Confronting the New Gatekeepers of Experimental Political Science

Clayton Webb*
Department of Political Science
University of Kansas
Lawrence, KS 66045
Email: webb767@ku.edu

Danielle Lupton
Department of Political Science
Colgate University
Hamilton, NY13346
Email: dlupton@colgate.edu

Abstract

Political science has witnessed a veritable explosion in the use of experiments. In response, authors, reviewers, and journal editors have sought to devise best practices for experimental work by promoting (i) a bias against the use of convenience samples, (ii) for the inclusion of multiple experiments per study, and (iii) for the requirement of pre-registration. However well-intended, these efforts only serve to increase the cost of experimental research. Without any clear scientific basis, journals are starting to require pre-registration plans exclusively for experimental research; researchers aggrandize the use of nationally representative and other non-convenience samples, even when there is no inferential target; and editors and reviewers often expect multiple experiments in a single paper, when only one is necessary for causal inference. We argue these norms are not only theoretically bankrupt, but they also serve to exacerbate institutional inequalities and elitism in the discipline. We highlight the problems of these norms and provide a set of recommendations to head off these bad practices.

*Corresponding Author. Authors are listed on the basis of author rotation. Both authors contributed equally.

Experimental research in political science has become a pay-for-play enterprise. The currencies are time and money. Researchers with more of these resources have a *relatively* easy time publishing experimental work, while researchers with less of these resources are finding it harder to publish their experimental findings. While critics may argue this is true of any form of research, we contend that experimental political science is becoming more elitist than any other domain of the discipline due to an emerging set of (non-scientific) norms about what constitutes good experimental research.

We identify three norms of experimental research that are increasingly advocated by editors, reviewers, and even authors: the bias against convenience samples; the need for multiple experiments per study; and the requirement that experimental studies (and only experimental studies) be pre-registered. While we do not dispute the idea that the discipline would benefit from a conversation about what constitutes appropriate application of methods, we argue these standards serve as de-facto gate keeping mechanisms for experimental research. This is especially concerning as these new norms evolved from a series of half truths and folk logics. These new "*standards*" are not grounded in sound scientific theory. The problems these norms are meant to address manifest in other research methods; yet, these norms are only applied to experimental work.

Our argument has three parts. We maintain that: (i) the problems these norms are ostensibly intended to resolve are overstated, (ii) these norms do not fix the underlying problems they are intended to address, and (iii) the only thing these new "*standards*" for research do is make it more difficult for scholars to conduct experimental research. These barriers may marginally increase the quality of some experimental studies, but they make it unnecessarily harder for most scholars to publish experimental research. Nationally representative samples (and many other non-convenience samples) are extremely expensive. Fielding multiple experiments per study costs time and money. Pre-registration can add substantial amounts of time to the research process.

The costs imposed by these non-scientific "*standards*" are changing the landscape of political science. We hypothesize that these new norms skew the experimental research published in major political science journals toward elite institutions. While there might be reasons to expect schol-

ars from elite institutions to be more productive than scholars from non-elite institutions, elitism in experimental research is more pronounced because observational and normative research lacks comparable research costs.

We test our proposition by analyzing the research designs used in *The Journal of Politics* (JOP), *The American Journal of Political Science* (AJPS), and *The American Political Science Review* (APSR) from 2010 to 2021. Non-experimental research serves as the counterfactual. We compare the average institutional rank of authors who published experimental research in these journals to the average institutional rank of authors who published non-experimental research over the same window. If these new norms for experimental research are benign, we would not observe any differences in average institutional rank based on research design. Yet, as these special requirements for experimental research are not only ill-conceived and ill-considered, but also inimical to the impartial system of study and review that is the basis of our discipline, we observe statistically significant differences in institutional rank based on experimental vs. non-experimental research. These trends are consistent across journals and subfield.

This slide toward elitism is not the result of some macabre conspiracy by elite faculty to cumber the academic endeavors of their colleagues at less elite institutions. These norms evolved from an earnest desire among scholars from elite and non-elite institutions to improve the quality of experimental research. Unfortunately, our collective failure to consider the scientific bases of these norms has had deleterious consequences. The goal of this study is to explain the problems with these new norms, highlight the obstacles they create, and appeal to the better judgment of reviewers and editors who should have a vested interest in keeping our discipline fair and honest.

1 Contemporary Experimental Political Science

An experiment is a research design where one randomly assigns values of the independent variable (the treatment) to participants and controls the levels of the treatment they receive (Kellstedt and

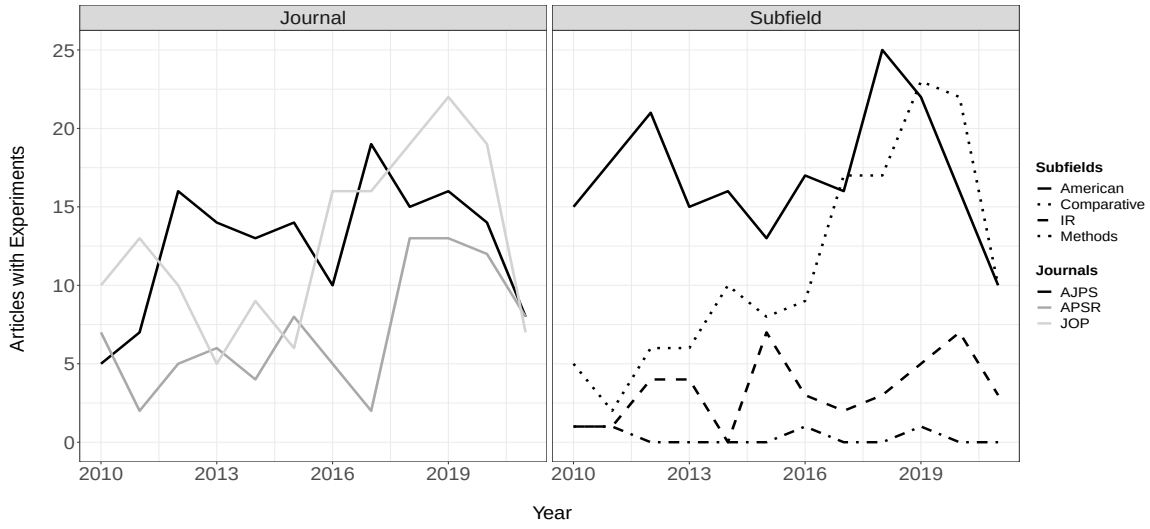
Whitten 2018). Control and random assignment are the key features that separate experiments from other research designs. In an observational study, one samples the data generating process (DGP) and tries to detect covariation between their dependent variable(s) and their variable(s) of interest. This often involves controlling for covariation that might confound inferences. In an experiment, one controls the DGP. Control and random assignment all but eliminate concerns about confounding because, by design, no confounder can be correlated with the randomly assigned treatment.

Survey experiments have been used to study a variety of phenomena. They are especially useful for understanding: (i) public perceptions, attitudes, and opinions; (ii) the mechanisms of decision-making; and (iii) the microfoundations of political processes. International Relations (IR) scholars employ experiments to understand audience costs (e.g, Levendusky and Horowitz 2012), foreign policy preferences (e.g, Guisinger and Saunders 2017; Kertzer and Zeitzoff 2017; Lupton and Webb 2022), and the influence of public opinion on foreign policy (e.g., Avdan and Webb 2019; Tomz and Weeks 2020). American politics scholars use experiments to analyze political behavior (e.g., Broockman and Green 2014; Saha and Weeks 2022), public preferences (e.g., Blankshain, Glick and Lupton 2023; Bulut and Yildirim 2023), and political psychology (e.g., Jost, Baldassarri and Druckman 2022). Comparative politics scholars use experiments to examine elections (e.g., Ichino and Schündeln 2012), institutional strategies (e.g., Gottlieb 2017), democratic backsliding (e.g., Fossati, Muhtadi and Warburton 2022), and media effects (e.g., Green et al. 2020).

In recent decades, political science has witnessed increased interest in the use of experiments. To catalogue this trend, we conduct a large scale survey of the three most prominent journals in political science—the AJPS, the APSR, and the JOP—between 2010 and 2021. Of the 2,229 articles in our sample, 388 employ at least one randomized experiment.[1] Figure 1 presents the annual count of experimental articles, broken down by journal (left panel) and subfield (right panel). For articles from the AJPS, we employ data obtained from the editors classifying each article's subfield. For

[1] A study is classified as experimental if it includes a randomized experiment. Monte-Carlo experiments, natural experiments, and quasi-experiments are not treated as randomized experiments.

Figure 1: APSR, AJPS, and JOP Experiments Over Time



Note: Annual counts of experiments over time, across subfield and journal.

articles from the APSR and the JOP (for which such data were not available), we classify subfield based on the content of each article.

Figure 1 shows the increased use in experiments over the first 10 years of our sample. Across all three journals and three of the four subfields, the number of experiments in 2019 is higher than 2010. The AJPS published the most experiments in the first half of the sample, but there is a marked increase in experiments published in the JOP between 2016 and 2019. Experimental designs are especially popular in American and Comparative politics during the sampling window. While there are less in IR, we still observe an upward trajectory across the subfield. The only outlier is methodology, which had few experiments.

The other discernible trend in Figure 1 is the precipitous decline in the number of experimental studies in the last two years of the sample. Across all three journals, there is a noticeable drop in the number of experiments in 2020, and an even steeper decline in 2021. This dramatic decline is also evident across subfields. The number of experimental papers in American politics drops from 22 in 2019 to 10 in 2021 (-55%), in Comparative politics from 23 to 10 (-57%), and in IR from 5 to 3 (-40%).

What can explain this decline? Perhaps we learned all we could with randomized experiments between 2010 and 2019 or perhaps the sudden change in trajectory reflects increased interest in other methods. The more likely explanation is that experimental research suddenly and surprisingly became more difficult to publish. In the next section, we discuss what we believe to be the genesis of this wane in experimental research: new "*standards*" for what constitutes "*good*" experimental political science that have almost no bases in scientific theory.

2 The New Gatekeepers

There are three bad norms stunting the development of experimental political science: (i) the edict that studies should not employ convenience samples, even when there is no coherent inferential target in the population; (ii) a similarly arbitrary requirement that there must be multiple experiments per study for that study to be sufficiently rigorous; and (iii) the bidding that experimental studies *in particular* must be pre-registered. We argue none of these norms have sound scientific bases. In what follows, we consider each norm, evaluate their rationales, and highlight their flaws.

2.1 The Bias Against Convenience Samples

The first norm we interrogate is the bias against the use of convenience samples in experimental research. This is often, but not always, exemplified by a preference for nationally representative samples. This bias focuses on avoiding the use of samples that are relatively inexpensive and less time consuming to collect: student convenience samples and online convenience samples.

Introductory statistics texts often stress the importance of sampling for inferential statistics. In these novice formulations, the practice of inferential statistics is expressly about the use of sample statistics to make inferences about population parameters. A *population* is the universe of all possible cases, or individuals, that could be included in a sample. A *sample* is a subset of a population. *Population parameters* are the quantities of interest; examples include a population mean (μ),

the difference between two population means ($\mu_1 - \mu_2$), or a population regression parameter (β). The sample statistics are the analogues from the sample: the mean ($\bar{x}$), the difference between two means ($\bar{x}_1 - \bar{x}_2$), and the sample regression coefficient ($\hat{\beta}$). When one is using a sample statistic to make an inference about a population parameter, the population parameter is the *inferential target* of the analysis.

The ideal samples for inferential statistics, in theory, are random probability samples. A *random probability sample* is a sample where each element of the population has an equal likelihood of being selected, and the selection of any one element does not influence the selection of any other (Kellstedt and Whitten 2018, 144). This formulation is highlighted in introductory texts because it provides a useful basis to discuss concepts like selection bias, measurement error, and sample uncertainty. In practice, a random probability sample is impractical, if not impossible, to collect.

There are theoretical and practical limitations that render random probability sampling a statistical pipe-dream. The primary theoretical limitation concerns access. Unless one posits an especially narrow sampling frame, it is almost impossible to guarantee that every member of a population has an equal likelihood of being sampled. Consider, for example, the population of voting eligible United States (U.S.) adults. How can one obtain a list of all people who are eligible to vote? If one had such a list, how could one contact them? Not all voting eligible U.S. adults have computers, cell-phones, or landlines. Even if they did, most people do not respond to surveys. There are sampling strategies one can use to obviate some of these problems, but the necessity of these compromises highlights the fanciful nature of a random probability sampling.

The practical limitation to random probability sampling is cost. According to the National Science Foundation's solicitation for the 2024 American National Election Study (ANES), the survey will cost $14 million to administer over four years. Overcoming the aforementioned challenges requires a lot of time and effort. This is why placing questions on the ANES, or comparable surveys, is so expensive. Experimental treatments tend to cost more than conventional survey

questions because they take more time. As most scholars cannot access these kinds of resources for most studies, they typically rely on cheaper alternatives: quota samples and convenience samples.

A *quota sample* – or stratified sample – is a non-probability sample drawn so features of the sample match features of a target population. The researcher selects strata, or "quota controls," relevant to their outcome variable; the sample is selected to be proportionate to the population on these strata (Yang and Banamah 2014, 33). Common strata include age, sex, race, education, and partisanship. Firms like Dynata, Qualtrics, and YouGov maintain online panels of participants that can opt into studies for pay. For a (considerable) fee, these firms provide samples matching requested strata. When researchers report using a "*nationally representative sample*," this is usually what they mean, not random probability samples.

A *convenience sample* is a non-probability sample selected based on availability. Undergraduate students (Lupton 2019) and participants recruited through Amazon's Mechanical Turk (MTurk) (Mullinix et al. 2015) are examples. Convenience samples are not costless. Some are able to recruit undergraduates through pools, but many must recruit students piecemeal. This, at a minimum, takes time and can have financial costs if students receive remuneration. While MTurk samples are not free, MTurk affords access to large samples at a fraction of the cost. For our purposes, the essential distinction is the difference between convenience samples and non-convenience samples. Convenience samples are comprised of student samples, online convenience samples, and other relatively inexpensive samples. Non-convenience samples are relatively expensive quota samples and prohibitively expensive random probability samples.

There are three other types of non-convenience samples that warrant attention: field samples, lab samples, and elite samples. A *field sample* is a sample used in a field experiment. A *field experiment* is "a research method that uses some controlled elements of traditional lab experiments, but takes place in natural, real-world settings" (Lee 2023), ostensibly for the purpose of promoting realism and external validity (Gerber and Green 2012). Field experiments vary in terms of

their "fieldness," including setting, authenticity of treatment, participants, context, and outcome measurement (Baldassarri and Abascal 2017, 43).

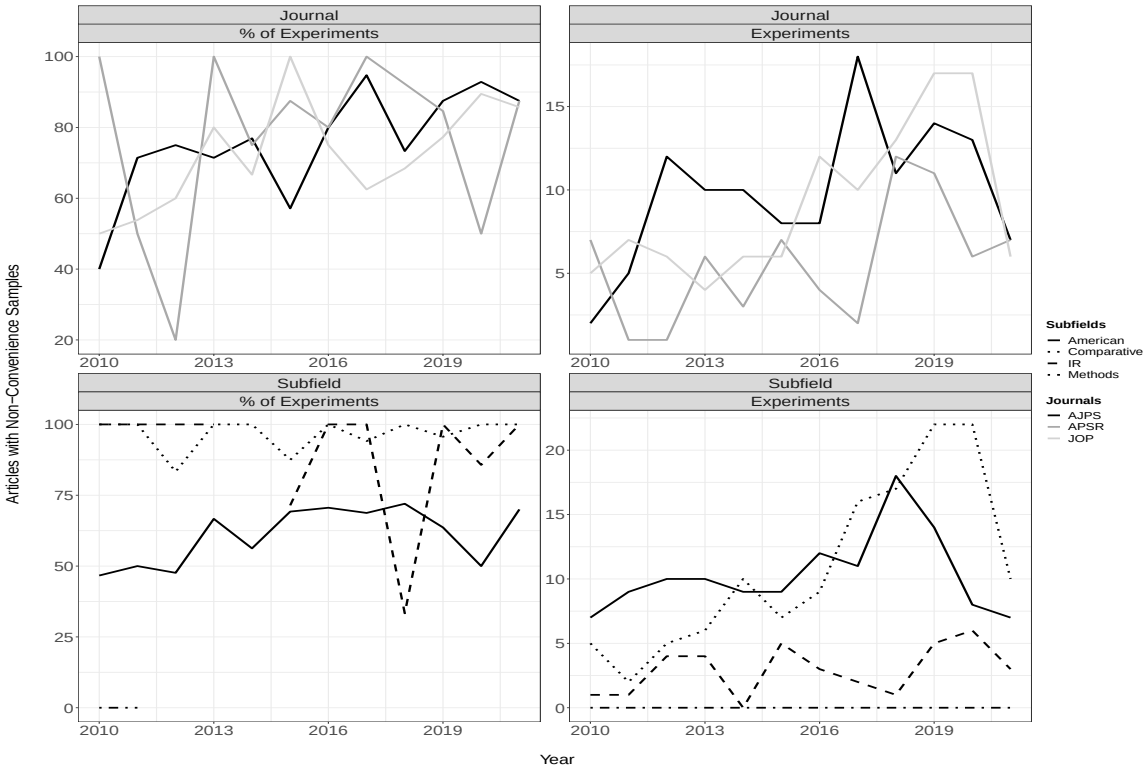
Lab samples often rely on undergraduates or volunteers from local communities. While some consider lab samples convenience samples, the costs associated with lab samples are large. There is the fixed investment required to establish the *lab*. Lab samples also take more time to acquire and participants are paid for the physical presence.[2]

Finally, *elite samples* draw from "elite" populations. They can take more time to acquire because they are difficult to access. Elite samples can range from elected officials (Grose, Malhotra and Parks Van Houweling 2015; Flavin and Hartney 2017; Zelizer 2019) to leaders in specific fields (Zhu and Shi 2019) to individuals with specific skills or backgrounds (Hafner-Burton, LeVeck and Victor 2016; Dietrich, Hardt and Swedlund 2021), just to name a few. The use of elite samples is often justified on the grounds that one can only adequately understand decision-making, or other processes, by solely studying elite populations (Dietrich, Hardt and Swedlund 2021).

We group these samples with other non-convenience samples for the sake of our discussion because they are all more costly than true convenience samples. The irony of field samples, lab samples, and elite samples is that almost none of them are representative of any population. Many field experiments do not aim to be representative; they "often take place in specific communities" and/or "rely on the voluntary participation of subjects" (Baldassarri and Abascal 2017, 44). Most elite samples are groups of convenient elites. Yet, the derision toward the use of non-representative student and online convenience samples does not apply to non-representative field and elite samples. A student sample from a lab is no more representative that a student sample from a classroom (Grose 2021). It may be less representative because individuals need to physically show up at a specific location and time. That these samples pass muster where typical convenience samples do not belies the fact that there often aren't principled bases for rejecting convenience samples. There

[2]The exception would be situations where undergraduates are required to participate for credit.

Figure 2: APSR, AJPS, and JOP Experimental Samples Over Time



Note: Annual counts and % of experiments relying on nationally representative samples and other non-convenience samples over time, across subfield and journal.

is only a sense amongst reviewers that you should get credit for having something special, paying for something expensive, or doing something hard. These are not a scientific criteria.

Figure 2 highlights the evolving bias against convenience samples in the big-three political science journals. There are four panels in Figure 2. The top (bottom) panels break down the variation in sample by journal (subfields). The panels on the right show the raw counts; the panels on the left express these counts as percentages of the experimental papers published annually.

The panels on the left of Figure 2 demonstrate the bias against convenience samples. For two of the three journals, there is a clear upward trend in non-convenience samples. In 2010, 40% of AJPS experimental articles and 50% of JOP experimental articles used non-convenience samples. By 2021, both rose to just below 90%. This is where the percentage of APSR experimental articles

lies at the end of the sample as well, but we cannot say that it increased because 100% of APSR experimental articles relied on non-convenience samples in 2010.

We see similar patterns across subfields. In 2010, just under 50% of American politics experimental articles employed non-convenience samples. By 2021, this rose to nearly 75%. The bias against non-convenience samples is even more pronounced in Comparative politics, where all, or nearly all, experimental articles rely on non-convenience samples. The trend in IR is more volatile but the broader trend across IR indicates that non-convenience samples are becoming more widely used; 100% of IR experimental articles published in 2010, 2016, 2019, and 2021 employed non-convenience samples.

There is one outlier among the sub-fields: methodology. There are no methodology papers published in the top three journals over the sampling window that rely on non-convenience samples. This may be because there are few methodology articles about experiments, but there is another explanation. The reason that political methodologists do not seem to share the bias toward non-convenience samples is because there is no methodological reason for this bias to exist!

We posit that most experimental studies do not require, or even benefit from, the use of expensive samples. This follows from the sampling concepts introduced earlier. Random probability samples – and their second-rate cousins, nationally representative samples – are only useful for bona fide inferential analyses: analyses where the sample statistics are aimed at *specific* inferential targets. When pollsters collect data about support for prospective nominees, for example, their inferential targets are the proportions of voting eligible U.S. adults. When the Institute for Social Research conducts their survey on consumer sentiment, their goal is to measure national perceptions of the economy. Many political science experiments do not profess goals so lofty.

This observation is not new. Consumer research and marketing scholars have been conscious of this point for decades. Calder, Phillips and Tybout (1981, 197) draw a distinction between "*theory application research*" aimed at testing theoretical arguments beyond the research setting and "*effects application research*" that aims to isolate the substance of specific treatment effects

in real world situations. A convenience sample is sufficient for theory application research. If the theory is generally applicable, one should be able to observe evidence consistent with the theory in any setting: with students in a classroom or with convenience samples online. As Sarstedt et al. (2018, 4) explain, "in theory application research, sample representativeness is of little concern," because "the aim is not to generalize findings to a real-world situation, but to examine an effect in a particular research setting." Effects application research, in contrast, seeks to "obtain findings that can be applied directly to the situation of interest" Calder, Phillips and Tybout (1981, 198).

Many experimental studies in political science can be classified as theory application research. In their study of public perceptions of terrorism, for example, Avdan and Webb (2019) pose hypotheses about the way people, generally, perceive terrorist attacks in foreign countries. Experiments like this, aim to "capture basic psychological processes that may be safely assumed to be invariant across populations, experimental context or subject awareness" (Gerber 2011, 116). A convenience sample is sufficient for this kind of inference. Indeed, these sorts of samples have been judged sufficient for research on reputation formation (Renshon, Dafoe and Huth 2018), how international political knowledge shapes perceptions of domestic political institutions (Huang 2015), and perceptions of the use of domestic force (Blankshain, Cohn and Lupton 2024). That these studies have landed in prominent political science journals suggests that the bias is neither universally held nor consistently applied. But rather than alleviating concerns about the observed bias, the non-uniformity highlights its arbitrary nature.

The absurdity of the bias is put in stark repose when one considers how the inferential aims of such studies would need to be modified to necessitate the expense of a nationally representative sample. Avdan and Webb (2019) use fake news stories, about fake terrorist attacks, by fake terrorist organizations, that did not take place, at single points in time. What would the inferential target for this kind of experiment be? A population treatment effect? It is true that they could have used a nationally representative sample to discern what the population difference would have been if everyone in the U.S. would have read the same fake news stories about the same fake terrorist

organizations at the same point in time. But they would have learned nothing especially interesting about the world and they would have wasted tens-of-thousands of dollars in the process.

The bias against convenience samples is not merely misguided, it is damaging. If this requirement for non-convenience samples were applied in the past as it is applied today, it would disqualify some of the most important experimental work ever published. For example, in their classic experiment on voter turnout, Gerber and Green (2000) use a sample of 30,000 registered voters in New Haven, Connecticut. This is hardly representative of the entire U.S., a point they highlight in the conclusion.

2.2 Multiple Experiments per Study

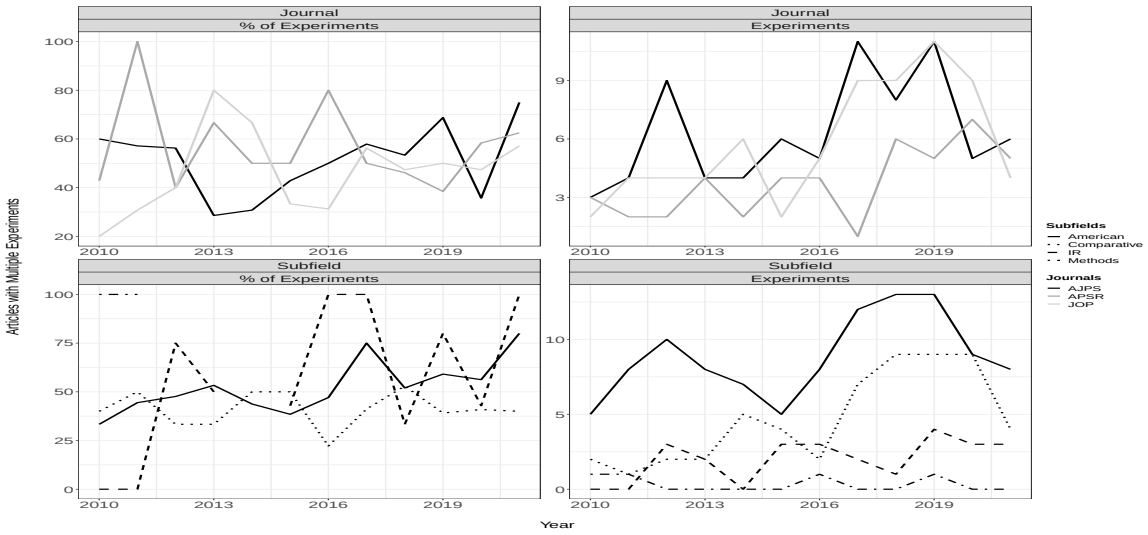
Another trend in experimental research is the proliferation of multi-survey/multi-experiment studies. Figure 3 shows the number of multi-experiment studies published over our sampling window is increasing over time, across journal and subfield.[3] Over a quarter (25.8%) of the experimental articles in our sample had three or more experiments. The maximum number of experiments published in a single article was 19, with the mean hovering around 2.

There is absolutely no scientific principal to support the idea that one *must* have multiple experiments per study to draw valid conclusions. Kellstedt and Whitten (2018) outline the standard precepts of knowledge production through theory building and hypothesis testing: (i) develop a causal theory; (ii) formulate hypotheses; (iii) select an appropriate design and empirical test; (iv) evaluate the hypothesis; (v) evaluate the causal theory in light of the hypothesis. Notably absent from this list is: (vi) repeat step (iii) an arbitrary number of times until you run out of money.

There are different rationales for conducting multiple experiments. Not all of them are valid. A good reason to conduct the same experiment multiple times is external validity (McDermott

[3]Methodology is an outlier because there are only four years in the data where there were methodology papers about experiments.

Figure 3: APSR, AJPS, and JOP Articles with Multiple Experiments Over Time



Note: Annual counts of experiments over time, across subfield and journal.

2002a). If one conducts an experiment using one sample in Raleigh, NC, a second in Los Angeles, CA, a third in Ann Arbor, MI, another from MTurk, and finds statistically equivalent effects in each, the researcher is demonstrating that the result is not an artifact of the sample (e.g., Lupton 2019). Cook and Campbell (1979) refer to this as "*deliberately sampling for heterogeneity*," and they recommend this approach as a means of establishing the external validity of a result when probability sampling or quota sampling is impossible or impractical. This rationale is *not* offered by any of the experimental studies in our sample.

Two less reasonable, yet more common, rationales for conducting multiple experiments are robustness and exploration. The former presents multiple experiments as a means of demonstrating that inferences are robust to incidental changes in the treatment, for which their may or may not be any good reason to expect the treatment effects to change. Guisinger and Saunders (2017), for example, employ nine different experiments to evaluate their hypotheses about elite cues and issue context. The authors argue that the same partisan hypothesis should apply regardless of whether the participant reads about China Currency, the China Pivot, the WTO, the International Center for the Investment Disputes, Cap and Trade, Climate Security, Iran, Syria, or the International

Criminal Court. Exploration is a euphemism for testing conjectures not explicitly proposed in the theory. Avdan and Webb (2019), for example, propose a theory about the physical and personal proximity of foreign terrorist attacks on public perceptions of terrorism. After testing their theory, they pontificate about the possibility that religion is a possible causal mechanism and test this conjecture using a separate experiment.

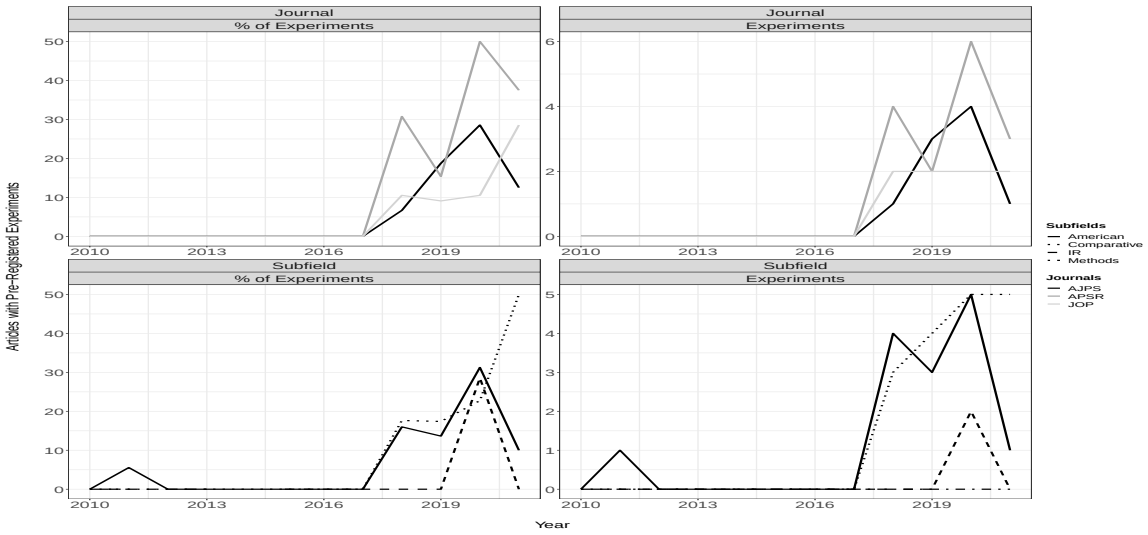
Much of this move toward multi-experiment studies seems a consequence of ignorance on the part of reviewers and negligence on the part of editors. Reviewers frequently ask authors to engage in additional experiments without a clear theoretical basis. A reviewer might say: "You found X, but why didn't you include Z in your treatments?" While there are cases where experimental designs are flawed, the Z in question is usually a curiosity more than an urgent concern. Editors, whose time is increasingly taxed, acquiesce to these requests and leave it to the authors to either (a) pay the high costs of fielding the additional experiments or (b) risk having their work rejected. The norm metastasizes when reviewers and editors begin to hold authors to this new standard of multiple experiments based on precedent; as reviewers and editors see more papers that include 3, 4, 5, or more experiments, they begin to expect multiple experiments. The result is what we observe in our sample data: a majority of papers with multiple experiments.

2.3 Pre-registration

The final norm we consider is pre-registration. *Pre-registration* is the practice of documenting one's analysis plan prior to data collection and posting that document publicly as a means of tying one's hands in the data analysis stage of a project. As Logg and Dorison (2021, 18) explain, "the researcher specifies the design, sample size, hypotheses, and planned statistical analysis of a study or a data set." These *pre-analysis plans* (PAPs) can then be submitted with the manuscript for review, ex-ante, and posted online ex-post.

Figure 4 shows the annual count of pre-registered experiments published over time, by journal and subfield. Across our sample, only 38 (or 9.7%) were pre-registered. Pre-registration of

Figure 4: APSR, AJPS, and JOP Pre-registered Experiments Over Time



Note: Annual counts of pre-registered experiments over time, across subfield and journal.

experiments is a relatively contemporary phenomenon. Pre-registration started to become more common around 2017, but the upward trajectory beyond that point highlights the evolution of the norm. Comparative and American politics are responsible for the initial wave; IR follows after. We observe a decline in pre-registered experiments after 2019. This, of course, reflects the general decline in experiments. The exception is Comparative politics, which accounts for the largest proportion of pre-registered experiments (16.3%). Pre-registration increases from 2017 onward, and stays constant through the end of the sampling window.

Beyond the observational evidence presented in Figure 4, several editorial teams at major journals have taken to encouraging and/or requiring pre-registration of experiments. The JOP will no longer review non-pre-registered experiments and has set explicit requirements for what must be included in PAPs. *International Studies Quarterly* (ISQ) recently announced a similar policy. Other journals do not maintain strict requirements, but promote pre-registration as an informal norm. The APSR requires authors who did pre-register their studies to submit their PAP along with the manuscript. The AJPS is more flexible, only suggesting authors include a PAP if they believe it would be helpful for reviewers.

Much of the policing of this norm happens informally. Individual reviewers will pass judgement on studies that are not pre-registered, even when it is not a requirement of the journal where the piece was submitted. Proponents of pre-registration are lobbying editors and reviewers to enforce stricter pre-registration standards in an effort to further codify the norm (Scoggins and Robertson 2024; Mize and Manago 2022).

The nominal goal of pre-registration is the promotion of transparency (Monogan 2013; Green 2022). Logg and Dorison (2021, 18) describe pre-registration as being part of the "*open science revolution*," a movement in the social sciences over the past decade, calling for greater transparency in how science is conducted. They liken this movement to the scientific revolutions described by Thomas Kuhn. Green (2022), in a similar vein, extols the virtues of pre-registration and notes that the credibility advantage of pre-registration is only conferred when authors stick closely to their original PAP.

There are three problems that the increased transparency associated with pre-registration is ostensibly meant to resolve. First, pre-registration is meant prevent the selective modification and reporting of statistically significant results, "*p-hacking*" (Head et al. 2015, 2). The logic is straightforward. If one is required to describe the models they will estimate in advance, they cannot iteratively revise their models until they arrive at a desired result (Monogan 2013, 24). In this way, pre-registration gives readers greater confidence that reported results are not "cherry-picked from a broader set" of results that were not consistent with the hypotheses (Logg and Dorison 2021, 19). The purpose of the PAP, therefore, is to constrain the "experimental and analytical freedom" of the researcher to prevent such nefarious practices (Hitzig and Stegenga 2020, 24).

The second problem pre-registration is meant to resolve is "*HARKing*," hypothesizing after results are known (Logg and Dorison 2021, 19). The concern is that researchers present exploratory analyses derived from collected data as confirmatory analyses, which tests propositions devised prior to data collection. Such behavior is born out of research incentives. Journals tend not to report null results, and editors tend to favor the publication of novel theoretical findings. These

twin pressures create incentives for scholars to fish for statistically significant effects in their data, developing explanations for the correlation ex-post. If the author is required to publish their theory, hypotheses, and research plan prior to data collection, HARKing is impossible.

Finally, proponents argue that pre-registration reduces publication bias. Publication bias exists because editors and reviewers prefer studies where the results match the hypothesized effects. If journals are willing to guarantee the publication of results from pre-registered studies, the bias towards the publication of statistically significant results is eliminated (Monogan 2013, 23).

Of the norms discussed in this paper, the requirement that *only* experimental studies should be pre-registered is the most ridiculous and repugnant. HARKing and p-hacking are not problems exclusive to experimental studies! In fact, p-hacking and HARKing are harder to achieve in experimental studies. By design, experimental treatments are orthogonal to all the other variables that can be collected in a survey. Including control variables can increase the efficiency of estimates, but subtle correlations cannot be exploited to produce spurious correlations because the variable of interest in an experiment is the *randomized* treatment (Mutz, Pemantle and Pham 2019). HARKing in an experimental study is complicated for a similar reason. With an experiment, the nature of the independent variable is determined ex-ante. While we won't go so far to say that p-hacking and HARKing are impossible in experimental research, we note that both practices are much easier to envisage with observational surveys like the ANES, or large macro-panel data sets that can included hundreds of country-year variables. If p-hacking and HARKing are possible with observational studies, why aren't authors of observational studies required to pre-register?

The scientific poverty of these special standards for experimental research is underlined by the fact that they are completely out of step with the literature on pre-registration. Proponents of pre-registration do not argue that pre-registration is *only* valuable for experimental studies. Rather, they argue this norm should be adopted broadly, including observational work. In his example of pre-registration, for instance, Monogan (2013) registers an observational study that hypothesizes an effect prior to an election. Coffman and Niederle (2015, 87) find that PAPs are only helpful

in reducing p-hacking in situations where a study is "the only attempt of a hypothesis," which is uncommon in social sciences. It is for these reasons that Samii (2016) proposes PAPs as a solution to these problems with observational studies, notably *not* restricting PAPs to experimental research.

To be clear, pre-registration is not the problem. In a world where all research – observational and experimental – was subject to the same pre-registration requirements and/or journals were willing to accept articles for publication based on the theory and PAP, the transparency benefits could be realized. Such a system would reduce publication bias; pieces with PAPs that pass review would be published regardless of whether the results were statistically significant. This is the state of affairs that most early proponents of pre-registration had in mind (e.g., Monogan 2013).

The problem with pre-registration of experimental studies is that the norm, as it was originally intended, has not been fully embraced by journal editors or reviewers. Only experimentalists are required to pre-register. Yet, and with relatively few exceptions, their results will only be published if the results are statistically significant and consistent with their theoretical argument(s). This perverse state of affairs imposes all the costs of pre-registration but provides none of the benefits. The norm maintains the bias toward the publication of statistically significant results while making experimental research more time consuming.

2.4 Philosophical Inconsistencies

The preceding discussion outlines the scientific poverty of each of the new gatekeeping practices in isolation. But the absurdity of these norms becomes all the more obvious when they are considered in concert with one another.

There is a philosophical tension between the logics impelling people to conduct multiple experiments and those that drive pre-registration. The standard approach to experimental design is deductive: a theory and hypotheses are posed a-priori, and the experiment is designed to evaluate the theory and hypotheses. Fielding an auxiliary experiment based on the results from a previous experiment is an inductive exercise. Much of the rationale behind pre-registration is wrapped up in

the idea that this kind of knowledge production is fatally flawed. Either the entire scope of a project must be planned in advance or it is reasonable to expect analysts to conduct multiple experiments to demonstrate the robustness of their theoretical expectations and/or probe the causal mechanisms revealed by the results, but it cannot be both!

The logics that impel pre-registration also collapse in on themselves in practice. Requiring experimental studies to be pre-registered encourages HARKing. Many scholars now pre-test their experiments on convenience samples prior to formal pre-registration. If the experiment fails to yield the desired/expected outcome, either the experiment or the theory and hypotheses are adapted to the experimental findings. If the pre-test isn't pre-registered, it is impossible to discern whether the pre-test was used to evaluate the quality of the experimental treatment or whether the theory and analysis were adapted ex-post. One could, for example, conduct an experiment and use exploratory statistical techniques – like factor analysis or extreme bounds analysis – to identify potential moderators for treatments. Having found a (potentially robust) moderator, an explanation can then be developed and passed of as *Capital-T-Theory* ex-post. As Pham and Oh (2020, 167) note in their scathing critique, such practices mean that pre-registration merely provides the "illusion of transparency." The extent of this practice is difficult to ascertain because there are no established norms regarding the reporting of these initial pre-tests nor of including them in PAPs. This practice also exacerbates the time and treasure costs of experimental research.

These tensions highlight the reality behind the evolution these new "*standards*" for experimental research: the practical implications, competing pressures, and theoretical bases for these norms have not been thoughtfully considered. As scholars endeavor to satisfy one norm, they are forced to engage in practices the undermine the rationale of others. Which norms are most important to a particular editorial team or group of reviewers is impossible to know. An editorial team might require that all experimental studies be pre-registered, and a reviewer for the same journal may ask the authors to conduct an auxiliary analysis or field an auxiliary study to indulge some half-baked curiosity. These kinds of contradictions are not just difficult to manage; they carry real costs. Not just for individual scholars, but for the discipline as a whole.

3 The Cost: Reinforcing Elitism

This section lays out the the second half of our argument. Not only are the new gatekeeping norms intellectually bankrupt, these new norms are causing experimental research to become more elitist. As experiments get more expensive – in terms of time and treasure – it gets harder for scholars without resources to publish experimental research.

Like any economic enterprise, knowledge production involves the conversion of scarce factors into outputs. For political scientists, the primary factors are time and money; the outputs are publications. These factors of production are substitutable, to a point. If one has more money, one can pay for more experiments, pay graduate students to collect and code data, and/or pay for large nationally representative samples. If one has more time, one can go classroom-to-classroom enlisting undergraduates, submit PAPs, and submit multiple amendments. But not all producers of knowledge have the same factor endowments.

Scholars at elite institutions have access to more time and money. Universities with larger endowments are more likely to offer dedicated research budgets. Indeed, Farris, Key and Sumner (2023) find that new assistant professors (APs) at private doctoral granting institutions more frequently report receiving research money and course releases as part of their start up packages. The amount of research funds reported by APs at private R1s was staggering in comparison to other institutions. As Farris, Key and Sumner (2023, 79) note, "private doctoral-granting programs have the largest modal package and the one that is most obviously geared toward facilitating research. In addition to what public institutions offer, private institutions also tend to offer research assistants, book workshops, and research leave."

More elite universities also tend to be the schools that have large graduate programs. Graduate students serving as research assistants (RAs) reduce the time required by faculty to complete a project because RAs can collect data, code data, and help compose manuscripts. Graduate students serving as teaching assistants (TAs) also increase the amount of time scholars have for research. If one is not required to grade homework, proctor exams, or lead labs and discussion sections, one

has more time for research. RAs and TAs are time multipliers, and they exist at the institutions where research faculty tend to have the lowest teaching loads. The standard teaching load at most tier 1 research universities is 2 courses per semester; faculty with course releases have even lighter loads. To be sure, R1 faculty should not be ashamed of these resources, we enjoy them ourselves, but they do create inequalities.

We argue that the costs associated with the new norms make experimental research more elitist. They tax scarce resources. These taxes mean that experimental scholars at elite institutions have a disproportionate advantage over experimental scholars from other institutions. Scholars at elite research institutions will always be able to publish more than scholars at teaching institutions or liberal arts colleges. The academic contracts and publication requirements for different colleges and universities reflect these different priorities. The reason that experimental scholars at elite institutions have a disproportionate advantage is that the new norms exacerbate the resource divisions. They make time and money more important for experimental research than they are for observational research.

Nationally representative samples, targeted field samples, and elite samples are more expensive than convenience samples. Nationally representative samples become more expensive as sample sizes get larger and as researchers designate additional strata for representativeness. The bias for these non-convenience samples means scholars conducting theory application research must muster resources to pay survey firms. For senior scholars at elite institutions, this requirement is an inconvenience. For everyone else, this resource requirement represents a major barrier to entry. Faculty from less endowed universities are either forced to do less experimental research or avoid experiments altogether.

Fielding multiple experiments not only increases financial costs, it costs time. Experiments must be designed, submitted to institutional review, and programmed before recruitment. Student samples can take months if they are recruited piecemeal. Field samples and elite samples often have

high dropout rates. Even online samples can take time to collect, depending on the platform. And all these time costs are incurred before data analysis.

The primary cost of pre-registration is time. While posting one's PAP online does not cost money, it does increase the length of the research process. Ofosu and Posner (2023, 7) note that "PAP users' surveyed reported devoting a week or more to writing the PAP for a typical project, with 32% reporting spending an average of two to four weeks and 26% reporting spending more than a month. It is perhaps not surprising, then, that 34% of researchers said that writing a PAP delayed their project's implementation."

These costs are onerous in isolation; they are crushing in conjunction. The most obvious problem arises in terms of financial cost. If (the collective) we, are going to dispense entirely with convenience samples, and we are going to require 3, 4, 5 experiments per study, experimental research will be cost prohibitive for anyone who doesn't have access to a lavish research kitty. An experimental study that employs multiple samples can easily reach a stifling sum. For example, if 5 experiments are required for a study to be considered sufficiently rigorous, and each nationally representative quota sample cost $6,000 to $10,000, the study can easily exceed $40,000 – a sum that is out of reach for most researchers.

A skeptical reader might argue that one does not need a nationally representative sample for every experiment in one's study. What is the scientific principle to justify using a nationally representative sample for one experiment but convenience samples for the others? There either is a coherent inferential target, or their isn't.

Pre-registering multiple experiments takes more time. If there are any deviations from the PAP(s), cataloging these deviations takes even more time. Then, debating the justifications for these deviations with editors and reviewers takes even more time.

These costs overlap and accentuate one another. There is a class of scholars that can continue to afford to do experimental research and a (larger) class of scholars who cannot. Because the costs we highlight only apply to experiments, there is a natural way to test whether these costs materially

affect political science. The observational studies in our sample provide the counterfactual. If our argument is incorrect, and these new norms do not increase the relative costs of experimental work compared to non-experimental work, we should observe the same patterns of elite representiveness across research designs. If our argument is correct, we should observe a higher representation of scholars from elite institutions in experimental research. This contrast informs our hypothesis:

> H_E **-** ***Elitism:*** The proportion of scholars from elite institutions publishing experimental research should be higher than the proportion of scholars from elite institutions publishing non-experimental research.

4 Research Design

We evaluate our hypothesis by comparing the institutions of authors of experimental studies to the institutions of authors of non-experimental studies. We use all of the articles published in the AJPS, APSR, and JOP between 2010 and 2021. There were 2,229 articles, 388 include at least one experiment.

The dependent variable is the nature of the study: experimental or non-experimental. A study is classified as experimental if it includes a randomized experiment. All other studies are classified as non-experimental. The non-experimental studies provide the counterfactual; if the norms are benign, there should be no difference in institutional profiles.

The variables of interest are our measures of the authors' institutions. We identify each author's institutional affiliation at time of publication. We also collected data on graduate training from authors' curriculum vitaes (CVs). We coded the prestige of the x institutions using the US News and World Report Best Global Universities Rankings.[4] This ranking system includes 2,165 schools across the globe.

[4]Available at: https://www.usnews.com/education/best-global-universities/rankings.

There are valid reasons to be skeptical of this approach to measuring elitism. Diermeier (2023) cites problems with the way US News incorporates information about student aid and career attainment. Knox (2023) highlights how US News's decision to increase the weight on access and social mobility sparked controversy, especially among selective, private colleges.

Despite the conflict these rankings engender, they are fit to our purpose. The methodology incorporates elements relevant to elitism. Faculty resources, financial resources, and elite perception are three of the most important criteria in the ranking system. The largest component is the elite perception metric (20%), which is based on national surveys of presidents, provosts, and deans of admissions. In effect, the US News Rankings are an elite survey of the perceived quality of institutions, with objective weights added in for graduation rates and financial resources.

The rankings also seems to have face validity. The top colleges are Harvard (1), MIT (2), and Stanford (3). It seems reasonable to say that these schools are more *elite* than the University of Illinois (100), the University of Alberta (150), or UMASS Amherst (175). To be sure, all are fine institutions but the faculty salaries and tuition rates are much higher for the first three compared to the universities with lower rankings (higher values).

To the extent that there are problems with the ranking system, they should not have important implications for our analysis.[5] Most of the complaints about the rankings are quibbles over year-to-year adjustments. For example, Vanderbilt fell from 13th to 18th in the 2024 rankings; this prompted a public message by the Vanderbilt chancellor excoriating US News for their "lack of rigor or competence" (Knox 2023). These sorts of a adjustments may be irritating to those who see the stars of their home institutions and alma maters fall ever so slightly, but it is not likely to induce significant bias in our analysis. The only way this measurement error could induce bias is if the institutions of experimental political scientists were being systematically overrated or underrated. This seems unlikely. What is more likely is that the measure is noisy, in both directions (some

[5]We could have used the Carnegie Classifications, but this omits international universities and would be subject to the same potential measurement error.

institutions are ranked too high; some too low). If anything, this inflates our standard errors, making it hard to reject the null hypothesis.

We also code a series of control variables. In addition to institution rank, we code the race and sex of each author by going to their faculty and/or personal web pages. We also include a dummy variable for whether an author is tenured (at the Associate rank or equivalent and above).

The collaborative nature of political science complicates our analysis. The unit of measurement for the outcome variable (experimental vs. non-experimental) is article, but the ranking of institutions happens at the author-level. We navigate the potential hazards of this incongruity by relying on the robustness approach to inference: we apply a variety of different measurement strategies and report the results from each analysis to demonstrate that our results are not an artifact of our measurement choices or modeling decisions (Neumayer and Plümper 2017).

The outcome variable is binary, so we estimate a series of logistic (logit) regression models. The first three (*Naive*) models ignore the differences in the unit of analysis between the outcome and the variables of interest. *Naive 1* includes the institution and PhD rank of the authors, *Naive 2* includes information about each author, and *Naive 3* includes dummy variables to account for differences across subfields. The fourth model is a multi-level regression (*Mixed*) that allows the authors to be nested within their individual studies; this accounts for the non-independence of the individual authors. The final three models are simple logistic regressions where we explore alternative ways of dealing with the unit-of-analysis-incongruence. We estimate one regression model where we only include the first author of each study (*First*). We repeat this analysis for the second author (*Second*), and we conduct an analysis where the Institution and PhD ranks are averaged across the authors of each study (*Average*).

5 Results

Our results are presented in Table 1. The independent variables are listed in the first column; the logistic regression coefficients and standard errors are presented in columns 2 through 8. The bottom panel presents the sample size for each regression (N) and two measures of model fit (Akiake Information Criteria (AIC) and Bayesian Information Criteria (BIC)). The multi-level (*Mixed*) model has two additional elements in the bottom panel: the number of groups (articles) in the sample and the estimated variance between the groups ($\sigma_{Article}$).

The results presented in Table 1 are consistent with our elitism hypothesis. As the rank of the institution where the authors are located increases (becomes less elite), the odds of the authors used a randomized experiment decreases. This negative effect is consistent across measures of elitism (current institution rank and PhD rank) and robust to specification. The signs are negative and statistically significant across almost all models. The only exception is the coefficient on institution rank in the first author sample; the sign does not change, but the effect is not statistically significant.

Some additional effort is required to discern the substantive effects. The coefficients have an oblique metric, log-odds, and a one unit change in rank is a relatively small quantum given the variation in the data. To facilitate interpretation, we exponentiate the coefficients to generate the odds ratios and calculate the % change in odds for a one unit increase: $(e^{\hat{\beta}-1\times 100})$. We then calculate the % change in odds over different percentile changes in the variables of interest.

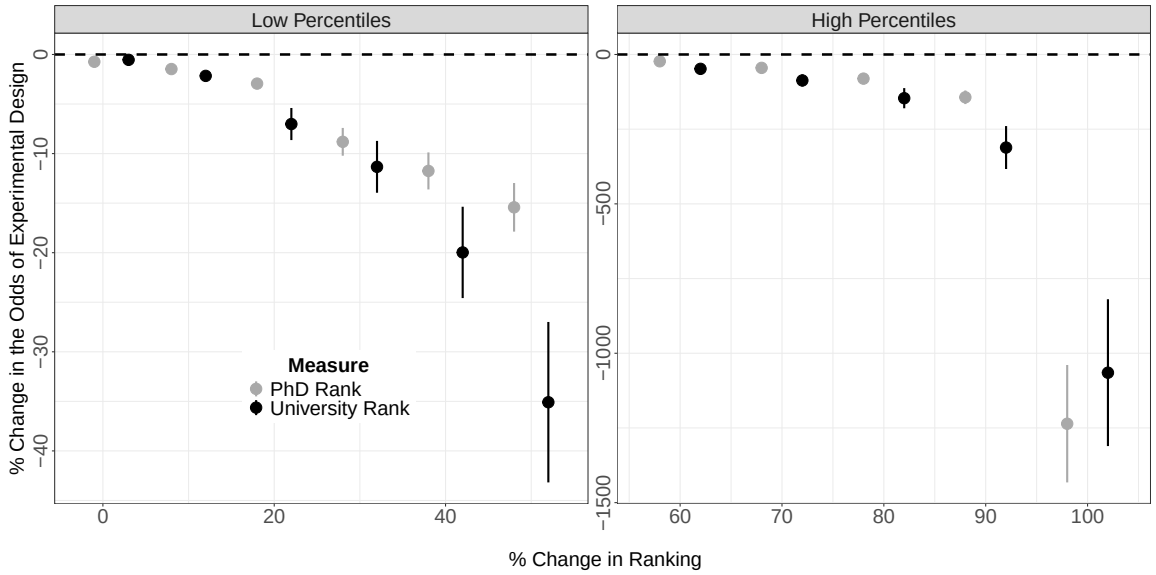
The substantive relationship between elitism and experimental design are shown in Figure 5. The effects beyond the median are so large that we had to split the odds ratios into two plots: low percentiles (left) and high percentiles (right). The points are the % change in the odds of experimental research; the vertical lines are 90% confidence intervals. The x-axes shows the % change from the minimum ranking (1 being the most elite school) to the lowest ranked institutions. We use the results from the mixed model to calculate these quantities because it has the best fit – according to the AIC and BIC – with the most data (N = 3,812) and the full complement of controls.

Table 1: University Rank and Experimental Research

	Naive 1	Naive 2	Naive 3	Mixed	First	Second	Average
Institution Rank	-0.000^{**}	-0.000^{**}	-0.000^{*}	-0.004^{***}	-0.000	-0.001^{**}	-0.001^{*}
	(0.000)	(0.000)	(0.000)	(0.001)	(0.000)	(0.000)	(0.000)
Phd Rank	-0.001^{***}	-0.001^{***}	-0.001^{***}	-0.006^{***}	-0.001^{*}	-0.002^{**}	-0.001^{**}
	(0.000)	(0.000)	(0.000)	(0.002)	(0.001)	(0.001)	(0.001)
Tenured		0.261^{***}	0.271^{***}	0.391	0.268^{**}	0.370^{**}	
		(0.081)	(0.083)	(0.313)	(0.121)	(0.146)	
Sex		0.026	0.093	-0.251	0.145	0.196	
		(0.098)	(0.100)	(0.407)	(0.142)	(0.179)	
Black		0.919^{***}	0.834^{**}	1.339	0.596	0.809	
		(0.324)	(0.331)	(1.210)	(0.513)	(0.584)	
Latino		-0.342	-0.255	-5.615^{*}	-0.446	-0.608	
		(0.231)	(0.234)	(2.997)	(0.365)	(0.489)	
Asian		0.346^{**}	0.366^{***}	-2.105^{**}	0.398^{**}	0.250	
		(0.136)	(0.140)	(0.975)	(0.196)	(0.261)	
Comparative			-0.466^{***}	-1.654^{***}	-0.559^{***}	-0.411^{***}	
			(0.090)	(0.436)	(0.132)	(0.158)	
International Relations			-0.249^{*}	-1.858^{**}	-0.531^{***}	-0.102	
			(0.134)	(0.789)	(0.200)	(0.229)	
Methodology			-0.897^{***}	-34.629^{***}	-0.910^{*}	-0.770	
			(0.330)	(8.578)	(0.545)	(0.550)	
Theory			-16.494	-5.010	-17.462	-15.548	
			(248.981)	(3.754)	(436.251)	(473.306)	
Constant	-1.227^{***}	-1.395^{***}	-1.124^{***}	-6.927^{***}	-1.189^{***}	-1.104^{***}	-1.350^{***}
	(0.050)	(0.073)	(0.084)	(0.347)	(0.119)	(0.148)	(0.075)
AIC	3927.289	3853.731	3682.268	1548.026	1760.226	1231.766	1989.280
BIC	3946.093	3903.746	3757.218	1629.222	1827.443	1293.016	2006.271
N	3898	3835	3812	3812	2001	1217	2129
Groups				2102			
σ_{Article}				157.580			

Logistic regression coefficients, standard errors in parentheses, $^{***}p < 0.01$, $^{**}p < 0.05$, $^{*}p < 0.1$.

Figure 5: % Change in University Rank and the Odds of Experimental Research



The effects presented in Figure 5 highlight the stark differences between elite institutions and non-elite institutions. An author at a university near the median of our sample (65) was 35% less likely to publish experimental research in the big three than someone at the most elite institution. That number climbs to an astonishing -1065% for people at the lowest ranked school (1973). Authors at elite universities are more than 10 times more likely to publish experimental articles. The effects are similar for PhD Rank: -15% for the median (21) and -1,235% for the maximum (1,683).

These results highlight the costs of the gatekeeping norms. Not only are these "*standards*" scientifically bankrupt, they re-enforce divisions between the haves and the have-nots in the academy. The magnitudes of the effects demonstrate how these norms are reshaping the discipline, to the benefit of some but to the detriment of most.

6 Auxiliary Analyses

We conducted a series of auxiliary analyses to probe the causal mechanisms behind the trend toward elitism in experimental political science. We argued that the new norms are making experiments more elitist because these norms impose significant costs. If these costs are real, they should have

observable implications beyond the main results. Due to space constraints, we present these results in the appendix and summarize the findings here.

We estimated a series of multinomial probit models to look at the relationship between elitism and sample choice. The outcome variable is a categorical outcome that classifies the sample as nationally representative, convenience, or other expensive samples (field experiments, elite samples, and lab samples). The variables of interest are the same measures of elite institutions used before, but we use the minimum and median ranks of the author(s) on a study. If a team of authors includes someone with a big research budget, that person can pay to field the expensive survey. We find evidence consistent with our elitism hypothesis. The results are presented in Section 2 of the appendix. The coefficient on minimum PhD rank is negative and significant. The scholar with the most elite profile is likely to have access to the most resources. As the minimum rank of a research team increases, the probability that the research team will use a nationally representative sample decreases. This effect depends on the median rank of the group, remaining negative and statistically significant for all groups where the median rank is in the top 200 schools.

We test our argument about multiple-experiments-per-study using a series of Poisson regressions. The outcome is the count of experiments in each paper. The variables of interest and controls are the same. We again find evidence consistent with our elitism argument. The coefficients for both measures of elitism are negative and statistically significant across all model specifications. Authors who graduated from and/or are working at the most elite institutions have more experiments in their papers. This, more then the other auxiliary analysis, highlights the impact of resource disparities on experimental research.

An analysis pre-registration war harder to conduct because so few studies in the sample were pre-registered, but we find evidence consistent with our argument. We compared the percentiles of Institution and PhD Ranks across pre-registered and non-pre-registered studies. The ranks of authors of pre-registered studies are lower (more elite) across all percentiles beyond the minimum. This is true for median and minimum institutional ranks. We estimated a series of logistic regressions

to subject these patterns to more rigorous testing. The effects of home institution are negative and significant; the PhD effects are negative.

These auxiliary analyses lend more confidence to our results and our arguments. While none of these auxiliary analyses are perfect, they all point to the same conclusion: faculty at elite institutions have an unfair advantage when it comes to experimental research. This is not because their ideas are better or their designs are more refined, but because they have access to more resources. While these disparities between elite and non-elite institutions have always existed, they are becoming more acute for scholars interested in conducting experiments because these new norms are only applied to experimental research.

7 Conclusions and a Call to Action: Rethinking the Norms of Experimental Research

A scientific community should have scientific standards for research. The rules for defining what constitutes good research should be based on sound reasoning and robust evidence. The following of such standards ensures the credibility, validity, and integrity of the research. When the standards that govern a scientific enterprise are not well thought-out, the discipline suffers.

This is the state of experimental political science today. Getting experimental research published has less to do with the quality of ideas and strength of designs, than with conforming to a set of norms for what constitutes "good" research. Not only are these norms scientifically bankrupt, we provide empirical evidence showing the policing of these norms unfairly tilts the playing field. Preregistration, the bias against convenience samples, and the arbitrary pressure to include multiple experiments per study all serve to privilege the work of scholars from elite institutions.

These norms are reshaping the landscape of experimental research and will have long-term consequences if they are not addressed. These costs are not just inconvenient; they threaten to undermine the future of experimental research through intellectual asphyxiation. A graduate stu-

dent may have 2 to 3 years after exams to publish and complete their dissertations. An Assistant Professor only has 5 years before they submit tenure files. These are the formative years of scholarly development. If young scholars don't have the time or money to do experimental research, they will either choose not to or fail out of the discipline. Thirty years from now we will be left with a geriatric group of well-heeled scholars who may find few people interested in reading their experimental research because most of the discipline will have elected to invest in alternative methodological specializations.

We conclude with a call to action. The only way to reform problematic norms is to develop new norms. Steps must be taken by editors, reviewers, and authors to mitigate the growing elitism in experimental research. We propose the following.

With respect to the use and misuse of nationally representative and other non-convenience samples, reviewers and editors should require authors to explicitly define their inferential targets and explain why their sampling strategies are necessary and/or sufficient to apprehend these targets. The bias against convenience samples persists because of misunderstandings about role of sampling in different kinds of research. If a coherent inferential target exists, as in effects application research, the author should be able to explain why a population weighted sample is required for inference. If there is no coherent inferential target, as in theory application research, a convenience sample will suffice. We do not go so far as to say that work lacking a sound basis for a costly sampling strategy should be rejected, but authors should be forced to acknowledge if and when their sampling methods do not add value. Requiring this sort of discussion would help resolve confusion and provide a bulwark against the propagation of this norm.

With respect to the pressure to include multiple experiments per study, reviewers and editors should require authors to offer a unique theoretical basis for each experimental design. If a hypothesis can be tested with a single experiment, only one experiment should be included in the study. If there is a meaningful theoretical reason to suspect that a treatment effect will not be consistent across some factor, that difference should be theorized, hypothesized, and incorporated into the

initial design. This requires forbearance on the part of reviewers and editors as well. Reviewers should not ask authors to field additional surveys or conduct auxiliary analyses where treatments are conditioned on observables, just to indulge half-baked curiosities about other factors or alternative measurements. The bar for these kinds of analyses should be as high as the bar for any other theory that would be the basis for an article submitted to a journal. If a reviewer believes they have a conjecture about a treatment effect that warrants additional study, they should be able to lay out that theory and field a survey of their own. If the theoretical basis for such a conjecture could not be the basis of a separate study, then it should not place demands on time and resources. Editors should explicitly intervene in these cases; authors should be told that additional studies are not required, and it should be made clear to reviewers that these kinds of demands are inappropriate.

One exception is deliberately sampling for heterogeneity. Authors should be allowed to make the case that conducting an experiment at different times, with different groups, in different settings is sufficient to establish the external validity of their treatment effects. But authors should still be required to explain their sampling decisions and explain why sampling for heterogeneity is the best approach to establish the external validity.

Finally, with respect to pre-registration, the arbitrary requirement that only experimental studies need to be pre-registered needs to die. It's all or nothing. If we, as a community, want to fully embrace transparency to avoid the iniquities of HARKing and p-hacking, so be it. But all studies where one *could* hypothesize after results are known or *could* manipulate regression analyses to produce statistically significant results need to be pre-registered, not just experiments. If we are not ready to incur the attendant costs, these requirements should be discarded.

We are not so naive to expect that these norms will be spontaneously abandoned. As with all barriers to entry, there are people invested in these gatekeeping norms; they will be reticent to let them go. So what can be done?

First and foremost, we should refuse to review for journals that have formalized these norms. Journals that impose arbitrary standards for experimental work and willfully ignore the negative consequences of those norms are not entitled to our time.

We also need to be more thoughtful about the standards we develop for research and be more judicious about how those standards are applied. Editorial teams should be explicit about all the standards and norms they intend to enforce and the rationale for these norms should be open to public scrutiny. There need to be opportunities for scholars negatively affected by the application of these norms to provide feedback, and editorial teams should be willing to revise their standards if their ratification or application cannot be defended. Importantly, the standards we develop for research should apply to all research designs equally.

References

Avdan, Nazli and Clayton Webb. 2019. "Not in my back yard: Public perceptions and terrorism." *Political Research Quarterly* 72(1):90–103.

Baldassarri, Delia and Maria Abascal. 2017. "Field experiments across the social sciences." *Annual review of sociology* 43:41–73.

Blankshain, Jessica, David Glick and Danielle Lupton. 2023. "War Metaphors (What Are They Good For?): Militarized Rhetoric and Attitudes Toward Essential Workers During the Covid-19 Pandemic." *American Politics Research* 51(2):161–173.

Blankshain, Jessica, Lindsay Cohn and Danielle Lupton. 2024. "I'm from the Government, and I'm Here to Help: Public Perceptions of Coercive State Power." *American Political Science Review* forthcoming:1–18.

Broockman, David and Donald Green. 2014. "Do online advertisements increase political candidates' name recognition or favorability? Evidence from randomized field experiments." *Political Behavior* 36:263–289.

Bulut, Alper and T. Murat Yildirim. 2023. "Elite influence on attitudes about gender egalitarianism: Evidence from a population-based survey experiment." *Political Behavior* 45(2):659–678.

Calder, Bobby, Lynn Phillips and Alice Tybout. 1981. "Designing research for application." *Journal of Consumer Research* 8(2):197–207.

Coffman, Lucas and Muriel Niederle. 2015. "Pre-analysis Plans Have Limited Upside, Especially Where Replications Are Feasible." *Journal of Economic Perspectives* 29(3):81–98.

Cook, Thomas and Donald Campbell. 1979. *Quasi-Experimentation: Design and ANalysis Issues for Field Settings*. Boston, MA: Houghton MIfflin Company.

Diermeier, Daniel. 2023. "Why the new "U.S. News" rankings are flawed (opinion) — insidehighered.com." `https://www.insidehighered.com/opinion/views/2023/10/09/why-new-us-news-rankings-are-flawed-opinion`. [Accessed 16-08-2024].

Dietrich, Simone, Heidi Hardt and Haley Swedlund. 2021. "How to make elite experiments work in International Relations." *European Journal of International Relations* 27(2):596–621.

Farris, Emily, Ellen Key and Jane Sumner. 2023. ""Wow, I Didn't Know These Options Existed": Understanding Tenure-Track Start-Up Packages." *PS: Political Science & Politics* 56(1):75–81.

Flavin, Patrick and Michael Hartney. 2017. "Racial Inequality in Democratic Accountability: Evidence from Retrospective Voting in Local Elections." *American Journal of Political Science* 61(3):684–697.

Fossati, Diego, Burhanuddin Muhtadi and Eve Warburton. 2022. "Why democrats abandon democracy: Evidence from four survey experiments." *Party Politics* 28(3):554–566.

Gerber, Alan. 2011. "Field experiments in political science." *Handbook of experimental political science* pp. 115–40.

Gerber, Alan and Donald Green. 2012. *Field Experiments: design, analysis, and interpretation*. W.W. Norton.

Gerber, Alan S and Donald P Green. 2000. "The effects of canvassing, telephone calls, and direct mail on voter turnout: A field experiment." *American political science review* 94(3):653–663.

Gottlieb, Jessica. 2017. "Explaining variation in broker strategies: A lab-in-the-field experiment in Senegal." *Comparative Political Studies* 50(11):1556–1592.

Green, Donald. 2022. *Social Science Experiments: A Hands-On Introduction*. Cambridge University Press.

Green, Donald, Dylan Groves, Constantine Manda, Beatrice Montano and Bardia Rahmani. 2020. "A radio drama's effects on attitudes toward early and forced marriage: Results from a field experiment in Rural Tanzania." *Comparative Political Studies* p. 00104140221139385.

Grose, Christian, Neil Malhotra and Robert Parks Van Houweling. 2015. "Explaining Explanations: How Legislators Explain their Policy Positions and How Citizens React." *American Journal of Political Science* 59(3):724–743.

Grose, Christian R. 2021. "Experiments, political elites, and political institutions." *Advances in experimental political science* pp. 149–64.

Guisinger, Alexandra and Elizabeth Saunders. 2017. "Mapping the boundaries of elite cues: How elites shape mass opinion across international issues." *International Studies Quarterly* 61(2):425–441.

Hafner-Burton, Emilie, Brad LeVeck and David Victor. 2016. "How Activists Perceive the Utility of International Law." *The Journal of Politics* 78(1):167–180.

Head, Megan, Luke Holman, Rob Lanfear, Andrew Kahn and Michael Jennions. 2015. "The extent and consequences of p-hacking in science." *PLoS biology* 13(3):e1002106.

Hitzig, Zoe and Jacob Stegenga. 2020. "The Problem of New Evidence: P-Hacking and Pre-Analysis Plans." *Diametros* 17(66):10–33.

Huang, Haifeng. 2015. "International knowledge and domestic evaluations in a changing society: The case of China." *American Political Science Review* 109(3):613–634.

Ichino, Nahomi and Matthias Schündeln. 2012. "Deterring or displacing electoral irregularities? Spillover effects of observers in a randomized field experiment in Ghana." *The Journal of Politics* 74(1):292–307.

Jost, John, Delia Baldassarri and James Druckman. 2022. "Cognitive–motivational mechanisms of political polarization in social-communicative contexts." *Nature Reviews Psychology* 1(10):560–576.

Kellstedt, Paul and Guy Whitten. 2018. *The fundamentals of political science research*. Cambridge University Press.

Kertzer, Joshua and Thomas Zeitzoff. 2017. "A bottom-up theory of public opinion about foreign policy." *American Journal of Political Science* 61(3):543–558.

Knox, Liam. 2023. ""U.S. News" rankings changes spur complaints and apologies — insidehighered.com." `https://www.insidehighered.com/news/admissions/traditional-age/2023/09/22/us-news-rankings-changes-spur-complaints-and-apologies`. [Accessed 16-08-2024].

Lee, Tori. 2023. "Field Experiments Explained." *UChicago News* .

Levendusky, Matthew and Michael Horowitz. 2012. "When backing down is the right decision: Partisanship, new information, and audience costs." *The Journal of Politics* 74(2):323–338.

Logg, Jennifer and Charles Dorison. 2021. "Pre-registration: Weighing costs and benefits for researchers." *Organizational Behavior and Human Decision Processes* 167:18–27.

Lupton, Danielle. 2019. "The External Validity of College Student Subject Pools in Experimental Research: A Cross-Sample Comparison of Treatment Effect Heterogeneity." *Political Analysis* 27(1):90–97.

Lupton, Danielle and Clayton Webb. 2022. "Wither Elites? The Role of Elite Credibility and Knowledge in Public Perceptions of Foreign Policy." *International Studies Quarterly* 66(3):sqac057.

McDermott, Rose. 2002a. "Experimental methods in political science." *Annual Review of Political Science* 5(1):31–61.

Mize, Trenton and Bianca Manago. 2022. "The past, present, and future of experimental methods in the social sciences." *Social Science Research* 108:102799.

Monogan, James. 2013. "A case for registering studies of political outcomes: An application in the 2010 House elections." *Political Analysis* 21(1):21–37.

Mullinix, Kevin, Thomas Leeper, James Druckman and Jeremy Freese. 2015. "The generalizability of surve y experiments." *Journal of Experimental Political Science* 2(2):109–138.

Mutz, Diana C, Robin Pemantle and Philip Pham. 2019. "The perils of balance testing in experimental design: Messy analyses of clean data." *The American Statistician* 73(1):32–42.

Neumayer, Eric and Thomas Plümper. 2017. *Robustness tests for quantitative research*. Cambridge University Press.

Ofosu, George and Daniel Posner. 2023. "Pre-Analysis Plans: An Early Stocktaking." *Perspectives on Politics* 21(1):174–190.

Pham, Michel Tuan and Travis Tae Oh. 2020. "Preregistration Is Neither Sufficient nor Necessary for Good Science." *Journal of Consumer Psychology* 31(1):163–176.

Renshon, Jonathan, Allan Dafoe and Paul Huth. 2018. "Leader influence and reputation formation in world politics." *American Journal of Political Science* 62(2):325–339.

Saha, Sparsha and Ana Catalano Weeks. 2022. "Ambitious women: Gender and voter perceptions of candidate ambition." *Political Behavior* 44(2):779–805.

Samii, Cyrus. 2016. "Causal empiricism in quantitative research." *The Journal of Politics* 78(3):941–955.

Sarstedt, Marko, Paul Bengart, Abdel Monim Shaltoni and Sebastian Lehmann. 2018. "The use of sampling methods in advertising research: A gap between theory and practice." *International Journal of Advertising* 37(4):650–663.

Scoggins, Bermond and Matthew Robertson. 2024. "Measuring transparency in the social sciences: political science and international relations." *Royal Society Open Science* 11:240313.

Tomz, Michael and Jessica Weeks. 2020. "Public opinion and foreign electoral intervention." *American Political Science Review* 114(3):856–873.

Yang, Keming and Ahmad Banamah. 2014. "Quota sampling as an alternative to probability sampling? An experimental study." *Sociological Research Online* 19(1):56–66.

Zelizer, Adam. 2019. "Is Position-Taking Contagious? Evidence of Cue-Taking from Two Field Experiments in a State Legislature." *American Political Science Review* 113(2):340–352.

Zhu, Boliang and Weiyi Shi. 2019. "Greasing the Wheels of Commerce? Corruption and Foreign Investment." *The Journal of Politics* 81(4):1311–1327.