Of Bias and Blind Selection: Pre-registration and Results-Free Review in Observational and Qualitative Research[1]

Alan M. Jacobs

Department of Political Science
University of British Columbia
C425-1866 Main Mall
Vancouver, BC V6T 1Z1, Canada
alan.jacobs@ubc.ca

January 2019

[1] Prepared for publication in Elman, Colin, John Gerring, and James Mahony (eds.), *The Production of Knowledge: Enhancing Progress in Social Science*, forthcoming with Cambridge University Press. Fabio Resmini, Pascal Doray-Demers, and Alexander Held, provided excellent research assistance. This paper has benefited from helpful comments on previous drafts by Henry Brady, John Gerring, Alexandra Hartman, Macartan Humphreys, Eddy Malesky, and Tom Pepinsky and from participants in the Political Science Department Research Seminar at University College London.

Abstract

There is good reason to worry that the body of empirical findings appearing in social scientific journals are heavily skewed in favor of positive or "strong" results. Many scholars have argued that publication bias can be effectively countered via mechanisms that force empirical tests to be selected in a manner blind to their results. Particular attention has focused on two forms of "blind selection": study pre-registration and results-blind review. In the case of pre-registration, researchers choose their tests blind to those tests' results; in the case of results-blind review, reviewers and editors consider manuscripts that describe study procedures but are silent about outcomes. Advocates of pre-registration have, however, devoted remarkably little attention to thinking through whether or how these mechanisms might be used outside the narrow worlds of experimental and prospective research. This paper examines how principles of blind-selection might operate for observational and retrospective research, both qualitative and quantitative. The paper first shows, via a systematic review of published qualitative and mixed-method work, that publication and analysis bias are as great a threat to the validity of qualitative as to quantitative published inferences. Turning then to potential solutions, the paper identifies three features of a research design on which the use of pre-registration or results-blind review depends: the prior unobservability of the evidence, the independence of new evidence from old, and the precision of test-specification. Through a review of published observational studies, the paper shows that much observational – and, especially, qualitative – research displays features that would make for credible pre-registration and results-blind review. The paper concludes that pre-registration and results-blind review have the potential to enhance the validity of confirmatory research across a broad range of empirical methods, while also raising the profile of exploratory analysis by making it harder for researchers to cloak induction in the mantle of testing.

Can published results be believed? As the editors of this volume note in their introductory chapter, there is good reason to worry that the body of empirical findings appearing in social scientific journals represents a heavily biased draw from the population of results that have in fact been realized, or could be realized, by researchers. This bias is widely understood to derive both from a preference among reviewers and editors for strong over weak or null results and from strategic efforts by authors to generate and report the kinds of results mostly likely to be published (Ioannidis 2005; Gerber and Malhotra 2008b, 2008a; Gerber, Malhotra, Dowling, and Doherty 2010; Humphreys, Sanchez de la Sierra, and van der Windt 2013; Franco, Malhotra, and Simonovits 2014; Nyhan 2015; Dunning 2016). This bias is increasingly considered to be a serious obstacle to the production of reliable social knowledge.

What to do? Social scientists have in recent years devoted increasing attention to two promising responses to publication bias: the pre-registration of study designs and analysis plans and results-blind review.[2] These responses, while different in important respects, both rely on a logic of what we might call *blind selection.* In the case of pre-registration, researchers choose their tests blind to those tests' results; in the case of results-blind review, reviewers and editors consider manuscripts that describe study procedures but are silent about outcomes. These mechanisms address the problem of post hoc selection for positive or "strong" results by occluding information about test results from the actor making the selection. Blind-selection procedures thus reduce opportunities for the cherry-picking of results and provide researchers with a credible means of distinguishing *testing* from unconstrained exploration. The case for pre-registration and results-blind review has been made extensively across the social sciences (see, e.g., Gerber and Malhotra 2008a; Wagenmakers, Wetzels, Borsboom, Maas, and Kievit 2012; Humphreys, Sanchez de la Sierra, and van der Windt 2013; Monogan III 2013; Miguel, Camerer, Casey, Cohen, Esterling, Gerber, Glennerster, Green, Humphreys, Imbens, Laitin, Madon, Nelson, Nosek, Petersen, Sedlmayr, Simmons, Simonsohn, and Van der Laan 2014; Nyhan 2015; Findley, Jensen, Malesky, and Pepinsky 2016).

To date, however, investigation of both the problem of publication bias and the promise of possible solutions has been methodologically bounded. The examination of publication bias has been limited to the scrutiny of quantitative results, via for instance the search for discontinuities around the critical values associated with the 0.05 significance level (as in, e.g., Gerber and Malhotra 2008a; Gerber, Malhotra, Dowling, and Doherty 2010), and the problem is frequently conceptualized as selection for studies and analytic choices that yield statistically significant results. We thus have little sense of the degree to which published *qualitative* research represents a cherry-picked collection of findings. Likewise, pre-registration and results-blind review have until very recently been discussed and practiced almost exclusively in connection with experimental or otherwise prospective analysis. Advocates of these approaches have devoted remarkably little attention to thinking through whether or how they might be used outside the worlds of experimental or otherwise prospective research, domains where their logic operates most simply. Yet, the vast majority of empirical (including quantitative) work in the social sciences falls outside these categories: we social scientists are mostly observationalists studying events that have occurred outside our control, in the past.

[2] A third response, replication, is dealt with at length in Part IV of this volume.

There has been little sustained effort, however, to examine whether or how blind-selection procedures might operate for retrospective hypothesis-testing work, whether qualitative or quantitative.

The aim of this chapter is to extend the consideration of publication bias and its potential fixes onto new methodological terrain. It does so by examining the extent of publication bias in qualitative research and by exploring the promise of pre-registration and results-blind review for enhancing the validity and credibility of non-experimental research more broadly. Arguments for pre-registration of qualitative (Kern and Gleditsch 2017; Piñeiro and Rosenblatt 2016) and, more broadly, observational research (e.g., Burlig 2018) in the social sciences have been gradually emerging. The present chapter seeks to contribute to this discussion in three distinctive ways: (1) through an empirical assessment of the degree of publication bias in qualitative political science research; (2) by elaborating an analytic framework for evaluating the feasibility and utility of pre-registration and results-blind review for any form of confirmatory empirical inquiry; and (3) by marshalling evidence on the practicability of these tools for qualitative and quantitative observational political science research, given the forms of data currently in use.

The analysis proceeds in several steps. Section II begins with a general account of the problem of publication bias, including a conceptual disaggregatation of the phenomenon into two components: gatekeeping bias (the biased selection of studies by editors and reviewers) and analysis bias (the biased selection of tests for reporting by researchers). While the two processes are related, the distinction is key to understanding the different ways in which results-blind review and pre-registration might enhance the production of knowledge: the first by reducing gatekeeping bias and the second by limiting the scope for analysis bias. Gatekeeping bias itself also bears disaggregation: it seems that editors and reviewers do not just prefer positive or strong over null or weak results; they also appear to prefer confirmatory over exploratory analysis. Researchers thus face incentives not merely to report strong results but to report *test* results, rather than inductively generated insights.

Following a brief review of evidence of publication bias in quantitative research, the chapter zeroes in on the nature and scope of the problem in qualitative scholarship. The first question, in such an examination, is whether qualitative research is a potential site of such bias. Publication bias refers to a skew in the set of reported empirical assessments of truth propositions about the world; it is a concept that is thus typically applied to confirmatory, as opposed to exploratory, research. Given the important role of exploration in much qualitative inquiry, we might wonder whether the concept of publication bias is a meaningful one in this domain. It is clear that key developments in process-tracing methodology over the last 15 years have involved the elaboration of the logic of empirical testing in qualitative inference. But how much qualitative research in fact seeks to test hypotheses, as opposed to generating inductive theoretical insights from case evidence? I report the results of a detailed review of the stated intellectual goals of a large set of recent qualitative journal publications, which reveals that a large share of qualitative researchers set out with the explicit intention of assessing the validity of empirical propostions. The chapter, further, illustrates why it is that the opportunities for selective reporting of results – in essence, qualitative "p-hacking" – are just as great for case-oriented as for statistical analysis.

Do we have reason to believe, then, that published qualitative test results in fact represent a biased draw from the population of realizable results? The chapter provides systematic evidence—the first reported, to my knowledge—of strong publication bias in qualitative research. Across 58 qualitative and multi-method articles published in a set of leading journals in 2015, not a single study reports a null finding on a primary hypothesis, only one draws attention to an observation that cuts against the study's conclusions, and not a single mixed-method study finds discrepancies between the statistical results and the case evidence.

Reflecting on these results, I argue that the powerful biases affecting the body of reported empirical findings – for qualitative and quantitative research alike – both reflect and reinforce a breakdown in scholarly communication. The current state of affairs is one in which researchers are free to claim to be reporting the results of structured empirical tests, but in which few can *credibly* claim to be doing so. This is, moreover, not just a problem for hypothesis-testing: it also a communicative equilibrium in which the pervasiveness and intellectual contributions of exploratory analysis are likely to be massively understated.

In the next three sections, the chapter turns to a consideration of solutions grounded in the logic of blind selection, examining their promise beyond the realm of quantitative prospective research. Section III, drawing on large existing literatures on these mechanisms, provides an overview of how pre-registration and results-blind review can enhance the credibility of empirical claims. Section IV then examines to what kinds of research these mechanisms can be effectively applied. I argue that there is nothing instrinsic to the logic of either pre-registration or results-blind review that limits their applicability to experimental, prospective, or even quantitative inquiry. Rather than thinking about these devices in connection to broad methodological categories, we must identify the *features of the data* that allow for the credible pre-specification of tests.

I argue that the credibility gains that can be reaped from blind-selection are a function of three features of the empirical situation:

1. the *prior unobservability* of the test evidence,

2. the *independence* of new evidence from old, and

3. the *precision* with which tests can be pre-specified.

Through a further systematic review of recent observational studies – qualitative and quantitative – appearing in leading journals, I then show that much of the observational quantitative and, even more so, qualitative evidence underpinning published work displays features that would enable credible pre-specification of tests. Pre-registration and results-blind review, I contend, thus has the potential to enhance the strength and credibility of confirmatory research across a broad range of empirical methods. Further, while bolstering the integrity of test results, mechanisms of blind selection can help raise the profile and status of empirical exploration by making it harder for researchers to cloak exploratory work – a critical component of any progressive research agenda – in the mantle of testing.

The chapter's final substantive section considers a curious difficulty posed by results-blind review: in removing from reviewers the information they would need to choose with bias, it may make it difficult for them to choose at all. How exactly does one evaluate a study's contribution to knowledge without knowing what it has uncovered? I argue that this challenge is largely surmountable if reviewers assess submissions by asking *how much we should expect to learn* from a given empirical procedure. To illustrate the concept of expected learning, I propose a Bayesian operationalization that offers clear guidance in a results-free situation using the same belief sets that are required for a conventional, *ex post* assessment of results.

Running through this discussion is the notion that we should conceptualize the threat of publication bias and the effectiveness of solutions in continuous, rather than dichotomous, terms. Pre-registration and results-blind review may generally contribute less to the credibility of retrospective and qualitative research, where the opportunities for "peeking" at the data or adjusting analytical procedures *ex post* are greater. These tools undoubtedly operate most cleanly for quantitative experimental research. Even under less-than-perfect conditions, however, mechanisms of blind selection can substantially reduce the scope for cherry-picking, by both researchers and gatekeepers. They can also go a long way toward extracting the observational social sciences from an unhappy equilibrium in which many claim to be testing but few really are.

Skewed social knowledge

Social inquiry can generate new social knowledge only if its results are made public and widely disseminated. We have good reason to think, however, that the body of published social scientific results represents a skewed sample from the population of results in fact realized, or realizable, by researchers. The problem commonly termed "publication bias" can in fact best be separated into two distinct, if related steps in the process of knowledge dissemination at which actors have the opportunity to select empirical analyses based on their results, rather than on how much is learned. What I will term *gatekeeping* bias is the tendency of editors and reviewers to select stronger over weaker test results for publication. *Analysis* bias refers to the tendency of researchers to select stronger over weaker results for reporting.

Gatekeeping bias

Gatekeeping bias arises from the combination of two tendencies, one more commonly noted than the other:

Preferences for strong results. As commonly noted, editors and reviewers appear to favor strong results, which tends to mean results that are statistically significant. Franco et al (2014) provide striking evidence of gatekeeping bias by taking advantage of a known population of studies: those survey experiments conducted via Time-Sharing Experiments in the Social Sciences (TESS). Franco et al find that those studies that uncovered statistically significant treatment effects were considerably more likely to be published than those yielding insignificant effects.

Preferences for testing. This bias toward strong results operates jointly with an equally powerful, if less often noted, orientation toward *hypothesis testing* as the

dominant form of empirical analysis. In quantitative analysis, hypothesis testing tends to mean null hypothesis significance testing – in which the focus lies on the p-value, or the probability of having estimated a given non-zero parameter value if the true value were zero – at the expense of other modes of analysis, such as the estimation of effect sizes. More importantly, confirmatory analysis appears to be preferred over *exploration*, or the generation of theoretical insights or hypotheses through inductive engagement with the data. This situation creates incentives for scholars to claim to have subjected propositions to tests, even when they have arrived at their findings through relatively unstructured exploration.

Analysis bias

If gatekeeping bias describes a filter applied to manuscripts by reviewers and editors, analysis bias describes a filter applied to analytic strategies by researchers: given a choice of analytic strategies (e.g., statistical models, process-tracing tests), researchers seem to favor specifications of their tests that generate stronger, rather than weaker, results.

Analysis bias may in part derive from cognitive sources, such as confirmation bias or researchers' political or normative commitments to the hypotheses they are testing.

Yet it is likely that gatekeeping bias itself plays a large role: that, in the face of strong professional pressures to publish widely and well, researchers' analytic choices are shaped by strategic anticipation of editors' and reviewers' preferences for strong results.

The behavior we are concerned about is sometimes called "fishing": researchers choosing their empirical strategies in light of those strategies' results (Humphreys, Sanchez de la Sierra, and van der Windt 2013). These researcher-induced biases are given free rein by the broad range of choices that must be made in the course of empirical analysis: how to code variables or generate indices; which functional form to assume for the relationship of interest; what to assume about the probability structure of the errors; which variables to enter as covariates in a regression model; which interactions to include or exclude; which cases to include or drop. Researchers thus have the option of choosing the estimation strategies that yield the results they, or publication gatekeepers, want. Note that the problem applies even when researchers do not in fact carry out multiple analyses – that is, if they *stop* trying additional tests and if that decision to stop is conditional on the results of initial tests (i.e., stop if you find a significant result, see Gelman and Loken 2014).

The problem, as Humphreys et al (2013) explain, is that the probability of "finding" a relationship where none in fact exists (Type I error) rises with the number of attempts if these attempts are treated in isolation from one another. In particular, with k attempts, the probability of finding at least one significant result at the 95% level, even if there is no true systematic relationship, is $1–0.95^k$. Put differently, unless the number of attempts conducted by the researcher is reported and appropriately adjusted for, the reported p-value for a "significant" result is not in fact meaningful. In this situation, what may be reported as a test result is better understood as an insight drawn from empirical exploration.

Gerber and colleagues have provided the most striking evidence to date of bias in the results published in top political science journals. Among results appearing in the *American Political Science Review* and the *American Journal of Political Science*, Gerber

and Malhotra (2008a) find a massive discontinuity in reported Z scores at the critical value required to achieve a p-value below 0.05 – a pattern strongly suggestive of either analysis bias or biased selection by editors and reviews. Gerber et al (2010) report parallel findings across large number of political science journals in the literatures on economic voting and the effect of negative campaign advertising, while Gerber and Malhotra (2008b) find a similar pattern in top sociology journals. Christensen and Miguel (this volume) provide an extensive review of evidence of publication bias in several literatures in the field of economics.

Qualitative publication bias

While issues of gatekeeping and analysis bias have received the greatest attention in quantitative research communities, there is good reason to think that they strike qualitative research with equal force. First, as I demonstrate through a review of recent publications, much qualitative research is – like much quantitative research – purportedly oriented toward establishing the validity of empirical propositions, and is thus potentially susceptible to gatekeeping and analysis bias. Second, opportunities for "fishing" for positive or strong findings are just as available to qualitative as to quantitative researchers. Third, a systematic review of reported findings yields strong evidence of skew in the set of qualitative results appearing in major political science journals.

The confirmatory orientation of qualitative research. First, note that testing can be central to qualitative research. Indeed, over the last decade, there has been a broad move toward the conceptualization of case-study methods as procedures for testing explanations and theoretical claims. The literature on qualitative process tracing has developed increasingly sophisticated ways of thinking about different types of qualitative tests. Many qualitative methodologists have advocated the use of, and elaborated upon, Van Evera's (1997) typology of tests, which classifies empirical predictions according to the "uniqueness" and "certainty" of the implications being examined (see also Mahoney 2012; Collier 2011; Bennett 2010). This schema yields test types known as "hoop tests," "smoking gun tests," "doubly decisive tests," and "straw-in-the-wind tests," with differing consequences for inferences when passed or failed. More recently, a number of qualitative methodologists have reformulated process tracing tests in terms of Bayesian updating, an approach that involves the formation of beliefs about the probability that a hypothesis is true, conditional on the evidence (Beach and Pederson 2013; Bennett 2015; Humphreys and Jacobs 2015).

We might wonder, of course, whether qualitative researchers in practice seek to carry out confirmatory analysis. Perhaps the typical qualitative study seeks to empirically induce theoretical insights rather than to test – in which case any concerns about selective reporting of qualitative results would be largely moot. I examine this question through a systematic review of a sample of 94 qualitative and multi-method articles appearing in 2015 and/or 2016 in 8 highly ranked political science journals.[3] Articles were categorized

[3] The journals searched were: *American Political Science Review* (2015-2016), *British Journal of Political Science* (2015-2016), *Comparative Politics* (2015), *International Organization* (2015-2016), *International Security* (2015), *Perspectives on Politics* (2015), *Studies in Comparative International Development* (2015), *World Politics* (2015-2016). Both the 2015 and 2016 volumes were searched for those journals that had 5 or fewer qualitative or multi-method articles in 2015. To be included in the sample, the article must had made substantive use of qualitative evidence, amounting minimally to a full

based on the purposes to which the qualitative evidence was put. In the process of classifying the articles, it became clear that the standard dichotomy of theory-testing vs. theory-generating would be insufficient for capturing the range of modes of analysis on display. Many qualitative studies, for instance, do not explicitly set up their analyses as testing a hypothesis, yet deploy a body of evidence to establish the validity of a theory or an explanation. Moreover, a good deal of qualitative work claims to be using case evidence to illustrate the workings of a theory, rather to generate the theory or probe its validity. In total, I distinguish among four potential uses of qualitative evidence:

> **Testing:** I categorize as "testing" those articles that either (a.) explicitly claim to be using the qualitative evidence presented to test or evaluate the truth value of an explanation, theory, or other claim or (b.) are structured as a test in that they explicitly identify criteria of evidentiary assessment for a theory or explanation, such as its empirical predictions or its observable implications.
>
> **Substantiating:** Some articles mobilize evidence to confirm or establish the validity of a proposition about the world without explicitly describing their analysis as a test. I code as "substantiating" those articles deploy the qualitative evidence as backing for the truth value of a proposition without explicitly describing or unambiguously structuring the analysis as a test.
>
> **Illustrative:** "Illustrative" articles are those that do not involve explicit test-oriented features and that clearly characterize their use of the qualitative evidence as illustrating or applying a proposition, rather than establishing its truth value.[4]
>
> **Theory-generating:** A "theory-generating" article is one that explicitly frames the empirical analysis as serving to generate a set of conceptual or theoretical insights.

Note that *testing* and *substantiating* articles share an important feature in that both ask the reader to buy into a truth claim about the world on the basis of the evidence presented. While the former group is explicitly confirmatory, the latter group is implicitly so. In that key sense, concerns about selective use or interpretation of evidence arguably apply with equal force to both categories.

The results of the classification exercise are presented in Table 1. In this sample, 41.5 percent of articles explicitly claimed to have used the qualitative evidence presented to test the veracity of an explanation, theory, or other claim. An additional 35 percent use the qualitative evidence to substantiate a claim about the world without describing the analytic procedure as a test. Meanwhile, notwithstanding common characterizations of qualitative research as oriented toward theory-generation or illustrative analysis, fewer than a quarter of the articles in the sample use qualitative data for these purposes. Thus, while exploration and a back-and-forth between theory and evidence remains an important focus of the qualitative tradition, it is clear that the vast majority of qualitative political science research involves the use of evidence to establish the empirical validity

---

section devoted to qualitative analysis. Abstracts of these 94 articles were then read further to determine the use to which the article put the evidence. Where abstracts were unclear, the article text was inspected. Coding rules and the dataset with articles codings are available in the paper's online appendix.

[4] The terms "elucidate" or "plausibility probe" also triggered a classification as "illustrative," as long as the empirical analysis was not otherwise structured as a test (e.g., with the assessment of empirical predictions).

of a proposition about the world. The question of whether evidence and analyses have been presented in a full and unbiased fashion thus arises for a large share of qualitative scholarship.

Table 1. Use of qualitative evidence in 94 articles in 8 highly ranked political science journals (2015 and 2016)

Category	Articles	%
Testing	39	41.5
Substantiating	33	35.1
Illustrative	20	21.3
Theory-generating	2	2.1

Raw codings can be found in the chapter's online appendix at https://politics.sites.olt.ubc.ca/files/2018/11/Appendix-to-Tables-1-and-2-Qualitative-Testing-and-Publication-Bias.xlsx.

Opportunities for fishing in qualitative research. The opportunities for fishing are just as great with qualitative as with quantitative work. The problem of fishing in case study analysis can be illustrated by conceptualizing process-tracing tests in probabilistic terms (as is increasingly common, see, e.g., Beach and Pederson 2013; Bennett 2015; Humphreys and Jacobs 2015). Consider, for instance, the dynamic as it might play out in relation to what the process-tracing literature refers to as "smoking gun" tests (Van Evera 1997). Smoking gun tests are tests for which passage strongly strengthens a hypothesis, but failing only minimally weakens the hypothesis. Imagine the following procedure:

- We seek to test a primary hypothesis, H.
- We can do so by going looking for some number, k, of clues, K_i, $i \in [1, 2, \ldots k]$
- Suppose each of these clues would represent "smoking gun" evidence for H. In particular, for each K_i, $p(K_i|H)=0.2$, while $p(K_i|\sim H)=0.05$.
- Assume that each of these tests is independent of the others: observing the result of one test does not change the probability of observing the other clues conditional on H.[5]
- Suppose that we allow ourselves to search for $k=5$ such smoking gun clues.

[5] Of course, the *unconditonal* probability of observing a clue will change as we observe the results of prior tests since $p(H)$ will be updated. But, as discussed further below, the independence of evidence hinges on the independence of the *conditional* probabilities, $p(K_i|H)$.

Given this procedure, there's a nearly 1 in 4 chance ($1-0.95^5= 0.23$) that the hypothesis will pass one of the "smoking gun" tests even if *H* is false. Thus, the search for these 5 smoking guns is not nearly as demanding a test of the hypothesis as each test is individually.

Now, suppose that we allow our decision to report a test result to depend on what we find: we report only those tests that are passed. Unless we come clean about how many tests we have conducted, we run a considerable chance of reporting evidence for our hypothesis that is much less probative than it would appear.

While we cannot know for sure (absent a faithful record of all tests conducted by researchers), this sketch plausibly describes common practice in process-tracing research. Because failing a smoking gun test is generally understood to have minimal impact on the validity of a hypothesis, researchers likely tend to think of it as inconsequential if they neglect to report such a failure. As we can see, however, the cumulative impact of failing multiple smoking gun tests can in fact be quite large. Selective reporting of such test results can, thus, seriously undermine the integrity of qualitative findings. And, of course, the non-reporting of failed smoking-gun tests is a "best-case" form of fishing. Far more problematic would be the non-reporting of hoop-test results that run counter to the favored hypothesis.

One potential defense against fishing – the demands of the skeptical reviewer – is also less likely to be effective against qualitative than quantitative analysis bias. When assessing quantitative manuscripts, reviewers can usually readily imagine and ask for a set of robustness tests involving the study dataset or readily obtainable measures, allowing a direct empirical assessment of the sensitivity of results to test specification. For qualitative work, unless a reviewer is deeply familiar with the case(s) being examined, it may be much more difficult to imagine the evidence that might have been collected but was not reported. Reviewers of qualitative manuscripts based on intensive fieldwork are also less likely to demand to see evidence not already reported in the manuscript, given the hurdles to new data collection for such work and the difficulty of knowing what the universe of readily available clues would have looked like.

Further, to the extent that editors and reviewers prefer evidence of effects over non-effects, there is no reason to believe that gatekeeping bias would operate with any less force to qualitative than for qualitative journal submissions.

Empirical evidence of qualitative publication bias. The empirical question nonetheless arises: how comprehensively do qualitative scholars in fact report the results of the tests that they conduct? To address this question, I examined in greater depth the evidence and inferences presented in the same sample of articles analyzed above, summarizing the patterns of findings in Table 2.

Table 2. Patterns of results in 94 qualitative and mixed-method articles in 8 highly ranked political science journals (2015-2016)

Category	Articles (n=94)	Null result[6]	Undermining evidence[7]	Cross-method difference[8]
Testing	52[9] (30)	0	3	3
Substantiating	30 (0)	0	1	--
Illustrative	10 (0)	0	0	--
Theory-generating	2 (0)	0	0	--

Note: Number of mixed-method articles within each category noted in parentheses. Cross-method-consistency examined only for mixed-method articles. See text for details on sample-construction. Raw codings can be found in the chapter's online appendix at https://politics.sites.olt.ubc.ca/files/2018/11/Appendix-to-Tables-1-and-2-Qualitative-Testing-and-Publication-Bias.xlsx.

Like Table 1, Table 2 disaggregates the sample by mode of analysis, separating those articles that explicitly test from those that mobilize evidence to establish the truth value of a proposition, those that employ the evidence illustratively, and those that use the evidence to generate new theoretical insights. Consider the following three features of the pattern in Table 2:

1. An absence of null findings. Across the 94 articles examined, there is not a

[6] An article is coded as having a null result where there is no significant evidence presented for the central theoretical proposition of interest. Where an article tests multiple competing claims, not all advanced by the author, then a null result is coded only if there is no significant evidence presented for any of the claims (other than a null hypothesis) or if there is no significant evidence for any hypothesis advanced by the author. An article that focuses on testing a prior, well-established theory and finds no evidence for this theory does not qualify as a null result since the authors' incentives are to present evidence against this theory.

[7] Undermining evidence was evidence that cut against the article's main finding(s) or conclusion. This could have included the failure to find supporting evidence when such evidence was sought. Undermining evidence must have been noted in at least one of the following locations, indicating that the author acknowledges the undermining effect on the findings: the article's abstract, introduction, any passages summarizing the empirical analysis (e.g., at the beginning or end of the empirical section(s)), discussion section, or conclusion.

[8] Cross-method inconsistency meant some significant difference – between the qualitative and quantitative evidence presented – in the degree of support lent to the article's main argument or finding. Any difference between the quantitative and qualitative findings must have been noted in at least one of the following locations, indicating that the author acknowledges the inconsistency: the article's abstract, introduction, any passages summarizing the empirical analysis (e.g., at the beginning or end of the empirical section(s)), discussion section, or conclusion.

[9] For the present analysis, an article is classified as "testing" if any major component was set up as a test. Since all multi-method articles in the sample contain statistical tests, all multi-method articles are here coded as "testing."

single example of a null finding: of conclusions that directly undermine the primary explanatory or theoretical claim that the article develops.

It is possible, outside this sample, to identify prominent instances of published qualitative null findings. These include, for instance, Snyder and Borghard (2011)'s process-tracing test of audience-costs theory and McKeown's (1983) process-tracing test of hegemonic stability theory. Yet it is telling that these studies are not tests of theories devised or otherwise advanced by the study authors. Rather, they are tests of highly influential existing theories, a situation in which clear null findings were likely to be considered novel and important. The publication of these studies shows that null qualitative findings can be published, at least when they generate surprising insights. Yet the within-sample pattern – in which we find not a single null – suggests that these prominent examples are rare exceptions.

2. A near-absence of weak findings or mixed evidence. Not only are all findings positive, but nearly all of the evidence *points in the same direction*, at least as interpreted by the authors. A close reading of these publications revealed only 4 articles, across 94 cases, in which the authors explicitly drew attention to evidence that diverged from or represented weak support for their primary argument or result.[10] Even if we limit the analysis to articles that use the evidence in confirmatory fashion – via explicit testing or substantiating – just over 5 percent of articles point to a single piece of evidence that cuts against the main argument. It is worth pausing to reflect on this pattern: the presentation of qualitative evidence that all points in one direction is arguably tantamount to the presentation of a statistical result with little or no residual error – with *all* data points lying very close to or on the regression line.[11]

3. Very high consistency of findings across analytic approaches. We often find divergence between the quantitative and the qualitative findings generated by different researchers. This is not surprising: different bases of evidence and forms of analysis should be expected to commonly generate different findings. We might then ask to what degree researchers report divergent results across analytical when presenting *their own* multi-method inquiry. Across the 30 multi-method articles examined, all of which involved explicit testing, only 3 reported any substantive difference between the statistical and the qualitative findings.[12]

[10] Cross-method differences, in multi-method articles, were also counted as undermining evidence where the author explicitly points to the findings of one method as undercutting the findings of the other.

[11] As with null results, it is possible to think of examples of mixed findings outside the sample, such as Haggard and Kaufman's (2012) test of the link between inequality and regime change. Again, however, the authors are testing a prominent existing claim in the literature, rather than finding mixed evidence for their own claim. And, as for null results, this out-of-sample instance appears to be a rare exception.

[12] An expansive definition of "difference in findings" was employed, encompassing any noted difference in the substantive meaning of the qualitative and quantitative results. For instance, in one of the three articles, Hanson (2015), the cross-method difference involved the identification in the qualitative

> Greater variance in *between*-study than *within*-study findings should be a cause for concern: it suggests either that the qualitative and quantitative results reported jointly in mixed-method work are commonly not arrived at independently of one another or that the editorial process is selecting for high levels of consistency.

The problems of gatekeeping and analysis bias should equally concern qualitative scholars and consumers of regression coefficients. It is important, moreover, to note that the observed pattern of published qualitative and quantitative results is not consistent with a world in which researchers, reviewers, and editors are simply attaching greater weight to novel or surprising findings, whether positive or null. If tests were being selected for the new insight or learning that their results generated, then we would expect to see a far higher proportion of null and weak results (unless, that is, we are prepared to believe that the set of hypotheses being tested is largely composed of low-plausibility claims). In sum, across empirical political science, there is strong evidence that published results represent a highly skewed sample from the population of potential findings.

An impediment to scholarly communication

Bias in the publication of empirical test results poses a serious threat to the production of knowledge. The principal problem is not one of individual-level motives—i.e., there is no reason to believe it is driven by researchers' desires to mislead their audiences—but of misaligned incentives and skewed selection processes. An important knock-on effect of gatekeeping and analysis bias is a weakening of scholars' ability to communicate to one another what kind of empirical enterprise they are engaged in. Consider for instance a researcher, quantitative or qualitative, who has *in fact* designed her test procedures prior to examining the evidence. In a world of frequent fishing, that researcher currently has no clear way to credibly distinguish her results—particularly when those results are strong—from the selectively reported findings of her colleagues.

Down that path lies a deeply troubling equilibrium. Where research audiences can no longer distinguish real from apparent tests, the likely result is a predicament much like the well-known "market for lemons." In what we might call the "market for fish," research consumers have little reason to believe scholars' claims to have tested; and research producers in turn have little incentive to do anything other than fish. Indeed, all scholars face mounting pressures to fish as fishing becomes increasingly common practice, raising the "strength" of the average published result. The market for principled testing disappears.

Equally worrying, moreover, are the consequences for exploratory and theory-generating work. As social scientists, we often arrive at our empirical investigations with weak theoretical priors and with weak commitments to any given test as particularly decisive. And so we frequently let the evidence suggest to us or guide us toward a set of plausible propositions: we examine a wide range of model specifications, try out multiple measures of key concepts, or comb through case study evidence for insight into how a process unfolded. Our answers often emerge from, rather than precede, our encounter with the data. Exploratory analysis is broadly understood to be a critical step in the unfolding of a progressive research agenda, a key source of theoretical insight and

---

analysis of factors not considered in the quantitative analysis, but no suggestion that this discovery undermined the quantitative results themselves.

conceptual innovation. While both quantitative and qualitative research are often undertaken in exploratory modes, in-depth qualitative analysis is frequently considered an especially fertile source of theoretical inspiration.

Yet the current situation obscures, and thus serves to deligitimize, exploratory and theory-generating empirical work. The strong pro-testing tilt in publishing and professional norms incentivizes scholars to present results as confirmatory even when they were arrived at through a process of exploration. And critically, because editors, reviewers, and readers are poorly positioned to distinguish testing from exploration, there is little or no penalty for labeling the latter as the former. The consequence is, very likely, substantial over-claiming to have "tested" rather than explored. This outcome is undesirable not just because it distorts the corpus of test results. Equally problematic, it serves to further elevate testing above other, equally valuable stages in the research process. Despite its central role in the production of knowledge, exploratory empirical analysis has become a form of inquiry that dare not speak its name.

Two Institutional Responses

With mounting concern about gatekeeping and analysis bias in the social sciences, recent years have seen greatly increased attention to and experimentation with possible institutional responses. Two potential solutions have been at the center of debate about, and of efforts to counter, the problem: study pre-registration and results-blind peer review (see, e.g., Gerber and Malhotra 2008a; Wagenmakers, Wetzels, Borsboom, Maas, and Kievit 2012; Monogan III 2013; Miguel, Camerer, Casey, Cohen, Esterling, Gerber, Glennerster, Green, Humphreys, Imbens, Laitin, Madon, Nelson, Nosek, Petersen, Sedlmayr, Simmons, Simonsohn, and Van der Laan 2014; Nyhan 2015; Humphreys, de la Sierra, and van der Windt 2013; Findley, Jensen, Malesky, and Pepinsky 2016). While their logics are different, each of these devices operates *by concealing information about empirical results* from decision-makers at the point at which tests are selected for implementation and/or reporting: under pre-registration *researchers* select tests prior to seeing results, while under results-blind review *reviewers* and *editors* select manuscripts for publication prior to seeing results. Moreover, the two can be readily combined.

These two blind-selection devices are seen by many as promising response to problems of publication and analysis bias.[13] Yet they are typically seen as applicable to a fairly narrow slice of empirical social scientific research. The domain of application has been defined in different ways by different advocates. Most frequently, though, pre-registration and results-blind review are associated with *experimental* or otherwise *prospective* research, in which the events or outcomes to be analyzed have not yet occurred. Study pre-registration originated in the medical sciences as a system for posting protocols for randomized controlled trials (RCTs). As the framework has spread across disciplines, the tight link to experimentation and prospection has remained intact. One of the most prominent study registries in the social sciences – that operated by the American Economic Association – is open strictly to RCTs. The Election Research Preacceptance Competition, tied to the 2016 U.S. elections, while focused on the ANES's observational

[13] Though certainly not by all: for critiques, see, e.g., Tucker (2014), Laitin (2013), and Coffman and Niederle (2015).

data was purely prospective in character, accepting design submissions only prior to the outcome of interest (the election). Broadly speaking, discussion of pre-registration and results-blind review tends to assume that these devices have little or no relevance to retrospective observational research – a category that comprises the large majority of empirical research in political science. Moreover, the use and discussion of pre-registration and results-blind review have been almost entirely limited to *quantitative* research. Just as the examination of publication and analysis bias have been largely limited to statistical work, there has been virtually no discussion of pre-registration or preacceptance of qualitative work.[14]

In the remainder of this chapter, I will argue that the methodological categories within which discussion of pre-registration and results-blind review have largely been contained are a poor or only partial fit to the problem. Rather than thinking of these devices as constrained by general methodological category – whether experimental, prospective, or quantitative – we should think of their scope of application as being defined by those features of data and of tests that logically relate to the challenge of blind selection. In particular, we can assess the potential gains from pre-registration and/or results-blind review by asking three questions of a research design:

1. **Prior unobservability:** How easily could the study data have been observed (by researchers or reviewers/editors) prior to test-selection?

2. **Independence:** How independent are the new (study) data of old data?

3. **Precision:** How precisely can a test be specified prior to seeing the data?

There is no doubt that most experimental or otherwise prospective quantitative research readily meet the criteria of prior unobservability, independence, and precision; the common association of pre-registration and results-blind review with these forms of inquiry is perfectly reasonable. As I aim to demonstrate, however, other forms of test-oriented research – *especially qualitative research that focuses on testing hypotheses* – also frequently display qualities that make them promising candidates for blind selection.[15] As I will also contend, it is unproductive to conceptualize the plausibility of pre-registration and results-blind review in dichotomous terms: to think that a researcher either has or has not credibly and fully pre-specified her tests. Rather, we should conceive of these two devices as instruments for reducing bias, with the answers to the three

[14] Nyhan (2015), for instance, refers to results-blind review as a tool for "quantitative studies" (79). Findley et al (2016), editors of the special results-blind-review issue of *Comparative Political Studies*, note that they were open to qualitative submissions but received none. As of January 2017, I was able to identify only one qualitative study on the EGAP registry. The only explicit treatment of qualitative study pre-registration that I am aware of is in the report of an ad hoc Committee on Study Registration (on which I sat), formed in 2014 by three methods sections of the American Political Science Association: Experimental Political Science, Political Methdology, and Qualitative and Multi-Method Research (Bowers, Gerring, Green, Humphreys, Jacobs, and Nagler 2015).

[15] In the medical field, the advocacy and use of registries for observational research, both prospective and retrospective, has also been growing (e.g., Williams, Tse, Harlan, and Zarin 2010; Swaen, Carmichael, and Doe 2011). Burlig (2018) makes a case for the pre-registration of observational studies in economics.

questions above and the attendant gains to knowledge lying along a continuum. The question, then, is not whether blind selection can be applied to retrospective or qualitative research: the task, rather, is to assess, for a given study, how much the pre-specification of tests might enhance the credibility of findings, given the nature of the research design.

The remainder of this section briefly outlines the basic logic of pre-registration and results-blind review. Section IV elaborates the principles of prior unobservability, independence, and precision and examines how non-experimental qualitative and quantitative research are likely to fare against these criteria.

Pre-registration

With its origins in the medical sciences, study registration developed as a way of preventing selective reporting of results of trials testing the efficacy and safety of pharmaceuticals and other medical treatments. Pre-registration has, more recently, gained significant traction in psychology, economics, and political science, with registries hosted by the Center for Open Science (COS), the American Economic Association, and EGAP among the most advanced initiatives in the social sciences to date. [16]

Here, in broad outline, is how pre-registration works. Prior to observing the data (or, at least, the realization of the outcomes of interest), the researcher archives a time-stamped description of the research plan. In the most rudimentary form of registration, the researcher merely registers a study design. Registration advocates in the social sciences generally call for a more comprehensive form of archiving in which the researcher additionally pre-specifies how the analysis will be conducted in what is called a "pre-analysis plan" (PAP). In an illustration provided by Humphreys et al (2013), the PAP involves the specification of the measures to be collected, the general model specifications to be employed, the analysis code to be used for estimation, and mock tables showing how the results will be reported. In principle, a PAP could specify a procedure for estimating a quantity, such as an effect size, as well as for testing a point hypothesis.

Importantly, pre-registration does not bind the researchers' hands by blocking the implementation of unregistered analyses; rather, it precludes researchers from reporting *as tests* those analyses that were not specified in advance. The chief advantage of comprehensive registration is thus communicative: it provides a mechanism for researchers to credibly claim not to have fished their results, and for readers to distinguish testing or estimation from exploration (Humphreys, Sanchez de la Sierra, and van der Windt 2013). Once the data are available, the researcher can conduct and report analyses precisely as specified. Reviewers and readers can, in turn, be confident that tests or estimation procedures were selected in a manner blind to the results. Where researchers conduct analyses that deviate from those that were pre-specified—and, indeed, they may explore to any extent that they please—exploration is then easy for research audiences to identify as such.

By having researchers choose their tests blind to (i.e., prior to seeing) the results, pre-registration offers powerful protection against analysis bias: researchers' *post hoc* choices about which tests or estimations to report. In addition, by making it easier for research audiences to identify true tests, pre-registration makes it more difficult for

[16] See, respectively, https://osf.io/registries/; https://www.socialscienceregistry.org/; and http://egap.org/content/registration.

researchers to disguise exploration as testing, forcing scholars to call exploratory analysis by its proper name.

While the problem of "fishing" has often been associated with frequentist work (consider the term "p-hacking"), the logic of pre-registration bears no particular relationship to frequentism as compared to Bayesianism. As illustrated above, Bayesian logic of the sort often employed in process-tracing research is highly sensitive the problem of selective reporting. Fairfield and Charman (2015 and in this volume) nonetheless argue that Bayesian reasoning renders pre-registration redundant. They contend that this is so because it is always possible, using the rules of Bayesian probability, to "assess the reasonable degree of belief in a hypothesis," (2015, 10) given the data at hand and background knowledge, regardless of the sequence in which data were gathered and the hypothesis formulated. They conclude that the time-stamping of hypotheses is thus "not relevant to the logic of scientific inference" (10).

It is, of course, true that one can arrive at the same posterior beliefs, regardless of sequencing, as long as one applies Bayes' rule in a principled fashion to all of the evidence collected, taking fully into account data that support and data that undermine the hypothesis. This is, in fact, not a feature unique to Bayesianism: in a frequentist mode, the principled analyst can conduct multiple tests and then adjust the calculation of standard errors accordingly (see Christensen and Miguel, this volume, for a discussion of corrective procedures).

Yet the availability of principled analytic solutions in no way resolves the basic incentive and informational problems that generate analysis bias. Whether Bayesian or frequentist, the researcher *can* undertake appropriate analytic steps to incorporate all relevant information in assessing post-hoc or multiple hypotheses. The problem is that, under prevailing publication norms, she may have strong incentives not to. In the absence of some form of pre-specification of analytic procedures, moreover, it is difficult for readers to know if they have been shown the full set of empirical results or for the researcher to credibly claim that she has presented them. The problem that pre-registration seeks to solve, in other words, is not *per se* a problem of sequencing. It is, rather, a problem of credible signaling in a context of asymmetric information and skewed incentives, and sequencing serves as a means of generating that credible signal.

Importantly, pre-registration does not by itself counter *publication* bias (Nyhan 2015). A world in which all studies are pre-registered could still be a world in which those studies that are published represent a biased sample, as long as reviewers and editors continue to strong over weak or null results. In principle, however, study registration offers the possibility of compensating for publication bias. The more widely used registration becomes the more comprehensive record it represents of the analyses that researchers at some point planned to undertake, enhancing knowledge of the population of tests from which published studies are drawn. Moreover, if researchers additionally post-register the results of all pre-registered analyses regardless of publication outcome,[17] an unbiased record of the population of test or estimation results would emerge.[18]

[17] This would include indicating when a study was not completed or when data were not collected or analyzed as planned.

[18] As this complete record would also make publication bias more readily apparent, editors and reviewers might in turn begin to place greater value on weak or null findings.

Results-blind review

A second device for generating a less-biased body of knowledge about the world is results-blind review (see, e.g., Greve, Bröder, and Erdfelder 2013; Smulders 2013; Nyhan 2015; Dunning 2016; Findley, Jensen, Malesky, and Pepinsky 2016). Under results-blind review, reviewers and editors evaluate a manuscript and make a publication decision in the absence of information about the study's results. The submitted manuscript typically contains a framing of the research question, a review of the relevant literature, a statement of key hypotheses to be tested or quantities to be estimated, and a specification of the empirical strategy, including a specification of study protocols and analytic procedures. The study's results may or may not in fact be known to the authors at the time of submission; but in either case, the results are not included in the submission or made publicly available prior to an editorial decision.

While registration forces authors to choose test procedures before observing the results, results-blind review forces publication gatekeepers to choose studies before observing their results. Thus, while registration principally operates to minimize analysis bias, results-blind review attacks gatekeeping bias head-on. Moreover, if we believe that analysis bias principally derives from publication incentives, widespread use of results-blind review should also substantially reduce researchers' motivation to fish for strong results, and thus additionally cut powerfully against analysis bias.

For these reasons, political science journals have begun to experiment with results-blind review. A recent special issue of *Comparative Political Studies* contained three articles selected through a purely results-blind review process (Findley, Jensen, Malesky, and Pepinsky 2016). Additionally, via the Election Research Preacceptance Competition (ERPC), tied to the 2016 U.S. general election, 9 leading political science journals agreed to review and provisionally accept manuscripts using American National Election Study data based on a pre-registered study design.[19]

The ERPC illustrates one important feature of blind-selection mechanisms: that they can be employed in combination. In principle, results-blind review can be implemented in the absence of pre-registration: authors might have analyzed the data at the time of submission but omit the results from the manuscript. In practice, the availability of this option may encourage a form of adverse selection in which researchers over-submit studies with null findings for results-blind review, where they stand a better chance of acceptance than if evaluated with results. If reviewers in turn come to expect a disproportionate share of null findings among the results-blind manuscripts they are assessing, the purpose of results-blind assessment is defeated. Moreover, the integrity of the process may be undermined if the findings have been previously presented at conferences or posted online. For these reasons, some have advocated for the use of results-blind review in tandem with pre-registration, as in the COS's Registered Reports model (Nosek and Lakens 2014; see also Nyhan 2015).

[19] See program announcement at http://www.erpc2016.com.

Criteria for Blind Selection

Under what conditions can study pre-registration and results-blind review reduce bias? At the most basic level, these devices are oriented toward enhancing the credibility of tests and estimates. They are thus not applicable to research that seeks to inductively uncover empirical regularities or to derive theoretical inspiration from the data.

However, to what kind of test- or estimation-oriented research can blind-selection methods be usefully applied? The core logic of both pre-registration and results-blind review is one in which the test (or estimation) procedures are specified and selected without knowledge of their results so that positive or strong results cannot be favored. For pre-registration, it is the researcher who selects tests before knowing their results; for results-blind review, reviewers and editors evaluate proposed tests in ignorance of their outcomes. It is from this core logic that we can derive the criteria of prior unobservability, independence, and precision. Prior unobservability and independence minimize knowledge of test outcome at the time of test selection. Test-precision maximizes the visibility of any after-the-fact adjustments to data-collection or analytic procedures, helping audiences readily distinguish the test component of a study from its exploratory components. Let us examine each criterion in turn.

Prior unobservability[20]

The credibility of study pre-registration hinges on the credibility of the researcher's claim not to have observed a test's result prior to the registration of the pre-analysis plan. For this reason, study registration has been seen as a natural fit with experimental research. Experimental tests are purely prospective: by definition, they hinge on outcome data that *could not* have been observed at the start of the study, prior to the experimental manipulation. The experimental method thus allows researchers to demonstrably pre-specify test procedures prior to having knowledge of the results of those procedures. Study registration in the biomedical field was initially applied to randomized controlled trials; the AEA registry is strictly available to experimentalists; and the EGAP registry has been almost entirely used for experimental work. Somewhat less commonly, registration advocates have pointed to its uses for observational prospective research, in which the events to be analyzed have not yet occurred at the time of registration (for a prominent application, see Monogan III 2013).

Results-blind review, technically speaking, does not depend on researchers not having seen test results when submitting study designs for review; it merely requires that reviewers and editors be ignorant of those results. Yet, any results-blind review process will be more credible to the extent that the data required to implement the tests are not readily available at the time of review: it is less likely that reviewers could have seen, or could readily download and take a peek at, the data themselves in evaluating the results-free manuscript.

Thus, especially for pre-registration, timing is critical to bias-reduction. Yet the timing of relevance here is not in fact the timing of *outcomes* relative to registration – whether the research is experimental/prospective or retrospective – but the timing of *observation* relative to registration. Let us define the concept of *prior unobservability* as

[20] The discussion in this section draws conceptually on my contributions to Bowers et al (2015).

the credibility of the researcher's claim not to have seen the test data at the time of registration. In experimental and other prospective work (say, a study of an election that has not yet occurred), establishing prior unobservability is relatively straightforward, assuming it can be verified that registration occurred prior to the start of the study or outcome being observed.

What about observational work that is retrospective? Much observational analysis involves, for instance, the use of readily accessible data about past events: e.g., election studies or other surveys that have already been carried out and their data archived; cross-national indicators that have already been compiled and posted online; or historical cases that have already been well documented in the secondary literature. Where the data to be analyzed come from a dataset that already existed in analyzable form prior to registration, establishing the prior of unobservability of the evidence may effectively be impossible, and registration will be a weak tool for credibly minimizing the scope for fishing.

However, a credible case for prior unobservability can also be made for retrospective research wherever the data themselves are newly collected, newly published, or otherwise newly available, and hence (prior to that point) inaccessible to the researcher. Opportunities of this sort arise prior to the implementation or release of a new survey, the conducting of elite interviews with new subjects, the opening of a new archive or release of a new collection of historical records, or the discovery of new sources (e.g., from an archeological site). For instance, the researcher who plans to conduct a survey of civic activity may effectively be measuring behavior that has already occurred; but it will be relatively straightforward to document that the data could not have been observed before the survey was fielded.

A degree of prior unobservability also obtains for research involving the use of existing data that are costly or difficult to access, especially where the date of access can be established. This will often be the case, for instance, with documentary evidence that exist only in original form at a specific research site (e.g., an archive) access to which must be applied for, or with existing data the use of which is restricted or the release of which occurs at a specific point in time (e.g., by a private company or government agency that holds the data). It is worth noting that much qualitative research involves original data collection that requires the researcher to undertake a discrete activity – e.g., a visit to the archives or an elite interview – with a specific start date prior to which the evidence could not have been observed.

Prior unobservability in practice. To what extent does the kind of data that observational researchers, quantitative and qualitative, use in practice have features that lend themselves to plausible claims of prior unobservability? To examine this question empirically, I conducted a survey of the forms of evidence used in non-experimental articles appearing in 2015 in nine top political science journals. I examined all non-experimental articles, both qualitative and quantitative, in the *American Political Science Review, World Politics,* and *International Organization.* Since the numbers of qualitative articles in these journals was low, I further examined all qualitative articles appearing in 2015 in six journals that more commonly publish qualitative work: *British Journal of Political Science, Comparative Politics, International Security, Studies in American Political Development,* and *Studies in Comparative Political Development.* This search resulted in a sample of 61 observational quantitative and mixed-method articles and 52 purely qualitative articles.

In each article, each form of evidence (e.g., variable measure, cited source, etc.)[21] was coded into a category designed to distinguish among types of data according to the *barriers to observation*. The rationale is that higher barriers to observation offer firmer ground on which researchers can rest a claim that the relevant evidence was effectively unobservable at the time of study registration. All data used to measure outcome and explanatory variables or to observe features of a causal process (as in process-tracing studies) were coded. The types of barriers coded for were:

- a need to code data or place them in analyzable form
- restrictions on access to the data imposed by third parties
- a need to create the data from scratch, whether without or with interactions with human subjects
- the fact that the events being studied had not yet occurred at the time that data-collection procedures were devised (i.e., that the study is fully prospective).[22]

In mapping from barriers to the credibility of prior unobservability claims, the highest credibility was assigned to data derived from events that had not yet occurred at the time of a project's conception; these are data to which the researcher could not possibly have had access prior to registration (had the study been registered). The second-highest level of credibility was assigned to data that were freshly created by the researcher, yet I distinguish between data that had to be created via interactions with human subjects (who could, in principle, independently confirm the date of that interaction) and data that were created without such interactions. Finally, I attribute moderate credibility to pre-existing data that can be accessed only with the permission of a gatekeeper (e.g., an archive or a government bureaucracy). On the one hand, a data-gatekeeper could verify a date of access; on the other hand, because the data were pre-existing, the researcher could have acquired them through an alternate route (e.g., from a fellow researcher). This is, of course, not the only plausible ranking of the evidentiary categories by potential prior unobservability; the reader is invited to adjust the labels in the second column of Table 3 to see how results are affected.

[21] For mixed-method articles, only the quantitative evidence was coded as, in most cases, this represented the majority of empirical analysis.

[22] Burlig (2018) similarly identifies observational studies that involve new data-creation, restricted data, or prospective analysis as ripe for pre-registration.

Table 3. Potential prior unobservability in all observational empirical articles published in 9 leading journals in 2015

Data form	Potential credibility of a claim to prior unobservability	Quantitative articles employing data form (%) (N=61)	Qualitative articles employing data form (%) (N=52)	All articles employing data form (%) (N=113)
a: Pre-existing, analyzable form				
a1: Readily accessible (e.g., downloadable digital file)	Low	90.2	17.3	56.6
a2: Restricted (e.g., by govt. agency, corporation, another scholar)	Moderate	26.2	1.9	15.0
b: Qualitative secondary evidence (e.g., books or articles)	Low	18.0	76.9	45.1
c: Pre-existing, not in analyzable form (e.g., require coding)				
c1: Readily accessible (e.g., newspaper articles, public official documents)	Low-moderate	16.4	69.2	40.7
c2: Restricted (e.g., archival and unpublished)	Moderate-high	1.6	26.9	13.3
d: Newly created				
d1: Via human subjects (e.g., interviews, ethnography)	High	4.9	46.1	23.9
d2: Without human-subjects (e.g., mapping)	Moderate	0	0	0
e: Collected via a procedure devised prior to occurrence of phenomenon of interest (e.g., an election study)	Very high	0	0	0

Note: Data forms with moderate, high, or very high potential credibility shaded for ease of reading. Percentages do not add to 100 because a single article could contain multiple forms of data. Qualitative data forms that were used only minimally or peripherally to key claims were excluded. In 16.4 percent of quantitative articles, there was some variable for which the data source could not be determined from the text or supplementary materials. Raw article codings are available in the chapter's online appendix at https://politics.sites.olt.ubc.ca/files/2018/11/Appendix-to-Table-3-Prior-Unobservability.xlsx.

None of the studies reported in the sample of articles were in fact registered. Nor do we know whether any of these data were in fact observed by study authors prior to the choice of analytic procedures. The exercise is thus a purely counterfactual one, asking: if a study employing these data were pre-registered, how credible would we find the claim of prior unobservability? The results are reported in Table 3.

A number of interesting conclusions emerge from this analysis:

1. The data reveal the likely limits of credible pre-registration of observational work. Not a single study examined involved data offering the greatest potential claim to prior unobservability (*e*): in which the event being analyzed had not yet occurred when data-collection procedures were devised (i.e., purely prospective analysis). Just below a quarter, in total, involved the collection of brand new data via interaction with human subjects (*d1*). For 3/4 of these studies, then, claims to prior unobservability would be of moderate credibility at best.

2. There is nonetheless considerable scope for *at least* moderately credible claims of prior unobservability in quantitative observational research. Among quantitative articles, the opportunity lies mostly in the use of restricted data (*a2*, *c2*), either analyzable or requiring coding.[23] A small percentage involved fresh data-creation, generally via survey methods (*d1*). Nonetheless, about a third of quantitative articles used at least one form of data that would lend itself to a moderately or highly credible claim of prior unobservability.[24]

3. There is far greater scope for establishing prior unobservability for commonly used forms of *qualitative* data. A little over a quarter of qualitative articles use restricted data not yet in analyzable form (*c2*), while just under half draw on fresh data-collection from human subjects (*d1*). Moreover, examining the joint distribution of these data forms across articles (not shown in table) reveals that *63% of qualitative articles used a form of data that was either restricted or freshly created through engagement with human subjects* (*a2, c2,* or *d1*). These are forms of data for which some third party – such as an archivist, an elite interview subject, or a survey research firm – was likely involved in making the data available and thus could, in principle, verify researchers' claims about when data were collected. This analysis thus suggests a widely unacknowledged and (to date) unexploited comparative advantage of the typical qualitative study over the typical quantitative, observational study: the greater opportunity, given the nature of the data, for making credible claims to having executed a principled test.

It is also worth noting a further, non-trivial feature of pre-registration, regardless of data form: it puts the researcher on the public record making a claim not to have observed the data yet. This feature alone is likely to be constraining for most scholars, as

[23] For an example of a study that, in effect, rests on the restricted form of quantitative data to make claims to prior unobservability credible, see Neumark (2001).

[24] This calculation cannot, of course, be derived directly from Table 1; it comes from the joint distribution of data forms across articles.

it turns what is now implicitly condoned as common practice (examining the data first, then claiming to have tested) into outright fraud.

As this analysis makes clear, however, researchers' claims to the prior unobservability of their data will in many cases be less than ironclad. Importantly, *this is true even for experimental research:* readers ultimately need to trust or verify researchers' claims about when treatments were administered. Researchers pre-registering retrospective observational studies need to make arguments for and present evidence of the prior unobservability of their test data; and reviewers and other audiences will need to evaluate these arguments and evidence. Like most other forms of empirical uncertainty – such as a standard error – prior unobservability is thus a matter of degree, rather than a binary condition. A researcher's claims to having carried out a true test will be more credible to the extent that they can persuasively establish that they could not have seen the data prior to registration. There may always remain some doubt about this claim, and thus about the potential for analysis bias in reported findings. Yet a world in which some forms of retrospective research were pre-registered would undoubtedly be a world of more interpretable empirical findings than the world in which we are currently operating.

Independence

That observations have not yet been made does not mean that they are new. A second criterion for blind selection relates to the degree to which the test data are *independent* of observations that the researcher could have made prior to registration or that are generally available in the literature or historical record. If the test observations are unseen but could be largely predicted based on prior observations, then "pre-registered" or "results-blind" tests have in effect already been conducted, and researchers, reviewers, and editors know the results.

Let us use the term *observational independence* to refer to the novelty of the "new" data on which hypotheses are to be tested. One way to think about observational independence is to consider the difference between independent and dependent pieces of evidence. Suppose that I plan two tests of a hypothesis, each of which involves the search for a piece of evidence, respectively E_1 and E_2. Each test is of moderately high probative value: e.g., $p(E_i|H) = 0.7$ and $p(E_i|{\sim}H) = 0.3$. Assume, further, that E_1 and E_2 are fully independent of one another, conditional on our confidence in the hypothesis. Suppose that we carry out the first test and observe E_1. Then the search for E_2 has no less probative value than it did before we conducted the first test. It remains the case that $p(E_2|H) = 0.7$ and $p(E_2|{\sim}H) = 0.3$. Finding E_2 could still have a substantial upward effect on our confidence in the hypothesis.

On the other hand, suppose that E_1 and E_2 are highly correlated with one another such that seeing E_1 greatly increases our confidence that, when we look, we will also see E_2 (independently of the first observation's effect on our confidence in H). In that situation, once we have seen E_1, the search for E_2 no longer carries the probative value that it did before. One intuitive way to think about this is that we are now much more likely to see E_2 if H is false. Were we to ignore the dependence of the second piece of evidence on the first, we would thus be greatly overstating the power of the second test. Likewise, the strength of a pre-specified test hinges on the independence of the yet-to-be observed test data from those observations that have previously been made.

Importantly, the dependence of the two pieces of evidence that we are concerned about here is their correlation conditional on *p(H)*. In the above example, seeing E_1 increases our confidence in *H*, and through an increase in *p(H)*, also increases the probability of observing E_2. Yet the effect of prior evidence that runs via our updated confidence in the hypothesis does not weaken the probative value of the second test. The dependence of concern, rather, is the portion of the effect of seeing E_1 on the probability of seeing E_2 that is not explained by E_1's effect on *p(H)*.

The concept of observational independence also underlines a further sense in which broad methodological categories (e.g., prospective vs. retrospective) can be misleading when thinking about the plausibility of blind selection. In particular, observational independence is not established by the fact of prospectiveness. The fact the outcomes to be observed may not yet have occurred at the time of test-specification does not imply that those outcomes are independent of data that have already been seen. Past and future events often form an autocorrelated time-series. Consider, for instance, the prospective Election Research Preacceptance Competition involving designs submitted prior to the 2016 U.S. elections. Many patterns of political behavior that persist through November and shape the election result were already observable (e.g., in polls or media coverage) prior to the election.[25] The "prospective" study and analysis plans submitted may thus, in effect, be grounded in considerable information about the outcomes on which the tests will be based. To put the point differently, retrospective analysis will sometimes involve "newer" evidence than prospective analysis. A scholar seeking to test a hypothesis against data from an upcoming election may in principle be using "older" (more observationally dependent) observations than a scholar seeking to test a retrospective hypothesis about, say, the rise of English democracy through a text-analysis of candidates' campaign speeches from the 19th century.

Observational independence is not a simple concept to operationalize precisely. However, I want to suggest that researchers can likely make interpretable claims about the approximate novelty of the tests that they register – claims that research audiences can reason through and scrutinize. Consider the possibility of an ordinal classification scheme, roughly as follows:

- *Low independence.* The researcher has seen evidence that (conditional on a given level of confidence in the hypothesis) is highly predictive of the evidence being sought. Low independence would hold where, for instance, old and new evidence consist of:
    - Reports deriving from a common source.
    - Interviews with different individuals known to have common interests and common knowledge.
    - Variables known from external cases to be highly correlated with one another.

[25] Again, the concern here is not with persistent patterns that reflect systematic features of the phenomena of interest, but rather random processes that shape observed patterns in a manner that persists over time. Thus, hypotheses could be "fit" around random patterns observed before the election and then confirmed by future election data that are partly a function of the same random disturbances.

- *Moderate independence.* The researcher has seen evidence that provides a clue to the shape of the test evidence. For instance:
    - The researcher knows what arguments politicians made in public; will go looking to see what arguments they made in private.
    - The researcher knows the ethnicity of warring groups' leadership; will go looking for data on the ethnic composition of their rank-and-file.
    - Variables of interest are known from external cases to be moderately correlated with one another.
    - The test data come from observation of events that have not yet occurred (but that will plausibly be correlated with events that *have* occurred).

- *High independence.* The researcher has seen no evidence that is significantly predictive of the test evidence. The highest level of observational independence is generated by random assignment, which breaks any link between *ex ante* patterns across cases and the patterns against which the hypotheses are to be tested. For retrospective observational research, plausible situations of high observational independence might include:
    - The researcher goes looking to see if a particular consideration was raised in private meetings on an issue that was the topic of little public debate.
    - The researcher conducts interviews to find out which business executives attended a private meeting with ministers
    - The researcher collects data on variables that are known from external cases to be largely orthogonal to those measures already collected.
    - The researcher collects observational data prospectively, following an exogenous shock that was likely to have disrupted random patterns in the data.

More can certainly be said about how one might establish or demonstrate the independence of new from old evidence. As with prior unobservability, researchers will need to make a logical and empirical case for the novelty of their pre-specified tests, and research audiences will evaluate the persuasiveness of that case.

It is also worth noting that the difficulties of establishing or characterizing the prior unobservability and novelty of test evidence are not problems with pre-registration or results-blind review in particular. As the E_1/E_2 example above makes clear, these two issues are fundamental, under any circumstances, to interpreting the results of an empirical test. To describe what can be or has been learned from evidence always requires taking proper account of what has already been observed and how old evidence relates to new. What registration and results-blind review can do, however, is to force claims about learning from tests to be made explicit and their justification open to scrutiny by readers.

Further, it is likely that the challenges of prior unobservability and observational independence are in part endogenous to the institutional environment in which knowledge is produced. In a world in which blind selection of non-experimental tests was a live option, researchers would be motivated to creatively address these challenges. In an effort to make their tests more credible, researchers undertaking hypothesis-testing work

would have incentives to develop novel and increasingly credible procedures for establishing the prior unobservability and independence of their test data.

What about the iterative nature of much empirical research, both qualitative and quantitative? The fact that investigators often move back-and-forth, between testing and inductive discovery, presents no fundamental barrier to blind selection. Registration or results-blind submission need not take place prior to all empirical work on a project. Rather, tests can be registered or submitted for results-free review at any point in a project's development – as long as there is some discrete set of evidence bearing on rival theories that has not yet been observed and that cannot be readily predicted from that which has been observed.

Precision

Finally, the gains to pre-registration and results-blind review hinge on how precisely tests can be specified prior to seeing the data. What is at stake in precision is the amount of *post hoc* discretion available to the researcher: that is, the "wiggle room" available for defining details of a test or its interpretation *after* having seen the evidence. The greater the precision, the more effectively a pre-analysis plan can reduce bias and aid credible communication.

At one extreme – purely exploratory work, whether quantitative or qualitative – the researcher has little preconception of what will be found, and so meaningful tests cannot be stated in advance. At the other extreme – for quantitative confirmatory research – it will often be possible for the researcher to indicate in advance the exact measures and model specifications to be employed, including the code to be used for the analyses and mockups of the tables in which results will be presented (see, e.g., the example in Humphreys, Sanchez de la Sierra, and van der Windt 2013). Notably, the distinctions between experimental and observational (or between prospective and retrospective) research do not map well onto the precision criterion. Precision is primarily a function of the degree to which measurement and analytic procedures take, or can be reduced to, algorithmic form. Precise test pre-specification should generally be as feasible for observational (or retrospective) quantitative research as for experimental (or otherwise prospective) quantitative research.

Most confirmatory qualitative research is likely to occupy a middle ground when it comes to precision. Qualitative researchers generally cannot specify their tests as precisely as quantitative researchers; there is no qualitative equivalent to Stata code. One reason is that the analytic procedures applied to qualitative evidence in case-study work are usually non-algorithmic; [26] they typically involve some element of researcher interpretation. Related to this is the wide diversity of forms of evidence that researchers typically encounter in the course of undertaking a case study: documents of many kinds, statements of many kinds from different sorts of interview subjects, a motley collection of secondary sources and news reports. It will often, therefore, be difficult or impossible for qualitative scholars to anticipate in advance the precise form that observations relevant to their hypotheses might take. Given the variety of ways in which the particulars of qualitative observations might vary, it will be difficult in turn to specify in advance how those particulars may affect inferences drawn from the evidence.

[26] An important exception is, of course, forms of qualitative research that are algorithmic, such as Qualitative Comparative Analysis.

Qualitative data-search procedures may also be harder to specify in advance than quantitative sampling procedures, especially given the common difficulty in identifying an *ex ante* "sampling frame" for common forms of qualitative evidence-collection, such as archival research and elite interviewing. It is, in part, the high degree of prior unobservability of much qualitative evidence—the degree to which qualitative research involves going "deep" into a case—that makes precise pre-specification of tests more difficult.

A high degree of precision in test-specification will thus likely be out of reach for much qualitative work. Yet there is still quite a lot of scope for qualitative scholars to indicate in advance how they plan to evaluate their hypotheses. In a pre-analysis plan, qualitative scholars seeking to test explanations or theories should usually be able to state in advance:

- **Test propositions:** The proposition(s) to be examined
- **Empirical predictions:** A set of empirical predictions, deriving from the test proposition(s), to be tested in the case
- **Cases:** The case(s) being studied
- **Search procedure:** A description of where in the case(s) she will look for evidence (e.g., which archives, interviews with what kinds of subjects)
- **Criteria:** A characterization of the kinds of evidence or observations that, if found in those places, would be consistent or inconsistent with each prediction.[27]
- **Approximate mappings from results to inferences:** Some account of how satisfying or not satisfying each prediction would affect inferences. The degree of precision with which scholars would be able to derive these implications in advance is likely to vary. For many qualitative projects, it may not be possible to pre-specify meaningful numerical likelihoods given the wide variety of unanticipated ways in which the details of qualitative observations and their context may vary. In most situations, however, qualitative researchers should be able to provide a broad indication of how test results will affect findings. Researchers should be able to state in advance, in a general way, whether the satisfaction of a prediction would substantially or only minimally boost confidence in an explanation, and whether the non-satisfaction of that prediction would severely or only modestly discredit the explanation. Van Evera's (1997) four test types (e.g., hoop, smoking-gun, straw-in-the-wind, doubly decisive) may provide a useful set of categories for expressing these mappings.

These are little more than the elements of a research design that we routinely require our graduate students to specify in qualitative dissertation prospectuses prior to beginning fieldwork. And, of course, there are many different ways in which PAPs for qualitative studies might be structured. Piñeiro and Rosenblatt (2016), Kern and Gleditsch (2017), and Hartman, Kern, and Mellor (2018) each propose detailed templates for qualitative PAPs. Piñeiro, Perez, and Rosenblatt (2016) represents the earliest qualitative pre-registered PAP of which I am aware, while Christensen, Hartman, and Samii (2018) is, to

[27] This might include an indication of multiple kinds of evidence that could speak to the same prediction (i.e., via a logic of triangulation), allowing for the possibility of uncovering mixed evidence and thus the partial satisfaction of a prediction.

my knowledge, the first completed study containing a pre-registered qualitative analysis.[28]

Imagine, for the sake of illustration, a qualitative scholar who seeks to test a theory explaining tax cuts by British governments in the 20th century. The researcher might specify in advance the primary explanation that tax cuts are motivated by Keynesian aims. The scholar might then derive from this proposition the prediction that we should observe prominent mentions of the logic of Keynesian demand-management in records of deliberations over the tax cut. Having selected and specified a set of tax-policy episodes to be examined, the researcher might further indicate that she will look for mentions of Keynesian logic in all public statements by ministers reported in the daily press, records of cabinet meetings and correspondence between senior officials available in the National Archives, and statements by majority-party members of parliament recorded in Hansard. Further, the researcher could specify that a failure to find prominent mentions of a Keynesian logic would severely undermine the explanation, making the search for this evidence a "hoop test."

A specification of this kind would, of course, still leave some significant features of the test ambiguous. Just how prominent do mentions of Keynesian logic have to be for the test to be passed? How many actors have to mention it? What forms of words will count as the use of Keynesian logic? There would inevitably be some wiggle room. Yet, compared with current practice – involving no public pre-declaration of research plans – this basic specification would in fact pin down a great deal about what will count as a positive or a null finding. Moreover, pre-specification is useful even in the face of unanticipated observations or unexpected features of the evidence that affect its interpretation. The provision of a pre-analysis plan allows the reader to compare the researcher's interpretation of unexpected observations to the pre-announced tests and to arrive at her own judgment about the extent to the interpretation of the evidence is consistent with the analysis plan's broad logic.

As with prior unobservability and independence, precision will be a matter of degree, with higher credibility attaching to more crisply defined tests. And, as with the other two criteria, precision is likely itself to be shaped by the presence of blind-selection opportunities. In a world in which registration and results-blind review are options, the equilibrium outcome is likely to be one that substantially mitigates the problem of precision in qualitative testing. Audiences will be free to judge how much "wiggle room" the researcher has given herself in her pre-analysis plan, and to assess the credibility of the test result accordingly. In turn, qualitative scholars interested in hypothesis-testing will be incentivized to bind their own hands with clear pre-analysis plans. However, researchers will have incentives to *optimize*, rather than maximize, the level of precision in their PAPs: a vague plan will reduce the credibility of the test, while excessive precision raises the risk of error by excluding relevant evidence from consideration.

As Findley et al (2016) point out, there remains the danger of "hypothesis trolling" (12): the pre-registration of a large number of empirical predictions, thus allowing ample room for "fishing" for supporting evidence. Yet pre-registration – combined with a reasonable degree of reviewer vigilance – can itself serve as a powerful

[28] Hartman, Samii, and Christensen's pre-registration was gated at the time of writing but located on the Open Science Framework registry at osf.io/46r87.

check against the strategic multiplication of hypotheses. Reviewers can compare a set of reported results with the predictions specified in the pre-analysis plan, ensuring that results are presented on all predictions (or that an explanation is provided for missing results). Reviewers can additionally check that conclusions take into account the multiplicity of tests conducted. In a quantitative context, researchers can be expected to correct standard errors for the number of hypotheses tested, a procedure that imposes a "penalty" for each additional test. (For detail on multiple-comparisons corrections, see Christensen and Miguel, this volume.) Assessors of qualitative work can similarly ensure that authors have given due weight to failed tests and/or that supporting evidence is suitably discounted for the number of predictions that were pre-specified. While the appropriate norms of assessment will take time to develop, limited experience to date with blind selection in political science is at least somewhat encouraging. Reflecting on the review process for their special issue of *CPS*, Findley et al. report that "hypothesis trolling was specifically targeted and rejected by reviewers" (17). To the extent that assessors penalize the proliferation of predictions, researchers will face clear incentives to specify their tests judiciously.

The test-credibility space

If we imagine a three dimensional space defined by our three credibility criteria – prior unobservability, independence, precision – individual studies are likely to be located at different points in this space. Tests to be carried out using evidence from highly restricted archives, which are expected to reveal details of decisions about which little is known, may have moderate-to-high levels of prior unobservability and independence, but relatively low precision because so little is known about what might be found. A project that will administer a structured interview protocol to elite decision-makers might display high prior unobservability and high precision (given the closed-ended nature of the questionnaire, we know what form the data will take), but middling independence (as the interview subjects may have made numerous public statements about the decisions being studied). And a quantitative social-network study using Facebook data that can be accessed on the company's Menlo Park campus may display high test-precision (as the statistical models can be specified in advance), high observational independence (if there are no other ways of observing the network patterns being studied), and high prior unobservability (since dates of access can be independently confirmed).

The overall credibility of a claim that a set of tests were selected blind, then, depends jointly on these three qualities. We might think of credible blind-selection as a product of the level of prior unobservability, the level of independence, and the precision of the test-specification.[29] Under current practice for observational studies (no pre-registration and results-based review), we have little reason to attribute to a level of prior unosbervability above 0. The overall credibility of current observational tests (and of unregistered or results-based-reviewed experimental tests) should thus itself be

[29] Visually, in our 3-dimensional space, this product would correspond to the volume of the rectangular cuboid defined by the origin and the point marking the levels taken on for each of the three qualities.

understood to be 0.[30] Blind-selection has the potential to enhance credibility relative to the status quo whenever levels of prior unobservability, independence, and precision are all in positive territory.

Can Studies Without Results Be Evaluated?

Results-blind review, in one sense, seems the perfect antidote to both publication and analysis biases: it takes the results – whether strong, weak, or null – entirely out of the equation. Results-blind review is also logistically simpler than pre-registration as it does not require the researcher to demonstrate the prior unobservability of the data. Yet results-blind review is also difficult precisely because it hides results from editors and reviewers.

Reviewers and editors are practiced at judging the strength and importance of empirical findings once they see them. But how are they to judge the value of a study *absent* any empirical claims at all? Perhaps a strong result of a planned test would be enormously informative, but how can the submission be judged without knowing whether that outcome will be realized? Reviewers might well worry about how interesting a null outcome will be, given that it might be an indication that there is in fact no true effect or merely of a weakness in the research design. Findley et al (2016), in reviewing their experience with editing the results-blind issue of *CPS,* identify this quandary as a key stumbling block to the success of results-blind review.

Moreover, how should reviewers think about the plausibility of the proposed hypotheses? Should authors be penalized for testing hypotheses that already have a great deal of backing? Should they be rewarded for proposing novel hypotheses that have yet to be tested?

For results-blind review to work, those assessing study designs and analysis plans for publication must have a metric by which to judge the *ex ante* knowledge-generating potential of a study. I want to suggest that this problem is soluble. The relevant question, I would argue, is *how much we expect to learn from a study*: how much, in either direction, do we expect our beliefs to shift as a result of carrying out the proposed test? There are a number of ways in which we might operationalize this expectation, but any measure of expected belief change must take into account at least three features of a proposed test:

1. **Probative value.** The metric should take into account the probative value of the proposed test. Whether statistical or qualitative, proposed tests will vary in their sensitivity (how likely is the test to be passed if the theorized phenomenon is present?) and in their specificity (how likely is the test to be failed if the theorized phenomenon is absent?). We are likely to learn more from empirical tests with greater than with lesser ability to accurately distinguish among rival hypotheses. Judging the probative value of a proposed test involves an assessement of the core features of a research design, such as: the quality of the

[30] To be clear, this does not mean that the results presented are not the true results of the specified tests; it is that we have no reason to believe that the tests were selected without reference to the strength of the results.

proposed measures; the relationship between sample and theoretical scope; and the relationship between theory and empirical predictions, including the appropriateness of a statistical model, the credibility of a claim of exogeneity, or the probability of observing a clue under a hypotheses or its negation.

2. **Scope for belief change.** Hypotheses vary according to how much scope there is for a change in beliefs in either direction as a result of an empirical test. For hypotheses in which we have very high confidence, there is some scope for downward movement, but little for upward movement; and vice-versa for hypotheses in which we have little confidence. As Humphreys and Jacobs (2015) demonstrate, moreover, with Bayesian updating the scope for upward revision in confidence is greatest not for very implausible hypotheses but for those of moderately low prior probability; the scope for downward revision is, likewise, greatest for hypotheses of moderately high prior probability. Any assessment of expected value of a study must thus take into account how the plausibility of the hypothesis being tested affects the scope for learning.

3. **Likelihood of an impactful result.** Empirical tests often have an asymmetrical character, in which the finding of a relationship or a clue may have a greater or lesser impact on our beliefs than the failure to find the relationship or the clue. For instance, with a hoop test of a hypothesis, not finding the clue has a greater impact than finding it. Our measure of expected learning thus should in some way capture which of these two outcomes – the higher- or the lower-impact one – is more likely. And this, too, will be function of our prior confidence in the hypothesis. For instance, consider a hoop test of a theory in which we are highly confident. Failure of the hoop test would a substantial effect on our beliefs; however, the very fact that the hypothesis is very likely true makes it relatively *unlikely* that the hoop test will be failed and that significant learning will occur.

By way of illustration, consider a simple Bayesian operationalization of expected learning. This operationalization takes the above three considerations into account by employing the same beliefs that we routinely use for assessing the impact of test results. Assume, for the sake of exposition, a relatively simple situation in which a study can be understood as generating a single test of a hypothesis. The function is itself a simple operationalization of expected belief change: it takes the average of (1.) the absolute value of the shift in confidence in the hypothesis that would result from the passage of the test and (2.) the absolute value of the shift in confidence that would result from the failure of the test, weighting each potential shift by the probability of each outcome, given our prior beliefs. The inputs into this calculation are:

1. The prior plausibility of the hypothesis ($p(H)$)

2. The likelihood of a positive result if the hypothesis is true ($p(E|H)$)

3. The likelihood of a positive result if the hypothesis is false ($p(E|\sim H)$)

Assume, further, that we have labeled *H and ~H* such that $p(E|H) > p(E|\sim H)$, and so finding the evidence will always increase our confidence in *H*. We then get the following expression for expected learning (EL):

$$EL = p(E) * ((p(E|H) * p(H)pE - p(H)) + (1 - pE) * (pH - (1 - pEH) * p(H)1 - pE)$$

The first set of terms represents the absolute belief change resulting from a positive result (finding *E*), weighted by the probability of finding *E*, and the second set of terms represents the absolute belief change resulting from a null result (not finding *E*), weighted by the probability of not finding *E*. This simplifies to:

$$EL = 2 * p(H) * (p(E|H) - p(E))$$

where

$$p(E) = p(H) * p(E|H) + (1 - p(H)) * p(E|\sim H).$$

In Figure 1, we graph expected learning for four different kinds of tests to examine how this quantity depends on the probative value of a test and on the prior plausibility of the hypothesis.[31] I use Van Evera's (1997) four test types, with illustrative probabilities. While these test types are generally associated with process tracing, we can readily conceive of statistical results as having similar qualities. For instance, we might believe that a particular statistically significant correlation is very likely to be observed with a given sample size if a causal effect is present, while it may also be observed even if the effect is absent – rendering this statistical test, in effect, a hoop test. Similarly, we might believe a statistical test to be sufficiently specific that we are unlikely to observe a significant correlation if the effect is absent, though the test might well miss the effect even if it is present – generating a smoking gun test.

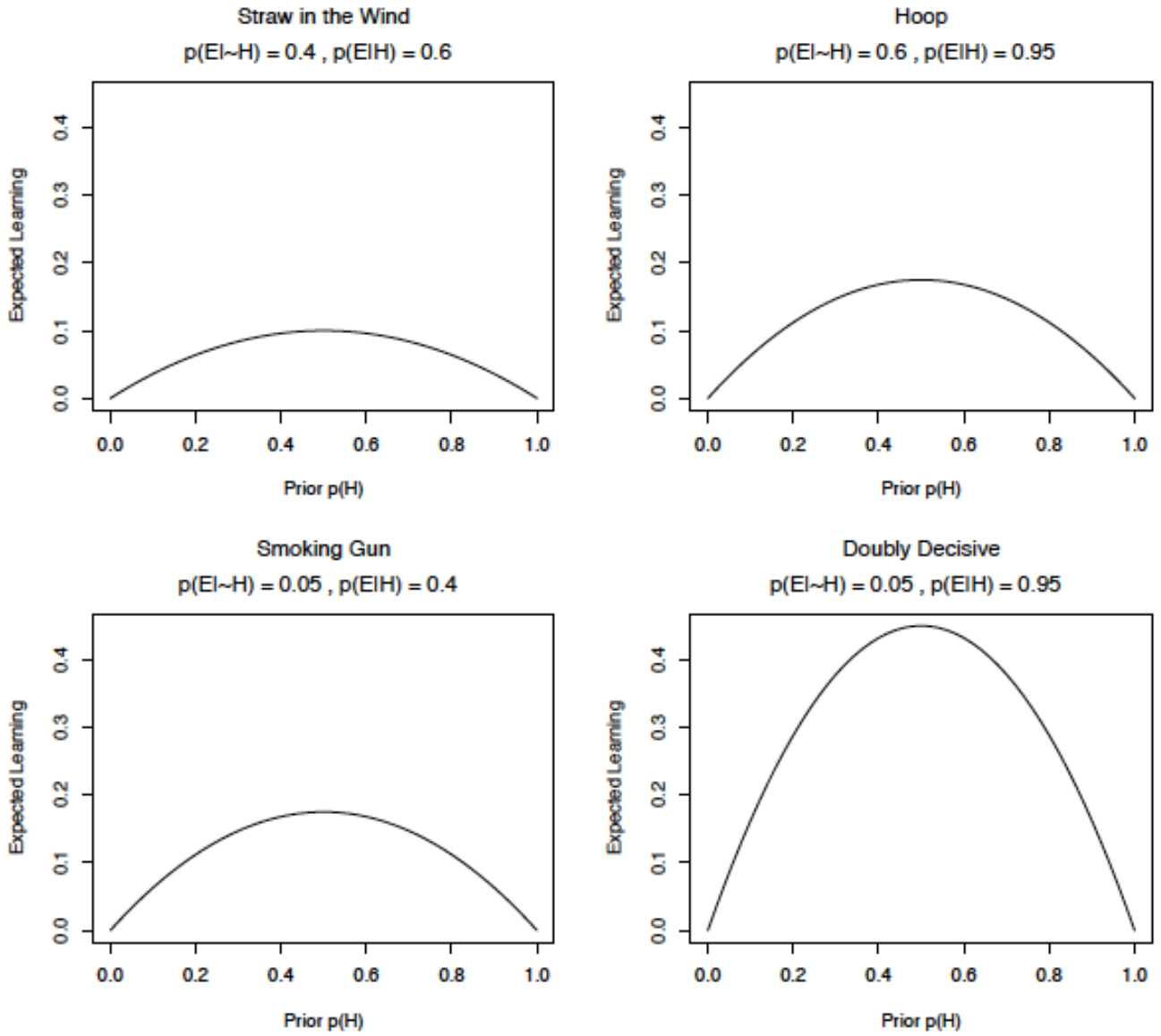
As can be seen here, a metric of expected learning of this kind has the potential to provide reviewers relatively straightforward guidance to the evaluation of manuscripts in the absence of results. Two principles of expected learning are readily apparent. First, unsurprisingly, the higher the probative value of the proposed test, the greater the expected learning. Better tests, more learning.

Second, and less obvious, it is for hypotheses of *middling* prior plausibility that the expected belief shift is greatest, regardless of the type of test. At first glance, this result may seem counterintuitive: shouldn't a hoop test, for instance, generate greater learning for *more* plausible hypotheses since a hoop test is most impactful on the downside? And, by the same reasoning, shouldn't a smoking gun test be most valuable for a less likely hypothesis? It is true that a hoop test has the greatest impact on beliefs *if* it is failed, and such a result would be more impactful for a hypothesis of higher probability.[32] Countervailing this effect, however, is the effect of $p(H)$ on how likely the

[31] Alternative conceptualizations of learning are, of course, possible. For instance, we might think of learning as a reduction in posterior variance rather than as a shift in beliefs.

[32] Though, as noted above, the maximum learning from a failed hoop test occurs at moderately high levels of $p(H)$, not at the highest level.

Figure 1: Expected Learning conditional on p(H), p(E|H) and p(E|~H)



Note: Figure shows how the expected learning for different types of tests depends on priors regarding the proposition.

hoop test is to be failed in the first place. The stronger the hypothesis, the less likely the hypothesis is to fail the hoop test, and thus the less likely we are to learn from the test. The same goes for the smoking-gun test and the unlikely hypothesis: the most impactful result, a passed test, is highly unlikely for a weak hypothesis.

There are other possible ways of conceptualizing and operationalizing the expected intellectual value of a study absent access to its results. The key point is that we can assess the likely contribution using the very same beliefs that reviewers and editors regularly use to evaluate studies *with* results. Assessing the substantive meaning of a test result requires us to form a view of the strength of the test: was it a hard (specific) test for the hypothesis or an easy (sensitive) one? And when evaluating manuscripts with results, we routinely use judgments about the plausibility of the hypothesis being tested, given prior knowledge – whether to judge the paper's novelty and contribution to the literature or to assess the plausibility of the findings.[33] These same beliefs simply need to be mobilized in a somewhat different way for judging manuscripts without results.

A standard like expected learning can also cut against a common concern about results-blind review: that it will lead to the publication of a large number of "failed" studies with uninteresting, null findings. Results-blind review would almost certainly generate a significant increase in the publication of studies that find little or no support for the main test proposition. And these findings would indeed be uninteresting to the extent that the hypotheses being tested were never plausible to begin with. However, if reviewers judge study designs based on how much we expect to learn from them, then studies with a substantial probability of generating null results will only pass muster to the extent that a null result would be *informative:* that is, to the extent that a null finding would undermine a hypothesis in which we had some prior confidence. Similarly, and just as importantly, an expected-learning standard would weed out studies likely to produce *positive* findings for hypotheses in which we already have strong confidence. If an "interesting" result is one that shifts our beliefs closer to the truth, then results-blind review arguably has a far greater prospect of delivering interesting findings by focusing assessors' attention on the core issue of what can be learned, rather than on the distraction of whether a study happened to be "successful."

Conclusion

The *status quo* is a world in which we are almost entirely unconstrained in our ability to choose "tests" conditional on their outcomes. And yet, for many forms of observational inquiry, we are capable of articulating what we are looking for before we know what we will find. This moment of empirical anticipation creates an opportunity for tying our own hands – whether as authors or publication gatekeepers – forcing ourselves to choose hypotheses and study procedures on their intellectual merits, rather than on the realization of favorable results.

Neither pre-registration nor results-blind review represents a panacea for the deep selection problems that plague confirmatory social-scientific research. As we move outside the realm of random assignment, assessing the contribution of blind-selection mechanisms to

[33] Interestingly, reviewers may be using their beliefs about the prior plausibility of the hypothesis to judge a paper's contribution in rather counter-Bayesian ways. When they judge a paper with a highly surprising finding as making a greater contribution than a paper with less surprising findings, they are of course ignoring the role of the prior in Bayesian updating. Bayes' rule tells us that a positive result for a very unlikely hypothesis should make us only marginally more confident in the hypothesis, even if the test is a highly discriminating one.

reducing bias requires a careful, complex assessment of the types of data and forms of tests being employed. In this assessment, there is room for subjective judgment and potentially divergent interpretations. I have sought to show, however, that these two devices can plausibly enhance the credibility of findings for a wide range of test-oriented research designs.

In closing, it is worth considering one possible interpretation of the argument that I have made. A common objection to pre-registration is that it will tend to stifle exploratory analysis and creativity in the generation of theory (e.g., Laitin 2013; Tucker 2014). Results-blind review might likewise be seen as a mechanism that favors confirmatory over exploratory research since it requires the clear specification of procedures in advance. More generally, the advocacy of the use of these mechanisms for a wide range of empirical methods could be read as a claim that the main thing that we should be doing as empirical researchers is implementing tests that have been specified prior to observation – that the scope for undisciplined exploration should be sharply limited. And even if it is not the intention of those advocating blind-selection mechanisms, the practical effect could be to constrain researchers to studying questions for which *ex ante* tests can be specified.

In fact, the case for expanding the use of pre-registration and results-blind submission runs precisely in the opposite direction. If the pre-specification of study procedures were to be widely *required* and imposed as a *binding constraint* – such that authors would always or usually be required to pre-announce tests and would be constrained to report *only* those analyses that were pre-announced – the result would indeed be to stifle empirical discovery and the conceptual and theoretical inspiration that it yields.

The likely effects of voluntary pre-registration and results-blind review are very different. As voluntary forms of hand-binding, these mechanisms allow for more credible communication between researcher and audience (Humphreys, Sanchez de la Sierra, and van der Windt 2013). They represent a powerful tool for the scholar who finds herself in a particular kind of research situation: one in which they seek to test a proposition and know how they want to go about that test before they implement it. Registration and results-blind review provide her with the means to credibly signal it is indeed a *test* that she has conducted. At the same time, scholars seeking to derive theoretical insights from their data or to examine questions for which the pre-specification of tests is impossible will be free to explore without constraint. The greatest benefit would accrue to research consumers, who will be much better positioned to tell the difference between the two.

Moreover, even when test procedures have been pre-specified, this specification need not be treated as constraining: in a non-binding model, researchers remain free to conduct and report analyses that were not announced in advance. What registration and results-free submission make possible is the credible distinction of the confirmatory from the exploratory elements of a set of research findings.

Rather than elevating the status of test-oriented work, wider use of blind-selection mechanisms would make it more difficult for researchers – doing quantitative or qualitative work – to claim to be testing or substantiating a claim when they are not. We would, then, all face greater pressure to identify the inductive sources of our insights and to give exploratory work its due in the process of producing knowledge.

References

Beach, Derek, and Rasmus Brun Pederson. 2013. *Process-Tracing Methods: Foundations and Guidelines*. Ann Arbor: University of Michigan Press.

Bennett, Andrew. 2015. "Appendix." In *Process Tracing: From Metaphor to Analytic Tool*, ed. A. Bennett and J. T. Checkel. New York: Cambridge University Press.

Bowers, Jake, John Gerring, Donald Green, Macartan Humphreys, Alan M. Jacobs, and Jonathan Nagler. 2015. "A Proposal for a Political Science Registry: A Discussion Document." Joint Committee on Study Registration, APSA's Organized Sections for Political Methodology, Qualitative and Multi-Method Research, and Experimental Research.

Burlig, Fiona. 2018. "Improving Transparency in Observational Social Science Research: A Pre-Analysis Plan Approach." *Economics Letters.* 168: 56-60.

Christensen, Darin, Alexandra Hartman, and Cyrus Samii. 2018. "Property Rights, Investment, and Land Grabs: An Institutional Natural Experiment in Liberia." Unpublished working paper. Available online at https://darinchristensen.com/publication/liberia-tenure/, accessed November 8, 2018.

Coffman, Lucas C, and Muriel Niederle. 2015. "Pre-analysis plans have limited upside, especially where replications are feasible." *The Journal of Economic Perspectives* 29 (3):81-97.

Dunning, Thad. 2016. "Transparency, Replication, and Cumulative Learning: What Experiments Alone Cannot Achieve." *Annual Review of Political Science* 19:541-63.

Fairfield, Tasha, and Andrew Charman. 2015. Bayesian Probability: The Logic of (Political) Science: Opportunities, Caveats and Guidelines. Paper presented at Annual Meeting of the American Political Science Association, September 3-6, San Francisco.

Findley, Michael G., Nathan M. Jensen, Edmund J. Malesky, and Thomas B. Pepinsky. 2016. "Can Results Free Review Reduce Publication Bias? The Results and Implications of a Pilot Study." *Comparative Political Studies* 49 (13):1667-703.

Franco, Annie, Neil Malhotra, and Gabor Simonovits. 2014. "Publication bias in the social sciences: Unlocking the file drawer." *Science* 345 (6203):1502-5.

Gelman, Andrew, and Eric Loken. 2014. "The Statistical Crisis in Science: Data-dependent analysis—a "garden of forking paths"—explains why many statistically significant comparisons don't hold up." *American Scientist* 102 (6):460.

Gerber, Alan S., and Neil Malhotra. 2008a. "Do statistical reporting standards affect what is published? Publication bias in two leading political science journals." *Quarterly Journal of Political Science* 3:313-26.

———. 2008b. "Publication Bias in Empirical Sociological Research: Do Arbitrary Significance Levels Distort Published Results?" *Sociological Methods & Research* 37 (1):3-30.

Gerber, Alan S., Neil Malhotra, Conor Dowling, and David Doherty. 2010. "Publication Bias in Two Political Behavior Literatures." *American Politics Research* 38 (4):591-613.

Gough, Ian. 2015. "The Political Economy of Prevention." *British Journal of Political Science* 45 (02):307-27.

Greve, Werner, Arndt Bröder, and Edgar Erdfelder. 2013. "Result-Blind Peer Reviews and Editorial Decisions." *European Psychologist*.

Haggard, Stephan, and Robert R. Kaufman. 2012. "Inequality and Regime Change: Democratic Transitions and the Stability of Democratic Rule." *American Political Science Review* 106 (3):495-516.

Hartman, Alexandra, Cyrus Samii, and Darin Christensen. (2017, October 20). "A Qualitative Pre-Analysis Plan for Legible Institutions and Land Demand: The Effect of Property Rights Systems on Investment in Liberia." Retrieved from osf.io/46r87 on Nov. 1, 2018.

Hartman, Alexandra, Florian Kern, and David T Mellor. 2018. "Preregistration for Qualitative Research Template." OSF. September 27. osf.io/j7ghv.

Humphreys, Macartan, and Alan M Jacobs. 2015. "Mixing Methods: A Bayesian Approach." *The American Political Science Review* 109 (4):653-73.

Humphreys, Macartan, Raul Sanchez de la Sierra, and Peter van der Windt. 2013. "Fishing, Commitment, and Communication: A Proposal for Comprehensive Nonbinding Research Registration " *Political Analysis* 21 (1):1-20.

Ioannidis, John P.A. 2005. "Why Most Published Research Findings are False." *PLOS Medicine* 2 (8):e124.

Kern, Florian G. and Kristian Gleditsch. 2017. "Exploring Pre-registration and Pre-analysis Plans for Qualitative Inference." Unpublished working paper. Available online at https://bit.ly/2NBXNrk.

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

McKeown, Timothy J. 1983. "Hegemonic Stability Theory and 19th Century Tariff Levels in Europe." *International Organization* 37 (1):73-91.

Miguel, E., C. Camerer, K. Casey, J. Cohen, K. M. Esterling, A. Gerber, R. Glennerster, D. P. Green, M. Humphreys, G. Imbens, D. Laitin, T. Madon, L. Nelson, B. A. Nosek, M. Petersen, R. Sedlmayr, J. P. Simmons, U. Simonsohn, and M. Van der Laan. 2014. "Promoting Transparency in Social Science Research." *Science* 343 (6166):30-1.

Monogan III, James E. 2013. "A Case for Registering Studies of Political Outcomes: An Application in the 2010 House Elections." *Political Analysis* 21 (1):21-37.

Neumark, David. 2001. "The Employment Effects of Minimum Wages: Evidence from a Prespecified Research Design." *Industrial Relations* 40 (1):121-44.

Nosek, Brian A. and Daniël Lakens. 2014. "Registered Reports: A Method to Increase the Credibility of Published Results." *Social Psychology* 45(3):137–141.

Nyhan, Brendan. 2015. "Increasing the Credibility of Political Science Research: A Proposal for Journal Reforms." *PS: Political Science and Politics* 48 (Supplement S1):78-83.

Piñeiro, Rafael, Verónica Pérez, and Fernando Rosenblatt. 2016. "Pre-Analysis Plan: The Broad Front: A Mass-Based Leftist Party in Latin America: History, Organization and Resilience." Posted at EGAP Registry, July 19, 2016, https://egap.org/registration/1989.

Piñeiro, Rafael, and Fernando Rosenblatt. 2016. "Pre-Analysis Plans for Qualitative Research." *Revista de Ciencia Política* 36 (3):785-96.

Smulders, Yvo M. 2013. "A two-step manuscript submission process can reduce publication bias." *Journal of Clinical Epidemiology* 66 (9):946-7.

Snyder, Jack, and Erica D. Borghard. 2011. "The Cost of Empty Threats: A Penny, Not a Pound." *American Political Science Review* 105 (03):437-56.

Swaen, Gerald M.H., Neil Carmichael, and John Doe. 2011. "Strengthening the reliability and credibility of observational epidemiology studies by creating an Observational Studies Register." *Journal of Clinical Epidemiology* 64:481-6.

Tucker, Joshua. 2014. "Experiments, preregistration, and journals." On OUPblog. http://blog.oup.com/2014/09/pro-con-research-preregistration/.

Van Evera, Stephen. 1997. *Guide to methods for students of political science*. Ithaca: Cornell University Press.

Wagenmakers, Eric-Jan, Ruud Wetzels, Denny Borsboom, Han L. J. van der Maas, and Rogier A. Kievit. 2012. "An Agenda for Purely Confirmatory Research." *Perspectives on Psychological Science* 7 (6):632-8.

Williams, Rebecca J., Tony Tse, William R. Harlan, and Deborah A. Zarin. 2010. "Registration of observational studies: Is it time?" *Canadian Medical Association Journal* 182 (15):1638-42.