

# Fishing

Macartan Humphreys\*

Raul Sanchez de la Sierra

Peter van der Windt

Columbia University

Columbia University

Columbia University

May 8, 2012

## Abstract

Social scientists generally enjoy substantial latitude in selecting measures and models for hypothesis testing. Coupled with publication and related biases, this latitude raises the concern that researchers may intentionally or unintentionally select models that yield positive findings, leading to an unreliable body of published research. To combat this “fishing” problem in medical studies, leading journals now require preregistration of designs that emphasize the prior identification of dependent and independent variables. However, we demonstrate here that even with this level of advanced specification, the scope for fishing is considerable when there is latitude over selection of covariates, subgroups, and other elements of an analysis plan. These concerns could be addressed through the use of a form of comprehensive registration. We experiment with such an approach in the context of an ongoing field experiment for which we drafted a complete “mock report” of findings using fake data on treatment assignment. We describe the advantages and disadvantages of this form of registration and propose that a *comprehensive* but *non-binding* approach be adopted as a first step to combat fishing by social scientists. Likely effects of comprehensive but non-binding registration are discussed, the principle advantage being communication rather than commitment, in particular that it generates a clear distinction between exploratory analyses and genuine tests.

---

\*Corresponding author: [mh2245@columbia.edu](mailto:mh2245@columbia.edu). Our thanks to Ali Cirone, Andy Gelman, Ryan Moore, and Ferran Elias Moreno for helpful comments. Our thanks to the Population Center at Columbia for providing access to High Performance Computing (HPC) Cluster. This research was undertaken in the context of a field experiment in DRC; we thank the International Rescue Committee and CARE International for their partnership in that research and the International Initiative for Impact Evaluation (3IE) for their support. Humphreys thanks the Trudeau Foundation for support while this work was undertaken.

# 1 Introduction

There is a growing concern regarding reporting and publication bias in experimental and observational work in social science arising from the intentional or unintentional practice of data fishing. The adoption of registries provides one possible response to the problem; but while the potential benefits of preregistration of research designs is easily grasped there has been essentially no adoption of the practice by political scientists. Moreover, even with agreement on registration there may be disagreement on what exactly should get registered and the implications of this practice. These choices are consequential, and merit a focused discussion in the discipline.

In this paper we provide a description of the scope for bias, an illustration of a candidate solution, and a proposal for the introduction of registration procedures in political science.

We begin with a short discussion of the fishing problem (Section 2) before reviewing the state of registration practice in Section 3. Section 3 also provides a set of simulation results that demonstrate the scope for fishing that might exist even in the presence of basic registration procedures such as those used in medical sciences. In Section 4 we discuss a limiting form of comprehensive registration in which there is full detailing of an analysis and reporting plan. We field-tested this approach to registration in the context of a randomized intervention on development aid that we are undertaking in Congo. At the risk of losing freedom during the analysis phase we draft a complete “Mock Report” of findings based on simulated data. As we describe below, this informal experiment revealed both benefits and practical difficulties that researchers may face when attempting comprehensive registration.

After reviewing the Mock Report and the main lessons that it generated, we propose an approach to registration for political science research in Section 5 that involves the creation of a nonbinding mechanism hosted by a third party and enforced by journals. After presenting the proposal we describe possible effects. First we focus on the informational effects of a nonbinding registration regime, arguing that nonbinding registration plays an important *communication* function independent of any *commitment* benefits. Second, considering that the institutional change we are recommending can generate a response in the researcher’s choices, we use a simple model to discuss the kinds of incentive effects (positive and negative) that might arise from a registration requirement. Section 6 concludes.

## 2 Fishing as a Reporting Problem

We begin with a very broad definition of data fishing. Say that there is a set of models  $\mathcal{M}(C) = \{M_1, M_2, \dots, M_n\}$  that could be examined in any given context  $C$ . We use “model”

in the broad sense to cover inquiries, specific tests, hypotheses, or procedures for assessing collections of hypotheses. By context we mean the setting in which data is gathered, which encompasses the cases used as well as the type of data available. Say then that each model is associated with a set of possible *conclusions* and that prior to the implementation of the research the conclusion to be drawn from model  $j$  in context  $C$  is a random variable. We say that a model is “fished” when the decision to report the model depends on the realization of the conclusion.

A few features of this definition are worth noting. First, under this definition both the models that are and are not selected for reporting suffer from the fishing problem, highlighting that the problem lies as much with those results that are not reported as with those that are. Second, the definition we give is deliberately broad in order to highlight that the problem of fishing is not tied to classic hypothesis testing. Thus for example an investigation may seek to assess simply how many sheep there are in a flock. One model might suggest counting the legs and dividing by four, another relies on a headcount. Fishing would occur if the researcher reported only whichever method yielded the largest estimate. Third, the use of complex models that determine how to make use of submodels conditional on the data does not constitute fishing if the models themselves are reported independent of results. For example the researcher might examine the complex model that selects the maximum number generated by the headcount and footcount approaches. Readers might not care for this approach and worry that it is biased upwards, but the result is not fished. This leads to the fourth and most important point. We deliberately distinguish the fishing problem from the problem of assessing whether models are good or not and in doing so we locate the fishing problem in reporting practices (which results end up being reported?) and not in the mode of analysis (for example, whether multiple hypotheses are examined and how any complications arising from this multiplicity are addressed).<sup>1</sup> This distinction is important for assessing the role of registration independent of corrections that might be implemented to account for specification searches. For example consider the complex model in which an investigator proposes to gather 20 outcome measures and declare a treatment effective if a  $t$ -test rejects the null of no effect on at least one of them at the 95% level. Under our definition, this approach, which fails to address the multiple comparisons problem, is not fished; it is simply unusually vulnerable to Type 1 error. Conversely, it is also possible to address a specification search problem in a seemingly reasonable way and still fish; for example the researcher may have 20 measures for

---

<sup>1</sup>We note that our definition differs from data snooping as described by (White, 2000). For White, snooping occurs “whenever a given dataset is used more than once for purposes of inference or model selection.” In this account the problem is located not in the reporting but in the inference; snooping, like data mining, may be a valuable way of undertaking analysis and the challenge taken up by White is to assess not how to avoid snooping but when the results from snooping are reliable.

each of two families of outcomes and employ some method such as the Bonferroni correction in each family, but report only the case that yields the more interesting results.

The problem with fishing is simple: selecting what results get reported can induce bias. If for example classical statistical tests are reported only when they yield “significant” findings, then false positives will be overreported, true negatives will be underreported, and overall conclusions will be wrong.<sup>2</sup> But as highlighted above the problem extends beyond this type of testing and could obtain for example in approaches that focus on “Type S” errors, seeking to make statements of the form “ $\theta_1 > \theta_2$ , with confidence” (see for example Gelman and Tuerlinckx (2000)). This kind of concern has led some to worry that “most current published research findings are false” (Ioannidis (2005)). For evidence of the prevalence of these problems in political science, see Gerber et al. (2001) and Gerber and Malhotra (2008). The evidence provided by (Gerber et al., 2001) of  $p$ -values concentrating just shy of the 5% mark (with a drop in concentration just above it) are consistent with a problem of selective *reporting* and not simply inappropriate *adjustment* of tests to account for multiple comparisons.

The problem, though most often discussed in terms of reporting and publication biases, enters at other stages, if more subtly: during analysis (if for example researchers find themselves more drawn to analyze patterns in those parts of the data where the action is), discussions (if researchers select stronger specification as “preferred specifications” or interpret their findings *ex post*), at the time of submission (filedrawer bias), and in the attention subsequently paid to results by others (where more counterintuitive positive findings get picked up in syllabi or popular and social media).

We note that while fishing may be the product of outright fraud (Callaway, 2011) (or, as termed euphemistically by Glaeser (2006), “researcher initiative”), much more subtle mechanisms may also be at play. One possible cause may be the presence of a kind of inferential error that stems from the attempt to let the data speak before formal analysis. In particular, if more reliable tests are more likely to produce true positives, then one might correctly infer that of two candidate tests the one with the “significant” results is the more reliable one. In that case, *while the inference that the more significant measure is more likely the right one may be correct, basing inferences on the the significant measure only would not be.*<sup>3</sup>

---

<sup>2</sup>Selective reporting does not *necessarily* induce bias. For example if the conclusion of interest is a possibility result—say that that black swans exist—then the conclusion stands independent of the number of unreported failed tests.

<sup>3</sup>Say a researcher has prior belief  $p$  that a proposition is true. The researcher runs two tests,  $M_1$  and  $M_2$ ; each test yields a response which is either positive or negative; that is,  $R_i \in \{P, N\}$ . *Ex ante* the researcher does not know the quality of the tests but expects that with probability  $q$  they are high quality (in set  $H$ ), otherwise they are low quality (in set  $L$ ). Say that for both tests the probability of observing a positive result if the proposition is false is  $\phi$  and that the probability of observing a positive result is  $\psi^H$  if the test is high quality and  $\psi^L$  if it is low quality. Assume that  $\phi$  is low (for example  $\phi = 0.05$ ) and that  $\psi^H > \psi^L$ . Assume types and results are drawn independently. In this case it is easy to show that

### 3 Design Registration

A solution to the problem is to adopt a practice that has been promoted in medical fields (De Angelis et al. (2004)) and requires some form of preregistration of research design. Registration provides a way for researchers to prespecify a set of tests and it allows readers to assess the results of a test in the context of the family of related tests.<sup>4</sup>

#### 3.1 Registration in Medical Studies

In medical studies the shift in norms took place in 2004 when the twelve journals belonging to the International Committee of Medical Journal Editors (ICMJE) announced that they would henceforth require “as a condition of consideration for publication, registration in a public trials registry” (De Angelis et al. (2004)). The ICMJE elected to recognize registries only if they meet several criteria: The registry must be electronically searchable and accessible to the public at no charge; be open to all registrants and not for profit; and have a mechanism to ensure the validity of the registration data (De Angelis et al. (2004)). The focus then was, and to a large extent still is, on experimental studies although a number of journals now encourage (but do not require) the registration of observational studies on a WHO-compliant registry before they begin.<sup>5</sup>

At the time, only ClinicalTrials.gov – maintained by the US National Institutes of Health (NIH) – complied with these requirements. Since then, the WHO’s International Clinical Trial Registry Platform (ICTRP) has developed a network of both primary and partner registers. Primary registers are WHO-selected registers that comply with the WHO’s 20 points of minimal registration requirements. Important elements include 1) Unique trial number, 2) Research ethics review, 3) The medical condition being studied, 4) Description of interventions, 5) Key inclusion and exclusion criteria, 6) Study type, 7) Target sample size, 8) Description of primary outcome, 9) Description of secondary outcomes. Note that the method of analysis does not enter in the list although it may be described under measurement of outcomes.

There has been rapid growth in the last decade in the use of registration procedures. Before the ICMJE policy in 2004 ClinicalTrials.gov contained 13,153 trials (Laine et al. (2007)),

---

$Pr(M_1 \in H | R_1 = P, R_2 = N) > Pr(M_2 \in H | R_1 = P, R_2 = N) \leftrightarrow \psi^H > \psi^L$  The researcher would then be right to conclude that Measure 1 is more reliable than Measure 2. But it would be a fallacy to then base inferences on Measure 1 only. If Measure 1 is really high quality then the chance of a false positive is just  $\phi$ . However if the proposition were false the probability of seeing one positive and one negative result is much higher:  $2\phi(1 - \phi)$ .

<sup>4</sup>There are other benefits to registration, including the ability to assess the population of studies that are never completed or never published. Here however we focus on the benefits of *ex ante* specification of analysis plans.

<sup>5</sup>See, for example, Lancet (2010). A mechanism to report observational studies already exists on many registries. ClinicalTrials.gov, for example, has (as of 23 March 2012) 22,297 observational studies registered.

today<sup>6</sup> it has 123,184 registrations and receives over 50 million page views per month. As the ICMJE editors noted in 2007: “three years ago, trial registration was the exception; now it is the rule” (Laine et al. (2007)).

However, although registration is now the rule, the *quality* of registration in medical research varies substantially. In principle, ICMJE’s individual editors review the data in the registration fields when deciding whether to consider the trial for publication to ensure that fields are not empty or do not contain uninformative terminology. Yet this happens, and many entries in the publicly accessible ClinicalTrials.gov database do not provide meaningful information in some key data fields (De Angelis et al. (2005)). In a recent study (Revez et al., 2010) investigated 265 trials randomly selected from the ICTRP’s search platform for the quality of reporting in trial registries. Their study concluded that the quality of reporting of trial methods in registry records is overall poor. Mathieu et al. (2009) report that of 323 trials examined, only 46% had registered satisfactorily, with others missing outright (28%), registered after the study completion (14%) or with an incomplete description of the primary outcomes (12%). Of the 46%, 31% had evidence of deviations from research plans.

Even when the standards of these medical registries are met however, information about modeling decisions that are routine for analysis in political science is absent (such as subgroup analysis and the covariates to be included in the analysis). Indeed, an examination of the guidelines posted by different registries reveals that guidelines are silent on these points, allowing in principle, leeway for researchers at the analysis stage (see Table C1 in the appendix).

### 3.2 Registration in Political Science and Economics

Although the idea behind registration—that tests should be constructed before they are implemented—adheres closely to standard approaches to research design in political science (see for example King et al. (1994), p23) and despite occasional calls for registration from political scientists (Gerber et al., 2001; Gerber and Malhotra, 2008), actual registration has not taken off as a practice in the discipline.<sup>7</sup>

No major political science journal requires registration; there is no central registry for political science studies, and we know of almost no major studies in political science or economics, experimental or observational, that have been registered. We undertook a search of the fifteen major registries to understand to what extent studies in political science have been registered using these mechanisms. We found no evidence that political scientists are mak-

---

<sup>6</sup> Accessed: 23 March 2012

<sup>7</sup> For related calls in development economics see Rasmussen et al. (2011) and Duflo (2007).

ing use of these registries. In addition we searched in three major political science journals (APSR, AJPS and PA) for any article that had either the name or the abbreviation of one of these registries.<sup>8</sup> We found no articles matching these search terms. To contrast this with the medical field, a search for “clinicaltrials” —the leading registry —yields 243, 446 and 448 hits for BMJ, Lancet, JAMA (See Online Appendix C: Table C1). In addition a search of the WHO centralized database (International Clinical Trials Registry Platform (ICTRP)) for studies containing “politics”, “political”, or “institutions” in the title yielded seven unrelated hits (all for “institutions”), compared to a JSTOR search of political science article titles which yielded 57,776 hits for these terms.<sup>9</sup>

A handful of studies in political economy have involved a form of self-registration in which researchers post their designs in some public space. Strikingly in most of these cases researchers sought to provide considerable detail on estimation strategies. Casey et al. (2011) and Finkelstein et al. (2010) for example archived detailed analysis plans with the Jameel Poverty Action Lab after project start but before analysis. Casey et al. (2011) included an item by item reference between tests and survey elements, and Finkelstein et al. (2010) contained empty tables indicating the exact structure of reporting. In political science King et al. (2009) published an article describing their research strategy and providing a general statement of the specifications to be examined; Monogan III (2010) posted in advance a detailed research design for assessing the effects of policy stances on 2010 election results.

### 3.3 Scope for fishing

Even registration procedures as adopted in medicine can still provide researchers with substantial leeway in practice. How much does this matter? It is well appreciated that if researchers are free to select their dependent or independent variables, then with a reasonably large pool of independent random outcome variables there is a reasonable probability that at least one will show up as significant (Sterling, 1959). In particular, with  $k$  measures the probability of a significant relation at the 95% level is  $1 - 0.05^k$  (see first panel of Figure 1). If results are reported on all  $k$  measures this may be treated as a multiple comparisons problem. But if results are only provided on the significant variable this is a fishing problem. Medical registries seek to combat this problem by ensuring that treatment and outcome variables are clearly specified in advance.<sup>10</sup>

---

<sup>8</sup>When a research design is registered this is generally recorded in the article, including its registration number.

<sup>9</sup>Accessed: 23 March 2012.

<sup>10</sup>The indications provided by the WHO Trial Registration Data Set (Version 1.2.1) on this point are that for each primary outcome, researchers should provide “the metric or method of measurement used (be as specific as possible).”

In fact, however, describing in advance the key treatments and outcome measures only partially addresses the scope for fishing. At the time of analysis, researchers make many other decisions regarding the use of covariates, the precise definition of variables, the examination of subgroups, and the precise statistical model to be used. In the language of our definition of fishing, each of these decisions results in a different model that may be examined and the selection of which models to present may give researchers discretion to produce the results they want.

To examine the scope for fishing we focus on classical hypothesis testing and conduct a suite of simulations to assess the chances that a researcher can produce false positive results given different types of discretion.<sup>11</sup> In each case we assume a simple data generating process in which  $n$ -dimensional  $x$  and  $y$  vectors are independently drawn from a standard normal distribution. For each draw we then examine the pool of results that a researcher could report given different forms of discretion and select positive results whenever possible. The share of simulations in which positive results are obtained is then our measure of the scope for fishing.<sup>12</sup>

We begin by examining the effects of discretion over the choice of covariates. In principle, the lack of prespecification can give rise to *complete* discretion over reported results in the sense that for any  $z$ ,  $y$ , and  $\hat{b}$  there exists a  $z$  such that a regression of  $y$  on  $x$  controlling for  $z$  yields a statistically significant estimate  $\hat{b}$ . In particular  $z = y - \hat{b}x$  will yield the desired result. However whether a researcher can find such a  $z$  is another matter. The second row of Figure 1 shows that for small data sets ( $n < 50$ ) there is in general considerable scope for fishing when researchers have discretion over covariates, but that this discretion declines rapidly with the size of the data set. With 200 observations, restricting attention to a model with only one covariate allows only a 12% chance of successful fishing given 1000 random covariates to choose from. In general if researchers can employ more than one covariate they have considerably more discretion since for  $k$  possible covariates they have  $2^k$  potential models to choose from; a number that increases rapidly with  $k$ . We estimate that for an  $n$  of 25, access to 16 random covariates yields a successful fishing probability of 56% when these covariates can be entered simultaneously; this falls to 15% with an  $n$  of 100.

A second approach is to fish by focusing on sections of the data. For example, a researcher might find no positive effects overall but notice that there are significant effects among male subjects. In this case the researcher might report results for this particular way of subdividing the data, or may even restrict attention and discussion to this subgroup. The third row in

---

<sup>11</sup>We note that independently Simmons et al. (2011) examined the effects of discretion in selecting dependent variables, sample size as a function on results to date, adding one covariate and/or an interaction, and reporting only subsets of experimental conditions.

<sup>12</sup>In this analysis we focus on two sided tests and fish for significance at the 95% level. For more detail on the methods used see the accompanying replication file.

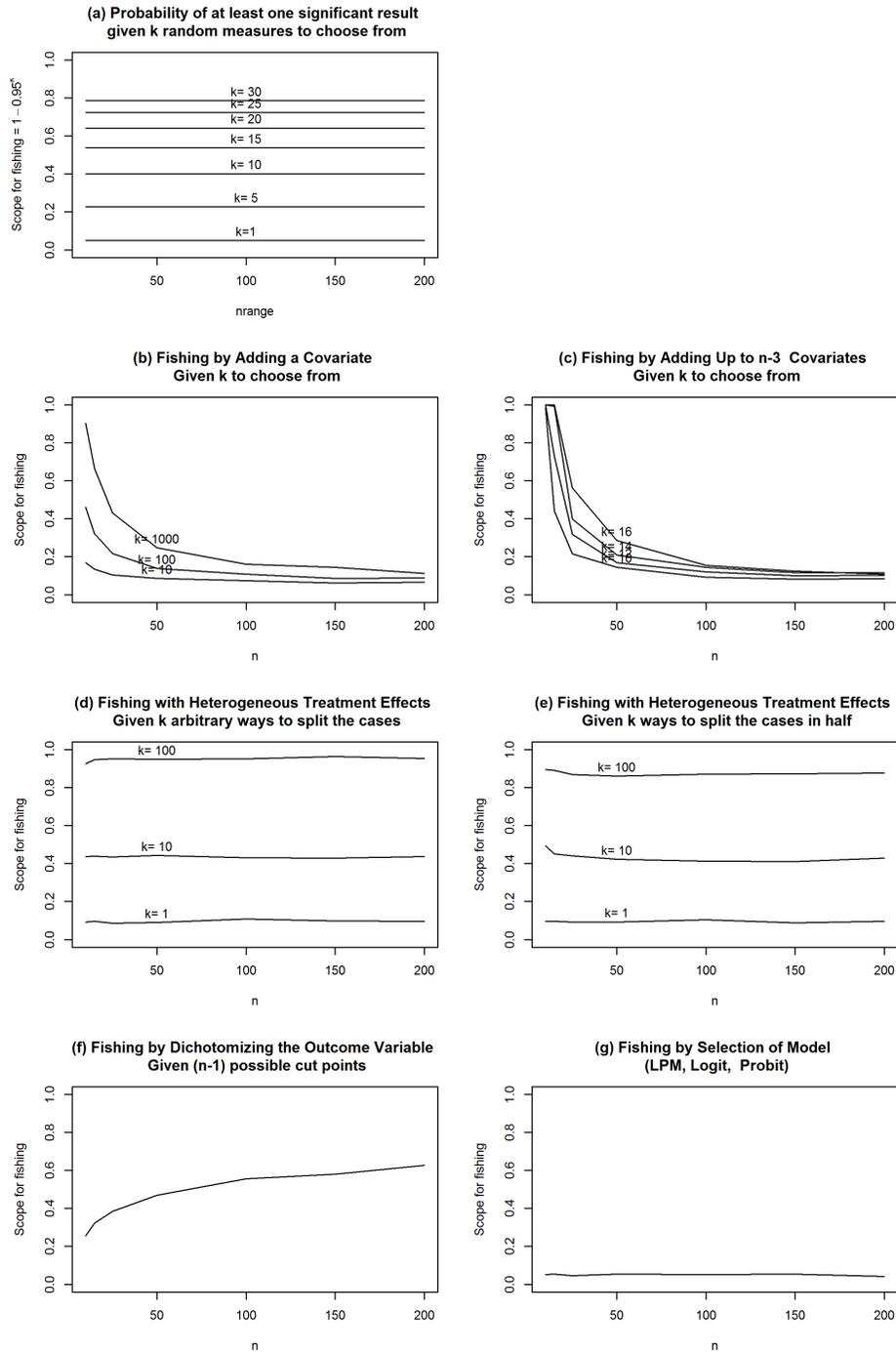


Figure 1: Each subfigure shows the chances with which researchers can produce false positives given different ways of altering analysis plans *ex post*. Each point in panels (b), (d), (e), (f) and (g) is based on 4,000 simulations. Points in panel (c) are based on 1,450 simulations for  $k=10, 12$  and  $14$  and 514 simulations for  $k=16$ .

Figure 1 focuses on this form of fishing for the cases where researchers have  $k$  arbitrary (and independent) ways of dividing the data in two groups of any size or into two groups of equal size. In both cases we find considerable scope for fishing and that the scope does not decline appreciably with sample size. Fishers face a 40% chance even restricting to 10 arbitrary ways of splitting the 200 observations in two.

A third approach to fishing is to select cutoffs for dichotomizing outcome variables. For example, a variable may take levels of the form “No education”, “Primary school education”, or “Secondary Education or above” and researchers, perhaps uncomfortable interpreting averages for such variables, might elect to dichotomize them into outcomes such as “Some education” or “Secondary education or above.” This choice of cutoff provides latitude and the lower left panel in Figure 1 suggests that this latitude can greatly increase the scope for fishing – moreover, the latitude increases with larger datasets (which since we are examining continuous outcomes variables provide more potential cutoff points). With  $n = 200$  a researcher has a 62% chance of getting significant results. There are many other ways of transforming variables and some of these may also provide similar latitude.

Finally researchers often enjoy some latitude in selecting models. There are many such choices available. The lower right panel in Figure 1 shows that for binary variables, discretion over the use of linear, logit or probit models does not provide latitude for fishers.

We have seen that these different areas of discretion sometimes provide considerable latitude for fishing even in situations in which the basic hypotheses and the general outcome of interest is predefined. We now discuss a strategy for eliminating this latitude.

## 4 Comprehensive Model Registration: The Mock Report Approach

A simple solution to these problems of data fishing that is consistent with standard ideas of theory testing in political science is the following. Write down in advance two complete versions of an article, with identical hypotheses but with two different analysis sections to be used in the event of positive or negative findings. Place them in envelopes, and wait for the data to come in to determine which envelope should be sent out for publication. This comprehensive prespecification of analysis and interpretation would completely remove the scope both for fishing for results and for reinterpreting tests only after the results are in.

But this proposition may be more problematic than it seems at first blush. We field-tested this kind of comprehensive prespecification in the context of the study of a major randomized intervention on development aid that we are undertaking in Eastern DR Congo, writing in

advance a complete mock report of our findings before we had any. As we describe below, this experiment in mock report writing, though we think unique in the discipline, falls short of the ideal described above in a few ways. Nevertheless it highlights the promises and pitfalls of this approach, which may shed light on the slowness to adopt registration procedures in political science.

We began work on a field experiment to examine the effects of a Community-driven Reconstruction program (Tuungane) in the Democratic Republic of Congo (DRC) in late 2006. Community-driven reconstruction programs are a major vehicle for delivering international development aid. Of interest to students of the political economy of development, they seek not only to deliver aid but to do so in a way that alters structures of local governance through the development of local institutions of accountability. These programs are now the subject of a number of experimental studies by political scientists and economists (Beath et al. (2011), Fearon et al. (2009), Casey et al. (2011)).

These studies seek to address questions of social scientific interest but are also closely tied to policy debates. They are often undertaken together with development agencies in order to help inform development strategies and they involve manipulations that affect allocations of benefits to hundreds of thousands, sometimes millions, of people. The scale of the projects, the many stakeholders involved—including both the populations and the international organizations—renders the onus to generate credible results particularly acute.

For the Congo study the core set of hypotheses were developed prior to program launch, shared with and agreed to by stakeholders (the research team, the implementing organization, and donors), and included in the research team’s submission for Institutional Review Board (IRB) approval.<sup>13</sup> They were not, however, submitted to a public registry.

As highlighted in section 3, even with such advanced identification of hypotheses there is still considerable scope for reporting biases to enter through the choice of measure and specification. The scope for fishing in studies like ours is greatly enhanced by the complexity of the treatment and outcome measures: given the large number of measures and room for discretion, the field for fishing can be wide.

To reduce these risks we developed in January 2011 a detailed analysis plan linking each hypothesis to specific items in our surveys and other data collection instruments, described the general class of specifications we would use for all analyses, and then produced code to produce the core analyses.

We then generated a “mock report” using real outcome data gathered in the first months of data collection but fake treatment data – fake in that we purposefully “scrambled” our

---

<sup>13</sup>See [http://www.columbia.edu/~mh2245/DRC/DRC\\_DESIGN.pdf](http://www.columbia.edu/~mh2245/DRC/DRC_DESIGN.pdf).

treatment indicator, thereby deliberately employing a false indicator of whether a village took part in the program or not.<sup>14</sup> As a result, we developed the analysis code and selected measures prior to actually looking at the results.<sup>15</sup> Once completed, we shared the mock report with our implementation partners, so that we could agree on a final design without reference to results, and protect the research from biases that could emerge from the real data. We then posted the mock report online and circulated it to the EGAP network of researchers and practitioners working in this area.

What did we learn from the exercise?

#### 4.1 Loss of latitude

Drafting the mock report forced us to take early action on the plethora of small decisions that can provide the latitude for fishing (intentional or not) of the form described in the section 3. Producing final tables requires pinning down which outcomes to use to test which hypotheses. But this exercise also removed latitude in other ways. To illustrate some of the choices that resulted from the exercise we provide summaries of Tables from the Mock Report in Table 1 below.

First, it required absolutely precise definitions of dependent variables and choices over dichotomizations. For instance, when measuring inequality one faces a wide array of possible options; in Mock Table 23 (see the first row of Table 1) we opted for the standard deviation of benefits received. Second, in some cases we dichotomized dependent variables (for example, to determine whether a set of decision making procedures should be classed as being participatory or not); in Mock Table 6 we elected two distinct dichotomizations, both of which we committed to reporting in the final report. Third, it forced a selection of covariates. Although in later analyses we are committed to including a set of prespecified covariates, we opted for the simpler analysis without covariates for the main policy report. For some tests however covariates were envisioned as particularly important and included without reference to how they alter the significance on the coefficient of interest. Thus in Mock Table 15 we were interested in the use of complaint mechanisms by populations; but we wanted to condition the effect on whether populations had something to complain about. Fourth, the exercise forced a selection of subgroup analyses. Whereas in most cases we were interested in average treatment effects, for some outcomes gender breakdowns are of particular policy interest—this was the case for

---

<sup>14</sup>See [http://cu-csds.org/projects/postconflict-development-in-congo/20110307\\_drc\\_registration](http://cu-csds.org/projects/postconflict-development-in-congo/20110307_drc_registration)

<sup>15</sup>In fact, the idea of conducting analysis using fake treatment data – rather than working on a subset of data—came somewhat late to us, and so unfortunately for a small initial set of measures the first drafts of tables used actual, though of course very incomplete, data. Data was gleaned from the initial stages of data collection in only two out of the 4 provinces slated for the evaluation and represented between about 2% and 15% of the data on various measures.

example for school attendance outcomes, reported in Mock Table 35.

Beyond the loss of latitude in constructing individual tests there is also a loss in latitude in determining how to summarize the results. In our Mock report, we summarized all results in a single table that reported one-sided tests of the strength of support received by the hypothesis (not shown here). In the spirit of one-sided testing the summary does not report whether there are ‘negative and significant’ results, although subsequently we did elect to flag suggestive evidence of adverse effects. As discussed below, the report also does not provide a summary of how we will interpret different *patterns* of results—for example, would the project be considered a success if there were positive outcomes in one set of measures but not another?

Measure	Average outcome	Treatment Effect	SE	N	Notes
Benefits Spread		3.24	3.54	11	Mock Table 23: Mean deviation of benefits distributed. Based on item QR 3.
Projects selected by elections	0.38	-0.05	0.09	124	Mock Table 6:
Projects selected by elections or lottery or consensus	0.77	0.01	0.08	124	Effect on Project Selection Mechanisms. Based on item B33.
Total # of Private Complaints <i>conditional</i> on quality		-0.21	0.90	62	Mock Table 15: Effect on Citizen Complaints. Based on item QR 26.
Days of school attendance (Overall)	2.87	-0.2	0.82	154	Mock Table 35:
Days of school attendance (Girls)	2.35	0.19	1.10	75	Attendance in previous 2 weeks.
Days of school attendance (Boys)	3.39	-0.62	1.21	79	Based on items QF 7 and QF 14.

Table 1: Sample of results from our DRC “Mock Report”. Different tables required making different decisions regarding how to define measures, how to condition, and how to subset.

## 4.2 Reflections from field-testing the mock report model

Beyond providing a stronger basis for assessing the credibility of future findings, there were a number of immediate and clear benefits from the exercise. Two stand out: first was the ability to reach agreement with partners on exactly what results should be reported and what will constitute evidence of success in different areas. In an environment in which consumers of research have strong interests in the outcomes, locking in a design at this level of detail provides strong protection against the risks of pressures to alter questions in light of findings.<sup>16</sup> A more practical benefit is that this sort of exercise helps to identify flaws and gaps in measure design before it is too late (although for this purpose the analysis would have more usefully been run prior to any data collection).

This exercise also revealed difficulties however. We focus on three.

<sup>16</sup>Ultimately our partners requested two additional measures subsequent to our posting of the mock report, although these requests were still not based on any information regarding outcomes on the measures.

### 4.2.1 Too many potential findings

The first problem that the exercise highlighted was that the ideal described above of writing down different versions of a report as a function of different patterns of findings can be entirely impractical. Drafting two versions of a write up may work in the case where a study seeks to test a single, simple hypothesis but for more complex designs the universe of potential findings is too large to make this feasible. For example with  $n$  related hypotheses, and results on each hypotheses coming out as positive or negative there are  $2^n$  possible patterns of findings; if one admits positive, negative, and indeterminate findings there would be  $3^n$  possible patterns— with fifteen hypotheses one would be busy drafting over fourteen million versions.<sup>17</sup> The solution to the problem of too many potential findings might be to focus on simple questions only—or forsake the interpretation of patterns in favor of a focus on single items. The problem revealed here however is a deep one, but one we think is common in political analyses and may stem in part from the complexity of the subjects we study: even though we specify a set of individual hypotheses *ex ante* we nevertheless often engage in post hoc theorization to make sense of the whole collection of the findings. This problem is quite dramatically exposed in the Congo experiment by the difficulty of writing up an interpretation of a random set of findings. An advantage of the mock report in our case was that it clarified that whereas we could reasonably register how we would interpret results from tests of individual hypotheses our interpretation of overall patterns should be rightly thought of as *ex post* theorizing.

### 4.2.2 The loss of the option value of waiting

A second problem that this exercise revealed is the loss of the option value of waiting. A common, perhaps standard, sequence of research in political science is to generate a general theory, gather data, and only then develop specific measures and tests, with much of the hard thinking and creative work taking place during the analysis phase. A registration process switches the order of the last two steps.<sup>18</sup> In doing so, researchers lose an option value of waiting. The idea of an option value of waiting is that various types of uncertainty get resolved over time, including uncertainty about the best methods for analyses or the state of substantive knowledge in the area. Moreover, opportunities also arise during implementation; for example researchers may gain new insights or it may become possible to collect more measures or access more data than previously envisioned.

---

<sup>17</sup>Note the problem here is not the well-known multiple comparisons problem since the interest here is in how to interpret a collection of independent findings rather than how to assess the validity of particular findings given a collection of findings.

<sup>18</sup>The International Committee of Medical Journal Editors requires registration to happen prior to enrollment of subjects; this is a very natural deadline, but one that we missed by about three years.

There are other reasons why risk might make researchers reluctant to change the sequence of research. First there may be risk over whether projects will be implemented: researchers may be reluctant to invest in detailed analysis planning and write up in a risky environment in which projects often fail. Second there may be uncertainty over whether a project makes sense. For a complex process some analyses may be contingent on others; for example tests intended to address the question of why  $A$  caused  $B$  presuppose that  $A$  does indeed cause  $B$ , something that might not be known at the start of a study.

A strict commitment to registered designs might come at the cost of giving up these options. But, as we will argue below, a registration requirement need only create a distinction between pre-analysis and post-analysis hypotheses, it need not preclude post-analysis hypothesizing. Thus for example in a study of health policy in Oregon, Finkelstein et al. (2010) implement a number of tests that were not originally included in their design—but they highlight when and where they do this.

### 4.2.3 The difficulty of specifying analyses without access to data

A third class of problems, similar to the loss of the option value of waiting, relates specifically to features of the final data. The analysis undertaken may ultimately depend on features of the data being analyzed. For example, it is a common practice to undertake balance tests and then control for variables that appear unbalanced (see Mutz and Pemantle (2011) for arguments against this approach). If one were to use such an approach one could not specify the precise controls to be used in advance. An alternative approach is to specify a set of controls and introduce them whether or not there is balance, although in fact whether or not this can be implemented may depend on the data (in the extreme case a covariate may be collinear with treatment). Similarly it is common to dichotomize or otherwise transform data using information about the distribution of variables, which may not be available at the time of registration.<sup>19</sup>

In these cases a solution is to specify at a more abstract level the procedure for determining the model specification. In line with the discussion in section 2, the separate models can be combined into a complex model that specifies the conditions under which different procedures are employed. But that solution may not be possible in general if the set of potential features of the data is not known. Indeed, many things may go wrong that can lead to model changes

---

<sup>19</sup>In our mock report we did use such distribution information in some instances to define variables; without using information regarding the relationship between treatment and outcomes. However this approach has problems of its own, in particular the distribution of outcome variables contains information about treatment effects—for example if an outcome variable has zero variance then estimated treatment effects will also be zero. If such information on distributions is to be used it is preferable from this perspective to use data from one group only (for example, the controls) or to use data on independent variables only.

in the analysis phase. *Ex ante* one may not know whether one will suffer from non-compliance, attrition, missing data, or other problems such as flaws in the implementation of randomization, flaws in the application of treatment, errors in data collection, or interruptions of data collection. Any of these possibly unanticipated features of the data could require fixes in the analysis stage. In each particular case one could in principle describe precisely how to handle different data structures, but in the absence of an off-the-shelf set of best practices for all these issues, such efforts towards complete specification is likely to be onerous.

Again, a registration requirement need not prevent changes in analysis plans in response to new information, rather it serves to highlight when and where they occur. Thus in their study of Sierra Leone, Casey et al. (2011) dropped some measures from analysis due to weaknesses in measures; but when this occurred, these cases were noted by the researchers.

## 5 Recommendation and Implications

### 5.1 Proposal

Past research has found strong evidence of data fishing and/or publication bias in political science (Gerber and Malhotra, 2008). We have demonstrated here that data fishing may remain a concern even when researchers prespecify basic hypotheses and dependent and independent variables. This is a problem for both experimental and observational work. Registration procedures as employed in medical fields impose relatively limited constraints on researcher latitude in specifying analysis and interpretation plans post analysis; a mock report approach however removes this latitude entirely. But a mock report also comes with costs if it prevents researchers employing analyses that deviate from prespecified plans but with good reason.

How should political science respond to these concerns? We believe that the correct model has the following five elements:

1. *Scope*: Preregistration is adopted as a norm at least for all experimental and observational research that (a) claims to provide a test of a theory or hypothesis and (b) employs data not available to researchers at the time of registration;
2. *Comprehensiveness*: Preregistration of key tests should be *comprehensive* in the sense of providing full analysis plans, for example in the form of mock tables or through the submission of mock data and analysis code; interpretations of patterns of findings should be included in registration insofar as possible;
3. *Deviations*: Deviations from registered plans should not preclude publication but should be indicated by authors; differences between results from the modified plans and the

registered plans should be reported where possible;

4. *Incentives*: The registry or registries should be housed by a third party; leading journals in the discipline should ‘recognize’ the registry and provide recognition for research that is registry compliant;
5. *Maintenance*: In addition to providing a repository function, a registry should also be managed: it should ensure basic criteria are met and maintain records of summary results of analyses, including a recording of whether the research was successfully undertaken in the first place.

We discuss each of these elements in turn:

**Scope.** The proposal has two key scope conditions. First our proposal is for studies—or parts of studies—that claim to be engaging in hypothesis testing. The intention of this limitation is to exempt inductive theory generation and exploratory analyses for which registration may in practice be impossible. Indeed a very large share of research in political science is not about hypothesis or theory testing, but rather theory generation. Thus, Collier, Seawright, and Munck note that “the refinement of theory and hypotheses through the iterated analysis of a given set of data is an essential research tool” for qualitative researchers, and that quantitative research “routinely involves an iterated, partly inductive, mode of research” (Collier et al., 2010). We think that Collier and colleagues are correct on this characterization of research practice; the concern, however, is that within these studies it can often be difficult to determine whether particular components are intended as tests or as explorations. Registration procedures could help delineate these activities.

Second, the proposal focuses on studies for which the data is not available to the researcher at the time of registration. While this includes both experimental and observational work that involves original data collection, it does not cover much historical analysis except insofar as they employ fresh data. The reason is that in these cases theories are often generated with considerable knowledge of data structures and the advantage of registration is less clear.

**Comprehensiveness.** Our discussion of the Congo mock report as well as our simulation exercises highlight the value of comprehensive prespecification of tests. Full clarity over analysis plans could be provided by posting dummy replication data and code at the time of registration. For instance code could be provided that generates the anticipated data structure, possibly combining existing and simulated data, and additional code could then provide the exact mapping from data structure to output (for example as an Sweave file).

**Incentives.** In practice the procedures we advocate require some kind of institutions and enforcement mechanisms that are presently absent. We recommend the creation of an

independent registry hosted by an institution or institutions with a long time horizon such as a professional association, the National Academy of Sciences, or a major library. The housing problem is very distinct from the compliance problem however. For political science we believe that a set of journals should take the lead on compliance, as was done by medical journals, in promoting registration.<sup>20</sup> Medical journals made registration mandatory, from a given date forward. This approach, we believe, is not likely to be implemented in the short run in political science, and moreover there may need to be a learning phase in which a suitable structure for a registry for political science research is developed. In the short term we recommend the creation of a registry coupled with (a) voluntary adoption of registration by researchers and (b) some form of recognition by journals of research as registered, perhaps as part of the review process or by using a certification of which articles report registered designs and whether deviations occurred.

**Deviations.** The proposal also seeks to highlight the informational value rather than the constraining value of registration. Although we have argued for the merits of comprehensive registration, our experience and that of others (Casey et al. (2011) and Finkelstein et al. (2010)) suggest that registration requirements should not preclude the alteration of designs. Value still remains in using registration to distinguish between anticipated analyses which are clearly not informed by outcome data, and unanticipated ones, which may be. We propose that when such deviations arise they be highlighted and the effects on results reported, so that readers can assess the merits of the alterations and the credibility of findings. In the study by Mathieu et al. (2009), of 23 studies that deviated from design, 19 did so in ways that led to significant findings. While this number seems high, explicit discussion of such deviations could allow a determination of whether this pattern reflects fishing biases or improvements in sensitivity of tests.

**Maintenance.** As advocated by others (Rasmussen et al., 2011) the benefits of registries would be augmented by a maintenance of records of research output. This is especially important for research that is not completed and/or not published in order to combat filedrawer and publication bias (independent of any analysis bias). Other maintenance tasks could include removing items that do not meet basic registration requirements and reporting to journals whether submitted research is consistent with the registered design (or correctly indicating if not).

We close this section with a note on discovery. Nothing in this proposal is intended to limit discovery processes. We note first that it is possible to register research that has a strong discovery component. Some discovery problems can be framed as multiple comparisons

---

<sup>20</sup>An alternative is to make funding of various forms conditional on registration.

problem or specification searches. A researcher interested in the hypothesis that a treatment affects economic wealth may be unsure of which of  $k$  measures is most accurate. Rather than precommitting to the most promising measure the researchers could examine all measures and use various approaches to deal with the multiple comparisons problem in the prespecified analysis, such as using an index of measures (the approach adopted by Kling et al. (2007) for example), statistical adjustments, such as the Bonferroni or Šidák corrections (see also Benjamini and Hochberg (1995) for methods to control the false discovery rate), or multilevel approaches (Gelman et al., 2009). Similarly a researcher might examine multiple models and register the procedure for selecting among them. For such approaches White (2000) provides a “reality check” for testing the hypothesis that the best model developed during a specification search is no better than a benchmark model (see also Hansen (2005) and related work). Or researchers could engage in partitioning data sets into training data and testing data and register the procedure *ex ante* or register between training and testing.

Second, less formal approaches to discovery may not be amenable to registration but, as with deviations, this need not preclude publication in any way. Indeed within a single article there may be components that have been registered and others that have not and among the registered components there may be elements that involve deviations and others that do not. The critical thing is the development of signposting so that readers have clarity regarding the status of different types of claims.

What are the likely effects of such a proposal? While the aim is to limit data fishing, such a proposal may have implications for the motivations of researchers and the types of research that is undertaken. We discuss the informational benefits of such a proposal, and other possible effects next before concluding.

## 5.2 Incomplete Compliance

In response to the class of difficulties that researchers in political science are likely to face our proposal allows for incomplete compliance with registered designs. What would be the implications of a system that allowed incomplete compliance?

To gain some insight into how gains may arise from the informational effects of registration even in the absence of commitment effects we consider a setting in which researchers’ analysis decisions are not affected by registration requirements (and publishers’ decisions are not affected by these or by the nature of results). We again use a classical hypothesis testing framework. Assume that some proposition,  $P$ , is true with probability  $q$ . Say that some orthodox test of the proposition would yield a positive result with probability  $p_T^o$  if in fact  $P$  were true and  $p_F^o$  if  $P$  were false. Say now that some share  $\beta$  of researchers engage in data fishing

and favor models that obtain positive results. Say that for such an enterprise, researchers would generate positive result with probability  $p_T^f$  if in fact  $P$  were true *and the orthodox test did not yield a positive result* and  $p_F^f$  if  $P$  were false and the orthodox test did not yield a positive result. Consistent with the low value placed on compliance we assume that these researchers publish noncompliant research when they can generate positive results, but if not they publish using the orthodox model. Say that the remaining  $1 - \beta$  do not engage in fishing but they do engage in learning and may seek, with probability  $\gamma$ , to change their analysis plan post-registration in response to new learning. Unlike the fishers, these researchers do not alter plans as a function of results, but rather for reasons exogenous to the results, for example, the generation of better methods or a realization that the data has previously unanticipated features unrelated to treatment effects. These researchers would then employ a modified test that generates positive result with probability  $p_T^m$  if in fact  $P$  were true *and the researchers decided to alter the analysis plan* and  $p_F^m$  if  $P$  were false and the researchers decided to alter the analysis plan. For this analysis we assume that  $p_F^o = p_F^m$  and  $p_T^o < p_T^m$ , that is the modified test is equally *specific* but more *sensitive*.

With these motivations, we can calculate the share of published results that will be positive (or, the probability that a result will be positive) and the positive and negative predictive values of the tests. Critically, in the case of partial registration we can estimate these rates conditional on whether research falls into a registration compliant group (no deviation from registered plans). Figure 2 illustrates these quantities for particular values of  $p_k^j$  and allows for a comparison of outcomes under different regimes; the quantities are also provided in the Appendix.

A number of features stand out.

1. We find that the probability that a positive published result is false in the complier group is the same as the probability that a positive result would be false under a binding registration regime. Moreover this rate does not depend on the share of fishers ( $\beta$ ) or the likelihood of discovery of better methods ( $\gamma$ ). In both cases the likelihood that a positive result is false is simply  $\frac{(1-q)p_F^o}{qp_T^o + (1-q)p_F^o}$ .
2. For any positive number of fishers ( $\beta > 0$ ), the probability that a negative published result is false in the complier group is *lower* than the probability that a negative result would be false under a binding registration regime if and only if fishing after a negative result is more likely to generate positives if the proposition is actually true ( $p_F^f < p_T^f$ ). This results in a disproportionate removal of false negatives from the complier group.
3. The probability that a negative published result is false in the noncomplier group is *lower*

than the probability that a negative result would be false under a binding registration regime if the modified tests had the same false positive rate but lower rate of false negatives ( $p_F^o = p_F^m$  and  $p_T^m > p_T^o$ ).

4. If there are no researchers who engage in fishing ( $\beta = 0$ ), then results in the complier group are identical to what would arise under binding registration, but results in the noncomplier group would be more reliable. This illustrates that binding registration comes at a cost: it prohibits modifications to the analysis plan that are justified for reasons exogenous to the results. The gains from the reduction in fishing that a binding registration would generate must compensate for the loss in research quality that it induces by imposing a constraint on reasonable modifications.

However we also find:

1. The probability that a result will be positive is higher in the noncomplier group than would be the case under binding registration. The increased probability is partly due to an increased number of false positives if  $p_F^f > 0$ .
2. For  $\beta > 0$  the probability that a result will be positive is also higher in the *complier* group than would be the case under binding registration. This is because fishing turns results that might otherwise have been negative into noncompliant positive findings.

Finally, note that the requirement that deviating researchers publish analyses based on both prior plans and modified plans would allow still finer inference as well as a recovery of what would occur under a binding scheme.

In our proposal we recommended that in the first instance registration processes be not simply nonbinding, but also voluntary. How might the voluntary nature of registration affect interpretations of published results? For simplicity consider a voluntary but binding mechanism. We bracket until the next section a discussion of changing incentives of researchers. For such a voluntary mechanism if the registration decision is unrelated to unobserved features of a study question then, conditional on study type, the error rates and share of positive findings among the pool that registers *would be the same as would occur under mandatory registration*, independent of how potential fishers select in or out of this pool. However, the collection of findings among non-registered studies would differ from the findings that would obtain in the absence of any registration procedures by virtue of the fact that registering researchers select out. If researchers that select to register are also those least likely to engage in fishing then the reliability of the pool of non registered studies would become more suspect.

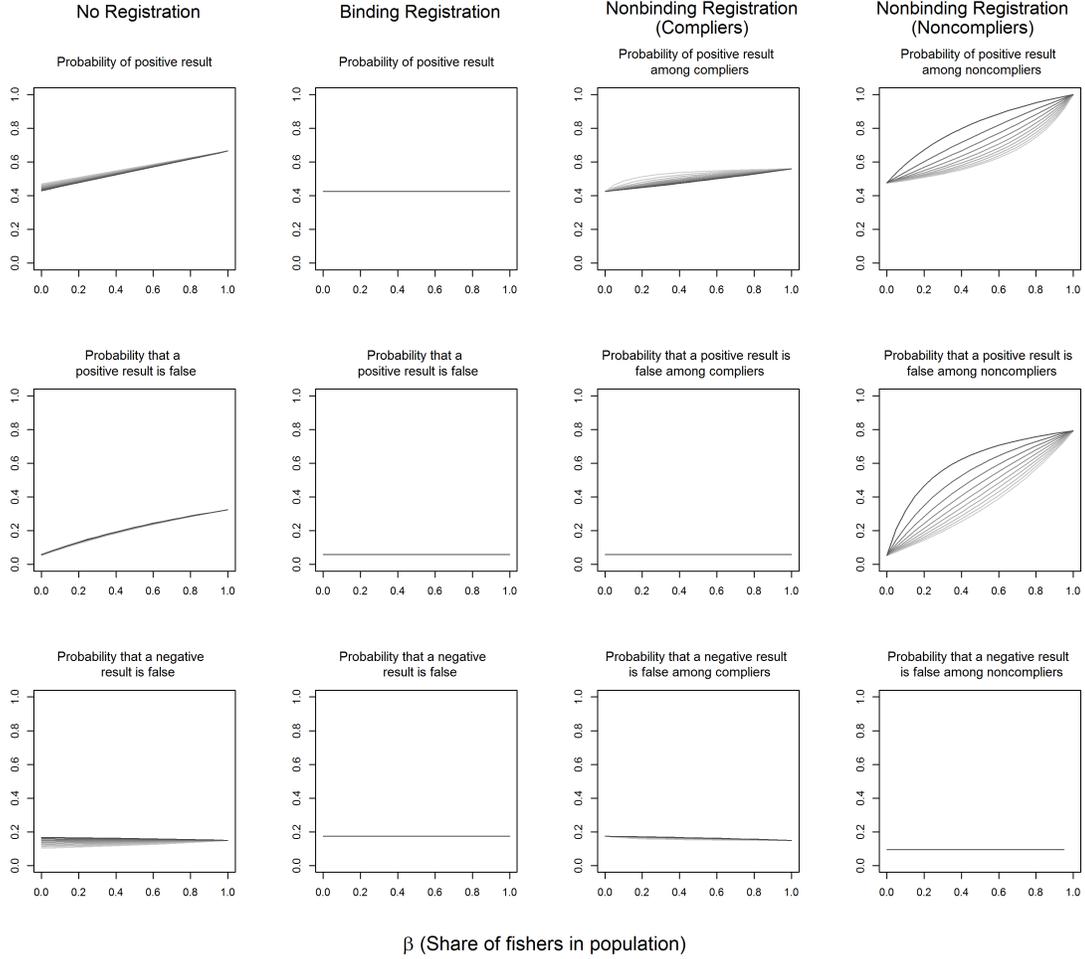


Figure 2: Findings and errors given three different types of registration compliance regimes: no registration, binding registration, and nonbinding registration (yielding compliant and noncompliant groups). In each graph the darkest curve indicates the case with  $\gamma$  low (.1) and the lightest curve indicates the case with  $\gamma$  high (.9). Parameters are set here to:  $p_T^o = 0.8$ ,  $p_T^m = 0.9$ ,  $p_T^f = 0.5$ ,  $p_F^o = p_F^m = 0.05$ ,  $p_F^f = 0.4$ . and  $q = 0.5$

The overall conclusion is that benefits from registration can still be achieved even with an imperfect compliance mechanism and voluntary mechanisms because the mechanism allows readers to condition interpretations of findings on whether or not there was registration and/or compliance. This ability to make conditional assessment would be impossible without registration. For some results the inferences would be as reliable or more so than under a binding regime. We emphasize, however, that the analysis above is undertaken under the hypothesis that the incentives of researchers do not change. One could imagine instead that, in order to separate themselves from the successful fishers, a set of nonfisher researchers decide to comply with preregistered designs even though they believe they have a better method

*ex post*. It is also possible that under such conditions fishers too would seek to comply; in that case although the nonfishers would be using suboptimal results, the outcomes would be identical to those under a full compliance model. In short, given the changes in incentives that might result from registration, the results above should be regarded as benchmark quantities. In the next section we turn to discuss ways that incentives might also change as a result of registration rules.

### 5.3 Incentives

Publication rules alter researchers' incentives before the analysis phase, and could substantially affect their choices about the type and quality of research they choose to undertake. We provide a simple framework to help think through how registration might affect some research choices at a stage earlier than analysis.

Say a researcher's prior belief about a treatment effect,  $\beta$ , is distributed uniformly over  $[\bar{\beta} - a, \bar{\beta} + a]$  for  $\bar{\beta} > 0$ . Assume that  $\bar{\beta} < a$  and so although researchers expect positive effects, negative effects are possible. The researcher is considering the implementation of a study with  $n$  subjects to test the null hypothesis that  $\beta \leq 0$ . Let the (convex) cost to the researcher of implementing the study be  $c(n)$ , the value to the researcher be  $\bar{u}$  if the null is rejected, and  $\underline{u}$  if not. If the true treatment effect is  $\beta$ , then the probability of a positive result is the power of the test, which we denote as  $\pi(\beta, n)$ . The researcher decides both whether to undertake the study and the optimal  $n$  for the project.

How might registration rules alter the optimal choice of  $n$  and the pool of research undertaken?

Consider first the choice of  $n$ . The researcher chooses the  $n$  that maximizes:

$$U(n) = \int_{\bar{\beta}-a}^{\bar{\beta}+a} (\pi(\beta, n)\bar{u} + (1 - \pi(\beta, n))\underline{u}) f(\beta) d\beta - c(n) \quad (1)$$

One can show that  $U(n)$  is a concave function for the range of parameters examined here.<sup>21</sup> Let  $n^*$  denote the solution to this problem. Comparative statics then give  $\frac{\partial n^*}{\partial \underline{u}} < 0$ .

Hence, the greater the value placed on a negative, the smaller the optimal sample.

Consider next the decision whether to undertake the study in the first place. Suppose the researcher undertakes a study if and only if  $\bar{\beta} > \bar{\beta}^*$ , where  $\bar{\beta}^*$  is an expected effect size that satisfies:  $V(n(\bar{\beta}^*), \bar{u}, \underline{u}, a, \bar{\beta}^*) = 0$ . Here  $V$  is the expected utility of a study given that the researcher chooses the optimal  $n$  as discussed above. One can show that  $\frac{\partial \bar{\beta}^*}{\partial \underline{u}} < 0$ , i.e. that the mean of the expected prior value of  $\beta$  required by the researcher to undertake the study

<sup>21</sup>See online appendix for proofs of propositions in this section.

is lower when null results are valued more.<sup>22</sup>

The effect of registration depends then on the effect on the value of a null result  $\underline{u}$  (and opposite statements hold for  $\bar{u}$ ). How might registration affect the value of a null finding? Under one logic, registration should decrease the value of a negative finding for the simple reason that it prevents the use of fishing to turn a negative finding positive through clever model manipulation (in the phrase associated with Coase, to torture the data until nature confesses). In this case we should expect to see that registered studies will be larger than the same studies would have been without registration, since researchers do not want to risk coming up dry. The pool of studies undertaken is also affected as with lower  $\underline{u}$  researchers will forgo seeking evidence for relations deemed *ex ante* to be less probable.

Registration may also however have the opposite effects if it increases the value of negative findings; this might arise for example because of greater scope for combining registered data in meta studies rises or because of norm change in the discipline away from the search for significant relationships.

Registration may also alter the set of studies undertaken through restrictions placed on the statements of hypotheses. If there is no registration, the researcher can accept any significant finding as a positive and claim it was a one-sided test of a hypothesis she can justify *ex post*. Let  $\pi^+(\beta, n)$  be the power of the test with null  $b \leq 0$  and  $\pi^-(\beta, n)$  be the power of the test of the null  $b \geq 0$ . The researcher maximizes:

$$\underline{u} + (\bar{u} - \underline{u}) \int_{\bar{\beta}-a}^{\bar{\beta}+a} (\pi^+(\beta, n) + \pi^-(\beta, n)) f(\beta) d\beta - c(n) \quad (2)$$

With registration however the researcher can either implement a two sided test of the null that  $\beta = 0$  (which as described by Gelman and Tuerlinckx (2000) and others is often a hypothesis that is known to be false *ex ante*) or select one of the two one sided tests to implement. In either case the researcher has less “power” than under the flexible model and this may reduce the set of studies to be undertaken at the margin.

Under registration, if researchers select a “side” they then choose  $n$  to maximize:

$$U = \underline{u} + (\bar{u} - \underline{u}) \int_{\bar{\beta}-a}^{\bar{\beta}+a} \pi^+(\beta, n) f(\beta) d\beta - c(n) \quad (3)$$

The solution to this problem yields an  $n$  that is smaller than that for the problem without registration since with registration the researcher only gains from increasing sample size when  $\beta$  is of the predicted sign. The importance of this effect of registration depends on expected effect size and is more important for studies in which the prior distribution is more evenly

---

<sup>22</sup>See appendix.

distributed around 0.

Hence, the introduction of registration norms can alter the type of studies undertaken. If it reduces the value of null findings, registration can lead to larger studies as researchers require a higher likelihood of obtaining a positive finding. In some cases, by limiting the availability to select the direction of the test conditional on the value of the test, it can also lead to smaller studies. Registration can also lead to the selection of what gets studied and lead to the eschewing of testing less *ex ante* plausible relationships. In a context in which counterintuitive findings might motivate new studies, this disincentive effect could reduce innovation.

## 6 Conclusion

We have reviewed arguments for more or less detailed forms of registration of social science research. While the medical sciences have made major steps towards such institutional developments, we argue that more specific forms of registration would help reduce the scope for data fishing in political science. We demonstrate that with the latitude that researchers enjoy, the room for unintentional or intentional fishing is very large even when researchers prespecify hypotheses and dependent and independent variables. We experimented with a limiting form of “comprehensive registration” in which we sought to write up an entire research report prior to accessing data. We believe that there are clear advantages to doing this, and that doing so will bolster confidence in our actual results when they come in. But doing so also reveals risks of loss of latitude and also costs of comprehensive registration that may not be welcomed by all researchers. Most obvious, besides the burden of registration, is the concern for loss of freedom to incorporate new information into analysis plans.

In light of the risks from researcher latitude and the possibly adverse effects of a registration mechanism that prevent flexibility in analysis, we have proposed the adoption of a detailed registration mechanism hosted at a third party institution and endorsed (and verified) by journals. As our discussion and proposal highlights, the importance of registration is to serve as a communication mechanism rather than a commitment mechanism. A detailed preregistered design can provide a mechanism for researchers and readers to distinguish between three different sorts of results: those that were executed according to predetermined specifications, those registered and that deviated on grounds that may be defended by researchers, and those that were not preregistered at all and for that reason should be interpreted as speculative. The middle category constitutes something of a gray zone in which analysis may stay true to the intent of the registered design but the defense of the details of implementation must be provided *ex post* rather than *ex ante*. This gray zone may be large and admit ambiguity, but

at least with a registration procedure it will be clearly gray.

This proposal, if implemented, could have large effects on research in the discipline. We have tried to sketch some possible impacts here but there are likely many other effects that could be at least as important and that we have not considered.<sup>23</sup> The introduction of registration norms could for example result in an unwarranted altering of the primacy given to analysis of new data relative to the analysis of preexisting data with registration norms possibly leading to increased costs associated with the gathering of new data (or alternatively it could lead to increased confidence in analysis based on new data).<sup>24</sup> Or public registration could lead to concerns over the poaching of research designs (or alternatively the strengthening of claims to unimplemented designs). There may also be broader implications for the publication process, for example registration may require a change in journal procedures since if editors want reviewers to be able to provide substantive guidance to researchers, the review process might also need to shift to before data collection. Indeed in principle journals could provisionally accept pieces for publication in advance on the basis of registered designs in order to limit the publication biases that could still persist even if our proposal is adopted. Finally there may be other gains that could be achieved through the development of a registration mechanism. For example if a registry were to publish hypotheses *ex ante* it might also be able to gather expectations about the outcomes of the research from other researchers, thereby generating a database of disciplinary priors. In addition there may be learning to assess what format can handle the array of methodologies, including quantitative and qualitative, that is employed in political science research and the extent to which registered designs should be publicly accessible prior to publication of results. These considerations suggest that a period of experimentation with the structure of a registry may be the critical next step.

---

<sup>23</sup>And we note that none of the procedures proposed can rule out outright fraud. But we believe it plausible that most data fishing is not fraudulent but rather driven by a general permissiveness towards consulting data to generate tests (for a counter view see Glaeser (2006)).

<sup>24</sup>We thank a reviewer for pointing out these possible effects.

## References

- Beath, A., F. Christia, and R. Enikolopov (2011). Elite Capture of Local Institutions: Evidence from a Field Experiment in Afghanistan. *Working paper*.
- Benjamini, Y. and Y. Hochberg (1995). Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *Journal of the Royal Statistical Society. Series B (Methodological)* 57(1), 289–300.
- Callaway, E. (2011, November). Report Finds Massive Fraud at Dutch Universities. *Nature* 479(7371), 15.
- Casey, K., R. Glennerster, and E. Miguel (2011). Reshaping Institutions: Evidence on External Aid and Local Collective Action. *Working paper*.
- Collier, D., J. Seawright, and G. L. Munck (2010). The Quest for Standards: King, Keohane, and Verba’s Designing Social Inquiry. In H. E. Brady and D. Collier (Eds.), *Rethinking Social Inquiry: Diverse Tools, Shared Standards* (2nd ed.), Chapter 2, pp. 33–65. London: Rowman & Littlefield Publishers.
- De Angelis, C. D., J. M. Drazen, F. A. Frizelle, C. Haug, J. Hoey, R. Horton, S. Kotzin, C. Laine, A. Marusic, A. J. P. Overbeke, T. V. Schroeder, H. C. Sox, and M. B. Van Der Weyden (2004). Clinical Trial Registration: A Statement from the International Committee of Medical Journal Editors. *New England Journal of Medicine* 351(12), 1250–1251.
- De Angelis, C. D., J. M. Drazen, F. A. Frizelle, C. Haug, J. Hoey, R. Horton, S. Kotzin, C. Laine, A. Marušić, A. J. P. Overbeke, T. V. Schroeder, H. C. Sox, and M. B. V. D. Weyden (2005). Is This Clinical Trial Fully Registered? A Statement From the International Committee of Medical Journal Editors. *Canadian Medical Association Journal* 172(13), 1700–1702.
- Duffo, E. (2007). Using Randomization in Development Economics Research: A Toolkit. In T. P. Schultz and J. A. Strauss (Eds.), *Handbook of Development Economics*, Chapter 61, pp. 3895–3962. Elsevier.
- Fearon, J. D., M. Humphreys, and J. M. Weinstein (2009). Can Development Aid Contribute to Social Cohesion after Civil War? Evidence from a Field Experiment in Post-Conflict Liberia. *American Economic Review: Papers & Proceedings* 99(2), 287–291.

- Finkelstein, A., S. Taubman, H. Allen, J. Gruber, J. P. Newhouse, B. Wright, and K. Baicker (2010). The Short-run Impact of Extending Public Health Insurance to Low Income Adults: Evidence from the First Year of the Oregon Medicaid Experiment. *Working paper*.
- Gelman, A., J. Hill, and M. Yajima (2009). Why We (Usually) Dont Have to Worry About Multiple Comparisons. *Working paper*.
- Gelman, A. and F. Tuerlinckx (2000). Type S Error Rates for Classical and Bayesian Single and Multiple Comparison Procedures. *Computational Statistics* 15, 373–390.
- Gerber, A. and N. Malhotra (2008, October). Do Statistical Reporting Standards Affect What Is Published? Publication Bias in Two Leading Political Science Journals. *Quarterly Journal of Political Science* 3, 313–326.
- Gerber, A. S., D. P. Green, and D. Nickerson (2001). Testing for Publication Bias in Political Science. *Political Analysis* 9(4), 385–392.
- Glaeser, E. L. (2006). Researcher Incentives and Empirical Methods. *Working paper* (DP2122).
- Hansen, P. R. (2005). A Test for Superior Predictive Ability. *Journal of Business & Economic Statistics* 23(4), 365–380.
- Ioannidis, J. P. A. (2005, August). Why Most Published Research Findings are False. *PLoS medicine* 2(8), 696–701.
- King, G., E. Gakidou, K. Imai, J. Lakin, R. T. Moore, C. Nall, N. Ravishankar, M. Vargas, M. M. Téllez-Rojo, J. E. H. Avila, M. H. Avila, and H. H. Llamas (2009, April). Public Policy for the Poor? A Randomised Assessment of the Mexican Universal Health Insurance Programme. *Lancet* 373(9673), 1447–54.
- King, G., R. O. Keohane, and S. Verba (1994). *Designing Social Inquiry: Scientific Inference in Qualitative Research*. NJ: Princeton University Press.
- Kling, J. R., J. B. Liebman, and L. F. Katz (2007). Experimental Analysis of Neighborhood Effects. *Econometrics* 75(1), 83–119.
- Laine, C., R. Horton, C. D. Deangelis, J. M. Drazen, F. A. Frizelle, F. Godlee, C. Haug, P. C. Hébert, and S. Kotzin (2007). Clinical Trial Registration: Looking Back and Moving Ahead. *Canadian Medical Association Journal* 177(1), 7–8.
- Lancet (2010, January). Should Protocols for Observational Research be Registered ? *The Lancet* 375(9712), 348.

- Mathieu, S., I. Boutron, D. Moher, D. G. Altman, and P. Ravaud (2009, September). Comparison of Registered and Published Primary Outcomes in Randomized Controlled Trials. *JAMA : the Journal of the American Medical Association* 302(9), 977–84.
- Monogan III, J. E. (2010). The Immigration Issue and the 2010 House Elections: A Research Design. *Working paper*.
- Mutz, D. and R. Pemantle (2011). The Perils of Randomization Checks in the Analysis of Experiments. *Working paper*.
- Rasmussen, O. D., N. Malchow-Møller, and T. Barnebeck Andersen (2011). Walking the talk: the need for a trial registry for development interventions. *Journal of Development Effectiveness* 3(4), 1–29.
- Revez, L., A.-W. Chan, K. Krleza-Jerić, C. E. Granados, M. Pinart, I. Etxeandia, D. Rada, M. Martinez, X. Bonfill, and A. F. Cardona (2010, January). Reporting of Methodologic Information on Trial Registries for Quality Assessment: a Study of Trial Records Retrieved From the WHO Search Portal. *PloS One* 5(8), 1–6.
- Simmons, J. P., L. D. Nelson, and U. Simonsohn (2011, October). False-Positive Psychology: Undisclosed Flexibility in Data Collection and Analysis Allows Presenting Anything as Significant. *Psychological Science*.
- Sterling, T. D. (1959). Publication Decisions and Their Possible Effects on Inferences Drawn from Tests of Significance—Or Vice Versa. *Journal of the American Statistical Association* 54(285), 30–34.
- White, H. (2000). A Reality Check for Data Snooping. *Econometrica* 68(5), 1097–1126.