the School of Global Policy and Strategy at University of California, San Diego. Her research interests include public opinion about foreign policy, the political economy of foreign aid, democracy assistance and elections, and experimental methods.

Yael Zeira is Croft assistant professor of political science and international studies at University of Mississippi. Her primary areas of interest are political conflict and violence, Middle East politics, and field research methods. In addition to *Comparative Political Studies (CPS)*, her research is forthcoming in *The Journal of Conflict Resolution*.

Original Article

Comparative Political Studies
1–35

DOI: 10.1177/0010414015626436
cps.sagepub.com
SAGE

Can Politicians Police Themselves? Natural Experimental Evidence From Brazil's Audit Courts

F. Daniel Hidalgo[1], Júlio Canello[2], and Renato Lima-de-Oliveira[1]

Abstract

To enhance government accountability, reformers have advocated strengthening institutions of "horizontal accountability," particularly auditing institutions that can punish lawbreaking elected officials. Yet, these institutions differ in their willingness to punish corrupt politicians, which is often attributed to variation in their degree of independence from the political branches. Taking advantage of a randomized natural experiment embedded in Brazil's State Audit Courts, we study how variation in the appointment mechanisms for choosing auditors affects political accountability. We show that auditors appointed under few constraints by elected officials punish lawbreaking politicians—particularly co-partisans—at lower rates than bureaucrats insulated from political influence. In addition, we find that even when executives are heavily constrained in their appointment of auditors by meritocratic and professional requirements, auditors still exhibit a pro-politician bias in decision making. Our results suggest that removing bias requires a level of insulation from politics rare among institutions of horizontal accountability.

[1]Massachusetts Institute of Technology, Cambridge, USA
[2]Instituto de Estudos Sociais e Políticos, Rio de Janeiro, Brazil

Corresponding Author:
F. Daniel Hidalgo, Massachusetts Institute of Technology, 77 Massachusetts Avenue, Room E53-470, Cambridge, MA 02139, USA.
Email: dhidalgo@mit.edu

Keywords
Latin American politics, corruption, accountability

Introduction

Elections are the defining institution of democracy, yet disappointment with electoral competition's capacity to reliably produce the rule of law is widespread (Collier, 2011; Fukuyama, 2011; Hayek, 1960). Similarly, Madisonian solutions, such as the separation of powers between independently elected executives and legislatures, have frequently failed to foster robust oversight of state functions (O'Donnell, 1994; Morgenstern & Manzetti, 2003). This disappointment has led scholars and policy makers to argue for the creation of institutional arrangements that can compensate for the failures of legislatures to ensure that officials, particularly members of the executive branch, govern within the bounds of the law. Disillusionment with standard institutional solutions has lead to increasing attention to the creation and functioning of *unelected* institutional bodies designed to oversee the state and sanction lawbreaking by the elected branches.

Among the most common non-elected institutional solutions proposed for constraining the state are "auditing agencies" or formally independent bodies tasked with monitoring government compliance with the law and, in many cases, sanctioning non-compliance. Multilateral agencies such as the World Bank and the Inter-American Development Bank argue that these agencies can be "an essential instrument for development, promoting good governance by improving public sector management" (Dye & Stapenhurst, 1998, p. 10). Prominent theoretical analyses of good governance suggest that horizontal accountability necessitates "state agencies that are authorized and willing to oversee, control, redress, and if need be sanction unlawful actions by other state agencies" (O'Donnell, 1998, p. 19). Furthermore, empirical analyses of how the revelation of government corruption affects political accountability (e.g., Ferraz & Finan, 2008) hinge on the credibility of the auditing institutions, which produce the information in the first place.

Of course, the degree to which these agencies actually are able and willing to confront elected officials who break the law differs greatly across contexts (Santiso, 2009). To explain this variation, scholars have emphasized—among other factors—the importance of institutional design (Diamond, 2002; Moreno, Crisp, & Shugart, 2003). Of particular importance are the rules governing how the unelected officials charged with monitoring the state are chosen, particularly the degree to which the process is shielded from political considerations. Yet, in stark contrast to the vast literature on the institutional rules governing legislatures and executives, empirical assessments of the

rules structuring audit agencies and related agencies of horizontal accountability are relatively few.[1]

In this article, we study how the rules governing auditor selection affect the outcome of audits and the extent to which these outcomes are politically biased. Specifically, we take advantage of two unique institutional features governing state-level auditing institutions in Brazil that create natural experimental leverage to test the link between selection rules and audit outcomes. First, state-level audit courts (ACs) are composed of councilors who are selected by one of a variety of possible procedures: (a) appointed by the executive with few restrictions, (b) appointed by the legislature with few restrictions, (c) appointed by the executive where the nominated member must be a career bureaucrat, and (d) professional "substitute" auditors who are not appointed by the electoral branches. In general, and as we discuss in detail below, these selection rules create two sets of auditors: professional bureaucrats and professional politicians. Second, annual audits of government agencies and subnational governments are assigned by *random* lottery to each of the councilors. These two institutional features create variation in the types of officials that are tasked with identifying and punishing malfeasance but remove the potential for confounding induced by strategic selection by the auditors of government actions to investigate. Thus, the research design allows for robust causal inferences on the relationship between official type and decision making in investigations of government lawbreaking.

Overall, we find that auditors appointed by the political branches with few restrictions are more reluctant to punish local governments than career bureaucrats. Although the average difference between bureaucrats and politicians is modest, there is substantial heterogeneity by the partisan affiliation of the mayor under scrutiny: Politician auditors are substantially more lenient toward mayors belonging to the party that appointed them than politicians belonging to other parties. Career bureaucrat auditors are heterogeneous as well: We find that even when governors are heavily constrained in their choices by the requirement to appoint career civil servants, *appointed bureaucrats* are less likely to punish politicians when compared with *unappointed bureaucrats* who are not selected by the executive. The answers we obtain have important implications for institutional design, as appointed politicians and appointed bureaucrats—even when granted strong tenure protections—behave quite differently from unappointed bureaucrats when tasked with ferreting out corruption and lawbreaking. Ensuring consistency of decision making and the removal of political bias from the application of the law, according to our results, may require a level of insulation from politics rare among institutions of horizontal accountability.

Auditing Institutions and Horizontal Accountability

Audit institutions such as Brazil's ACs are quite heterogeneous organizations that vary both on how the information they generate is used and how they are structured (Santiso, 2009; Speck, 2011). Most generally, audit institutions are unelected public agencies tasked with generating information about state activities that can be used for a variety of purposes by policy makers, bureaucrats, and the broader public. A primary function of this information is to provide actors—such as legislatures, public prosecutors, and voters—an evidentiary basis for punishing lawbreaking (Schedler, 1999). Another common use for information generated by audit institutions is to identify inefficiencies and otherwise poor performance in policy implementation, which can be used by policy makers to reform government processes. In some cases, audit institutions can directly sanction lawbreakers, but generally these agencies are dependent on other actors such as public prosecutors, courts, and voters to punish misconduct.

The heterogeneity in auditing agencies' goals and capacities is reflected in variation in institutional organization. While some audit institutions are organized around a chief auditor, others are headed by a collegial body or panel of councilors, as is the case of Brazil's ACs (Santiso, 2009). Another dimension of variation, which we examine empirically, is the relationship between the audit institution and the political branches.

The degree to which audit institutions or any bureaucracy in a democracy fulfill their intended role is linked to their relationship with the elected branches and the relationship of the elected branches with each other (Moe, 1984). Of chief importance is institutional *independence*, that is, the degree to which the selection and survival in office of the institution's agents is controlled by elected officials (Wood & Waterman, 1991). On one extreme of no independence, a chief auditor may be unilaterally appointed by the executive and serves at his or her pleasure. In this case, the chief executive might prefer to select an agent interested in ferreting out deviations of the bureaucracy from the executive's preferred policies, but who also show little interest in the exposure of politically damaging lawbreaking by the executive himself or his allies. At the other extreme of high independence, auditors may be given life tenure by a committee of experts with no formal links to elected officials. Auditors picked under such an arrangement are presumably more willing to confront executive lawbreaking.

Lack of independence does not imply that auditors cannot generate useful information and sanction wrongdoing, but standard delegative models predict that their behavior will be aligned with the preferences of the electoral authorities that control their selection and persistence in office (Calvert, McCubbins, & Weingast, 1989). Audit agencies are often beholden to legislative majorities, for example, and thus likely to be biased in favor of officials

belonging to the majority party or coalition. Yet, even these legislature-beholden audit agencies may be quite willing to expose malfeasance by the executive, particularly during periods of divided government. Of course, executive dominance of the legislature through partisan ties or patronage is not uncommon, so even nominal independence from the executive may be undermined by cross-branch collusion.

Our research design enables us to test the empirical relevance of the predictions that arise from delegative models of separation of powers when applied to agencies of horizontal accountability. Randomization of cases to councilors and a dependent variable that is comparable across units gives us an unusually strong opportunity to test our proposed hypotheses. Furthermore, in contrast to the existing empirical literature that has relied on cross-national comparisons (Blume & Voigt, 2011) and cross-sectional observational studies (Schelker & Eichenberger, 2010) potentially confounded by unmeasured factors, we can compare the behavior of different types of officials in a common institutional setting. These design features allow us to observe the degree to which politicians on the AC behave similar to bureaucrats, appointed or unappointed, when judging other politicians, and thus assess how much political incentives distort political accountability.[2]

The answers we obtain have important implications for institutional design, for if politicians behave very differently from bureaucrats when tasked with ferreting out corruption and lawbreaking, the case for insulation of auditors from the elected branches may be considerably strengthened. If, however, politicians do not exhibit bias toward other politicians, then this would suggest that concerns over political influence via the appointment process are exaggerated or overcome by other institutional factors.

The article proceeds as follows. First, we provide institutional background on Brazil's state ACs and the annual auditing process of municipalities' government accounts. In the next section, we provide background information and delineate several testable hypotheses drawing from the judge effects and inter-branch delegation literature. In the subsequent sections, we detail our research design, present basic characteristics of our data, and present our results. Finally, we conclude with a brief discussion of the theoretical implications of our results.

Audit Courts in Brazil

Audit institutions in Brazil follow the AC model, where the court acts as a quasi-judicial authority with an independent budget and staff, but headed by ministers or councilors (*conselheiros*) nominated by the political branches.

Both federal and state constitutions mandate that the ACs aid the national and state legislatures in overseeing public sector spending and programs by providing independent and professional assessments of compliance with the law. A chief advantage in studying the Brazil's state ACs is that they are collegiate bodies composed of councilors who are selected under different decision rules that imply varying levels of dependence on the elected branches.[3] The legal framework in the 1988 constitution grants the state legislature the authority to nominate four out of seven councilors on the court, as well as mandating that two councilors be professional auditors or public prosecutors. In general, to fill the "bureaucrat" slots on the court, the governor must choose, alternately, a career auditor or a public prosecutor off of a list of three nominees presented by the AC.[4] In addition to the two bureaucrat appointments, the executive can only choose one councilor unconstrained by technical requirements.[5] Independence of the councilors is further reinforced by the rule that they cannot be removed by the political branches and remain in office until a mandatory retirement age.

Every appointed councilor has to be vetted through a public hearing and win confirmation in the state legislature. Approval is by simple majority, the same process necessary to elect the president of the assembly. In Brazil, governors typically build multi-party coalitions by appointing party members to key executive positions, effectively building a majority in the local legislature (Abrucio, 1998; Santos, 2001). Consequently, councilors are normally candidates aligned with the governor and/or with the largest party in the assembly, usually representing the strongest member of the political coalition at the time of appointment. Although minority parties can propose candidates for the slots appointed by the legislature, those candidates still need to pass the bar of a simple majority. Minority victories only occur in rare cases of coordination failure between parties in the governing or majority coalition. In addition, because legislative minorities do not have the filibuster option in Brazil, opposition groups have little power in the nomination process.

When the court lacks regular councilors due to absences or retirement, unappointed bureaucrats (*Conselheiros-Substitutos or Auditores-Substitutos*) temporarily fill vacancies. Substitute auditors are career bureaucrats hired by a competitive and open selection procedure. Generally, substitutes are auditors who regularly prepare the evidence that form the basis of councilors' overall judgments. While serving as a substitute, an auditor enjoys the same prerogatives and salary as a regular councilor. A substitute can serve until the member returns or, in case of retirement or death, a new one is appointed. In a few cases, ACs hire auditors directly to serve as a substitute. As are summarized in Table 1, these rules thus create four types of

Table 1. Appointment Procedures for State AC Councilors.

Type	Appointed by	Restrictions	Number of positions
Executive appointed	Governor, with legislative approval	Minimal	1
Legislature appointed	Legislature	Minimal	4
Appointed bureaucrat	Governor, with legislative approval	Selected from a list of three public prosecutors	1
	Governor, with legislative approval	Selected from a list of three professional auditors	1
Unappointed bureaucrat	Not appointed	Only professional auditors	NA

Unappointed bureaucrats are substitutes that fill vacancies on the court. AC = audit court.

councilors: *executive appointed, legislature appointed, appointed bureaucrat*, and *unappointed bureaucrat*.

The ACs operate at the federal, state, and local levels. The federal AC (*Tribunal de Contas da União* or TCU) is responsible for investigating federal activities, including federal transfers to subnational governments and the operation of state-owned enterprises (SOEs). All 27 states have an analogous institution, designed to monitor each state government and all 5,570 of Brazil's municipalities. These state ACs (*Tribunal de Contas do Estado* or TCE) all have a similar overall structure, but vary substantially with respect to budget and staff size (Mello, Pereira, & Figueiredo, 2009).

The role played by politicians in the appointment of councilors—who will ultimately judge the accounts of other politicians—is a common source of criticism both in the press and in academic circles. A common charge is that councilors are selected through political influence irrespective of technical capacity. The perquisites of office—among them high salaries with tenure—are commonly treated as a reward for politicians approaching the end of their career, especially state deputies belonging to the legislative majority. According to a report prepared by the non-governmental organization (NGO) Transparency Brazil (Paiva & Sakai, 2014), based on an examination of all ACs in the country, 60% of councilors were elected politicians before being appointed to an AC. Another 17% are relatives of politicians and 20% faced or were convicted of criminal charges. Alston, Melo, Mueller, and Pereira (2005) claim that the greatest limitation of the Brazilian AC model is the appointment procedure for selecting councilors. Similar criticisms are made by Santiso (2009) and Speck (2011). Paiva and Sakai (2014) go as far as to

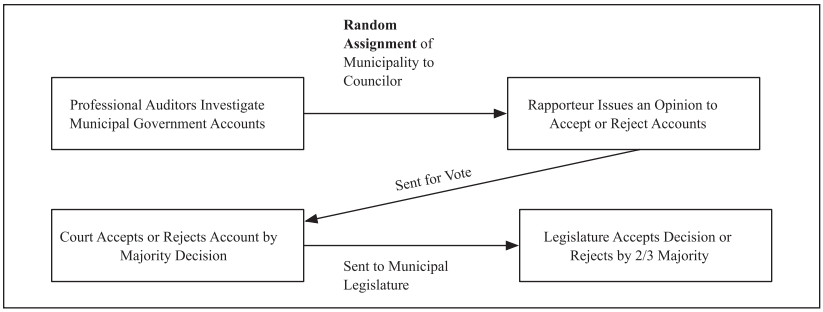

Figure 1. The municipality accounts auditing process.
This figure is a simplified representation of the accounts process, and details can vary by state.

say that ACs are designed not to work, arguing that politicians are appointed to neutralize the oversight role of the institution.[6]

Despite these criticisms, Pereira and Melo (2016) show that the information provided by court audits negatively affect the probability of municipal incumbent re-election when corruption is revealed, indicating that the activities of the courts are not as meaningless as some critics argue. Related research by Mello et al. (2009) shows that broader institutional factors, particularly volatility and political competition, affect the overall performance of the state courts. Specifically, states with higher levels of programmatic political competition are more likely to have professional auditors appointed to the court, as well as reject the annual accounts of the governor. Our research design allows us to directly test some of the mechanisms postulated by these authors, but we treat the broader institutional setting as fixed given that our comparisons are within states as opposed to across states.

One of the chief means by which ACs oversee state agencies is by annual audits (*prestação de contas*) of federal, state, and local governments. The ACs produce an overall recommendation to accept, accept with reservations, or reject the "accounts" of government entities with respect to compliance with the law. In this proposal, we focus on state ACs' adjudication of municipal accounts, which entails an examination of each municipality's execution of the budget, fiscal management, legality of contracts, procurement policies, fulfillment of mandated spending requirements, and related matters. This process is carried out in phases, where the first stage is a technical examination of each municipality's accounts by the professional auditing staff and the second stage is a deliberative process involving representatives of the public prosecutor's office and AC councilors. The overall process is illustrated in Figure 1. The recommendation of the technical staff and accompanying materials are given to

a randomly assigned AC councilor known as the "rapporteur" (*relator*) who adjudicates the case.[7] Because councilors receive technical assessment and evidence from the AC's permanent staff, the quality of evidence should be the same for all types of councilors. After a defense is presented, the rapporteur generates an opinion for adjudication by the court (or a subset of the court) on whether the municipality's accounts should be rejected, as well as any associated punishments. The court then decides by majority decision whether to uphold the rapporteur's opinion and notifies the municipal legislature about the result (known as *parecer prévio de contas*). The median time for a court to issue a decision is 2 years, though the process can drag out for many years.[8]

The final outcome of the audit process is an overall recommendation of approval, approval with accompanying recommendations for improved compliance with the law (approval with reservations), and rejection. Rejection of accounts, according to Mello et al. (2009), is the "most severe sanctions that the [Audit Court] can inflict on a mayor . . ." (p. 1228). The political ramifications of rejection can be severe: At the federal level, for example, the rejection of President Dilma Rousseff's accounts in 2015 was considered grounds for a possible impeachment. In addition to the negative political or electoral effects of the rejection, the court may set a fine, mandate reimbursements for financial losses due to irregularities, and even recommend civil and criminal prosecution. However, because the state ACs are not formally part of the judicial system, enforcement of these rulings are left to the public prosecutors and the courts. Enforcement can be blocked or delayed in the courts due to plaintiffs' extensive right to appeal, the complexity of statutes that govern public expenditures, and the courts' huge backlog of cases. Yet, despite inconsistent enforcement in the courts and as we discuss in the conclusion rejected accounts have become substantially more consequential in recent years due the passage of a law, which makes politicians with rejected accounts ineligible to run for elected office for 8 years.

Hypotheses

Because ACs are quasi-judicial institutions in a civil law legal system, councilor decision making is ostensibly constrained by legal procedure so as to produce consistent and predictable case outcomes. Yet in practice, whether a municipality's accounts are rejected or accepted can vary widely depending on the councilor assigned as rapporteur. Figure 2 displays the average rejection rate for councilors in our six-state sample (discussed below). As shown in the figure, some councilors reject the accounts of fewer than 5% of municipalities, yet others reject more than 70%. To account for this striking variation, we draw from the judicial politics literature linking judicial

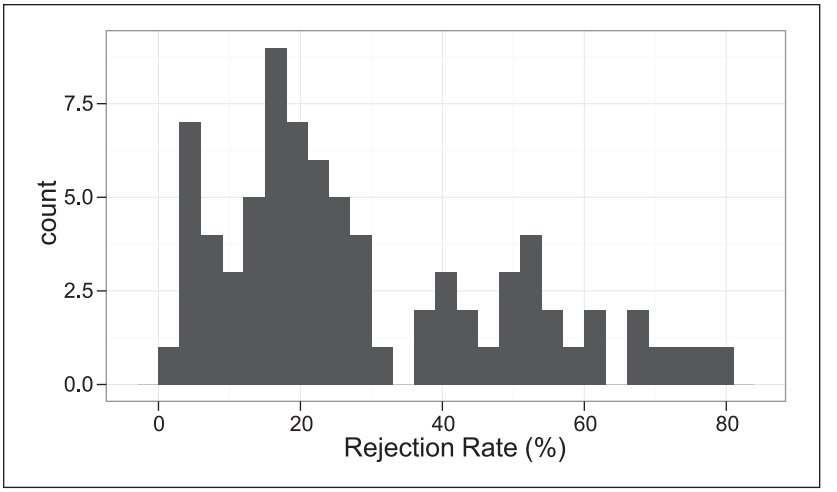

Figure 2. Variation in rejection rates.
Histogram shows the distribution of rejection rates of municipality accounts across 81 councilors in state ACs in six states over 10 years. Councilors who adjudicate fewer than 50 cases are omitted. AC = audit court.

identity—encompassing group affiliations such as gender (Boyd, Epstein, & Martin, 2010), ethnicity (Alesina & La Ferrara, 2014; Grossman, Gazal-Ayal, Pimentel, & Weinstein, 2016), or party (Pinello, 1999)—and case outcomes. Even in highly constrained legal environments, the judge effects literature finds that assignment to judges of distinct group identities can affect case outcomes both in individual-judge and panel settings. These group-based differences in judicial decision making are typically attributed to ideological differences correlated with group status, such as partisanship and ideology, or through in-group favoritism, as has been documented by examining differences in case outcomes when accused criminals are assigned to judges of the same or different ethnic group.[9] For the Brazilian case, Oliveira (2008) provides evidence that the professional background of Supreme Court Justices influence their behavior on constitutional cases.

When the decision is made by a single judge, it is obvious that judge identity such as partisanship or ethnicity can influence the ultimate outcome of the case. However, even in panel settings where a majority of judges must concur with a decision as is the case with ACs, the panel effects literature establishes two important reasons why the initial decision writer (i.e., the rapporteur) can powerfully affect case outcomes. First, the rapporteur has a first mover advantage due to the time and effort required to challenge, overturn,

and rewrite the initial decision. Given the large workloads of the ACs, dissenting majorities are unlikely to pay these costs unless the adverse outcome is consequential. Second, there is extensive evidence that courts tend to operate under a norm of consensus or "dissent aversion" when making routine decisions because judges seek to preserve collegiality by not challenging their colleagues' decisions (Fischman, 2015; Oliveira, 2012). In addition to maintaining collegiality, this phenomenon also reflects a norm of reciprocity, where judges decline to disagree with their colleagues, so as to avoid challenges to their own decisions in the future (Posner, 2010). Due to these two mechanisms, dissenters would likely only be willing to bear the costs of challenging the rapporteur in important cases, such as the adjudication of the accounts of the governor or mayors of major cities (we test this proposition below).

The primacy of the rapporteur in AC deliberations is evident in data on disagreements between court majorities and the rapporteur. We collected data on full court versus rapporteur decisions in four states[10] and found extremely low rates of disagreement: less than 2% in Maranhão, less than 1% in Pernambuco, 3% in Rio Grande do Sul, and less than 9% in Rio de Janeiro. These data indicate that by far the most important factor in determining court decisions is the recommendation of the rapporteur, which is consistent with empirical evidence on panel decision making in other settings. Lack of disagreement is not dispositive, however, because agreement might simply reflect the decisions of strategic rapporteurs who always recommend decisions that align with the majority. If being overturned is highly costly for the rapporteur and majorities can easily bear the costs of overturning the initial recommendation, then high rates of agreement would reflect the power of the majority to shape decision making.[11] This scenario is most likely in states with little political competition as most political councilors would be affiliated with a single party or group and can more easily coordinate (Mello et al., 2009). As such, we examine heterogeneity by a measure of the partisan diversity among political councilors to assess whether bureaucrat councilors are more distinct in their decision making in politically competitive states.

The chief distinction between our study and the judge effects literature is that instead of group identity such as ethnicity or gender, we focus on the institutional mechanism used to appoint the councilors. In line with standard models of delegation (e.g., Calvert et al., 1989), we hypothesize that distinct institutional procedures will be associated with councilor biases that comport with the political or career incentives of those with influence over the appointment. These associations arise because the governor or legislature will nominate councilors with biases that further their goals, under the constraint that the nominees must win consent from a majority of the legislature.

The bias in decision making could reflect strategic considerations by the councilors themselves. Although many politicians on the court are appointed at the end of their political careers, some return to electoral politics after their stint on the AC. For example, Weitz-Shapiro, Hinthorn, and Moraes (2015) find that retirements from the ACs tend to occur in election years, suggesting that a return to electoral politics is not rare. In addition, politicians on the court have been known to protect family members involved in politics,[12] as well as political allies. Because bureaucrat councilors generally are not involved in electoral politics, these considerations should not affect their decision making.

The differences in decision making between bureaucrat and political councilors could also arise due to variation in training and socialization. Bureaucrat councilors, for example, are more educated on average than non-bureaucrat councilors, which could lead them to evaluate technical evidence differently from their more political counterparts.[13] Bureaucrats appointed to the court will typically have served in the institution for many years and they will be more motivated by professional prestige and a desire to cultivate a reputation for technical expertise. Politicians with experience of governing, however, may be more sympathetic to the challenges faced by mayors in complying with complex bureaucratic regulations. Furthermore, former politicians on the court are more likely to be friends or acquaintances of the mayors they adjudicate, particularly co-partisans. These past relationships can consciously or unconsciously bias former politicians when assessing evidence of lawbreaking. From the point of view of models of delegation, the precise origin of individual councilors' bias is less relevant than whether or not the bias furthers the goals of those making the appointments.

What are the goals of governors and legislative leaders? With respect to adjudicating the accounts of municipalities, governors and legislators will wish to shield allied mayors from scrutiny and thus will not want their accounts to suffer rejection unless evidence of lawbreaking is pronounced. As is well established in the literature on Brazilian politics, mayors are important political actors who act as vote brokers and political operatives for gubernatorial and, in particular, legislative candidates (Bezerra, 1999; Mainwaring, 1999; Novaes, 2014). Candidates to state and national office invest considerable resources in cultivating mayors, as mayors have the extensive—often clientelistic—relationships with voters that are relied upon for votes on election day. Given the importance of currying favor with local politicians for the political careers of state-level politicians, the legislature and governor will, when unconstrained, likely nominate councilors who require a high standard of proof to reject the accounts of a mayor.[14] In contrast, bureaucrat councilors should be more interested in technical proficiency

and status within the institution, which makes them less likely to be sympathetic to the political interests of mayors. This yields our first hypothesis:

> **Hypothesis 1:** Municipal accounts adjudicated by governor- or legislature-appointed councilors will be rejected at lower rates than when adjudicated by bureaucrat councilors.

While we expect political councilors to be more favorable toward local governments than bureaucrat councilors, not all mayors will be treated equally. Although party attachments are more fluid in Brazil than some other established democracies, substantial evidence indicates that cross-level partisan ties are important for a range of outcomes including elections (Avelino, Biderman, & Barone 2012) and government transfers (Brollo & Nannicini, 2012).[15] As a result, we expect governors and legislatures to appoint officials who are sensitive to the interest of local co-partisans. State-level politicians will seek to forestall the negative electoral and financial consequences of account rejection for co-partisan officials by appointing councilors with biases that further their partisan aims. Although this bias is likely to be strategic or conscious, it need not be, as councilors may be unconsciously or implicitly biased toward co-partisans. Whatever the precise reason, governor-appointed councilors should be more reluctant to reject the yearly accounts of municipalities governed by mayors belonging to the party of the governor that appointed him than non-co-partisan mayors. A similar logic should pertain to legislature-appointed councilors, who should be particularly sensitive to the interests of the largest party of the state legislature.

> **Hypothesis 2:** Municipal accounts will be rejected at lower rates when the mayor belongs to the same party that selected the assigned governor- or legislature-appointed councilor.

In addition to partisan ties, the literature on Brazilian politics emphasizes the importance of multi-party electoral and governing coalitions in shaping executive–legislative relations. As such, councilors may be loyal to the constituent parties of the governor's coalition or the majority coalition in the state assembly, in addition to the specific party of the governor or the largest party in the legislature. Thus, we might expect legislature- and governor-appointed councilors to less frequently reject the accounts of mayors belonging to a party in the coalition that appointed the councilor.[16] In the case of governor-appointed councilors, the coalition of the governor should be most relevant, while legislature-appointed councilors will be responsive to the majority coalition within the state assembly.

Hypothesis 3: Municipal accounts will be rejected at lower rates when the mayor's party belongs to the coalition that selected the assigned appointed councilor.

Although the legal requirements for the two bureaucrat positions should substantially diminish the capacity of the legislature and executive to appoint councilors heavily biased toward their interests, it is still the case that the political branches have some discretion in which senior auditor or public prosecutor they appoint. Career bureaucrats are generally more interested in enhancing their prestige within their profession and organization and thus less attune to the interests of professional politicians, but there is likely some variation among career auditors or prosecutors in their propensity to punish mayors. As such, it is plausible that the governor and legislature would seek to appoint the most lenient of the potential bureaucrat councilors.[17] As explained above, however, a large proportion of cases in Brazilian states are adjudicated by unappointed bureaucrats (substitute councilors) who are members of the technical staff of the auditing institution and are not appointed by the political branches. As such, it is plausible that non-appointed bureaucrat councilors are, on average, even less sensitive to the interests of political actors than the appointed bureaucrat councilors selected by the governor. Under a similar logic, appointed bureaucrats should be more sympathetic to mayors who are co-partisans of the governor who appointed them than with mayors from other parties.

Hypothesis 4a: Municipal accounts adjudicated by appointed bureaucrat councilors will be rejected at lower rates than when adjudicated by unappointed bureaucrat councilors (substitute councilors).
Hypothesis 4b: Municipal accounts will be rejected at lower rates when the mayor belongs to the same party that selected the assigned appointed bureaucrat councilor.

In addition to our main hypotheses listed above, basic assumptions about the strategic logic of governors and legislatures also generate predictions about treatment effect heterogeneity. In particular, we expect political considerations to be especially important for the adjudication of the accounts of municipalities where the mayor is an important political actor in state politics. Although political importance can depend on a variety of factors, a good proxy is the size of the municipality, as mayors of larger municipalities can influence more voters and thus can be important allies for governors and legislators. As a result, we expect that the contrasts outlined in Hypotheses 1, 2, and 3 to increase in magnitude with the size of the municipality.

A similar logic applies to heterogeneity by year in the electoral calendar. Because of delays in adjudication, only audits that occur in the first or second year of a mayoral term are likely to be released in time to influence local elections, which occur every 4 years. As a result, audits of the first or second year are substantially more politically sensitive than audits in the third and fourth years. Because of this timing issue, we expect governor- or legislature-appointed councilors to be reluctant to reject the accounts of mayors in the first or second year of office, especially with respect to co-partisans or coalition partners.[18]

Research Design and Data

The common institutional rule across Brazil's ACs that the annual audits of government accounts are assigned by random lottery to councilors forms the basis of our empirical strategy. To take advantage of lottery, we collected 10 years (2000-2009) of municipal audit and councilor data from six Brazilian states: Bahia, Maranhão, Minas Gerais, Pernambuco, Rio de Janeiro, and Rio Grande do Sul.[19] These states are among the largest states in Brazil, containing about 40% of the country's population and 41% of its municipalities, and are heterogeneous with respect to economic and political characteristics. Maranhão, for example, has a GDP per capita of about US$3,500, whereas Rio Grande do Sul's is almost 3 times higher at about US$11,000. Politically, the states in our sample are also quite diverse: Maranhão is well known for its oligarchic politics (Cabral da Costa, 2006), whereas electoral politics in Minas Gerais and Rio Grande do Sul are highly competitive and structured around a stable left–right ideological divide (de Lima, 1997; Nunes, 2013; Santos, 2001). Given this economic and political diversity, our findings are likely to be broadly applicable to ACs throughout much of Brazil.

To classify councilors, we consulted a variety of sources, including news accounts and legislative debates. Preliminary information was obtained through ACs' and state legislatures' websites, consulting official documentation available online. To double check the data, we made formal requests to state ACs using their library system (when available) and Brazil's Freedom of Information Law, as well as sources from newspapers and magazines, official gazettes, interviews with the councilors themselves, and cross-referenced party affiliation data with records from Brazil's National Electoral Tribunal. Specifically, for each councilor, we collected information on year of appointment, branch of government that nominated him or her, prior party affiliation (when former politicians), governor's party at time of appointment, largest party in the state assembly when appointed, electoral coalition of governor when appointed, and if the councilor is a substitute, a bureaucrat or politician.

Our data set contains more than 22,000 cases, which encompasses more than 2,000 municipalities (see Table 2). Because a new randomization occurs every year, the unit of analysis is municipality–year. The average rate of rejection of municipal government accounts by the ACs is about 25%, but this overall average masks considerable state-by-state variation. In Rio Grande do Sul, the rejection rate is only about 8%, whereas in Maranhão, the rejection rate exceeds 60%. Data on treatment status are missing in a relatively small percentage of cases.

The distribution and number of councilor types can be found in the bottom panel of Table 2. We obtained biographical data on 93 different councilors and categorized them into five distinct types. The most numerous type is "legislature appointed," representing 40% of all councilors. Each councilor adjudicated an average of 231 cases. The substitute councilors, which we call "unappointed bureaucrat," are relatively numerous but six of the 23 substitutes observed in our sample adjudicated fewer than 25 cases. Note that substitutes were more active in the states of Maranhão, Minas Gerais, Pernambuco, and Rio Grande do Sul, so inferences involving this type of councilor are largely confined to these states.

To evaluate whether rejection rates are affected by partisan considerations, we created a binary variable that measures mayor-councilor partisan ties. In the case of municipalities assigned to governor-appointed councilors, this variable measures whether the mayor's party belongs to the party of the governor that appointed the rapporteur adjudicating the municipality's accounts. In the case of legislature-appointed councilors, this variable indicates that the mayor belongs to the largest party of the legislature at the time of the councilor's appointment. Among mayors assigned to appointed councilors (non-substitutes), we find a rate of mayor and councilor "co-partisanship" of 16%. For coding coalitions, we classified a mayor as sharing a coalition with the councilor if the mayor belonged to a party that was part of governing electoral coalition at the time of the councilor's appointment.[20] Data for coding party of mayors were obtained from the Supreme Electoral Court (*Tribunal Superior Eleitoral*).

Specification and Inference

We treat the natural experiment created by randomization of audits to councilors as a block randomized design. A separate randomization occurs in each state in each year and consequently each state–year pairing constitutes an experimental block. In three states—Pernambuco, Maranhão, and Rio Grande do Sul—the assignment lottery is restricted to prevent the same councilor from adjudicating the accounts of any given municipality 2 years in a row. To account for this restricted randomization procedure, we further stratify

Table 2. Descriptive Statistics.

Variable	Full sample	Bahia	Maranhão	Minas Gerais	Pernambuco	Rio de Janeiro	Rio Grande do Sul
No. of municipalities	2,257	417	217	853	183	91	496
No. of cases	22,542	4,170	2,170	8,530	1,830	910	4,932
% of cases rejected	24.9	26	66.3	20.5	45.3	14.1	8.3
% with missing treatment data	0.8	0.4	3.7	0.4	0.3	0.5	1.1
No. of councilors	93	12	11	18	23	9	20
No of governor appointed	20	5	3	4	3	2	3
No. of legislature appointed	37	4	4	7	9	6	7
No. of appointed bureaucrats	13	1	1	4	2	1	4
No. of unappointed bureaucrats	23	2	3	3	9	0	6

This table shows descriptive statistics on municipal audits (top panel) and councilor characteristics (bottom panel) for six states for the years 2000 to 2009.

municipalities in these states by the identity of the rapporteur in the previous year, which ensures that within each block, treatment assignment probabilities are equal. Controlling for these strata ensures that comparisons could not be confounded by cross-state or time-varying confounders. Randomization ensures that—in expectation—municipal-level differences cannot account for differences in rejection rates across councilors.

As is common in field and natural experiments, there is a degree of non-compliance with random assignment in three of our states where the rapporteur initially assigned to a given municipality does not adjudicate the case. In Maranhão, we observe a small degree of non-compliance due to vacation and retirement. In Pernambuco and Rio Grande do Sul, non-compliance is substantially larger as the initial randomization allocates cases only to appointed councilors, but in practice many cases are redistributed to substitute councilors. In Pernambuco and Maranhão, the second distribution of cases occurs via random lottery, whereas in Rio Grande do Sul, they are distributed to substitutes in order of seniority. Although the majority of redistribution to substitutes occurs due to vacation and retirements, strategic allocation to substitutes is also possible, which would possibly introduce bias. Fortunately, we observe the outcome of the initial randomization before the redistribution to substitutes, thus allowing us to still take advantage of the lottery as an instrument. To adjust for this non-compliance, we instrument[21] the endogenous treatment variable with assignment-to-treatment status as represented by this first stage equation:

$$T_i = \alpha_0 + \pi Z_i + \sum_{k=1}^{K-1} \mu_k B_{ki} + \varepsilon_i, \quad (1)$$

where T_i is a dummy variable for treatment status (e.g., municipality's rapporteur is a bureaucrat councilor) that varies at the municipality level, α_0 is the intercept, π is the effect of the instrument on treatment status, Z_i is an assignment-to-treatment indicator (e.g., municipality is randomly assigned to bureaucrat councilor), B_{ki} is a block dummy for the k th block, μ_k is the block effect, and ε_i is the disturbance term. The first stage is quite strong across all of our specifications with F statistics on the excluded instrument well over thresholds recommended in the literature.[22]

Our second stage estimating equation is as follows:

$$Y_i = \beta_0 + \tau T_i + \sum_{k=1}^{K-1} \gamma_k B_{ki} + u_i, \quad (2)$$

where Y_i is dummy variable for whether the accounts of the municipality are rejected, β_0 is an intercept, τ is the treatment effect, T_i is a treatment indicator, B_{ki} is block fixed effect for the k th block, γ_k is the block effect, and u_i

is the disturbance term. As is well known, two stage least squares identifies effects among "compliers," that is, municipalities that follow the treatment assignment.[23] In addition to this basic specification, in the online appendix we also present covariate adjusted results, which we estimate by including a vector of pre-treatment variables. Standard errors are heteroscedasticity-consistent and clustered on the unit of randomization.[24]

Because we perform several hypothesis tests with different subsamples and treatment variables, conventional *p* values risk producing false positives due to multiple testing. To account for this possibility, we report for our main hypotheses—in addition to conventional confidence intervals, *p* values that account for multiple testing using the Westfall and Young (1993) bootstrap method. This method controls for the family-wise error rate (the probability that one or more true null hypotheses are rejected) but is less conservative than Bonferroni-like tests because the resampling procedure accounts for the dependence between *p* values across individual tests. In addition to our main results, we also report treatment effect heterogeneity by municipality characteristics, but we treat these as exploratory analyses and thus do not adjust these *p* values for multiple testing.

For analyses of partisan bias, it is important to account for the fact that the probability of assignment to treatment varies by party. For mayors belonging to minor parties that never successfully elected a governor or achieved a plurality in the legislature, for example, the probability of having a partisan tie to the councilor adjudicating their accounts is 0. Under the same logic, mayors will have a probability of assignment to treatment that is a function of the number of AC councilors serving that year appointed by governors or legislatures controlled by his or her party. To account for this issue, we include a full set of block by party fixed effects when estimating co-partisanship effects, which ensures that comparisons are made within party strata and any effects are not confounded by cross-party differences.[25]

An implication of random assignment is that pre-treatment municipality characteristics should not be systematically correlated with the type of councilor assigned to adjudicate the accounts of the municipality. To show that this is the case, in the online appendix we examine two contrasts: (a) whether a municipality is assigned to a political councilor (appointed without technical requirements) or a bureaucrat councilor and (b) whether the municipality is assigned to a councilor who shares a partisan tie with the mayor. On a range of covariates, including lagged values of the outcome variable, lagged values of the treatment variables, and political and socioeconomic characteristics, covariate balance is consistent with random assignment.

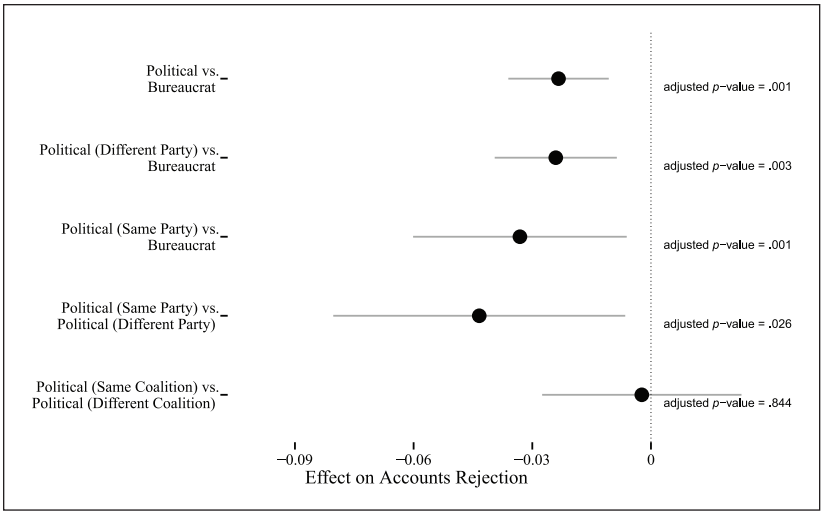

Figure 3. Political councilors versus bureaucrat councilors.
Point estimates and 95% confidence intervals are from a regression with block (first row) or block by party (second row) fixed effects. Confidence intervals based on standard errors clustered on unit of randomization, which varies by state. Mean of dependent variable in the full sample is 0.25; *p* values that adjust for multiple testing using the Westfall and Young step-down method are reported in the right margin.

Results

Recall that our first hypothesis posited that bureaucrat councilors would be comparatively more willing to punish mayors than political councilors appointed under less restrictive procedures. To evaluate this hypothesis, we compare the average probability of rejection of municipalities assigned to political councilors (governor appointed or legislature appointed) with those assigned to bureaucrat councilors, be they appointed or unappointed. Estimates of the causal effect of assignment to a political councilor as the rapporteur for the municipality's accounts can be found in the first row of Figure 3, along with multiple testing adjusted *p* values. This estimate supports Hypothesis 1, as we find that being assigned to a political councilor decreases the probability of rejection by about 0.023, which amounts to about 9% of the average rejection rate in the sample (0.25). Although supportive of the hypothesis, the point estimate is rather small and suggestive of only modest differences in bias between the two types of councilors.

Hypothesis 2 predicts that political councilors will be biased toward mayors with whom they share partisan ties. To test Hypothesis 2, we separate the

sample of mayors assigned to political councilors by whether or not the rapporteur of the municipality's accounts was appointed by a governor or legislature of the same party as the mayor. According to Hypothesis 2, the contrast in rejection rates between political councilors and bureaucrat councilors should be greatest when the mayor and the political councilor share a partisan tie. As evidenced by the coefficients in the second and third rows of Figure 3, our estimates are consistent with this expectation. In row 2, we compare municipalities assigned to governor- or legislature-appointed councilors appointed by a party other than the mayor's party with those assigned to career civil servants. This estimate represents a statistically significant increase in the probability of rejection by about .024. Even without a shared partisan affiliation, politician councilors punish mayors at greater frequencies than their bureaucrat counterparts, though the difference remains rather modest.

When the treatment group is mayors assigned to councilors *with* a partisan tie (row 3 in Figure 3), the coefficient increases substantially to a statistically significant .033. The magnitude of this effect is more politically meaningful than previous estimates given that it represents about 13% of the average rejection rate in the full sample. Assignment to a councilor with a shared partisan affiliation imparts a distinct advantage to mayors, on average.[26]

Next, we directly compare rejection rates among mayors assigned to political councilors with whom they share a partisan background to assignment to political councilors without a partisan link. In other words, this comparison holds councilor type constant by dropping cases adjudicated by bureaucrat councilors and examining only those municipalities assigned to governor- or legislature-appointed councilors. As shown in row 4 of Figure 3, partisan ties matter substantially. Conditional on assignment to a political councilor, being assigned a rapporteur who was appointed by a co-partisan governor or legislature reduces the probability of rejection by 0.04 when compared with those with accounts adjudicated by councilors appointed by another parties. The estimate for assignment to a councilor who shares an electoral coalition in row 5, however, is small and insignificant. This null result indicates that partisan interests are more potent and enduring than shared interests among coalitional partners, likely owing to the ideological heterogeneity and short-term duration that characterizes most winning electoral coalitions.[27]

Overall, Hypothesis 2 is supported by our data.[28] In fact, partisan bias is substantially larger than politician–bureaucrat differences, indicating that differences in partisan interests, on average, are more substantively important for the outcome of decisions than differences in professional background.

Heterogeneity by Municipality and Court Characteristics

As discussed in the "Hypotheses" section, the existing literature would suggest that the differences between politicians and bureaucrats would be especially pronounced in politically sensitive government accounts, especially if councilors acted strategically when adjudicating cases. In an exploratory fashion, we check whether the political-bureaucrat difference, as well as partisan bias, is larger when councilors adjudicate the accounts of larger municipalities and or when adjudicating accounts that could affect local election outcomes (see Table 3).[29] Larger municipalities, as classified by whether they are larger than the median municipality in the state, are more politically important, and thus, one could expect larger treatment effects. We find the opposite: The difference between political and bureaucrat councilors disappears when adjudicating the accounts of larger municipalities (column 1).[30] For partisan bias, there is no difference by size of municipality (column 4). Similarly, we expect that political councilors would be more lenient when adjudicating the accounts from the first 2 years of a mayor's term because these audit results would be more likely to be published in time to affect the next election. Against expectations, we find no heterogeneity for the politician–bureaucrat contrast (column 2), nor for partisan bias (column 5). These results provide suggestive evidence that politician councilors may not be particularly strategic and that differences between politician and bureaucrat members of the court are more likely due to differences in socialization or taste-based biases. That said, our measures of political sensitivity are only rough proxies, so these results should be interpreted with caution.

Next, we test the hypothesis that the composition of the court itself may play an important role in moderating the distinction between bureaucrat and politician councilors. As argued by Mello et al. (2009), bureaucrat councilors will be less likely to punish governments when they operate in a politically monolithic court out of fear of reprisals from allied councilors. In a more politically diverse setting, coordination among politician councilors will be less likely and bureaucrat councilors will have a freer hand to implement their preferred outcome. To test this, we classified each state–year as whether it was above or below the median in the diversity of partisan backgrounds of the politician councilors.[31] As shown in column 3 of Table 3, we find strong support for this hypothesis. In less diverse courts, bureaucrat councilors punish mayors at the same rate as political councilors, whereas in more diverse courts, bureaucrat councilors are much more likely to reject accounts. In fact, in more politically diverse courts, the effect of assignment to a bureaucrat councilor is a 0.08 increase in the probability of an accounts rejection, which is 4 times the magnitude of our full sample estimate.

Table 3. Heterogeneity by Municipality and Court Characteristics.

	Dependent variable				
	Accounts rejected				
	(1)	(2)	(3)	(4)	(5)
Political	−.038*** (.009)	−.014 (.010)	−.005 (.007)		
Political × Large Municipality	.030** (.012)				
Political × Early Period		−.018 (.013)			
Political × Diverse			−.080*** (.016)		
Political (same party)				−.048** (.020)	−.045 (.030)
Political (Same Party) × Large Muni				.007 (.022)	
Political (Same Party) × Early Period					.002 (.039)
Block fixed effects	X	X	X		
Party × Block Fixed Effects				X	X
Observations	20,786	21,045	21,045	12,718	12,923

In specifications with the "Political (Same Party)" variable, sample is restricted to accounts assigned to political councilors. "Large Municipality" is a dummy variable indicating that the municipality has a population larger than the median municipality in its state. "Early Period" is a dummy variable indicating that the accounts being adjudicated are for one of the first 2 years of a mayor's term. "Diverse" is a dummy variable indicating that the accounts were adjudicated in a court with above-the-median political diversity, as measured by the partisan backgrounds of political councilors. Main effects are omitted or perfectly collinear with block fixed effects.
*$p < .1$. **$p < .05$. ***$p < .01$.

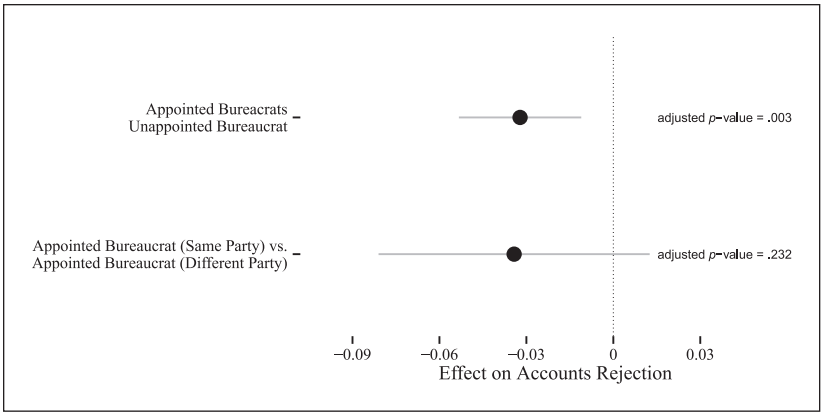

Figure 4. Unappointed versus appointed bureaucrat councilors.
Point estimates and 95% confidence intervals are from a regression with block (first row) or block by party (second row) fixed effects. Confidence intervals based on standard errors clustered on unit of randomization, which varies by state. Mean of dependent variable in the full sample is 0.25. *p* values that adjust for multiple testing using the Westfall and Young step-down method are reported in the right margin.

Do Appointed Bureaucrats Differ From Unappointed Bureaucrats?

Up until this point, we have found consistently negative, albeit mostly modest, effects of assignment to political as opposed to bureaucrat councilors. Grouping together appointed and unappointed (substitute) bureaucrats into a single category, however, may mask substantial differences between the two types of councilors. Because appointed bureaucrats are selected by the governor and approved by the legislature, appointed councilors—even if they are professional civil servants—may be chosen precisely because they tend to be favorable to politicians, particularly co-partisans of the governors. If so, the difference in rejection rates between appointed and *unappointed* councilors should be negative. Before discussing results, it is important to note that our inferences about unappointed bureaucrat councilors have more limited external validity than previous estimates. In Rio de Janeiro, Pernambuco, and Rio Grande do Sul, the substitute councilors are not eligible to be assigned cases in the initial randomization. As a result, for these states, we have no instrument for assignment of unappointed bureaucrats and consequently these states do not contribute to our estimates.

As evidenced by the top row of Figure 4, we find that assignment to bureaucrats appointed by the governor lowers the probability of rejection by

a statistically significant .034. To make a direct comparison with the results from Figure 3, assignment to a political councilor, as opposed to an unappointed bureaucrat, decreases the probability of rejection by .046, which is twice the size of the estimated treatment effect when the comparison group is a mix of both types of bureaucrats. When the political councilor was appointed by the same party of the mayor whose accounts he or she is judging, the effect size is a substantial, −0.093, which is almost 40% the full sample mean. These latter two estimates were not pre-specified in the analysis plan, but nevertheless suggest that there are very large benefits for mayors whose accounts are assigned to a partisan ally when the alternative is an *unappointed* career civil servant.

Overall, these results indicate that the difference between political councilors and bureaucrat councilors reported in Figure 3 is driven by the unappointed bureaucrats.[32] Indeed, when unappointed bureaucrats are removed from the sample, the difference between politicians and bureaucrats is a statistically insignificant .002 (SE = 0.007). Although not pre-specified, this estimate indicates that with respect to their observed behavior, appointed bureaucrats are closer to political councilors than to unappointed bureaucrats. In fact, we find some, albeit weaker, evidence that appointed bureaucrats are biased toward the parties of the governors who appointed them. When we compare rejection rates among municipalities assigned to an appointed bureaucrat with a partisan tie versus assignment to a bureaucrat without a partisan tie (row 2 of Figure 4), we find an imprecisely estimated difference of −0.034. Surprisingly, even heavily restricting the choice set of the executive does not prevent the selection of politically biased councilors.

An explanation for this result could be that the constraint faced by the executive when choosing bureaucrats may be less restrictive than it first appears or that the composition of the executive's choice set is subject to political manipulation. The formal criteria governing the composition of the list of three senior auditors or public prosecutors eligible for appointment by the governor emphasize seniority and "merit," but the actual process often involves an internal election. This internal election process could allow for executive influence over the final composition of the list through partisan and other political ties with the councilors who serve as voters. Or more generally, the career bureaucrats who expend effort on "winning" this internal selection process may be more willing to strike political bargains than those who do not.

However this list is constructed, if there is sufficient variation in the degree of pro-politician bias among the civil servants in the choices available to the governors, then the executive may succeed in choosing a bureaucrat substantially more aligned with his interests than the average civil servant. Some

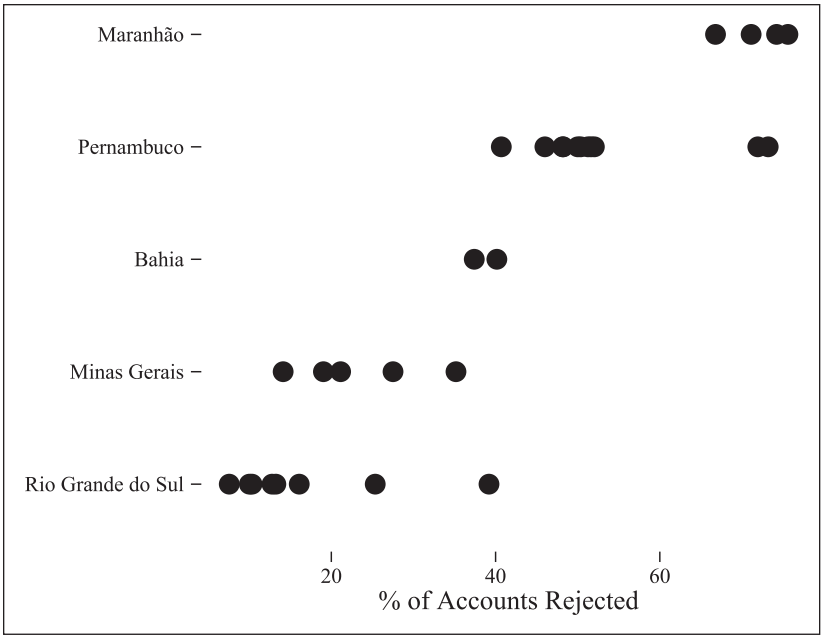

Figure 5. Rejection rates of unappointed bureaucrats.
Each dot represents the average rejection rate of one unappointed bureaucrat.

exploratory evidence for this is presented in Figure 5, which plots the distribution of account rejection rates by unappointed bureaucrats, with each dot representing one bureaucrat.[33] As is evidenced by the figure, in most states there is considerable variation in the willingness of unappointed bureaucrats to punish mayors, as proxied by the average rejection rate. Given that the auditors who act as substitutes are frequently the same civil servants eligible to be chosen by the governor, this variation indicates that executives often will have the option of choosing relatively lenient bureaucrats even when constrained to choose one among a menu of three options.

Further evidence that governors do succeed in choosing civil servants who are less likely to punish politicians can be found in Table 4. This table classifies unappointed bureaucrats into two categories: those who would eventually be appointed by the governor to fill a position on the accountability court and those who would never be appointed (as of 2010). As the table demonstrates, civil servants chosen by the governor have rejection rates (before appointment) that are meaningfully lower than those bureaucrats never selected to be formally on the court. This roughly 8-percentage-point

Table 4. Comparing Eventually Appointed and Never Appointed Bureaucrats.

Type	% accounts rejected	No. of councilors	No. of accounts
Never appointed	45	27	5,066
Eventually appointed	37	8	2,560

This table shows the rejection rates of career auditors who were either eventually or never appointed by the governor to the audit court.

difference is consistent with the hypothesis that governors strategically choose the most pro-politician choice available to them, which then produces the small difference between appointed bureaucrat councilors and politician councilors.

Conclusion

Despite the general consensus that institutions of "horizontal" accountability matter for reigning in government malfeasance, there is relatively scant evidence on the question of how these institutions can be designed to best fulfill their promise. In this article, we study how selection procedures affect the propensity of auditors to punish government officials. Our empirical analysis suggests that constraining those who appoint the auditors matters for subsequent behavior, as auditors appointed by relatively lax procedures and who tend to be politicians, are relatively less likely to punish subnational officials than career civil servants. Yet, even career civil servants appear to exhibit some bias toward politicians when they are appointed by the political branches. This finding calls for more research on the relationship between elected officials and the bureaucratic staff of accountability agencies, particularly on how civil servants—despite strong tenure protections and meritocratic promotion criteria—behave as political actors and respond to political incentives.

These results also have implications for the increasing reliance on unelected bodies such as auditing institutions and judicial courts to "correct" failures of the electoral process to select honest and competent public officials. Brazil is a case in point. The passage of the so-called "Clean Slate" law in 2010 created a new rule that bans politicians from holding elected office for 8 years after their accounts are rejected by a state or federal AC. In 2014, for instance, the public prosecutor's office sued to prevent almost 500 candidates from running for office, with the majority of challenges attributed to a rejection of accounts.[34] This law—even if inconsistently enforced—has dramatic consequences for the importance of these auditing institutions as their power over the careers of politicians has sharply increased. Yet, our results indicate that the decision of these

courts is partly a function of the partisan identity of the politician facing judgment. Given that a politicians' career is now on the line, losing the "lottery" of case assignment can have enormous consequences for an elected official. From a normative perspective, it is troubling that factors apart from the merit of the case such as party can have such serious consequences.

But perhaps of greater concern are the implications for voter welfare. Although a fuller theoretical analysis is needed, the growing power of the accountability courts will plausibly lead politicians to increase their efforts toward obtaining favorable judgments from them. If the interests of the courts are well aligned with the interests of the voters, such a change may incentivize better performance from politicians. If, however, the goals of the court do not perfectly align with those of the voters, then politicians may sacrifice some effort to comply with court demands that otherwise would be spent pleasing the electorate. Such a shift could have troubling implications even if the court's decision making was free of political considerations. For example, accountability agencies may be more interested in formal compliance with the letter of the law, rather than policy innovations tailored to the needs of the electorate. Strengthening accountability courts could have the unintended consequence of increasing the conservativism and sluggishness of local governments, as fear of inadvertently breaking the law could paralyze policy making and innovation.

Even more troubling, however, is the possibility that partisan decision making by court councilors incentivizes local officials to follow the priorities of their governor or legislative majority rather than their local constituency. In such a scenario, increasing the power of agencies of horizontal accountability may end up *undermining* electoral accountability and political responsiveness.

Acknowledgments

The authors thank the following people: Scott Desposato, Rebecca Weitz-Shapiro, Marcos Nóbrega, and Eric Kramon. For help with data, the authors thank the staff of audit courts in Bahia, Minas Gerais, and Pernamubco. The authors also appreciate input provided by participants at workshops at University of California, Berkeley and the Harris School at University of Chicago.

Declaration of Conflicting Interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) received no financial support for the research, authorship, and/or publication of this article.

Notes

1. Important exceptions include Mello, Pereira, and Figueiredo (2009); Santiso (2009); and Blume and Voigt (2011).
2. It is important to stress that in our design, selection procedures are not exogenously assigned, but rather cases are assigned to different types of officials selected under different procedures. Although we can examine the correlation between official-type and decision-making behavior that is unconfounded by characteristics of the cases, we cannot necessarily attribute differences in behavior to differences in selection procedures per se.
3. All councilors must meet general requirements: older than 35 and less than 65 years of age; moral standing and "unblemished" reputation; legal, accounting, economic, and financial or public administration knowledge; and more than 10 years of experience in a profession related to auditing. Despite these legal provisions, it is often the case that the importance or meaning of reputation, specialized knowledge, and experience is interpreted liberally, and thus, these restrictions are of little practical importance. We can find instances of former journalists, physicians, and dentist serving in audit courts (ACs), as well as several councilors with criminal charge or under judicial investigation.
4. Typically, this list (known as the *lista tríplice*) is formed by AC's councilors following rules of seniority and merit. As a result, high performing and long-tenured bureaucrats should be favored in the selection process. However, it is possible that internal politics in some instances play some role in the composition of the list. Furthermore, the timing of appointments is not strictly regulated and governors have been known to delay appointing bureaucrats to the AC. These delays and related controversies have led to appointments being frequently contested in court.
5. Of course, in some instances, governors might appoint highly qualified bureaucrats to their "unconstrained" slots, even though they are not required to do so. In our analyses below, we focus on the appointment mechanism as opposed to the actual qualifications of the appointees because judgments about professional qualifications are likely to be subjective.
6. It is not difficult to find examples of former politicians serving as councilors involved in corruption scandals with charges of influence peddling, money laundering, and receiving kickbacks. Robson Marinho, a long serving councilor in the São Paulo AC, for example, was removed from office in 2014 by judicial decision after being convicted for receiving bribes to favor a multinational company with state-owned enterprise (SOE) contracts. In a more extreme case, the councilor Luiz Eustáquio Tolêdo was convicted of murdering his wife in 1989, but kept his position in the Alagoas AC. He served a 6-year sentence where he was allowed to work during the day.
7. The details of the accounts process vary by state. In some states, the rapporteur is randomly assigned before the auditors investigate the municipal accounts. In addition, the public prosecutor advises the rapporteur in arriving at a decision. In some states, the final decision of the court is made by a panel of three councilors (known as a *câmara*) rather than the full court.

8. The length of adjudication is itself subject to political considerations as councilors may seek to delay a final decision until after elections or otherwise slow the process. As we show in the online appendix, political councilors, on average, issue decisions in less time than bureaucrat councilors and there is little evidence of partisan bias.
9. Judicial decision making in these types of situations has been modeled as a process of Bayesian updating after receiving a private, noisy, signal about the true state of the world, such as the guilt or innocence of an accused criminal (Alesina & La Ferrara, 2014; Iaryczower & Shum, 2012). Group-based bias in these models is parameterized as a weight placed on the costs of mistakenly convicting an innocent criminal versus mistakenly exonerating a guilty criminal. Under this framework, judge effects emerge due to group-based differences in the weight placed on each type of mistake.
10. Unfortunately, we were unable to obtain similar data from Bahia and Minas Gerais.
11. This possibility seems relatively unlikely given the large within-state variation in average rejection rates by councilor, which indicates little conformity within ACs.
12. For example, the president of the Maranhão AC faced accusations that he used his position to pressure mayors to support his son's 2014 run for a seat in the state legislature.
13. Relying on data on education from published curriculum vitae (CV), for example, we found that bureaucrat councilors are substantially more likely to have more than one undergraduate degree than other types of councilors.
14. This assumption is a simplification, as politicians may have other goals such as increasing the quality of public services by combating corruption. Governors, who are less dependent on the votes delivered by individual mayors than legislators, may be especially interested in punishing particularly corrupt local officials to encourage economic development. The extent to which this is true would reduce the chances of confirming our hypothesis.
15. In some states, however, parties are quite weak and party switching is common. As a consequence, party labels may be relatively uninformative in these contexts, thus making it less likely to find evidence of partisan bias. In the online appendix, we provide state-specific estimates of partisan bias.
16. A problem with studying state-level governing coalitions is that they tend to change dynamically over time, and thus, identifying precisely which parties are part of the governing coalition at any particular point in time can be error prone. Furthermore, collecting data on the precise composition of the governing coalition, particularly coalitions in operation decades ago when some of the councilors were appointed, is quite challenging. Due to these constraints, we focus on the governor's electoral coalition, which should be correlated with the governing coalition and for which data are more easily available.
17. It is also possible that auditors or prosecutors could make an agreement with the governor to favor his political allies in exchange for a permanent position on the court, though we lack evidence on any such arrangements.

18. Because of term limits that only allow two consecutive terms, timing effects should be especially pronounced in first terms. That said, audits could still be politically relevant in second terms as incumbents may seek to elect a co-partisan successor.
19. These particular states were chosen out of a combination of considerations; specifically data availability, size of the state, and regional diversity. Size of the state was important because statistical power to detect treatment effects depends on the number of municipalities, which are generally more numerous in populous states. Audit data were collected via web scraping of the ACs' public databases of cases. For Pernambuco, we were unable to obtain the original randomization for years before 2003. As a result, instrumental variable estimates drop these years.
20. For governor-appointed councilors, we classify mayors as being aligned with a councilor if the mayor belongs to a party that was formally party of the gubernatorial electoral coalition of the governor in power when the councilor was appointed. We use second round coalition, unless no second round occurred. For legislature-appointed councilors, we classify mayors as being aligned with a councilor if the mayor belongs to a party that was formally part of the electoral coalition of the largest party in the legislature when the councilor was appointed.
21. Intent to treat estimates are reported in the online appendix.
22. With respect to the *monotonicity* assumption necessary to identify average causal effects among compliers, there is no reason to believe that assignment to a particular type of councilor would induce a municipality to be endogenously assigned to the opposite treatment status. This is especially the case in Pernambuco and Maranhão where municipalities are re-randomized when cases are re-assigned.
23. As pointed out by Angrist (1998), two stage least squares with block fixed effects is not a consistent estimator for the complier average treatment effect, but rather is consistent for a precision weighted complier average treatment effect. In the online appendix, we also show results when employing a consistent estimator of the complier average treatment. See Lin (2013) for a discussion of consistent estimation of experimental treatment effects with covariate adjustment.
24. The unit of randomization in four of the six states is the municipality–year, as a distinct randomization occurred each year. In Rio de Janeiro (see *Deliberação 221*, January 30, 2001) and Pernambuco (e.g., see *Portaria* 438/2008), however, municipalities were assigned to groups and these fixed groups are randomized to councilors each year. The composition of these groups is rather haphazard suggesting that the correlation in outcomes within groups should be rather low, and thus only minimally affect precision. Nevertheless, for these states, we cluster our standard errors on group–year to account for the process of randomization. As a robustness check presented in the online appendix, we also show results where standard errors are clustered on municipality without regard to group or year.
25. Because units that have a 0 or 1 probability of assignment to treatment are effectively dropped from the sample when estimating partisan bias, inferences under

this design, are not necessarily applicable to all municipalities in the six states we study. Out of the 2,257 municipalities in our sample, 1,887 municipalities have a positive probability of a party match in at least 1 year.

26. In the online appendix, we show how this effect varies by party. We find that parties sometimes identified as more traditional and clientelistic (Partido de Partido do Movimento Democrático Brasileiro [PMDB] and the Partido da Frente Liberal / Democratas [PFL/DEM]) drive this result, whereas more councilors appointed by more programmatic parties (PT and PSDB) show no partisan bias.
27. This null result may also be due to measurement error, as electoral coalition may be a poor proxy for governing coalitions.
28. One might question why estimate in row 4 is not equal to the difference between the estimates in rows 2 and 3 (i.e., the difference-in-differences). In a block randomized experiment where all units have a positive probability of receiving each treatment condition, this would indeed be the case. In our case, however, within some party-block strata, there is a 0 probability of being assigned to either the bureaucrat or one of the partisan alignment treatments. This issue arises mostly because of the alternation rule used in some states, which mandates that municipalities not be assigned to the same rapporteur 2 years in a row. As a consequence, the sample of municipalities which contribute to treatment effect estimates differs somewhat across rows 2, 3, and 4. In row 4, the effective sample is compromised of units in strata where units have a positive probability of the "Political (Same Party)" or "Political (Different Party)" treatments, which is not exactly equivalent to the sample that contributes to the estimates in rows 2 and 3.
29. In the online appendix, we also present state-specific and party-specific treatment effect estimates.
30. A post hoc explanation for this unexpected finding might be that decisions on accounts for larger municipalities receive more scrutiny and consequently councilors feel more constrained when adjudicating these accounts. We have no direct evidence on this point, however.
31. We operationalize political diversity by computing the proportion of political councilors who belong to the largest party represented on the court in each year, where partisanship is measured by the party of the governor or legislature that appointed them.
32. Because the number of appointed bureaucrats is relatively few, this inference is somewhat more sensitive to the behavior of individual councilors, and thus, one should be cautious on the external validity of this conclusion. In the online appendix, we show the sensitivity of our estimates to dropping individual councilors from the data set. Comparisons involving appointed bureaucrats are indeed more sensitive to omission or inclusion of particular individuals.
33. In Rio de Janeiro, substitutes never adjudicate cases, so no data are available for this state.
34. Press release by the federal prosecutor's office. Accessed on September 15, 2014: http://noticias.pgr.mpf.mp.br/noticias/noticias-do-site/copy_of_eleitoral/eleicoes-2014-mpf-impugna-mais-de-4-mil-candidatos-sendo-500-pela-lei-da-ficha-limpa

Supplemental Material

The online appendix is available at https://mfr.osf.io/render?url=https://osf.io/mwn5h/?action=download%26mode=render

References

Abrucio, F. L. (1998). *Os Barões da Federação: Os Governadores e a Redemocratização brasileira* [Barons of the Federation: Governors and Democratization in Brazil]. São Paulo: Editora Hucitec.

Alesina, A., & La Ferrara, E. (2014). A test of racial bias in capital sentencing. *American Economic Review*, *104*, 3397-3433.

Alston, L., Melo, M., Mueller, B., & Pereira, C. (2005). *Who decides on public expenditures? The political economy of the budget process in Brazil* (Unpublished mimeo). Washington, DC: Inter-American Development Bank.

Angrist, J. D. (1998). Estimating the labor market impact of voluntary military service using social security data on military applicants. *Econometrica*, *66*, 249-288.

Avelino, G., Biderman, C., & Barone, L. S. (2012). Articulações Intrapartidárias e Desempenho Eleitoral no Brasil. *Dados*, *55*, 987-1013.

Bezerra, M. O. (1999). *Em Nome das "Bases": Política, Favor e Dependência pessoal* [In the name of the "bases": Politics, favor, and personal dependence]. Relume Dumarà.

Blume, L., & Voigt, S. (2011). Does organizational design of supreme audit institutions matter? *European Journal of Political Economy*, *27*, 215-229.

Boyd, C. L., Epstein, L., & Martin, A. D. (2010). Untangling the causal effects of sex on judging. *American Journal of Political Science*, *54*, 389-411.

Brollo, F., & Nannicini, T. (2012). Tying your enemy's hands in close races: The politics of federal transfers in Brazil. *American Political Science Review*, *106*, 742-761.

Cabral da Costa, W. (2006). *Do "Maranhão Novo" ao "Novo Tempo": A Trajetória da Oligarquia Sarney no Maranhão." Fundação Joaquim Nabuco* (Working paper). Retrieved from http://www.fundaj.gov.br/images/stories/observanordeste/cabral2.pdf

Calvert, R., McCubbins, M. D., & Weingast, B. R. (1989). A theory of political control and agency discretion. *American Journal of Political Science*, *33*, 588-611.

Collier, P. (2011). *Wars, guns, and votes: Democracy in dangerous places*. New York, NY: Random House.

de Lima, O. B., Jr. (org.). (1997). *O Sistema Partidário Brasileiro: Diversidades e Tendências—1982-1994* [The Brazilian party system: Diversity and tendencies, 1982-1994]. Rio de Janeiro: Editora Fundação Getulio Vargas.

Diamond, J. (2002). *The role of internal audit in government financial management: An international perspective* (IMF working paper). Washington, DC: International Monetary Fund.

Dye, K., & Stapenhurst, R. (1998). *Pillars of integrity: The importance of supreme audit institutions in curbing corruption*. Washington, DC: World Bank.

Ferraz, C., & Finan, F. (2008). Exposing corrupt politicians: The effects of Brazil's publicly released audits on electoral outcomes. *The Quarterly Journal of Economics*, *123*, 703-745.

Fischman, J. B. (2015). Interpreting circuit court voting patterns: A social interactions framework. *Journal of Law, Economics, & Organization*, *31*, 808-842.

Fukuyama, F. (2011). *The origins of political order: From prehuman times to the French revolution*. New York, NY: Farrar, Straus and Giroux.

Grossman, G., Gazal-Ayal, O., Pimentel, S. D., & Weinstein, J. M. (2016). Descriptive representation and judicial outcomes in multiethnic societies. *American Journal of Political Science*, *60*, 44-69. doi:10.1111/ajps.12187

Hayek, F. A. (1960). *The constitution of liberty*. Chicago, IL: University of Chicago Press.

Iaryczower, M., & Shum, M. (2012). The value of information in the court: Get it right, keep it tight. *The American Economic Review*, *102*, 202-237.

Lin, W. (2013). Agnostic notes on regression adjustments to experimental data: Reexamining freedman's critique. *The Annals of Applied Statistics*, *7*, 295-318.

Mainwaring, S. (1999). *Rethinking party systems in the third wave of democratization: The case of Brazil*. Stanford, CA: Stanford University Press.

Mello, M. A., Pereira, C., & Figueiredo, C. M. (2009). Political and institutional checks on corruption: Explaining the performance of Brazilian audit institutions. *Comparative Political Studies*, *42*, 1217-1244.

Moe, T. M. (1984). The new economics of organization. *American Journal of Political Science*, *28*, 739-777.

Moreno, E., Crisp, B., & Shugart, M. S. (2003). The accountability deficit in Latin America. In S. Mainwaring & C. Welna (Eds.), *Democratic Accountability in Latin America* (pp. 79-131). Oxford, UK: Oxford University Press.

Morgenstern, S., & Manzetti, L. (2003). Legislative oversight: Interests and institutions in the United States and Argentina. In S. Mainwaring & C. Welna (Eds.), *Democratic accountability in Latin America* (pp. 132-169). Oxford, UK: Oxford University Press.

Novaes, L. (2014). *Promiscuous politicians and the problem of party building: Local politicians as party brokers* (Working paper). Washington, DC.

Nunes, F. (2013). Os Determinantes dos Resultados de Soma Positiva em Minas Gerais e no Rio Grande do Sul. *Revista de Sociologia e Política*, *21*(47), 91-112.

O'Donnell, G. (1994). Delegative democracy. *Journal of Democracy*, *5*, 55-69.

O'Donnell, G. (1998). Horizontal accountability in new democracies. *Journal of Democracy*, *9*, 112-126.

Oliveira, F. L. (2008). Justice, professionalism, and politics in the exercise of judicial review by Brazil's supreme court. *Brazilian Political Science Review (Online)*, *3*, 93-116.

Oliveira, F. L. (2012). Supremo Relator: Processo Decisório e Mudanças na Composição do STF nos Governos FHC e Lula. *Revista Brasileira de Ciências Sociais*, *27*(80), 89-115.

Paiva, N., & Sakai, J. (2014). *Quem São os Conselheiros dos Tribunais de Contas* (Research report). São Paulo: Transparência Brasil.

Pereira, C., & Melo, M. A. (2016). Reelecting corrupt incumbents in exchange for public goods: Rouba Mass Faz in Brazil. *Latin American Research Review*, *51*(1), 88-115.

Pinello, D. R. (1999). Linking party to judicial ideology in American courts: A meta-analysis. *The Justice System Journal*, *20*, 219-254.

Posner, R. A. (2010). *How judges think*. Cambridge, MA: Harvard University Press.

Santiso, C. (2009). *The political economy of government auditing: Financial governance and the rule of law in Latin America and beyond*. New York, NY: Routledge.

Santos, F. G. M. (2001). *O Poder Legislativo Nos Estados: Diversidade e Convergência* [Legislative power in the states: Diversity and convergence]. Editora FGV. Retrieved from https://books.google.com/books?id=y1OEjHFzw3sC

Schedler, A. (1999). Conceptualizing accountability. In A. Schedler, L. Diamond, & M. F. Plattner (Eds.), *The self-restraining state* (pp. 13-28). Boulder, CO: Lynne Rienner.

Schelker, M., & Eichenberger, R. (2010). Auditors and fiscal policy: Empirical evidence on a little big institution. *Journal of Comparative Economics*, *38*, 357-380.

Speck, B. (2011). Auditing institutions. In T. Power & M. Taylor (Eds.), *Corruption and democracy in Brazil: The struggle for accountability* (pp. 127-161). Notre Dame, IN: University of Notre Dame Press.

Weitz-Shapiro, R., Hinthorn, M., & Moraes, C. (2015). *Overseeing oversight: The logic of appointments to Brazilian state audit courts* (Working paper). Retrieved from http://papers.ssrn.com/sol3/papers.cfm?abstract_id=2623788

Westfall, P., & Young, S. S. (1993). *Resampling-based multiple testing: Examples and methods for p-value adjustment* (Vol. 279). New York, NY: John Wiley.

Wood, B. D., & Waterman, R. W. (1991). The dynamics of political control of the bureaucracy. *The American Political Science Review*, *85*, 801-828.

Author Biographies

F. Daniel Hidalgo is an assistant professor of political science at the Massachusetts Institute of Technology (MIT). He works on political accountability and elections in the developing world, with a particular focus on Latin America.

Júlio Canello is a PhD candidate in political science at the Instituto de Estudos Sociais e Políticos, State University of Rio de Janeiro (IESP-Uerj), Brazil. His research interests include comparative judicial politics, horizontal accountability institutions, legislative studies, and coalitional presidentialism, particularly in Latin America. His work has appeared in journals like *Brazilian Political Science Review* and *Dados*.

Renato Lima-de-Oliveira is a PhD candidate in political science at MIT. He holds a BA in journalism (Universidade Federal de Pernambuco, Brazil) and an MA in Latin American studies (University of Illinois at Urbana–Champaign), and has previously worked as a reporter in Brazil. His research interests include the topics of development, natural resources management, and accountability.

Original Article

Comparative Political Studies
1–35

DOI: 10.1177/0010414015621072
cps.sagepub.com
SAGE

Banners, Barricades, and Bombs: The Tactical Choices of Social Movements and Public Opinion

Connor Huff[1] and Dominika Kruszewska[1]

Abstract

In this article, we use an experimental survey design to explore how the tactical choices of social movements affect public opinion about whether the government should negotiate with the movement and the bargains that should be struck once negotiations begin. In doing so, we test competing theories about how we should expect the use of tactics with varying degrees of extremeness—including demonstrations, occupations, and bombings—to influence public opinion. We find that respondents are less likely to think the government should negotiate with organizations that use the tactic of bombing when compared with demonstrations or occupations. However, depending on the outcome variable and baseline category used in the analysis, we find mixed support for whether respondents think organizations that use bombings should receive less once negotiations begin. The results of this article are generally consistent with the theoretical and policy-based arguments centering around how governments should not negotiate with organizations that engage in violent activity commonly associated with terrorist organizations.

Keywords

social movements, experimental research, terrorism, protest

[1]Harvard University, Cambridge, MA, USA

Corresponding Author:
Dominika Kruszewska, Center for European Studies, Harvard University, 27 Kirkland Street, Cambridge, MA 02138, USA.
Email: dkruszewska@fas.harvard.edu

Introduction

Some of the most prominent examples of contentious political activity in recent years—over five hundred thousand Catalan demonstrators marching through Barcelona dressed in their region's red and yellow, pro-Russian protesters occupying government buildings in Donetsk, Ukraine, and renewed bombing efforts by the New Irish Republican Army (IRA) in Northern Ireland—capture public attention not only because of their political goals but also because of the tactics the movements choose to employ in pursuit of these goals.[1] Indeed, the combination of the goals of a movement with the tactics it employed has played an important role in determining whether some of the most iconic social movements in history succeeded or failed. For example, innovations in repertoires of contention such as civil rights sit-ins and revolutionary barricades in France have been argued to have shaped the success of the movements that chose to employ them.[2] Moreover, social movements are cognizant of the impact their tactical choices have on both potential government action and public opinion. For instance, in 1980, in Poland during Solidarity's protest in the Lenin Shipyard, the Strike Committee imposed a strict ban on alcohol to ensure discipline and peacefulness of protest.[3] Similarly, Spanish indignados occupying Puerta del Sol in Madrid patrolled the tent city with the goal of preventing violence and vandalism. These actions were taken with the understanding that protest forms incompatible with the movement's nonviolent strategy could negatively affect the public image of the movement and jeopardize future success in achieving its preferred political outcome.

There are two main channels through which public opinion about the tactical choices of social movements can affect whether a movement succeeds or fails. First, public perception of a movement determines how much support it is able to attract. For example, the use of tactics that potential supporters deem unwarranted or unethical can have a negative effect on the ability of an organization to garner support.[4] Second, public opinion has an effect on the strategic behavior of governments in response to protest (Giugni, 1998).[5] Governments facing a social movement are forced to make a number of decisions including whether to repress the movement, whether to negotiate, and how much to concede if they do bargain. While we recognize that public opinion does not always directly translate into a specific government action, in democratic settings governments must be cognizant of how these decisions will be received by the public and the impact they will have on their future electoral prospects (Giugni, 1998).[6]

In this article, we use an experimental survey design to explore how the tactical choices of social movements affect public opinion about whether the

government should negotiate with the movement as well as what bargain should be struck once negotiations begin. As both the decision to negotiate and the outcome of bargaining are publicly observable, the decisions that governments make at each of these stages can have long-term electoral ramifications. This makes democratic governments uniquely susceptible to the influence of public opinion at each of these decision-making stages. Using an experimental design also provides the benefit of allowing us to directly measure the effect of different tactical choices on public opinion while holding constant the goals of the movement and the government they are facing. By doing so, we attempt to address one of the primary concerns about empirical tests in the social movements literature: The effect of the choice of tactics on public opinion is confounded by attributes specific to a particular organization such as the identity of the movement, sympathy for its goals, or the movement's institutional alliances. We also ask respondents to explain their answers. We then analyze these open-ended responses using a Structural Topic Model (STM)—a cutting edge text-analysis tool—to provide descriptive statistics exploring potential mechanisms through which the tactics chosen by protest movements affect public opinion. Doing so allows us to explore the extent to which the theoretical logic driving the hypotheses presented in the following sections permeates the thinking of respondents about the tactical choices of social movements and government action.

In designing the experiment and selecting our location, we take care to avoid situations in which respondents rely on easily accessible frames about prominent tactics or issue areas which might lead respondents to attribute the tactic they are assigned through treatment to a familiar organization. If this occurs systematically, we are no longer able to isolate a tactic-specific effect but instead are measuring a combination of a tactic with respondent frames about the goals of organizations. Given this concern, we vary the movement presented in the vignette between a movement pursuing independence and a movement whose goals are not specified. In addition, we administer our survey experiment in Poland. Doing so allows us to conduct our experiment in a democratic political system in which the government is accountable to the electorate and responsive to public opinion but where respondents are unlikely to rely on easily accessible frames about either the tactic or secession. This is in contrast to places such as the United States, with the recent Occupy movement, and the United Kingdom, with the long history of bombing by the Provisional IRA and currently the New IRA, where respondents likely have strong opinions about what the government should do about organizations adopting the tactics of occupation and bombing, respectively. While both the issue area and the tactical choices are realistic within Polish and European politics, Poland provides a low salience context which mitigates

concerns that respondents might systematically associate particular tactics or goals with a specific organization.

The findings of this article are threefold. First, we find that respondents are less likely to think that the government should negotiate with organizations using the tactic of bombing when compared with organizations using occupations or demonstrations. The findings hold at the $\alpha = .05$ level across three out of four specifications in which we vary the generality of the vignette and the way in which we measure the dependent variable.[7] Second, we find that there is not a significant difference in support for government negotiations with movements employing the tactics of an occupation and demonstration. Similarly, we find that there is not a significant difference in how much respondents think the government should be willing to concede to organizations that use the tactics of an occupation or demonstration once negotiations begin. Finally, we find mixed support for whether respondents think organizations that use bombing as a tactic should receive less once negotiations begin depending on the outcome variable and baseline category used in the analysis.

In general, the structure of this article and the ways in which we analyzed the experimental results are consistent with a previous version of the manuscript that went through the peer-review process results-blind.[8] This means that the article went through the process of being revised and resubmitted, and conditionally accepted, prior to fielding the experiment. Throughout this article, we explicitly detail where the analysis of results is either consistent with or departs from the pre-analysis plan. This article is part of a broader movement in social sciences seeking greater transparency in research.[9]

The remainder of the article is structured as follows. First, we present a theoretical framework that allows us to directly test competing arguments about how we should expect the tactical choices of movements engaging in contentious politics to affect public opinion. We use this framework to derive empirically testable hypotheses. Second, we present an experimental design that we used to identify how the tactics chosen by social movements affect public opinion. We focus on two critical junctures in the strategic decision making of governments—whether to negotiate and how much to concede in bargaining—during which public opinion is likely to be particularly salient to the leaders of democratic governments. Third, we discuss the theoretical and practical considerations that led to our selection of Poland as the location for our experiment as well as how we can expect the results presented in this article to generalize to other locations and issue areas. Fourth, we present our results. In general, the analyses presented in this section are consistent with our pre-analysis plan with the exception of the STM. In the final section, we conclude.

Tactical Choice, Public Opinion, and Government Bargaining

Prior research has identified a number of factors influencing the strategic decision of social movements to select a particular tactic. A range of work has demonstrated how the character of the political and social structure in which movements operate, and availability of resources, plays an important role in determining the tactical repertoires of social movements (Gamson, 1975; Kitschelt, 1986; McAdam, Tarrow, & Tilly, 2001). In contrast, others emphasize cultural over material factors. These authors highlight the importance of congruence between forms of action and the identity of the organization (Taylor & Van Dyke, 2004), as well as the ability of movement leaders to draw on symbols and discursive frames from the culture of the community in which they are embedded (Swidler, 1986). Finally, another strand of scholarship emphasizes the role of internal organizational structures (Polletta, 2012) and the influence of the process of professionalization and organization-building on the choice of tactics (Piven & Cloward, 1979; Staggenborg, 1988).

While the internal and external factors influencing the movement's tactical decision have been studied extensively, the effects of these choices on public opinion remain underexplored. In this article, we shift the focus away from the motivations behind the adoption of particular strategies to the impact that they have on whether the public supports the decision to negotiate with the movement and the extent of concessions the government should make once negotiations begin. Building directly on three types of theories explaining how the strategic choices of nonstate actors should affect public opinion during conflict processes, we derive a number of hypotheses which are tested directly in our experiment. The presentation of the three types of theories and derivation of hypotheses is unaltered from the prior version of the manuscript that went through the peer-review process results-blind.

The Benefits of Extremism

The first theory, explored primarily through the study of strikes and urban riots, posits that disruptive and violent tactics increase public support for a movement. This occurs because the use of violent tactics facilitates recognition of a movement as a legitimate claimant to an issue and increases its perceived ability to obtain concessions from the government (Astin, Astin, Bayer, & Bisconti, 1975; Bueno de Mesquita & Dickson, 2007; Gamson, 1975; Giugni, 1998; McAdam, 1983; Pape, 2003; Shorter & Tilly, 1974; Tarrow & Tollefson, 1994). In doing so, the movement is able to garner more

support and increases its probability of success. Similarly, research on civil wars argues that rebel groups that execute more terrorist attacks impose extreme costs on the state and undermine its ability to win the conflict. These costs have been shown to increase the likelihood that the rebel group will be invited to participate in negotiations and will obtain concessions in the bargaining process (Thomas, 2014). A parallel argument holds that the adoption of violent tactics acts as a costly signal of commitment to the political cause which improves the organization's ability to attract recruits and new sources of funding (Bloom, 2004; Bueno de Mesquita & Dickson, 2007; Pape, 2003). This reasoning has been used to explain spikes in public support following the incidence of both suicide bombings and terrorist attacks more broadly. While these arguments are drawn from studies of conflict processes across a wide range of organizations and movements, the logic driving them is clear: The public has a preference for more extreme tactics. We derive our first two hypotheses directly from this core intuition.

Hypothesis 1A: The more extreme the tactic, the more respondents believe that the government should negotiate with the movement.
Hypothesis 1B: The more extreme the tactic, the more respondents believe that the government should concede to the movement during negotiations.

The Benefits of Moderation

In contrast, a second theory about the relationship between the tactics of nonstate actors and public opinion asserts that moderation and nonviolence increase public support for the movement. There are a number of reasons this might occur. Shaykhutdinov (2010) asserts that nonviolence provides a moral advantage for the organization and allows the group to garner support without provoking the animosity or distrust fueled by violent conflict. This means that adopting more extreme tactics causes individuals that might otherwise be willing to support the movement to now oppose it. A related argument asserts that nonviolence can be both a moral and pragmatic choice (Sharp, 1973). That is, movements choose nonviolence because they realize that violence can have negative consequences including diverting attention from grievances, polarizing public opinion, and providing justification for governmental repression. The benefits of nonviolence have also been demonstrated by Stephan and Chenoweth (2008) in an analysis of aggregate data on major nonviolent and violent conflicts between nonstate and state actors. They show that nonviolent campaigns tend to be more successful and posit two explanations for this finding.[10] First, nonviolent methods strengthen domestic

and international legitimacy. Second, the public perceives violent groups as extremists "beyond accommodation." Thus, moderate strategies work in the group's favor, "enhancing their appeal and facilitating the extraction of concessions through bargaining" (Stephan & Chenoweth, 2008, p. 9). Another argument for why more moderate tactics might increase public support for a movement is presented by Abrahms (2012) who shows that the demands of terrorist organizations are not effectively communicated through their violent strategies. Instead, respondents infer radical political ends from extreme tactics which closes the bargaining space even where one could exist (Abrahms, 2012). In addition, the use of extreme tactics decreases the credibility of the terrorist organization's promise to demobilize if concessions are granted (Abrahms, 2013), which would likely decrease public support for negotiations. Based on these theories, we derive a set of alternative hypotheses in direct contrast to Hypotheses 1A and 1B.

Hypothesis 2A: The more extreme the tactic, the less respondents believe that the government should negotiate with the movement.
Hypothesis 2B: The more extreme the tactic, the less respondents believe that the government should concede to the movement during negotiations.

The No Concessions Policy

The third theory about how the tactics chosen by nonstate actors affect public opinion builds on the idea that governments should adopt a strict policy of "no concessions" when negotiating with terrorist organizations. This theory is driven by the claim that giving in to the demands of groups that adopt the most extreme tactics proves this type of violence to be effective and encourages its use in the future (Chellaney, 2006; Clutterbuck, 1992; Netanyahu, 1986). In this article, we explore whether this idea, pervasive throughout the academic and policy worlds, is also supported by public opinion about social movements that adopt extreme tactics often associated with terrorist organizations.[11]

Hypothesis 3A: When organizations adopt bombing as a tactic, respondents are less likely to believe that the government should negotiate with the movement relative to other possible tactics.
Hypothesis 3B: When organizations adopt bombing as a tactic, respondents are less likely to believe that the government should make concessions to the movement during negotiation relative to other possible tactics.

Experimental Design

In this section, we present an experimental design intended to test the hypotheses presented in the previous section. In the experiment, respondents are asked to read a vignette describing a social movement that is either pursuing regional independence or does not have a specified goal.[12] We manipulate two factors separately, giving us a 2 × 3 fully factorial design. In the vignette, we randomize (a) whether the goals of the organization (regional independence) are specified and (b) the tactics employed by the organization. Strategies appearing in the vignette are a demonstration, an occupation, and a bombing. The experimental design presented in the remainder of this section is the design proposed in the version of the manuscript that went through the peer-review process results-blind.[13]

We chose social movements pursuing regional independence for the following reasons. First, independence movements are active in democratic settings. Unlike revolutionary movements whose goals are often to overthrow a nondemocratic regime, the success of independence movements often hinges on their ability to pressure the government by attracting public support. Indeed, the importance of public support is embodied in the referendum processes—such as the vote following the negotiation of the Good Friday Agreement in Northern Ireland—that are often used to approve the bargains struck in negotiations between the movement and government. Second, independence movements use a wide range of tactics. This means that our treatment categories, from demonstrating in the streets to bombings, could all plausibly be employed in pursuit of the objectives of the organization. This is important as it allows us to hold constant the issue area for which the movement is fighting while ensuring that the scenario specified in the vignette is plausible. Third, unlike in cases of civil rights or abortion policy, independence is an issue for which individuals are unlikely to hold strong prior believes about the policy that should be implemented. This creates greater space for treatment effects to emerge as the bargains that can be struck, such as full independence or regional autonomy, are not as strongly shaped by the stances respondents already have on the issue. This occurs because the outcomes of independence movements, while topical and salient, are neither a partisan issue nor considered a basic human right in contemporary society. To summarize, in selecting the type of movement to use in our experiment, we searched for a movement that (a) operates in democratic settings, (b) could plausibly employ a variety of tactics in pursuit of its objectives, and (c) advocates for an issue that is salient but about which respondents are unlikely to hold strong beliefs prior to observing the organization's tactical choice.

While our main interest throughout this article is in tactical variation, randomizing between a vignette that specifies an organization pursuing independence and a more general vignette allows us to address two important potential concerns with our design. First, the inclusion of a condition with an organization with a specific goal is intended to assuage the concern that a more generalized vignette invites personal associations beyond researchers' control. In other words, when presented with a generalized vignette, respondents might rely on easily accessible frames and attribute the tactic they are assigned through treatment to a familiar organization that is fighting within a particular issue area. If this occurs systematically, we are no longer able to isolate a tactic-specific effect but instead are measuring a combination of a tactic with respondent frames about the goals of organizations. Indeed, our selection of Poland as the country in which to conduct the experiment is driven in part by our efforts to address this potential problem. The second concern is that the decision to focus on separatist movements limits the external validity of our experiment. While we still face the limitations on external validity common to most experimental studies,[14] presenting a generalized vignette allows us to make inferences more broadly than for only organizations pursuing independence. By randomizing between a general vignette and one that specifies an organization pursuing independence, we are able to address both these concerns in a way that directly tests the robustness of our findings.

In our randomization scheme, we allocate three quarters (approximately 1,500 respondents) to the vignette discussing a separatist organization and the other quarter (approximately 500 respondents) to the general vignette. We do this for two reasons. First, in the general vignette, we are only able to ask the question about whether the government should negotiate with the movement. As defining the goals of the organization is essential in asking about whether the government should make more or less concessions during negotiations, we are unable to ask these types of questions without providing additional information about the goals of the movement. Thus, respondents in the general movement category are only asked whether or not they think the government should be willing to negotiate with the organization. The second reason we divide the sample in this way is to increase power in the treatment category with the movement pursuing independence creating a larger space to allow treatment effects to emerge.

We chose the tactics of demonstration, occupation, and bombing for the following reasons. First, each of these tactics has been used in recent years by movements actively pursuing independence. Indeed, in 2014, there were marches and rallies for independence in Scotland and Spain, occupations of government buildings by pro-Russian separatists in Donetsk, Ukraine, and

attempted bombings by the New IRA in Northern Ireland. Second, each of these tactics involves a premeditated strategic choice by a movement rather than an action that arose spontaneously as a situation escalated. For this reason, we chose not to explore a range of violent activities, such as throwing Molotov cocktails, damaging property, and rioting. This allows us to directly explore our primary interest in how the tactical choices made by social movements affect public opinion. Third, demonstrations are used as a baseline tactic. Demonstrations are one of the most common forms of contentious political activity in democracies and, if registered with the authorities, they are legal. This provides a useful baseline that allows us to directly compare a common choice of movements against the more extreme tactics of occupation and bombing. Finally, the three tactics form a clear scale of extremeness, allowing us to adjudicate between hypotheses about the benefits of extremeness, benefits of moderation, and no concessions.[15]

To prevent idiosyncratic features of the vignette from driving the results, we also randomly vary the contextual variable of whether the movement's activity takes place in a foreign country.[16] This is intended to address concerns that respondents' perceptions about whether the experiment is occurring in Poland might vary with treatment status. For example, respondents might be more likely to believe that the experiment is occurring in Poland if they receive the treatment with the tactic of demonstration than the treatment of bombing. This becomes a problem if differential association about whether the movement is operating in Poland also sparks feelings of domestic pride or nationalism that also affect the willingness of respondents to think the government should negotiate or make concessions. Thus, we randomly inform half the participants that the movement is in a foreign country and for the other half do not specify the location.

Another possible concern about the vignette is that we do not directly specify how the tactic that is chosen relates to prior interaction between the government and social movement. For example, a movement might have tried a range of institutionalized and legal tactics and only resorted to bombing after failing to achieve its outcome through alternative means. We view an exploration of the influence of context and timing of the protest activity on public opinion as an exciting opportunity for future research. The vignette and experimental design presented here are intended to provide a foundation for exploring the links between tactical choices of social movements and public opinion and thus provide the most concise and direct test of the specified hypotheses possible.

To summarize, the vignette provides details about a movement pursuing independence and a generalized group with no details about their goals. Fixing the goals of the movement allows us to focus on groups operating in a

democratic setting that could plausibly employ a variety of tactics in pursuit of their objectives. Moreover, the use of the goal of independence allows us to focus on a salient issue for which public opinion clearly matters but where respondents likely do not hold a strong prior belief about what the government should do. Finally, we randomize between three tactical choices that have each been employed in recent years including demonstrations, occupations, and bombings. The vignette is as follows:

> *Please consider the following hypothetical scenario. Some parts of the description may strike you as important; other parts may seem unimportant. Please read the details very carefully. After describing the situation, we will ask your opinion.*
>
> *[A separatist movement//A social movement] [demonstrated in front of a government building//occupied a government building//bombed a government building] in the capital city [of a foreign country] today. [The stated goal of the movement is regional independence including the establishment of a new democratic government that would be responsible for overseeing the region's security and economic affairs. The region has its own language and culture.]*

After reading the vignette, respondents are asked several questions designed to test the hypotheses presented in this article. In the questions, we make a distinction between the two major decisions governments are forced to make when faced with a challenging social movement: whether to negotiate and what concessions should be made once negotiations begin. This distinction mirrors the two stages of the bargaining process that governments and social movements could plausibly engage in. The first question, which focuses on whether respondents think the government should negotiate, is a direct test of Hypotheses 1A, 2A, and 3A. These hypotheses present competing claims about how the tactical choices of social movements affect whether respondents should be more or less likely to think the government should negotiate with the movement. Respondents are asked the following:

> *Do you think that the government should negotiate with the movement?*
>
> - *Yes*
> - *No*

Following the first question, we utilize a conditional branching format to measure how strongly respondents feel about whether the government should negotiate with the movement. This provides us with a more fine-grained measure of how the tactical choices of the movement affect the strength of opinion about whether respondents think the government should negotiate with

the movement. We use the results of this question to create a 4-point scale varying from respondents feeling very strongly that the government should negotiate to feeling very strongly that the government should not negotiate with the movement. The follow-up question asks as follows:

> *You stated that you think the government [should//should not][17] negotiate with the movement. Do you feel very strongly about this, or not very strongly?*
>
> - *Not very strongly*
> - *Very strongly*

After answering these two questions, respondents are asked an open-ended follow-up in which they are requested to explain their response. This question is intended to provide the opportunity for an exploration of possible mechanisms through which treatment of a particular tactic affects variation in the outcomes. The open-ended question is as follows:

> *You stated that the government [should/should not] negotiate with the movement. Could you please type a few sentences telling us why you think the government [should/should not] negotiate with the movement?[18]*

The fourth and fifth questions directly test Hypothesis 1B, 2B, and 3B. In both of them, we explore how variation in the extremeness of the tactic employed by the social movement affects public opinion about the extent of concessions the government should make to the movement once negotiations have begun. The fourth question asks what the government should be willing to concede to the movement during negotiations with the options being either independence for the region, regional autonomy, or no concessions. For each of the options, we provide a brief description alongside the answer detailing what each outcome entails. We chose regional autonomy and independence for the region because these are some of the most common outcomes of separatist movements. Examples of regional autonomy following negotiations have occurred in Northern Ireland and Bangsamoro in the southern Philippines.[19] Examples of independence following separatist movements have occurred in Estonia, Latvia, and Lithuania, each voting for the establishment of a democratic republic independent of the Soviet Union. Finally, we allow for the option of no regional autonomy because even if this is not a feasible bargain for the government to strike within negotiations, it is possible that respondents still support solutions outside of the bargaining range. Incorporating the option of no regional autonomy gives participants the option of disagreeing with the government, and thus either pressuring the government not to concede or shifting the bargains that the government is

able to accept. To assess how variation in tactical extremeness affects public opinion about the concessions, the government should make during negotiations we ask respondents the following question:

> *The government and the separatist movement have entered into negotiations. Which of the following do you think the government should be willing to concede to the movement?*
>
> - *No regional autonomy. The central government will maintain full control over the region.*
> - *Regional autonomy. The regional government will control its own economic affairs but the region will remain part of the country.*
> - *Independence for the region. The region will establish a new democratic government that will be responsible for overseeing the region's security and economic affairs.*

While the above question is motivated by some of the most frequent resolutions to separatist movements it is a relatively coarse measure. The structure of the dependent variable might not allow enough space for treatment effects to emerge as respondents are forced into one of the three categories. This means that we might be unable to detect more subtle but nonetheless important differences between the tactical choices of a movement and their effect on public opinion about what the government should be willing to concede during negotiations. For example, it might be that respondents think that the government should be willing to concede more or less during bargaining but these concessions should all be made within the framework of a regional autonomy agreement. To address this possibility, we present respondents with an additional vignette and then ask a question that utilizes a continuous dependent variable.

In particular, the vignette provides information that the government and movement have settled on regional autonomy for the area under dispute. However, as part of the settlement, the two sides must still determine the extent of regional fiscal autonomy. We focus on fiscal autonomy for the continuous outcome as it is both a contentious and important component of negotiations over the devolution of power in Europe. The extent of fiscal autonomy a region should be granted has featured in debates in a diverse range of regions including Catalonia, Northern Italy, Bavaria, and Flanders. Financial themes also feature prominently in the politics of regional autonomy movements. In Belgium, the Flemish movement perceives the financial transfers between regions as draining its resources. Similarly, in Spain, Catalonia wants to reduce its contribution to a national system that redistributes portions of tax revenue to poorer regions of Spain.

The vignette used in our experiment builds directly on the negotiations between the Spanish government and regional representatives in Spain and focuses on the proportion of shared taxes that should be granted to the regions.[20] In particular, in the vignette, we present a simplified version of the negotiations that occurred in Spain in which the movement proposes that 100% of regional income taxes stay in the region while the government proposes that 50% should stay in the region and 50% should go to the central government.[21] The government's proposal of 50% mirrors the actual bargain struck by the Spanish government with the autonomous regions in Spain, while the movement's preferred outcome is full fiscal autonomy. This bargaining range has important advantages in that it is continuous and respondents are unlikely to hold strong prior beliefs about what types of fiscal autonomy arrangements should be struck during negotiations. As in the previous question, while the government proposes 50%, we allow respondents to enter any response from 0% to 100% to allow respondents to disagree with the government's decision to negotiate with the movement. The structure of the dependent variable allows us to construct a finer measure of how the tactical choices of movements affect the concessions respondents think the government should be willing to make during bargaining with the movement using a dependent variable that is both plausible and salient throughout Europe. Respondents are presented with the following additional information:

> *The government and the separatist movement have settled on regional autonomy for the area under dispute. As part of the settlement they must determine the extent of regional fiscal autonomy. The movement is proposing that 100% of regional income taxes stay in the region while the government is proposing that 50% should stay in the region and 50% should go to the central government.*

After being presented with this additional information, respondents are asked the following question:

> *What percentage of regional income tax revenues do you think should stay in the region?*
>
> [Sliding scale from 0% to 100%].

Case Selection

We used the following criteria in determining that our experiment should be conducted in Poland. The ideal country is one in which respondents do not systematically associate a particular tactic with a specific movement which

has already obtained concessions from the government. As we seek to avoid situations in which respondents rely on easily accessible frames and attribute the tactic they are assigned through treatment to a familiar organization, it is important that in the selected country, there are no prominent movements currently using that tactic. This means that places such as the United States, with the recent Occupy movement, and the United Kingdom, with the long history of bombing by the Provisional IRA and currently the New IRA, would not be ideal as respondents would likely have strong opinions about what the government should do about movements adopting the tactics of occupation and bombing, respectively. In these cases, we are no longer able to isolate a tactic-specific effect but instead are measuring a combination of a tactic with respondent opinions about the particular movement or organization with which they associate the tactic. For the same reason, we avoid conducting our experiment in places with a long history of separatist activity. This means that places such as Spain, with the prominent Basque National Liberation Movement, would also not be an ideal location. This is not to say that the tactics chosen by organizations within these countries do not affect public opinion. Indeed, we contend that they almost certainty do. Rather, in selecting the country in which to conduct the experiment, we are seeking a location in which the emergence of movements employing a range of tactics is plausible but respondents are unlikely to systematically associate a particular tactic with a specific movement.

In contrast to the United States, United Kingdom, and Spain, respondents in Poland are unlikely to hold strong associations of tactics with a particular organization. The population has also not been polarized over the issue area through highly contentious referenda like it has in the United Kingdom over Scotland or in Spain over Catalonia and is unlikely to have preconceived notions of the extent of autonomy that should be granted to regions with distinct cultural and linguistic heritage. Similarly, unlike countries in Southern Europe, Poland has not experienced the recent wave of anti-austerity protests or occupy movements, which alleviates the concern that fresh awareness of a massive mobilization incorporating a sizable proportion of the population and a large number of organizations with wide-ranging goals would strongly shape the beliefs of respondents about what the government should do. Though mostly ethnically homogeneous, Poland does have two minor autonomy movements: a Silesian Autonomy Movement and Kaszebsko Jednota. Both of these movements are small[22] and minor in the political scene.[23] Given this, we argue that most Poles are unlikely to hold strong prior beliefs about governmental policies toward secessionist movements.

Despite the fact that respondents in Poland are unlikely to hold strong associations between a particular tactic and a specific organization, Poland

contains a population cognizant of protest activity. Contentious politics in Poland has been characterized by large variation in the tactics employed by protest groups as well as the types of political and economic goals pursued. For example, the massive Solidarity protest wave, which swept through the country in the 1980s mobilized millions of Poles, exposing them to a range of protest strategies (Ash, 1999). In the years following the transition to democracy, many Poles chose to articulate their concerns and express their opposition to policy decisions through disruptive actions in the streets (Ekiert & Kubik, 2001). Large protest waves engulfed the country again in the late 1990s during the implementation of reforms of the budget sector and leading up to the accession to the EU in 2004. More recent protests have surrounded legislative and social debates, including the role of religious and ideological symbols in public space, abortion, and environmental issues. Demonstrations take place frequently and trade unions have remained active as organizers of strikes and other contentious activities next to new social movement organizations and single-issue groups. Historical and contemporary forms of protest have ranged from petitions, demonstrations, strikes, occupations, road and rails blockades to destroying property, and clashes with the police.

To field the survey, we partnered with the Polish branch of Taylor Nelson Sofres (TNS) Global, a company, which merged with Ośrodek Badania Opinii Publicznej (OBOP), one of the oldest public opinion survey groups in Poland and in the region. The survey was administered to a nationally representative sample of approximately 2,000 Polish adults. Respondents were sampled using random-quota sampling.[24] The interviews were computer-assisted personal interviews, conducted in person by TNS interviewers with the use of mobile computers instead of pen-and-paper questionnaires.

The structure of the experiment in conjunction with the selection of Poland has important implications for external validity. Throughout the design of the project, we attempted to focus on an issue area that is important but has relatively low salience in Poland. This was done in an effort to prevent respondents from associating the tactic they read about in the vignette with a particular organization which could then influence their opinion about the bargaining process with the government. However, this decision also has important implications for the generalizability of our experiment. Because we designed our experiment to be in a location, where an issue is important but respondents do not have a strong prior about the movement type or its goals, we should be able to generalize to equivalent contexts for a host of different social movements. This means that our findings are likely to be most relevant when we observe either nascent organizations or issue areas on which the public does not have consolidated views. Given that a wide range of social movements emerging around the world are making similar tactical

choices, we hope to provide a framework from which researchers can continue to explore how the tactical choices of both violent and nonviolent movements affect public opinion about their organization and goals.

Results

In this section, we present and interpret the results of the experiment as specified in our pre-analysis plan. In doing so, we seek to explicitly detail where our analyses are either consistent with or depart from the pre-analysis plan. The results of the experiment are generally consistent with Hypothesis 3A. That is, for the question asking whether respondents think the government should negotiate with the movement, there is a negative and statistically significant difference between an occupation and a bombing and not a statistically significant difference between a demonstration and an occupation. We find tentative support for Hypothesis 3B—that is, whether respondents think organizations that use bombing as a tactic should receive less once negotiations begin—depending on the outcome variable and baseline category used in the analysis. The remainder of this section presents the findings for each of the questions in turn.

Figure 1 presents the results for the question asking whether the government should negotiate with the movement. Consistent with our pre-analysis plan, responses to the first question were analyzed by comparing each tactic against the less extreme category. That is, we compared the tactic of demonstrating against occupying, and occupying against bombing. This yielded two difference-in-proportions estimates where positive values indicate that the more extreme tactic caused an increase in support for negotiations with the movement. The results are consistent with Hypothesis 3A in which there is a statistically significant difference between an occupation and a bombing and not a statistically significant difference between a demonstration and an occupation. The use of the tactic of bombing by separatist organizations decreases support for the government entering negotiations by 6.5% from a baseline of 57.4%. This difference becomes even starker when comparing the results for the more general vignette which provides information about a social movement without specifying the goals of the organization. The use of the tactic of bombing by a social movement decreases support for the government entering negotiations by 28% from a baseline of 70.1%. In contrast, there is not a statistically significant difference in the responses for individuals reading about social movements using occupations as a tactic when compared with the baseline category of demonstrations.[25]

Figure 2 presents the results for the follow-up branching question asking respondents how strongly they feel about whether the government should

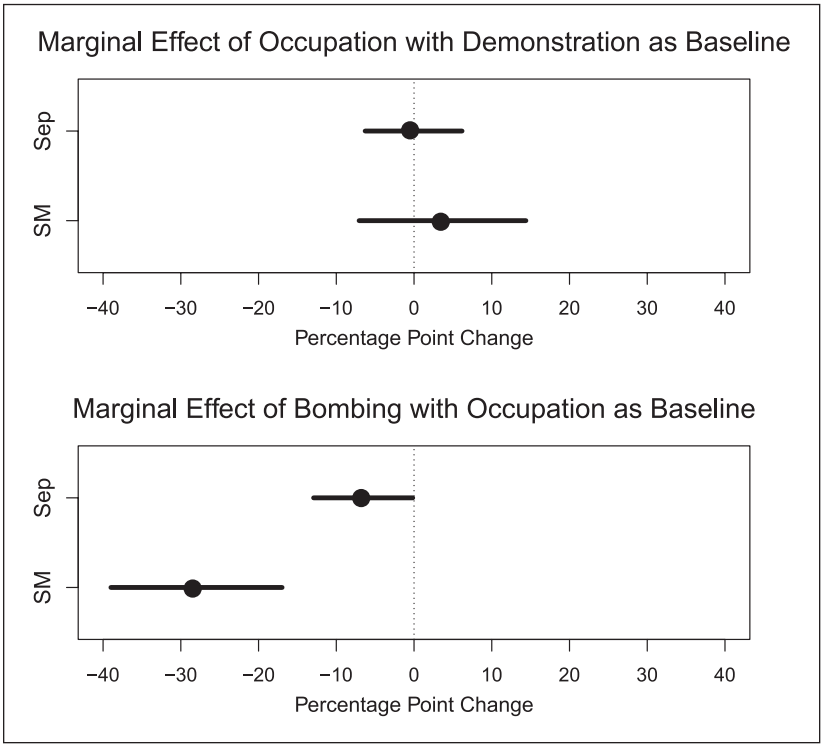

Figure 1. The percentage point change in support for negotiations by tactical choice.

The top panel shows the percentage point change in support for organizations that use occupations with demonstrations as the baseline category. The bottom panel shows the percentage point change in support for organizations that use bombings with occupations as the baseline category. Results comparing bombing with occupations are statistically significant at the $\alpha = .05$ level for both separatist organizations and SMs. SM = social movement.

negotiate with the movement. The results of this question were used to create a 4-point scale varying from respondents feeling very strongly that the government should negotiate to feeling very strongly that the government should not negotiate and subsequently analyzed by comparing each tactic against the less extreme category. As was specified in our pre-analysis plan, estimates were obtained using regression where positive values indicate that the more extreme tactic caused an increase in the strength of the belief that the government should negotiate with the movement. Results comparing bombing with occupations are statistically significant at $\alpha = .05$ for social movements.

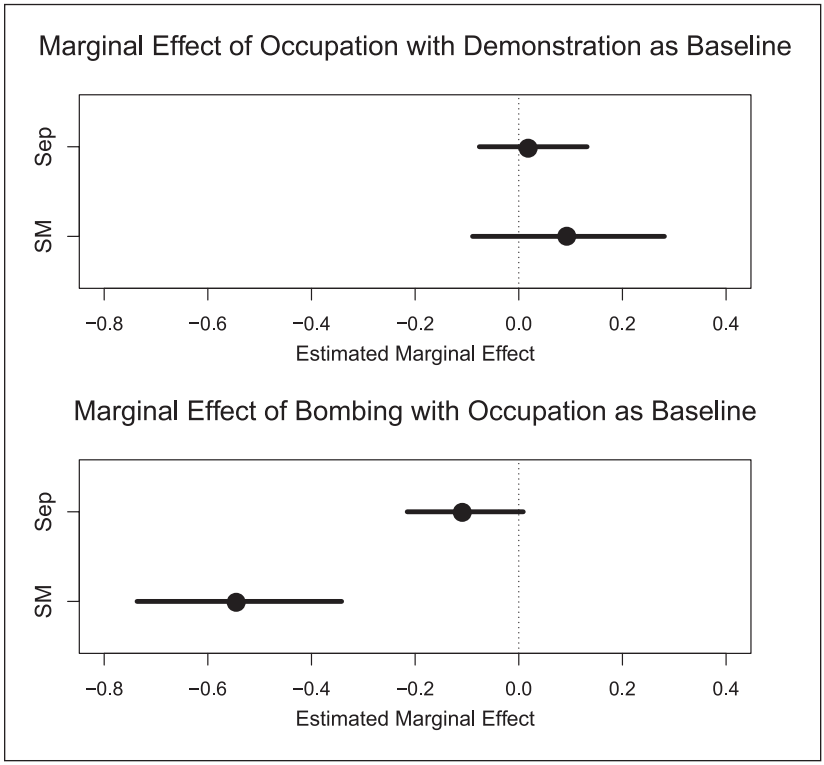

Figure 2. The marginal effect of tactical choice on support for negotiations with the movement, using a 4-point scale measuring how strongly respondents feel the government should/should not enter negotiations.
The top panel shows the marginal effect for organizations that use occupations with demonstrations as the baseline category while the bottom panel shows the marginal effect for organizations that use bombings with occupations as the baseline category. Results comparing bombing with occupations are statistically significant at $\alpha = .05$ for SMs. However, for separatist organizations the results are not statistically significant at $\alpha = .05$ (though they are at the $\alpha = .1$ level). SM = social movement.

However, for separatist organizations, the results are not statistically significant at $\alpha = .05$ though they are at the $\alpha = .1$ level.

To summarize, the results for the first two outcome questions are generally consistent with Hypothesis 3A in which we expect a statistically significant difference between organizations that use bombings when compared with occupations, paired with a null finding when comparing organizations using occupations relative to demonstrations. The findings hold at the $\alpha = .05$

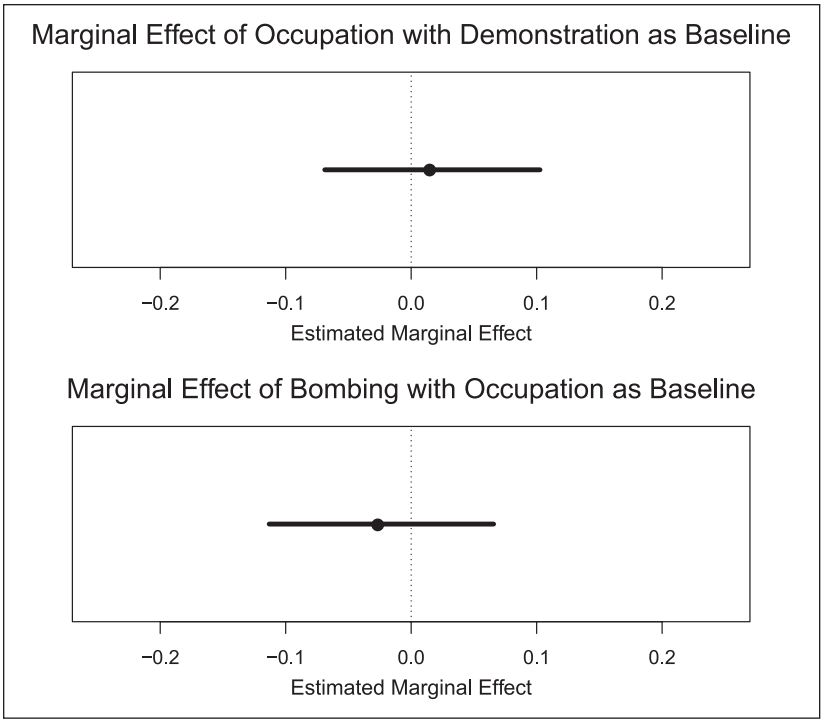

Figure 3. The marginal effect of tactical choice on support for concessions once negotiations have begun with the dependent variable being either full independence, regional autonomy, or no regional autonomy.
The top panel shows the marginal effect for organizations that use occupations with demonstrations as the baseline category while the bottom panel shows the marginal effect for organizations that use bombings with occupations as the baseline category. We fail to reject the null hypothesis at the α = .05 level.

across three out of four of the model specifications discussed in our pre-analysis plan.

Figure 3 presents the results for the question asking respondents what bargain the government should strike with the movement, with options ranging from no regional autonomy to regional autonomy and full independence. The results are statistically indistinguishable across all treatment categories. That is, there is not a statistically significant difference between what respondents think the government should be willing to concede in negotiations when comparing organizations using the tactics of bombings and occupations, and occupations and demonstrations.

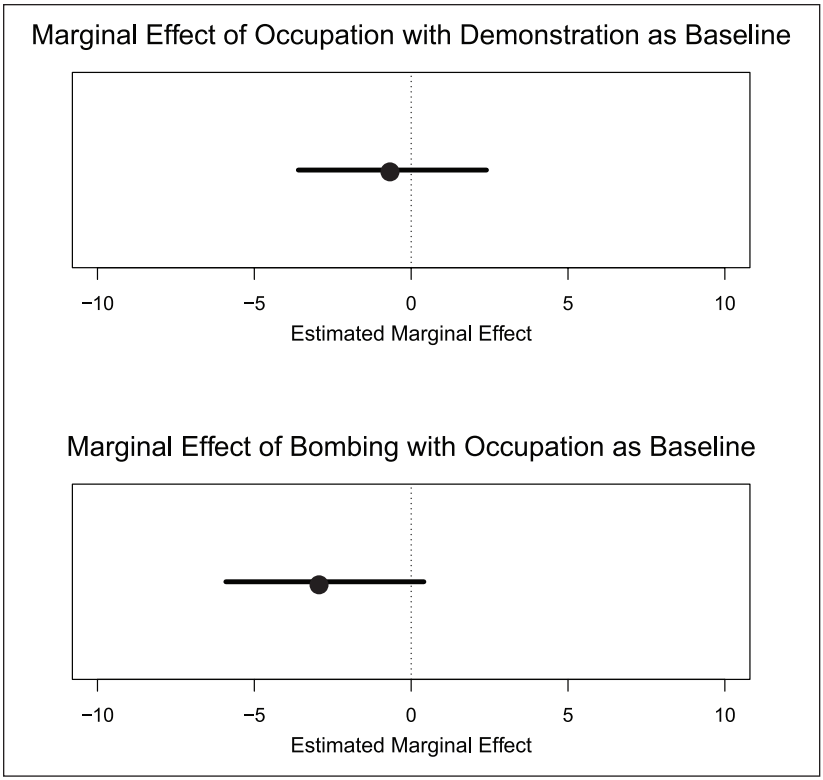

Figure 4. The marginal effect of tactical choice on support for tax concessions. The top panel shows the marginal effect for organizations that use occupations with demonstrations as the baseline category while the bottom panel shows the marginal effect for organizations that use bombings with occupations as the baseline category. We fail to reject the null hypothesis at the α = .05 level (though results are statistically significant at the α = .1 level).

However, as discussed in the "Experimental Design" section, this type of finding was possible due to the coarse nature of the dependent variable. That is, forcing respondents into one of these three categories might make it so that we are unable to detect more subtle but nonetheless important differences between the tactical choices of a movement and their effect on public opinion about acceptable bargains. We attempted to directly address the possibility that this would occur in the results-blind version of our manuscript. Figure 4 presents the results of a question in which respondents were asked about the extent of tax concessions the government should make to the region during negotiations over the terms of the region's fiscal autonomy. When the results

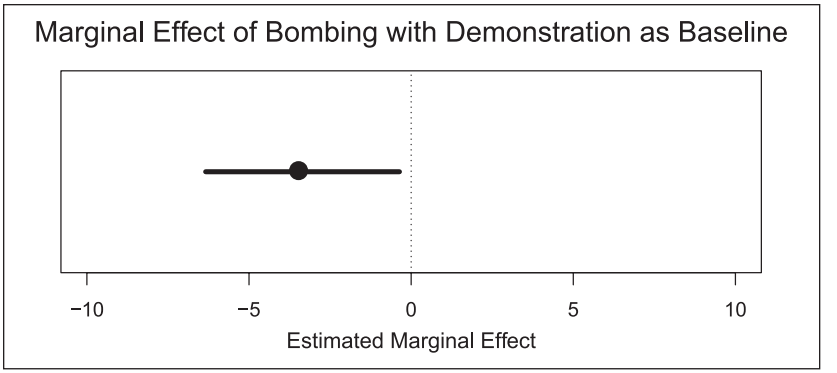

Figure 5. The marginal effect of the use of bombing with demonstration as the baseline category on support for tax concessions.
The results are now statistically significant at the $\alpha = .05$ level.

are analyzed in accordance with our pre-analysis plan—comparing bombing against an occupation and testing for statistical significance at the $\alpha = .05$ level—we fail to reject the null hypothesis that there is no difference between the tactics of occupation and bombing.

Interestingly, the results would have been statistically significant had we specified our pre-analysis plan in one of two slightly different ways. First, if we had instead specified that we would be conducting significance tests at the $\alpha = .1$ rather than the $\alpha = .05$ level, the results would be statistically significant. In particular, when comparing the tactic of occupation against bombing we obtain a p value of .088. Second, if we had set the baseline category to be a demonstration, rather than an occupation, the results would have been statistically significant at the $\alpha = .05$ level with a p value .028. Figure 5 presents the marginal effect of a bombing using demonstrations as the baseline category.

The results of the open-ended question, which asked respondents to explain why they thought the government should or should not negotiate, were analyzed using a STM.[26] A STM is an unsupervised topic model that builds directly on the Latent Dirichlet Allocation model.[27] In these types of unsupervised learning models, topics are inferred by the model rather than assumed before the analysis.[28] In STM, a document is represented as a mixture of topics where each word within a given document belongs exactly to one topic. Each document can then be represented as a vector of proportions that denote the fraction of words in that document that belong to each topic.

The STM innovates on previous unsupervised topic models by allowing for the inclusion of covariates of interest—such as whether respondents read about an organization using the tactic of a bombing—into the priors for the document-topic proportions and topic-word distributions. Each open-ended response is a mixture of topics where the researchers can incorporate relevant covariates where they might expect variation in the proportion of topics. Analysts can subsequently examine differences across included covariates. In the context of a survey experiment, the STM can be used to estimate the effect of a treatment embedded within the survey on text written by survey subjects.

Unlike in prior models, in the STM topic proportions (θ) can be correlated. The prevalence of the topics can be influenced by some set of covariates X through a standard regression model with covariates where $\theta \sim \text{LogisticNormal}(X_\gamma, \Sigma)$. In this article, these covariates include the treatment status to which respondents were assigned, either a demonstration, occupation, or bombing, and we use the STM to explore how responses change for each of the relevant treatment classes. For each word (w) in the open-ended response, a topic (z) is drawn from the response-specific distribution. Conditional on that topic, a word is chosen from a multinomial distribution over words parameterized by β. β is formed by deviations from the baseline word frequencies (m) in log space $(\beta_k \propto \exp(m + k_k))$.

To summarize, the STM allows each document to have its own prior distribution over topics defined by relevant covariates of interest to the researcher. This is in contrast to other mixed-membership models, such as Latent Dirichlet Allocation (LDA), which do not incorporate relevant covariates and instead share a global mean. The use of these covariates allows researchers to build relevant covariates directly into the model and make better inferences about relevant quantities of interest. It is important to note that while we stated that we would analyze the results of the open-ended questions using a STM in the pre-analysis plan, we did not specify the details of its implementation. This was necessary given the need to substantively interpret the output of the STM to determine the number of topics in the analysis as well as the labels that would be applied to each topic.

As discussed in the pre-analysis plan, the goal of the STM analysis is to allow us to explore the extent to which the theoretical logic driving the hypotheses permeates the thinking of respondents. Given the finding that respondents are less likely to think the government should negotiate with movements that choose bombing as a strategy when compared with demonstrations or occupations, we focus on exploring what theoretically

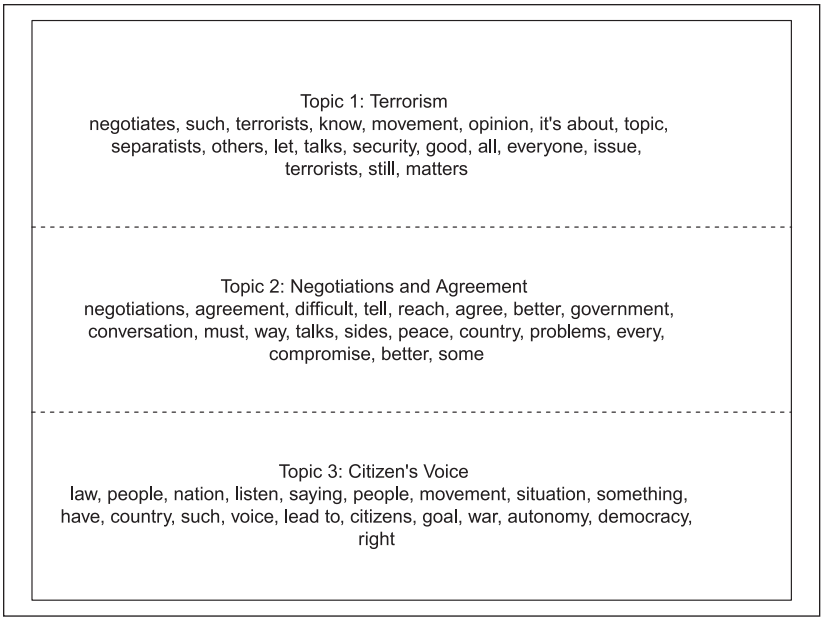

Figure 6. Words associated with each topic (English translation).
STM model with bombing as a covariate and respondents in all experimental conditions included in the data set. STM = Structural Topic Model.

separates bombing from other less extreme tactics. Thus, in estimating the STM we compare responses for individuals that read about movements that use the tactic of bombing against the responses of individuals that read about movements using tactics of demonstration and occupation pooled together as the baseline category.[29] Doing so allows us to explore why respondents that read about movements employing the tactic of bombing are less likely to think that the government should negotiate. Prior to analysis, we used standard text preprocessing conventions such as removal of punctuation, numbers, and stopwords.[30] No other covariates were included in the analysis.[31] The model was run with three topics.[32]

Figure 6 summarizes each of the three topics with the English translation of the top 20 most probable and exclusive words.[33,34] We inferred the following topic labels from these words: "Terrorism," "Negotiation and Agreement," and "Citizen's Voice."[35] The first topic is characterized by words such as terrorists, security, and separatists. An exemplar response states as follows: "Because one cannot negotiate with criminals," expressing the resistance to

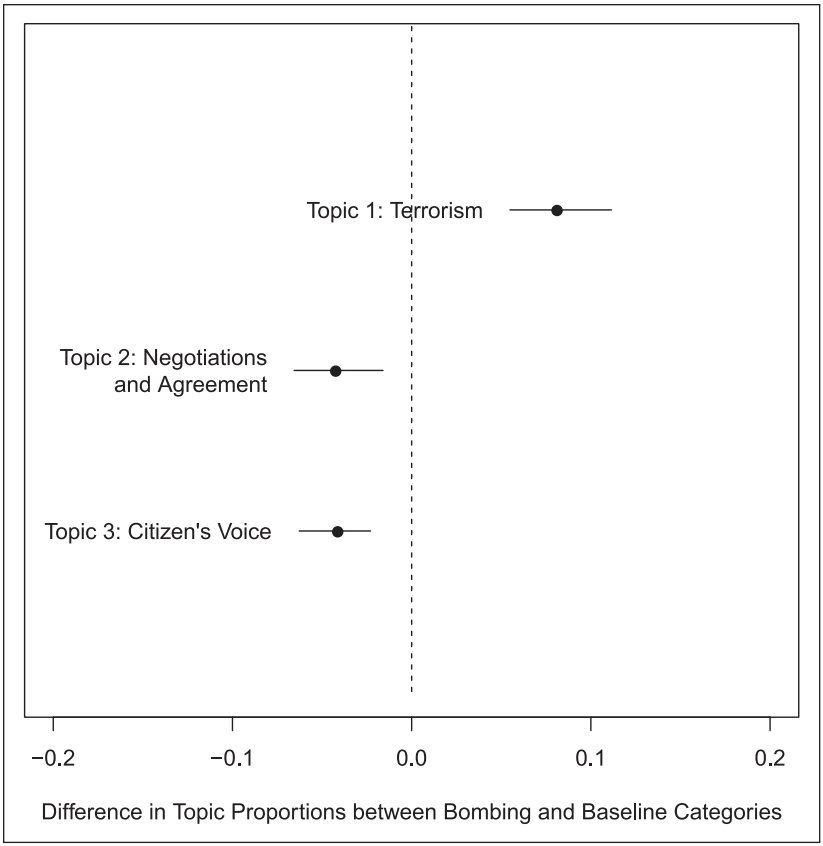

Figure 7. The effect of tactical choice on topic prevalence in open-ended responses with bombing as treatment and all other tactics as the baseline category. 95% confidence intervals.

entering dialogue with movements using violent means. The second topic conveys a more conciliatory attitude with top words such as agreement, reach, compromise, and peace. A response with the highest proportion of words drawn from the topic begins with "negotiations are better than war." Finally, words associated with the third topic seem to stress citizenry including words such as voice, rights, democracy, and autonomy. An exemplar response from the third topic states that "The government should listen to people's opinion."

Whether respondents read about an organization using the tactic of a bombing also affected topic prevalence in their written responses. Figure 7

presents difference-in-means estimates with 95% confidence intervals for the effect of reading about an organization using the tactic of bombing on the proportion of response dedicated to Topics 1, 2, and 3. For example, on average, reading about a movement that adopted bombing in pursuit of its goals increased the proportion of response discussing the topic we labeled "Terrorism" by 8%.

In general, the results of the STM are consistent with the logic behind Hypotheses 3A and 3B, which was derived from the logic that governments should adopt a strict policy of "no concessions" when negotiating with terrorist organizations. Respondents that read about movements using bombing as a tactic were more likely to provide responses with high proportions of words associated with conflict and terrorism. Considering that the words terrorists or terrorism were never used in the survey, their emergence in the topic model provides an insight into the mechanism driving the results we observe. In particular, organizations that use bombing as a tactic are more likely to be inferred to be terrorists, and respondents are subsequently less likely to think that the government should negotiate with terrorist organizations. Further research could explore whether this finding holds across a range of violent tactics or whether there is something unique about the use of bombings that invokes the idea that the movement is a terrorist organization.

Conclusion

In this article, we use a survey experiment conducted in Poland to explore how the tactical choices of social movements affect public opinion about whether the government should negotiate with the movement as well as the bargain that should be struck once negotiations begin. The use of an experimental design makes an empirical contribution to a field that has relied mainly on observational and qualitative data. Our results show that public support decreases for both separatist organizations and social movements that adopt bombing as a tactic when compared against occupations and demonstrations. The analysis of open-ended responses with a STM demonstrates that an important mechanism through which this occurs is the association of extreme strategies with terrorism. This article also contributes to a long-standing debate in both academic and policy worlds on the benefits of nonviolent action by nonstate actors during conflict processes by showing that the public is less likely to be sympathetic to dialogue with organizations adopting the most extreme tactics in pursuit of their goals.

The results of this article by no means invalidate the strains of research from which we derived Hypotheses 1 and 2 but rather point to the need for both further theoretical and empirical research specifying the conditions under which we might expect more extreme tactics to have heterogeneous effects among different segments of the public as well as how we think about tactics on the border between moderate and extreme. For the first category of research, which focuses on how higher levels of violence lead to increases in support, it is possible that these increases are occurring within a small subset of the population that was perhaps already in favor of either the organization or their cause. This points toward the need for future research exploring how the organization's target pool of supporters feeds back into its tactical choices. For example, if individuals more tolerant of violent tactics comprise a minority of the pool of potential supporters, then organizations face a trade-off between becoming the legitimate claimant to an issue among a smaller subset of the population and a decrease in support among the population as a whole.

For the second category of research, which focuses on how more moderate tactics lead to increases in support, it could be that rather than conceptualizing the tactics of movements as falling along a scale of extremeness as done in this paper, tactics are instead thought of as either being extreme, or not. Thus, the findings of this article are actually consistent with this second category because occupations are not extreme enough to garner the negative response that we observe for bombings. We agree this might be the case, and think that this points to the need for further research exploring how the public responds to different potential tactics operating along the potential border between moderate and extreme. That is, if occupations are not extreme enough to garner the negative response associated with bombings, what tactics are? We view answering this question, as well as further exploration of how the strategic choices of social movements affect public opinion, as interesting and exciting areas of further research.

Appendix

Words Associated With Each Topic in English and Polish

In Table A1, we report the original output of the STM model in Polish with the English translation. Words with multiple possible meanings were translated with an eye toward context. For example, "ruchem" (a declension of "ruch") was translated as "movement" instead of "motion" or "zdanie" as "opinion," instead of "sentence."

Table A1. Top Words Associated With Topics 1, 2, and 3 in Polish and With Their English Translation.

Topic 1		Topic 2		Topic 2	
Polish	English	Polish	English	Polish	English
negocjuje	negotiates	negocjacje	negotiations	prawo	law
takie	such	porozumienia	agreement	ludzi	people
terrorystami	terrorists	trudno	difficult	państwo	nation
znam	know	powiedzieć	tell	słuchać	listen
ruchem	movement	dojść	reach	powiedzenia	saying
zdanie	opinion	dogadać	agree	ludźmi	people
chodzi	it's about	lepiej	better	ruchu	movement
temat	topic	rządu	government	sytuacji	situation
separatystami	separatists	rozmowa	conversation	cos	something
innych	others	musi	must	maja	have
niech	let	sposób	way	państwie	country
rozmawia	talks	rozmów	talks	coś	something
bezpieczeństwo	security	stron	sides	głosu	voice
dobre	good	spokój	peace	doprowadzić	lead to
wszyscy	all	kraj	country	obywatele	citizens
wszystkimi	everyone	problemy	problems	celu	goal
sprawa	issue	każda	every	wojny	war
terroryści	terrorists	kompromis	compromise	autonomii	autonomy
jeszcze	still	lepsze	better	demokracji	democracy
spraw	matters	jakieś	some	racji	right

The Location of the Incident: Foreign or Domestic

Figure A1 shows that there are heterogeneous treatment effects depending on whether we specify the bombing to have occurred in a foreign country. When we do not specify the location of the bombing, respondents are significantly less likely to think that the government should enter negotiations with the government. In contrast, when we specify that the bombing occurred in a foreign country, there is not a significant difference in whether respondents think the government should negotiate with organizations using bombings when compared with occupations. Indeed, we only observe a 2.8% associated decrease in support for negotiations from a baseline of 57%. This is in marked contrast to the 10.5% decrease when the location is unspecified. We leave further theorizing of how the public views violent and separatist movements in locations within and external to their own country as an interesting area of further research.

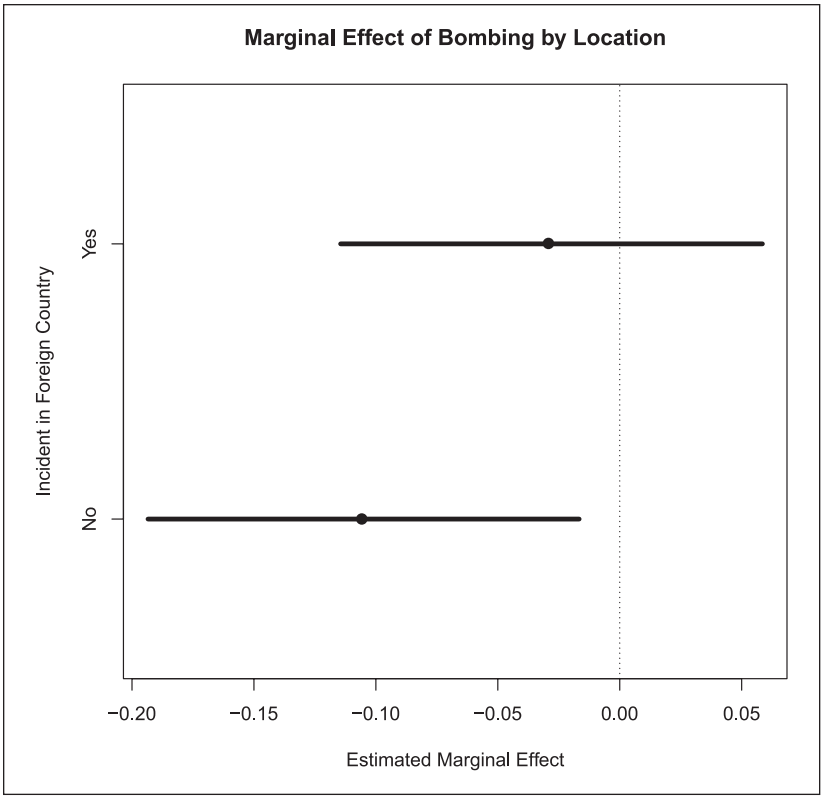

Figure A1. Heterogeneous treatment effects depending on whether the vignette specified that the separatist movement occurred in a foreign country.
The results demonstrate that while there is a significant different between occupations and bombings when the location of the incident is unspecified, this effect goes away when the incident is specified to have occurred in a foreign country.

Acknowledgments

For their excellent feedback throughout the development of this project we thank Mauricio Fernandez Duque, Joshua Kertzer, Audrey Latura, Christopher Lucas, David Romney, Anton Strezhnev, Kai Thaler, Dustin Tingley, Kris-Stella Trump, Ariel White, three anonymous reviers, and the editors of CPS.

Author's Note

The authors are listed in alphabetical order and contributed equally. Any remaining errors are their own. Replication files are available in the Data Archive on Dataverse http://dx.doi.org/10.7910/DVN/DMMQ4L

Declaration of Conflicting Interests

The authors declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: We thank Michael Findley and the University of Texas at Austin, the Harvard Experiments Working Group, and Dustin Tingley, for generously contributing funding to this project.

Notes

1. This view has been pervasive throughout the academic literature. For example, sociologists Taylor and Van Dyke (2004) state that "the tactics of protest used by social movements are so integral to popular views of social movements that sometimes a movement is remembered more for its tactics than for its goals" (p. 263).
2. For an account of how factors within the control of a movement affect the probability of success, see Gamson (1975). For a discussion of the impact of revolutionary barricades in the Age of Revolution in France, see Traugott (1993, 2010). For a discussion of the importance of tactical innovation in increasing the bargaining leverage of the civil rights movement, see McAdam (1983).
3. For a discussion of the decision to ban alcohol and a more thorough account of events during Solidarity's mobilization, see Ash (1999).
4. This was clearly demonstrated when public support swung dramatically against the Real Irish Republican Army (RIRA) after the bombing of Omagh in Northern Ireland. The bombing caused outrage and shock throughout both Catholic and Protestant communities in Northern Ireland and reaffirmed their commitment to peace negotiations. For a detailed discussion of the impact of the Omagh bombing, see Dingley (2001).
5. For how domestic politics constrains government actions in international diplomacy, see Putnam (1988).
6. A number of authors have shown how public opinion impacts legislative change. See, for example, Burstein (1979), Burstein and Freudenburg (1978), Page, Shapiro, and Dempsey (1987). Weeks (2008) argues that even nondemocratic leaders can be held accountable by domestic audience. An interesting area of further research would be to explore how the tactical choices of social movements affect public opinion in nondemocratic regimes.
7. When we measure how strongly respondents feel about whether the government should negotiate with a separatist movement, which is a slightly different outcome variable than our main quantity of interest, this finding is not statistically significant at the $\alpha = .05$ level.
8. The results-blind version of the manuscript can be found on the Experiments in Governance and Politics website: http://egap.org/registration/1257
9. For a broader discussion of the benefits and costs of a results-blind peer-review process, see Findley, Jensen, Malesky, and Pepinsky (in press). For other articles

that have gone through the peer-review process results-blind, see Bush, Erlich, Prather, and Zeira (in press); Hidalgo, Canello, and Lima-de-Oliveira (in press).

10. This issue is explored in even greater detail in their book length project Chenoweth and Stephan (2011).
11. It is important to note that throughout this article, we focus on the use of the tactic of bombing rather than whether a given movement is deemed a "terrorist organization." For a discussion of the evolution of the definition of terrorism, see Schmid and Jongman (1984).
12. Before being presented with the experimental portion of the survey, respondents are asked a number of standard questions about their background.
13. It should be noted that there were significant changes to the experimental design between the initial version of the manuscript and the final version that was conditionally accepted. An important advantage of going through the peer-review process prior to fielding the experiment is that it allows the researcher to make substantial revisions at the design stage, unlike in the standard peer-review process, which takes place after data collection and analysis have already been finalized.
14. These concerns generally focus on whether and how respondents in the sample generalize to other populations.
15. As we discuss in the conclusion, rather than creating a scale of extremeness as we do in this article, further research could explore how the tactical choice between violent and nonviolent tactics affects public opinion. Pursuing this question would involve increasing the number of treatment categories by including both more violent and more nonviolent tactics.
16. The approach used in this article is consistent with how prior research deals with potentially idiosyncratic features of vignettes. See, for example, Tomz (2007).
17. Whether this sentence says "should" or "should not" varies to match the respondent's answer to the previous question.
18. The wording used to ask respondents to explain their answer in a few sentences is consistent with prior research exploring potential mechanisms through open-ended survey questions (Tomz, 2007).
19. The Comprehensive Agreement on Bangsamoro is a peace agreement signed in March 2014 between the government of the Philippines and the Moro Islamic Liberation Front. The agreement establishes an autonomous Bangsamoro.
20. In particular, we model the structure of the bargaining process for the extent of fiscal tax autonomy after the 2009 Spanish arrangement with the autonomous communities, which increased the share of the national pool of personal income tax revenues assigned to the regions from 33% to 50% (Blöchliger & Vammalle, 2012; Boletin Oficial del Estado. ley 22/2009, 2009).
21. In Spain, taxes levied in all regions are collected into a common pool and then shared between the central government and the autonomous communities. The recent agreement is a result of negotiations over a range of fiscal and financial matters, but for the purposes of the experiment, we focus on the aspect most useful for our theoretical question of interest and simplify it to mimic political rhetoric surrounding comparable issues throughout Europe.
22. Only about 2% of population identified itself as Silesian and 0.6% as Kashub in 2011 according to Central Statistical Office of Poland (2011). The German

minority, about 0.6% of the Polish population, is exempt from the electoral threshold of 5% to facilitate electoral representation. Upper Silesia enjoyed brief autonomy in the interwar period (Dembinska, 2012), including the collection of taxes and public fees, with only a share (determined based on a formula) of taxes going to the central government for national services (Bialasiewicz, 2002) but no similar arrangements were made in post-war period for any region.

23. The Silesian Autonomy Movement has run candidates in local elections, organizes an annual Autonomy March, and initiated a petition gathering approximately 124,000 signatures for the recognition of Silesians as an ethnic group. The Kashub claims are mostly centered on cultural issues.
24. Quota sampling is a widely used sampling technique which requires interviews to fill a quota of respondent specific attitudes by certain classifying variables. For more on random-quota sampling, see Smith (1983). Taylor Nelson Sofres (TNS) Global uses random-quota sampling based on geographic location. Due to the high correlation in Poland between geography and other potential moderating variables such as income and political ideology this sampling technique helps ensure balance on these important covariates.
25. We present heterogeneous treatment effects depending on whether we specify the bombing to have occurred in a foreign country in Figure A1 in the Appendix.
26. For an overview of the application of Structural Topic Models (STMs) to political science, see Lucas et al. (2015) and Roberts et al. (2014).
27. Latent Dirichlet Allocation is a mixed-membership model where each document is represented as a mixture over a set of topics (Blei, 2012; Blei, Ng, & Jordan, 2003). Each topic is a distribution over the words in the vocabulary which is learned, rather than assumed, in the model.
28. In contrast, when using "supervised" methods, the researcher defines the topics before analysis by hand-coding a set of documents into predefined categories. For a prominent introduction to supervised learning in political science, see Hopkins and King (2010).
29. The results of the STM are similar after excluding responses from individuals reading about an organization using the tactic of a demonstration from the baseline category.
30. The stopword list was customized to include additional uninformative but frequently occurring Polish words. The words were not stemmed as available text-analysis tools so far do not support Polish.
31. As specified in the pre-analysis plan, we leave the exploration of potential heterogeneous treatment effects to future research.
32. We chose to run a model with three topics after estimating the model varying number of topics (k) from 3 to 10 and a substantive interpretation of the results of each specification. Three to 10 topics are recommended as a useful starting range for survey experiments because of a low variety in the content of responses given to a focused question and we found that in our case, models with a number of topics higher than three produced less readily interpretable topics.
33. We present top words for each topic in Polish and English in Table A1 in the Appendix.

34. We use the simplified frequency-exclusivity scoring (FREX), which summarizes words with the harmonic mean of the probability of appearance under a topic and the exclusivity to that topic, providing the most semantically intuitive representation of topics (Roberts et al., 2014).
35. As a validation step, we also examined exemplar documents for each topic which included responses with the highest proportion of words drawn from the topic.

References

Abrahms, M. (2012). The political effectiveness of terrorism revisited. *Comparative Political Studies*, *45*, 366-393.

Abrahms, M. (2013). The credibility paradox: Violence as a double-edged sword in international politics. *International Studies Quarterly*, *57*, 660-671.

Ash, T. G. (1999). *The Polish Revolution: Solidarity*. New Haven, CT: Yale University Press.

Astin, A. W., Astin, H. S., Bayer, A. E., & Bisconti, A. S. (1975). *The Power of Protest*. San Francisco, CA: Jossey-Bass.

Bialasiewicz, L. (2002). Upper Silesia: Rebirth of a regional identity in Poland. *Regional & Federal Studies*, *12*, 111-132.

Blei, D. M. (2012). Probabilistic topic models. *Communications of the ACM*, *55*(4), 77-84.

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent dirichlet allocation. *The Journal of Machine Learning Research*, *3*, 993-1022.

Blöchliger, H., & Vammalle, C. (2012). *Reforming fiscal federalism and local government: Beyond the zero-sum game*. Paris, France: Organisation for Economic Co-Operation and Development.

Bloom, M. M. (2004). Palestinian suicide bombing: Public support, market share, and outbidding. *Political Science Quarterly*, *119*, 61-88.

Boletin oficial del estado. ley 22/2009. (No. 305). (2009). Jefatura del Estado. (Reference number BOE-A-2009-20375.)

Bueno de Mesquita, E., & Dickson, E. S. (2007). The propaganda of the deed: Terrorism, counterterrorism, and mobilization. *American Journal of Political Science*, *51*, 364-381.

Burstein, P. (1979). Public opinion, demonstrations, and the passage of antidiscrimination legislation. *Public Opinion Quarterly*, *43*, 157-172.

Burstein, P., & Freudenburg, W. (1978). Changing public policy: The impact of public opinion, antiwar demonstrations, and war costs on senate voting on Vietnam war motions. *American Journal of Sociology*, *84*, 99-122.

Bush, S., Erlich, A., Prather, L., & Zeira, Y. (in press). The effects of authoritarian iconography: An experimental test. *Comparative Political Studies*.

Central Statistical Office of Poland. (2011). Ludnosc wedlug identyfikacji narodowoetnicznych oraz ekonomicznych grup wieku w 2011 roku [Population according to national-ethnic identification and economic age groups in 2011]. Available from http://stat.gov.pl/

Chellaney, B. (2006). Fighting terrorism in southern Asia: The lessons of history. *International Security*, *26*, 94-116.

Chenoweth, E., & Stephan, M. J. (2011). *Why Civil Resistance Works: The Strategic Logic of Nonviolent Conflict*. New York, NY: Columbia University Press.

Clutterbuck, R. (1992). Negotiating with terrorists. *Terrorism and Political Violence*, *4*, 263-287.

Dembinska, M. (2012). (Re) framing identity claims: European and state institutions as opportunity windows for group reinforcement. *Nations and Nationalism*, *18*, 417-438.

Dingley, J. (2001). The bombing of Omagh, 15 August 1998: The bombers, their tactics, strategy, and purpose behind the incident. *Studies in Conflict and Terrorism*, *24*, 451-465.

Ekiert, G., & Kubik, J. (2001). *Rebellious Civil Society: Popular Protest and Democratic Consolidation in Poland, 1989-1993*. Ann Arbor: University of Michigan Press.

Findley, M., Jensen, N., Malesky, E., & Pepinsky, T. (in press). Transparency in the social sciences. *Comparative Political Studies*.

Gamson, W. A. (1975). *The Strategy of Social Protest*. Homewood, IL: Dorsey Press.

Giugni, M. G. (1998). Was it worth the effort? The outcomes and consequences of social movements. *Annual Review of Sociology*, *24*, 371-393.

Hidalgo, D. F., Canello, J., Lima, R. (in press). Can politicians police themselves? Natural experimental evidence from Brazil's audit courts. *Comparative Political Studies*.

Hopkins, D. J., & King, G. (2010). A method of automated nonparametric content analysis for social science. *American Journal of Political Science*, *54*, 229-247.

Kitschelt, H. (1986). Political opportunity structures and political protest: Anti-nuclear movements in four democracies. *British Journal of Political Science*, *16*, 57-85.

Lucas, C., Nielsen, R. A., Roberts, M. E., Stewart, B. M., Storer, A., & Tingley, D. (2015). Computer-assisted text analysis for comparative politics. *Political Analysis*, *23*, 254-277.

McAdam, D. (1983). Tactical innovation and the pace of insurgency. *American Sociological Review*, *48*, 735-754.

McAdam, D., Tarrow, S., & Tilly, C. (2001). *Dynamics of Contention*. Cambridge, UK: Cambridge University Press.

Netanyahu, B. (Ed.). (1986). *Terrorism: How the West Can Win*. New York, NY: Farrar, Straus and Giroux.

Page, B. I., Shapiro, R. Y., & Dempsey, G. R. (1987). What moves public opinion? *The American Political Science Review*, *81*, 23-43.

Pape, R. A. (2003). The strategic logic of suicide terrorism. *American Political Science Review*, *97*, 343-361.

Piven, F. F., & Cloward, R. A. (1979). *Poor People's Movements: Why They Succeed, How They Fail* (Vol. 697). New York, NY: Vintage Books.

Polletta, F. (2012). *Freedom is an Endless Meeting: Democracy in American Social Movements*. Chicago, IL: University of Chicago Press.

Putnam, R. D. (1988). Diplomacy and domestic politics: The logic of two-level games. *International Organization*, *42*, 427-460.

Roberts, M. E., Stewart, B. M., Tingley, D., Lucas, C., Leder-Luis, J., Gadarian, S. K., . . . Rand, D. G. (2014). Structural topic models for open-ended survey responses. *American Journal of Political Science*, *58*, 1064-1082.

Schmid, A. P., & Jongman, A. J. (1984). *Political Terrorism: A Research Guide to Concepts, Theories, Databases and Literature*. New Brunswick, NJ: Transaction Books.

Sharp, G. (1973). *The Politics of Nonviolent Action* (3 Vols.). Boston, MA: Porter Sargent.

Shaykhutdinov, R. (2010). Give peace a chance: Nonviolent protest and the creation of territorial autonomy arrangements. *Journal of Peace Research*, *47*, 179-191.

Shorter, E., & Tilly, C. (1974). *Strikes in France, 1830-1968*. Cambridge, UK: Cambridge University Press.

Smith, T. (1983). On the validity of inferences from non-random sample. *Journal of the Royal Statistical Society. Series A (General)*, *146*, 394-403.

Staggenborg, S. (1988). The consequences of professionalization and formalization in the pro-choice movement. *American Sociological Review*, *53*, 585-605.

Stephan, M. J., & Chenoweth, E. (2008). Why civil resistance works: The strategic logic of nonviolent conflict. *International Security*, *33*, 7-44.

Swidler, A. (1986). Culture in action: Symbols and strategies. *American Sociological Review*, *51*, 273-286.

Tarrow, S., & Tollefson, J. (1994). *Power in Movement: Social Movements, Collection Action and Politics*. Cambridge, UK: Cambridge University Press.

Taylor, V., & Van Dyke, N. (2004). "Get up, stand up": Tactical repertoires of social movements. In D. A. Snow, S. A. Soule, & H. Kriesi (Eds.), *The Blackwell Companion to Social Movements* (pp. 262-293). Malden, MA: Blackwell.

Thomas, J. (2014). Rewarding bad behavior: How governments respond to terrorism in civil war. *American Journal of Political Science*, *58*, 804-818.

Tomz, M. (2007). Domestic audience costs in international relations: An experimental approach. *International Organization*, *61*, 821-840.

Traugott, M. (1993). Barricades as repertoire: Continuities and discontinuities in the history of French contention. *Social Science History*, *17*, 309-323.

Traugott, M. (2010). *The Insurgent Barricade*. Berkeley: University of California Press.

Weeks, J. L. (2008). Autocratic audience costs: Regime type and signaling resolve. *International Organization*, *62*, 35-64.

Author Biographies

Connor Huff is a PhD candidate in the Department of Government at Harvard University. His research focuses on the causes and consequences of the strategic decisions of non-state actors. His dissertation explains why individuals in the same militant organization come to different conclusions about whether their organization should accept a settlement to a conflict.

Dominika Kruszewska is a PhD candidate in the Department of Government at Harvard University. Her research explores the relationships between social movements, the public, and the institutions and agents of the state. Her dissertation examines the effects of protest or activist origins on voter mobilization strategies and political program of new political parties in Europe.