Transparency, Replication, and Cumulative Learning: What Experiments Alone Cannot Achieve*

Thad Dunning

Travers Department of Political Science, University of California, Berkeley, California 94720-1950; email: thad.dunning@berkeley.edu

Erratum

Annu. Rev. Polit. Sci. 2016. 19:S1–S23

First published online as a Review in Advance on March 9, 2016

The *Annual Review of Political Science* is online at polisci.annualreviews.org

This article's doi: 10.1146/annurev-polisci-072516-014127

*This is a corrected version of an article published on May 11, 2016

Keywords

transparency, replication, cumulation, experiments

Abstract

Replication of simple and transparent experiments should promote the cumulation of knowledge. Yet, randomization alone does not guarantee simple analysis, transparent reporting, or third-party replication. This article surveys several challenges to cumulative learning from experiments and discusses emerging research practices—including several kinds of prespecification, two forms of replication, and a new model for coordinated experimental research—that may partially overcome the obstacles. I reflect on both the strengths and limitations of these new approaches to doing social science research.

INTRODUCTION: DO EXPERIMENTS SUFFICE FOR LEARNING?

The use of experiments in political science has grown markedly in recent years, paralleling trends in several other social-scientific fields.[1] In principle, the advantages of experiments are clear. First, they foster strong causal inference. Manipulation of experimental treatments allows researchers to isolate the impact of particular interventions, and random assignment ensures that comparison groups are identical up to random error—save for the presence or absence of a treatment. Other assumptions must hold, but manipulation and random assignment offer potent resources.

Second and more relevant to the themes of this article, experiments may offer several related benefits. (*a*) Because randomization obviates confounding, simple comparisons across treatment and control groups may suffice to establish a causal effect. In partial consequence, (*b*) experiments can be highly transparent, which is (*c*) helpful for those who wish to replicate experimental designs. Finally, (*d*) the ability to replicate simple, transparent, and credible designs should help researchers to cumulate knowledge across a set of related experimental studies. In summary, experiments are deemed valuable not only because they aid causal inference but also because they promote cumulative learning.

The case for these advantages seems strong in theory. To what extent does experimental research achieve these desiderata in practice? Consider the following difficulties:

1. *Simplicity*. In applications, analysts may rely on complicated parametric modeling assumptions that demonstrably violate the chance properties of random assignment.
2. *Transparency*. Researchers may conduct numerous statistical tests but only report some of them; and publication bias may prevent dissemination of null effects. The distribution of published effects can therefore be highly misleading.
3. *Replicability*. Researchers often work independently, pursuing research questions that interest them; and professional incentives militate against replication. In consequence, broad conclusions are sometimes based on a single pioneering study.
4. *Cumulation*. In consequence of the challenges of simplicity, transparency, and replicability—as well as other factors, such as incomparable interventions and outcome measures across related studies—knowledge may not effectively cumulate across a set of related experimental studies.

Together, these points suggest important barriers to learning from experimental research. Erroneous analysis and nontransparent reporting can undermine inferences from particular studies. Perhaps even more importantly, these challenges reduce the reliability of conclusions from *bodies* of experimental research. Of course, these difficulties apply to most empirical social science—especially, and perhaps with even greater force, to conventional observational research. Yet, to the extent that experiments have a special claim to produce credible inferences, their appeal is substantially reduced if they cannot produce cumulative learning from simple, transparent, and replicable designs.

What can be done to overcome these challenges? The solution seems to lie at least partially in the adoption of new standards and practices that complement experimental research—as well as observational research, as appropriate.

1. *Pedagogy*. The emphasis of methods research and teaching has begun to change. In some sense, this is a natural consequence of the experimental turn: The emphasis on design-based

[1] De Rooij et al. (2009), Humphreys & Weinstein (2009), Hutchings & Jardina (2009), Hyde (2015), McDermott (2002), and Palfrey (2009) have contributed previous *Annual Review of Political Science* articles on the growth of experiments in different areas of the discipline.

inference tends to reduce the role of complicated parametric models and privilege simplicity of analysis.

2. *Prespecification*. Several emerging practices have commanded attention: (*a*) study registration, (*b*) filing of preanalysis plans, and (*c*) results-blind review. A number of organizations now provide online, third-party registries, allowing researchers to post details about a study before it is conducted; some academic journals have even begun to review submissions on the basis of preanalysis plans rather than realized p-values. These practices promote transparency and may reduce publication bias but have also sparked debate about the conditions under which prespecification is appropriate and helpful.
3. *Third-party validation*. Although the privacy of data has long been a concern for internal replication—that is, the ability to reproduce results using materials available to original authors—journal policies requiring data sharing and even third-party verification prior to publication can ameliorate this problem. Some institutions and organizations are also placing heightened emphasis on external replication—the extension of experimental designs to new subject pools or empirical contexts.
4. *Coordinated experimental research*. The inaugural Metaketa initiative of the Evidence in Governance and Politics (EGAP) group, undertaken in conjunction with the Center on the Politics of Development (CPD) at the University of California, Berkeley, is an example of coordinated experimental research. Here, teams of researchers are funded to work in parallel on a predefined theme: the impact of information on political accountability. Participants develop consistent outcome measures and harmonize interventions across disparate contexts, to the extent possible. The research is also preregistered and subject to third-party verification, and it will be published in an integrated fashion, regardless of findings. The structure of this enterprise is intended to foster replication, reduce publication bias, and enhance cumulative learning.

My aim in this article is to review some of these practices and to assess what they can and cannot achieve. A major reason that these key challenges exist, as I describe below, likely relates to the structure of professional incentives in the field. Thus, a critical issue is the extent to which these new standards and practices are consistent with—or can effectively reshape—the ways that both young and established scholars build academic careers. If new approaches cannot get the incentives right, they are unlikely to foster cumulative learning.

Two clarifications are important in light of this article's title. First, my key contention is that random assignment and experimental manipulation per se do not guarantee simplicity, transparency, replicability, and cumulative learning. Experiments may certainly promote these important goals; yet, it is also clear that experiments alone are insufficient to achieve them. The question I address in this article is what ancillary methods, standards, and practices are needed to attain these objectives.

Second, the title also suggests broader points about what experiments can and cannot achieve, beyond questions of transparency, replication, and cumulative learning. Experiments are prized for their special capacity to provide valid causal inferences, yet a vigorous conversation persists about the conditions under which they serve broad social-scientific aims better than other methods. Certainly, experiments cannot readily reveal the effects of variables that researchers cannot manipulate. At some level, this is a feature, not a bug. There is no silver bullet in social science, and like other methods, experiments serve some purposes but not others. A relevant question, beyond the scope of this article, is whether and with what reliability other forms of research can answer substantive questions that experiments cannot. Others have extensively broached the topic of causal inference in observational studies, as have I in the case of natural experiments (see, e.g., Brady & Collier 2010, Rosenbaum 2002, Dunning 2012).

Although the new practices described in this article expand the reach of what experiments can achieve, they are very far from a panacea. An extreme illustration comes from the recent retraction of an article in *Science* (LaCour & Green 2014) on the effects of canvassing on attitudes toward gay marriage. In this study, households in certain Los Angeles precincts that voted for a state ballot proposition banning gay marriage were assigned to a control group (no canvasser), discussion with gay or straight canvassers about recycling (placebo control), or discussion with gay or straight canvassers about same-sex marriage. The study found long-lasting persuasive effects of 20-minute conversations with gay canvassers, stirring excitement among researchers who had believed attitudes to be relatively impervious to such interventions. This important experimental study, conducted jointly by an ambitious graduate student (LaCour) and a senior scholar who is one of the most thoughtful and innovative of political science's experimentalists (Green), seemed to achieve several of the desiderata described above. The experimental design was elegant and powerful; the analysis was preregistered; the replication data were publicly posted; and the study built on more than a decade of experimental research on the effects of canvassing on attitudinal and behavioral outcomes.[2] The only problem is that the "data" were apparently fabricated by the graduate student who was the junior coauthor on the project.[3]

For reasons discussed below, the procedures described in this article unfortunately provide imperfect protection against such outright fraud. Even in this worst-case scenario, however, movement toward these practices was helpful—not least, in making possible the discovery of the deception. In the conclusion, I reflect further on this example and what it may teach us about the power and perils of the new practices I describe in this article. In the next section, I describe four challenges for experimental research in more detail. I then turn in the third section to new standards and practices, discussing the strengths and limitations of what amounts to a new model for doing social-science research.

FOUR CHALLENGES FOR EXPERIMENTAL RESEARCH

Simple Analysis

One important question is how to bolster the simplicity, clarity, and credibility of statistical analyses in experiments. Of concern here is not only the complexity of the analysis but also the link between modeling assumptions and the key chance element in an experiment—to wit, random assignment. Fortunately, a leading model for statistical and causal analysis in experiments justifies highly simple and tractable data analysis. The less appealing alternatives, which often appear in applications, do not reflect such simplicity.

Consider a factorial experimental design adapted from Mauldon et al. (2000), in which teenage mothers are randomized to receive (*a*) financial incentives to stay in high school, (*b*) case management, (*c*) both financial incentives and case management, or (*d*) neither. The outcome is whether each mother graduates from high school. **Table 1** reports data for one group of teen mothers, who were not enrolled in school at the start of the experiment.

How should one analyze these data? The most natural choice is to estimate causal effects by subtracting graduation rates in the control condition (cell A) from those in each of the treatment

[2] In fact, only the second of two experiments reported in LaCour & Green (2014) was preregistered, although LaCour appears to have made a fraudulent effort to retroactively "pre"-register the first experiment (Singal 2015).

[3] LaCour has not admitted fraud, but the evidence appears overwhelming. At the least, LaCour did not make payments to survey respondents, as stated in the published article, and he did not receive funding from sources identified in the publication. For discussion of irregularities, see Broockman et al. (2015).

Table 1 Boosting graduation rates of teenage mothers: a factorial experiment (adapted from Mauldon et al. 2000, p. 35)[a]

		Case management	
		No	**Yes**
Cash incentives	No	(A) 10.5%	(B) 9.0%
	Yes	(C) 14.8%	(D) 19.7%

[a]Percentages indicate high school graduation rates. N = 521. Results are shown for teenage mothers who were not in school at program entry. Cells are labeled A–D to clarify the discussion in the text.

conditions (cells B–D). Thus, we find that case management by itself provides no boost in graduate rates: $9.0 - 10.5 = -1.5\%$.[4] However, financial incentives by themselves increase graduation rates by almost 50%: $14.8 - 10.5 = 4.3\%$. Finally, receiving both interventions together nearly doubles graduation rates: $19.7 - 10.5 = 9.2\%$. Inspection of the table therefore suggests an important interaction effect.

This data analysis is straightforward—and, crucially, it can be justified under a causal and statistical model that is a highly persuasive depiction of the data-generating process. In brief, suppose each mother graduates or does not when assigned to each of the four conditions; these are potential outcomes.[5] This leads to a missing-data problem. If all mothers are assigned to receive financial incentives only, we observe the average graduation rate for this condition—but we do not see average potential outcomes in the other three conditions, so we cannot assess average causal effects. Suppose, however, that we represent each mother by a ticket in a box, and each ticket has four values on it—one for each of her potential outcomes. Thus, we have a box with 521 tickets in it, one for each mother in this study. If each mother is assigned with equal probability to each of the four experimental conditions, it is *as if* each of the cells of **Table 1** is a simple random sample from this small population of 521 tickets. Under the model, standard errors for the average difference between any two cells can be calculated using a conservative formula, identical to the standard error for the difference of means of two independent samples (Freedman et al. 2007, A32–A34, n. 11; Dunning 2012, appendix 6.1). Normal approximations give confidence intervals and p-values; alternately, researchers may use distribution-free randomization tests. Display of the results can be tabular or graphical.

The model is critical, because the analogy to random sampling from the study group justifies estimators of average causal effects. The key principle is that averages of random samples are unbiased estimators for averages in the population from which they are drawn.[6] Crucially, the model is also closely connected to the actual chance process being studied: There is a strong analogy between sampling tickets at random from a box and assigning teen mothers at random to treatment conditions. Note several features of the model. The observations are not independent and identically distributed (i.i.d.); if the first mother goes to the first treatment group, that changes

[4]Under the box model discussed in the next paragraph, the difference is not statistically significant. With approximately 130 mothers assigned to each condition, the estimated standard error for the difference between the two cells is about 3.7 percentage points.

[5]Thus, here we assume the Neyman (or Neyman-Rubin-Holland) causal model; the key objective however is to discuss statistical inference under this model.

[6]The simulation in the Appendix demonstrates unbiasedness of difference-of-proportions estimators as well as the accuracy of the conservative formula for the standard errors.

the probability of treatment assignment for the subsequent mothers.[7] Also, the distribution of the potential outcomes in the box changes as we go. Here, the distribution of sample statistics is governed only by the values of the tickets in the box—not by some assumed parametric distribution. Statistical inferences are made about the box, not about some ill-defined population from which the box was ostensibly "drawn."[8]

Contrast these features with a canonical parametric model such as logistic regression—a common choice for analyzing experimental data with dichotomous outcomes. Thus, let $C_i = 1$ if mother i is assigned to case management and $F_i = 1$ if she is assigned to financial incentives; if she graduates from high school, $Y_i = 1$. According to a latent-variables formulation of the logistic regression model,

$$Y_i = 1 \quad \text{if} \quad \alpha + \beta_1 C_i + \beta_2 F_i + \beta_3 (C_i * F_i) + u_i > 0, \qquad 1.$$

where u_i is a random variable drawn from the standard logistic distribution; the u_i are assumed to be i.i.d. across subjects. The interaction term $C_i * F_i$ is the product of the two dichotomous treatment variables. Using the symmetry of the logistic distribution, we can rewrite Equation 1 as

$$\text{Prob}(Y_i = 1) = \Lambda(\alpha + \beta_1 C_i + \beta_2 F_i + \beta_3 (C_i * F_i)), \qquad 2.$$

where Λ is the standard logistic distribution function. The regression model says that the observations are independent—and so is the probability that $Y_i = 1$ for each i.

The model is again critical, because it defines the parameters of interest—e.g., α, β_1, β_2, and β_3—and suggests natural estimation strategies, given the model. Yet, is it closely connected to the data-generating process? It seems not to be. For example, why is there a latent variable u_i, and why are its realizations i.i.d.? Why are the probabilities independent across i? Finally, why does the logistic distribution function come into play? Random assignment does not imply any of these properties of the statistical model. And the assumptions are quite different from the box model, where observations are neither independent nor identically distributed—and there is no latent variable u_i in the picture.

There are several costs of estimating the parametric model in Equation 1, relative to the simple comparison of percentages in **Table 1**. With logistic regression, common estimators may not relate to well-defined quantities. For example, Freedman (2009b) shows that a commonly used estimator derived from a logit fit does not accurately estimate the differential log-odds of success. Moreover, the modeling strategy makes analysis both unnecessarily complex and prone to mistakes. For instance, rather than properly calculating the effect of, e.g., exposure to case management alone as $\Lambda(\alpha + \beta_1) - \Lambda(\alpha)$, analysts might be tempted to take derivatives of Equation 2 to get $\frac{\partial \text{Prob}(Y_i=1)}{\partial C_i} | (F_i = 0) = \beta_1 \lambda(\alpha + \beta_1)$, where λ is the density of the standard logistic distribution. Because C_i is dichotomous, this formulation of the "marginal" effect makes little sense. To be sure, there can be convergence between model-based estimators and the simple comparison of percentages in **Table 1**. In the Appendix, I show by simulation that certain logistic regression estimators are equivalent to the differences of percentages and thus are unbiased for average causal effects. These points of convergence should not obscure the important differences between the logistic regression model and the box model as representations of the data-generating process. Note also that this is a very simple setting for the regression model: The only right-hand-side

[7] This is true when the number of units to be assigned to each group is fixed in advance of the experiment, which is sometimes called a fixed-margins design.

[8] In some contexts, such as survey experiments, the study group is itself a random sample from a larger population. For field and lab experiments, this is the exception rather than the rule.

variables are indicators for randomly assigned treatments. In more complicated settings—e.g., when pretreatment covariates are included—further issues arise.

The larger point of this discussion is that the analytic procedure is not dictated by the existence of an experiment—and some modes of analysis may take us far from simple, design-based approaches. Researchers may be surprised to learn about the variety of modeling approaches that have been used or recommended for analysis of experimental data. For example, analysts have even proposed using set-theoretic comparative methods (STCM)—all variants and extensions of Qualitative Comparative Analysis (QCA) (Ragin 1987, 2000)—for policy analysis and impact evaluation, including experiments. For discussion and a review of applications, see Tanner (2014); for reasons why using STCM is a bad idea, see Collier et al. (2015). In summary, the existence of random assignment and experimental manipulation does not suffice to guide the choice of analytic procedure. Analysts must think in each application about the match between their model and the true chance process that generates the data. The assumptions are crucial because statistical inferences are only as good as the model under which they are made.

Transparent Reporting

In experimental and nonexperimental research alike, a second major concern is publication bias: the tendency of journals to publish "significant" results, leaving apparent null effects to languish in working papers (or never to be written up or reported in the first place). The evidence for this problem is now quite extensive. Gerber et al. (2001) showed that in a large body of published voter mobilization studies, there is a negative relationship between the estimated effects of interventions and the size of the experimental study group. This is ipso facto evidence of publication bias because these variables should be independent—unless publication requires crossing the threshold for statistical significance, and thus a smaller study necessitates a larger effect size to be published.[9] Gerber & Malhotra (2008), surveying articles published in the *American Political Science Review* and *American Journal of Political Science*, show that the distribution of z-statistics jumps up very discontinuously at 1.96—the threshold using normal curve approximations for statistical significance at the 0.05 level. Their technique is also related to diagnoses of p-hacking developed by Simonsohn et al. (2014). Malhotra (2014) compares published findings using data from the Time-sharing Experiments for the Social Sciences (TESS) to a record of applications for TESS filed before data collection and finds that null effects tend to go unpublished—and often the results are never even written up.

There exist at least two theories about the source of this problem. First, journal editors and reviewers may tend to favor statistically significant results for publication. Second, analysts may tend, either consciously or unconsciously, to adjust their empirical strategies until they obtain statistically significant results—which they write up without disclosing the many other tests that did not cross thresholds for statistical significance. Of course, these two theories are not unrelated and are not mutually exclusive; analysts who engage in "fishing" or p-hacking presumably do so in anticipation of the reaction of editors and reviewers to null effects.

Whatever the ultimate source of publication bias, there are two major problems for inference. First, it undermines the interpretation of significance tests. The point of significance testing is to distinguish chance variation from real effects. Yet, under the null hypothesis of no effect, nominally

[9]Of course, it might be that scholars choose higher-powered designs—those with larger sample sizes—precisely when they anticipate smaller effects. However, such careful preplanning based on heterogeneity in expected effect sizes appears unlikely in the surveyed literature.

significant estimated effects arise in one out of twenty independent tests. If the test for which $p < 0.05$ is the only one reported, then it is not in fact the case that $p < 0.05$; the probability of one such test statistic in a family of hypothesis tests is substantially higher. The issue is that the reporting does not reflect the multiple comparisons being made. In particular, the choice of what test to report is conditioned on the observed p-value.

Second, null effects can and should be informative, especially in the context of an overall research program. As the saying goes, a null effect is not a null finding. To be sure, there can be many sources of null estimated effects, from a true lack of a causal relationship to the theoretically or substantively irrelevant details of a study's execution. It is also the case that failure to reject a null hypothesis does not imply proving the null. Yet, the finding of no estimated effect in a high-powered study should be able to teach us something about the phenomenon in question. Even more importantly, the tendency of social-scientific inferences and policy recommendations to be drawn from a small number of highly visible published studies showing important treatment effects—while less visible studies that suggest null effects remain unpublished—leads to a distorted view of the likely effects of interventions. Publication bias therefore undermines the possibility of cumulating knowledge from a series of related studies.

Publication bias is an issue for experiments as much as for other kinds of research designs. Indeed, experimental researchers often gather measures for multiple outcome variables, and they can calculate effects for multiple subgroups. This raises the possibility that the effects of chance variation can be substantially understated if only the statistically significant p-values from these multiple comparisons are reported. The existence of random assignment and experimental manipulation does not guarantee disclosure of multiple hypothesis tests, nor is it necessarily a safeguard against publication bias.

Critical Appraisal

Closely related to the issue of publication bias is the challenge of reporting results in a way that allows critical evaluation by third parties—and provides a bulwark against researchers' mistakes in data analysis. To err is human (as the author of this article regretfully acknowledges), and simple mistakes are all too easy to make in a long and often complex research process. When data are private and key details of experimental designs are not transparent, third parties cannot readily assess the credibility of inferences drawn in any particular study.

There are two components to the problem of critical appraisal. The first relates to what has sometimes been called internal replication.[10] Looking at a research report, a third-party analyst might ask: Can I replicate the results if I use the author's own replication dataset and code? Stepping back, if I have access to the raw unprocessed data but not a polished dataset or replication code, can I duplicate the results? Finally, do additional analyses using supplementary data or observations that were available to the original analyst tend to confirm or reject the analyst's account? These questions provide successively more demanding criteria for replication, yet all are considered "internal" in the sense that what is at issue is the veracity and freedom from error of the original researcher's own analysis.

The challenges to internal replication are familiar from observational studies, where the record is not encouraging. McCullough et al.'s (2006) attempt to replicate research in a journal with

[10]Hamermesh (2007) calls this "pure" replication. Clemens (2015) uses the term "verification test"; his "reproduction test" might also be consistent with my usage of the term internal replication.

a policy of mandatory data archiving found that data for only 69 of 193 articles were in fact archived—and only 58 of these 69 had both data and code present (pp. 1101, 1105). The authors then attempted to replicate 62 of these 69 studies but could only reproduce results for 14 (or 23%) completely. Other attempts have yielded similar results. For observational and experimental studies alike, perhaps the most basic barrier to internal replication is private data. Despite the editorial policies in leading political science journals that require or strongly suggest posting of replication files upon publication—policies I discuss further under Third-Party Validation—the practice is still very far from universally followed. The inability to access data thus remains a substantial barrier to open science.

A second component of critical evaluation involves what may be called external replication. In contrast to internal replication, the objective is to introduce a new intervention—ideally, one that is identical to the original study's—sometimes using the same subjects, or a new sample of subjects drawn from the same population as in the original study. In this case, replication may allow estimation of the same target parameter as in the original study, only using an independent draw of data.

In principle, experimental research offers attractive possibilities with respect to external replication—given that experimental manipulation offers the opportunity to introduce an intervention anew. In practice, however, external replication seems almost as rare in experimental as in observational research. The reasons are likely many, but one key factor may be weak career incentives to replicate. Once a high-visibility study showing an effect of some intervention has been published, studies that suggest weaker or null effects tend not to appear in high-profile journals, if they are published at all (a point related to the previous discussion of publication bias). By contrast, the professional rewards of "planting the flag"—e.g., assessing the effects of a new intervention for the first time—appear much more substantial. Together, the difficulties of conducting internal and external replications both hamper critical evaluation of experimental research results.

Cumulative Learning

The difficulties of publication bias and external replication suggest another, broader challenge: the problem of the cumulation of knowledge. Several advocates for experimental methods suggest that replication and extension of experimental designs are the most reliable route to cumulative learning.[11] Thus, for example, the only way to evaluate the external validity of an experimental result is to repeat the design in a new context.

In practice, however, the barriers to replication identified above can substantially undermine this route to knowledge. Consider the recent literature on the role of community monitoring in improving service provision in developing countries. This literature rests on a fairly plausible theory: Service provision involves principal–agent problems in which multiple principals (citizens) face informational and collective action problems that prevent them from effectively monitoring the performance of agents (bureaucrats, politicians, and other service providers). Accordingly, increasing the capacity of citizens to monitor the work of doctors, teachers, and other public employees should lead those service providers to boost their performance, improve the quality of services, and ultimately improve health, educational, and other outcomes.

[11] Banerjee & Duflo (2009, p. 160) note that "to address . . . concerns about generalization, actual replication studies need to be carried out. Additional experiments need to be conducted in different locations, with different teams."

One high-profile study by Björkman & Svensson (2009) does indeed find striking effects of community monitoring in the health sector. The authors present results of a randomized field experiment in Uganda, in which "localized nongovernmental organizations encouraged communities to be more involved with the state of health service provision and strengthened their capacity to hold their local health providers to account for performance" (Björkman & Svensson 2009, p. 735). Not only do Björkman & Svensson find greater involvement in monitoring and higher effort among health workers in treatment communities, they also report "large increases in utilization and improved health outcomes," e.g., a reduction in child mortality of approximately 33%. These effects are remarkable—so large as to spur some questioning of their plausibility, as discussed below. Yet the study has had an important policy impact, sparking international aid donors and others to promote community monitoring.

However, other studies lead to considerably less optimism. For example, Olken (2007, p. 200) presents a randomized field experiment on reducing corruption in Indonesian village road projects. Although "top down" government audits substantially reduced missing expenditures, increasing "grassroots participation" in meetings regarding spending on road projects—i.e., community monitoring—had no average impact on corruption. Lieberman et al. (2013, p. 69) report a randomized experiment in which Kenyan parents were provided with "information about their children's performance on literacy and numeracy tests, and material about how to become more involved in improving their children's learning." However, these authors find "no discernible impact on either private or collective action," nor any impact on educational outcomes. Other studies have similarly shown mixed or null effects (e.g., Banerjee et al. 2008).

So, is community monitoring effective? And what is one to make of these contrasting results? There are many possible conclusions. The interventions reported in these studies differ in many ways. So does the nature of the service being provided in each case—and thus, *inter alia*, the outcome variables. Perhaps the most natural, though difficult to validate, inference is that "it depends": Interventions to increase community monitoring sometimes do improve service provision and sometimes do not, but in ways that may depend substantially on the political or socioeconomic context. In the view of philosophers such as Cartwright, mechanisms are closely tied to external validity and thus to cumulative learning: If we understand *why* X produced Y in context Z, we may form a better prediction of whether X will also produce Y in context W—by considering whether the relevant enabling mechanisms in context Z are likely to be operative in context W (see Cartwright & Hardie 2012). However, evidence on context is hard to come by. It depends on the ability to identify mechanisms, which may be only slightly less challenging in experimental research than in observational research. In the context of community monitoring, it is difficult to know how much difference in effects contextual distinctions across the studies may produce. Thus, despite the striking findings of Björkman & Svensson (2009), it would be difficult to walk away from the broader literature with a conclusion that "community monitoring is effective"—or an understanding of the conditions under which it works.

These examples of research on community monitoring also raise other issues, such as the publication and reporting biases discussed above. The size of the Björkman & Svensson study is small (clustered assignment of 50 public health dispensaries, 25 in treatment and 25 in control), but the large effects the authors report are statistically significant; indeed, given the small sample size, large estimated effects would be necessary to cross the threshold for statistical significance. Thus, perhaps reflecting the thrust of Gerber et al.'s (2001) findings for the voter mobilization literature, here a published study with a small study group features a strikingly large estimated treatment effect. Moreover, this paper has become a key reference in a literature with several unpublished papers showing null effects. Thus, one single high-profile study has tended to drive

policy conclusions—while ancillary or unpublished results in other studies draw less attention and generate less impact.

Overall, this discussion highlights the challenges that can arise when the goal is cumulative learning. All of the examples on community monitoring are randomized experiments—and thus in each case the credibility of inferences drawn from each single study may be high. Yet relying on "naturally occurring" replication of experimental research may not suffice to promote cumulative learning because, among other reasons, the interventions and outcome measures are so different across disparate contexts.

WHAT REMEDIES CAN WORK?

What methods, standards, or practices might help overcome the challenges discussed in the previous section?

Recommending solutions depends on a theory of the problem at hand. There may be many reasons for the failures sketched in the previous section. A crucial consideration in each case appears to be the existing structure of professional rewards and the incentives that they generate among researchers. Thus, (*a*) there may be a tendency in the profession to privilege complex analysis over simplicity. As I discuss next, this may be shifting, with helpful consequences for the clarity and credibility of statistical analyses. However, (*b*) the existence of publication bias still penalizes research showing null effects, and (*c*) professional incentives to engage in either internal or external replication appear to be quite weak. By contrast, (*d*) the rewards of being the one to plant the flag are substantial, which may inhibit cumulative learning in a variety of other ways.

If this theory is correct, any efforts to confront these challenges must get the incentives right—that is, effective solutions must take into account the relevant institutional and behavioral constraints. In this section, I describe several emerging practices in political science that may help meet these challenges. The solutions are often interrelated, with implications for more than one of the four desiderata sketched above. In each case, I offer some thoughts on their likely effectiveness.

Pedagogy

Perhaps the easiest of the four challenges to address—but one that can still be surprisingly hard—is the valorization of simple and clear analysis, founded in credible assumptions about data-generating processes. Thus, the clearest remedy for the difficulties discussed above in the Simple Analysis section may lie in the teaching of quantitative methodology in political science.

The traditional approach to teaching introductory PhD-level empirical/quantitative methodology begins with the linear regression model—perhaps after some introductory material on probability and sampling theory. This teaching of statistical and causal inference presumes the model; shows that, under the model's assumptions, ordinary least squares is the best linear unbiased estimator; and gradually introduces minor departures from the model, such as violations of sphericality assumptions on error terms. According to this approach, the major threat to causal inference is correlation between the error term and right-hand-side variables. To be sure, lack of such correlation is required for unbiased estimation under the model. Yet, the possibility that the linear model is an unhelpful approximation to the world—or that it tends to be used to draw statistical inferences in observational settings where well-defined chance processes may not actually exist—is rarely given more than superficial discussion. Critical scrutiny of the fundamentals of the model is left for later courses, if it occurs at all. For most students, this sequencing requires

suspending disbelief, so much so that an influential revisionist textbook makes the unusual request that readers instead "suspend belief" (Freedman 2009a, p. xii).

In several top departments, however, this conventional approach is no longer strictly followed. For example, the first course in the sequence may focus partly on analysis of experiments under the Neyman causal model—not because experiments are the only form of quantitative research that social scientists do, but because they allow ready introduction to statistical and causal inference in settings where modeling assumptions may approximately hold. Compared to conventional pedagogy, much more of the focus is also on the role and credibility of the assumptions—and on how strong research designs can sharpen the assumptions' observable implications and thus make them more amenable to empirical test. In contrast, topics such as estimation of generalized linear models—for instance, maximum likelihood estimation of probit or logit models—tend to come later in the sequence (if at all) and sometimes receive short shrift. Thus, along with the rise of new texts, there appears to be an important shift in emphasis in graduate instruction.[12] In this respect, graduate education in political science has arguably moved closer to statistics than to economics, where theoretical econometrics still has a strong hold (Angrist & Pischke 2015).

These pedagogical developments no doubt have important implications for applied experimental research. More broadly, the focus on research design and simplicity of analysis changes the tenor of what is seen as strong work—which speaks to the point that effective remedies should be incentive compatible, especially for young researchers building their careers. Importantly, none of this implies that statistical technique is unimportant: Validating design assumptions often requires advanced statistical or computational tools. Moreover, there are active debates among scholars working in the Neyman tradition, for instance, about the utility of adjusting for covariates in experimental analysis; recent work adapts insights drawn from the literature on regression adjustment in survey sampling to the experimental context (see, e.g., Lin 2013). The point is simply that recent methodological developments valorize simple empirical analysis in a very welcome way. Although there is further progress to be made, the new pedagogy weakens the reflex to turn to complex, unvalidated, and unwarranted modeling assumptions.

Prespecification

What practices might encourage transparent reporting and reduce publication bias? Among the most important is research prespecification, an idea that has attracted substantial recent attention.[13] In this subsection, I describe three approaches—study registration, preanalysis plans, and results-blind review—and then discuss their likely ability to reduce publication bias. Each involves recording some aspects of the research in advance of observing outcome data, yet these practices involve progressively more demanding departures from traditional approaches.

Study registration refers simply to documenting the existence of a study in advance of its execution. Thus, in principle, it allows description of a universe of planned studies—which provides a denominator against which one can assess the set of completed or published studies. To date, registration has been somewhat ad hoc, with several different organizations providing third-party

[12] For new texts, see, e.g., Freedman (2009a) in statistics, Green & Gerber (2012) and Dunning (2012) in political science, or Angrist & Pischke (2009, 2014) in economics.

[13] On the benefits of registration, see Humphreys et al. (2013), Miguel et al. (2014), or Monogan (2013, 2015); for some of the drawbacks, see Laitin (2013).

registration services.[14] Several political science journals now have a policy of encouraging study registration.[15] However, registration is typically voluntary, and the level of detail about the planned study varies greatly; for example, detailed preanalysis plans—considered next—may or may not be included.

Preanalysis plans describe the hypotheses and statistical tests that will be conducted, once outcome data are gathered. There is currently no strong standard for their form and content. Empirically, preanalysis plans involve greater or lesser specificity about the number and kind of tests. At one extreme of prespecification is Humphreys et al.'s (2011) approach of posting the complete analysis code with mock data, which allows analysts to simply run the code once the real outcome data are collected. This arguably represents best practice because reading the preanalysis plan leaves little guesswork as to what is intended in the analysis. Dunning et al. (2015) take a related approach, which is to post the analysis code with real outcome data but randomly reshuffled treatment labels.[16] The downside of the latter approach is that it requires more trust from readers, since the authors have access to the outcome data before posting the code. The upside is efficiency and accuracy: Unlike mock data, real data often have peculiar characteristics that may require adaptive analytic strategies, and data processing and other errors can be caught in advance of filing the preanalysis plan. Moreover, authors can make effective use of a sequence of amendments to pre-analysis plans, for instance, by posting an initial document with the full set of hypotheses and tests, and a subsequent document specifying further analysis details after initial data collection (e.g., Dunning et al. 2015). An interesting alternative without the downside may exist for studies with pilots: The code could be preregistered after analysis of pilot data, which are likely to resemble data collected later.

Finally, results-blind review—as the name implies—refers to the practice of reviewing a research report blind to the study's findings. Thus, referees would evaluate a journal submission on the basis of the interest and importance of the research question, the strength of the theory, and the quality of the empirical design—but not the study's p-values. Though still quite rare, the practice has been applied in several venues; for instance, a forthcoming special issue of the journal *Comparative Political Studies* will feature only articles reviewed in this results-blind way.

What is the likely impact of these three forms of prespecification on publication bias? Study registration (without preanalysis plans) enables measurement of publication bias by providing a denominator, the number of studies in a given area—but it seems unlikely to reduce the bias. Indeed, whatever the true source of publication bias, the mere fact of having announced a study's existence prior to its execution should not affect its chance of publication, conditional on the p-values. Consistent with this conjecture, Fang et al. (2015) find no evidence that the 2005 mandate of study registration in medical journals—which did not require detailed preanalysis plans or results-blind review—led to a reduction in publication bias.

With preanalysis plans, the likely impact is subtler and depends on whether the source of publication bias is (*a*) specification searches (fishing) on the part of authors or (*b*) the preferences of reviewers and editors for statistically significant findings. In principle, prespecifying the set of tests to be performed limits the scope for ex post specification searches or fishing for statistically

[14] As of July 19, 2015, the American Economic Association (AEA) registry has 413 studies in 71 countries (**https://www.socialscienceregistry.org**), while the Evidence in Governance and Politics (EGAP) registry has 178 designs registered since inception in March 2011 (**http://egap.org/design-registration/registered-designs/**).

[15] See, e.g., *Political Analysis*, **http://www.oxfordjournals.org/our_journals/polana/for_authors/general.html**.

[16] See preanalysis plans and amendments as protocol [87] 20140723AA at **http://egap.org/design-registration/registered-designs**.

significant effects; and preanalysis plans may allow meaningful adjustment for multiple statistical comparisons—without which the interpretation of nominal p-values may be undermined. Note that researchers have substantial latitude, both in selecting the mode of adjustment and in specifying the families of tests or hypotheses to which adjustment will be applied. For example, for the mode of adjustment, analysts may choose between Bonferroni corrections, the false-discovery-rate correction of Benjamini & Hochsberg (1995), and other alternatives. A complete preanalysis plan should therefore prespecify the mode of adjustment and thus limits the scope to condition adjustment on realized p-values. However, if journal editors and reviewers simply refuse to publish null estimated effects—perhaps because they find null effects uninformative—prespecifying the tests will not reduce publication bias.

By contrast, results-blind review does appear to offer an effective remedy for publication bias. It is impossible for reviewers and editors to condition publication decisions on the p-values if they do not know what the p-values are. As with preanalysis plans, results-blind review is an area of active development. One question is how to handle evaluation of the quality of the ultimate analysis; one answer may be to make acceptance after results-blind review conditional on successful execution of a set of additional analyses. For example, results-blind review of a natural experiment might require a placebo test with outcomes that should not be affected by the treatment. For journals conducting results-blind review, another question is whether to limit submissions to studies that have not yet been conducted (i.e., to permit submission of true preanalysis plans only) or also to allow reports on research that has already been conducted but are stripped of results for submission to reviewers. The *Comparative Political Studies* special issue allowed both types of submissions. This seems sub-optimal: As a reviewer, I might reasonably infer the existence of null results from the fact of submission to such a special issue (at least in a world in which not all journals are reviewed in a results-blind way). Allowing authors simply to strip out results thus seems to encourage a selection bias in the types of articles submitted for results-blind review.

Finally, a key issue is whether the importance or interest of a research question can indeed be evaluated absent the results. Some critics maintain that we do not learn much from proving an obvious hypothesis—which may certainly be correct. Nonetheless, it is also the case that deeming only counterintuitive or unexpected results worthy of publication is an important source of publication bias (as well as false negatives: If the true effect is null, this precept ensures that only estimated effects in the tails of the sampling distribution—those that falsely reject the null—will be published). Results-blind review may work best in settings where a range of outcomes, including possible null effects, might ex ante be deemed interesting and informative.

Results-blind review may not work for all forms of research, but it offers a powerful and interesting approach that scholars will continue to consider. Independent of publication bias, prespecification may also improve the quality of research because it forces ex ante consideration of a range of issues that can otherwise escape researchers' attention until too late. A vigorous discussion is emerging in the discipline about the types of research for which preanalysis plans are appropriate, about implications for scholarly creativity and the extent to which exploratory (not preregistered) analyses should be reported along with registered analyses, and about the usefulness of adaptive plans and the practice of filing amendments (before analysis of outcome data) as features of the particular context become clear during the research process (for a lucid discussion, see Laitin 2013).

Third-Party Validation

What policies or practices would make external assessment more effective? As discussed in the Critical Appraisal section above, perhaps the biggest barrier to internal replication is private

data. It is therefore modestly encouraging that a growing number of journals in political science now require the posting of replication data for published articles. To be sure, a recent survey (Gherghina & Katsanidou 2013) of replication policies at 120 peer-reviewed political science journals found that only 19 even *had* a policy. Yet, the journals with policies tended to be the discipline's most prominent, highest-impact venues. Among leading journals, the *American Journal of Political Science* has gone furthest, requiring independent third-party verification of reported results—essentially, internal replication using the author's own data and code. The current (as of July 2015) policy of the discipline's flagship journal, the *American Political Science Review*, states that "authors of quantitative or experimental articles are expected to address the issue of data availability. You must normally indicate both where (online) you will deposit the information that is necessary to reproduce the numerical results and when that information will be posted" (**http://www.apsanet.org/apsrsubmissions**). Thus, although it would be a mistake to paint too rosy a picture of data availability, the situation does appear to have improved.

However, the availability of replication data does not guarantee that it can be used effectively. Provision of all data collected for a study allows third parties to compare tests not reported in a published paper, a particularly useful exercise in combination with a preanalysis plan. Yet, in the survey of a data archive discussed in the Critical Appraisal section, data were provided for only about 36% of articles; and even for these, the completeness and intelligibility of data and code varied widely.

Moreover, data availability policies do not advance the goal of external replication. It appears that some researchers are prone to replicate their own work in new contexts, but third-party external replication of experimental research appears rarer. Why is replication so rare? A plausible hypothesis is that for individual researchers, the incentives to engage in replication are exceedingly weak. Research that assesses a new topic or question for the first time can be highly rewarded, whereas critiques can be difficult to publish. The problem is compounded for experiments by the high cost of external replication; experimental research is often expensive, especially survey and field (rather than lab) research. The likelihood that a researcher would spend scarce funding on replicating another researcher's work is thus exceedingly low—although examples certainly exist. Any effort to increase replication, it seems, must therefore address the funding model for social science research.

Coordination

If the broad problem is that scholars benefit more, professionally speaking, from publishing work deemed innovative or groundbreaking—and benefit less from publishing work that replicates existing findings—then one potential solution is to change the incentives researchers face. One way to do this is to fund new research in a manner that requires replication across contexts, ideally in a way that contributes to the cumulation of knowledge while also respecting the freedom of individual researchers to generate and test new ideas.

This is the approach taken in the inaugural Metaketa initiative of the Evidence in Governance and Politics (EGAP; **http://egap.org/metaketa**) group, undertaken in conjunction with the Center on the Politics of Development (CPD; **http://cpd.berkeley.edu/initiatives/egap-regranting**) at the University of California, Berkeley. In a pilot initiative, EGAP and the CPD partnered to fund coordinated field experiments in the substantive area of political accountability. Specifically, EGAP first called for short "expressions of interest" from researchers to allow identification of a topical focus. The second stage of the selection process requested project proposals related to this predefined theme. The Metaketa selection committee prioritized

identifying foci that were related to a body of previous literature, in which interventions were tested, scaleable, and portable, and in which researcher interest was high enough to generate a pool of similar projects.[17]

On this basis, the committee identified a focus motivated by the following questions: Why do people elect poor or underperforming politicians in developing countries? In particular, do voters lack the information they need to make informed choices—and if they are given better information, do they select "better" politicians? As in the case of community monitoring discussed above, there is cogent theory suggesting that informational problems undergird failures of political accountability. Yet, in previous studies the effects on electoral behavior of providing information about politicians' performance appeared quite mixed. Conflicting results may indicate that key factors function differently across different contexts—for example, the effect of informing voters about politicians' malfeasance or corruption may depend on what voters already know or believe. But they might also stem from distinctions of study design, from interventions and outcomes that differ across contexts, and from the expectation that researchers demonstrate "novel" results in each published study.

An overarching goal of the Metaketa initiative is to harmonize research questions, interventions, and outcomes across funded studies. Given the emphasis on innovation and originality in empirical research, a major question was how to make this venture intellectually and professionally attractive for researchers. The solution was to request proposals for studies that have at least two treatment arms in addition to a control group. In the first, "common" arm, which is harmonized across all studies and thus builds in replication across contexts, each study provides voters with credible information about politician performance and assesses effects on electoral behavior. This replication of interventions across studies is critical; for example, as discussed below, it prevents major conclusions being drawn from a single study showing large treatment effects.

The second arm then encourages researchers to develop distinctive interventions. For example, several funded proposals focus in their second arm on the effects of different forms of group-based provision of information, which may generate common knowledge among community members (i.e., "you know that I know that you know . . . that a politician is underperforming"), and contrasts that with the effects of individual provision of information in the common arm. This second arm thus also allows analysis of comparative effectiveness, by asking which type of intervention most powerfully shapes electoral behavior if the common intervention does not. This research structure leaves room for creativity on the part of each research team—which remains critical for social science and public policy—while also requiring replication of results.

On this basis, the Metaketa committee selected projects taking place in quite disparate contexts, including Benin, Brazil, Burkina Faso, India, Mexico, and Uganda (two projects).[18] Grantees preregister all individual analyses with detailed information on specification and adjustments for multiple comparisons; moreover, a plan of plans—that is, a preanalysis plan for a meta-analysis combining results of the different experiments—is also preregistered.[19] Principal investigators are also expected to make their data public in order to allow third-party internal replication prior

[17] The selection committee was composed of Thad Dunning (University of California, Berkeley), Guy Grossman (University of Pennsylvania), Macartan Humphreys (Columbia University), Susan Hyde (Yale University), and Craig McIntosh (University of California, San Diego).

[18] With an initial grant from an anonymous donor of $1.8 million and supplementary funding of $150,000, Metaketa made individual grants within the $150,000–$300,000 range. For a list of funded projects, see **http://egap.org/research/metaketa/metaketa-information-and-accountability**.

[19] For the meta-preanalysis plan, see protocol 127 at **http://egap.org/design-registration/registered-designs**.

to publication, which should reveal errors or discrepancies in the data, increasing the chance of reliable findings. And in addition to publishing any reports on their individual projects, researchers are expected to participate in integrated publications, such as a journal article summarizing the results of all studies or a volume that presents, in one place, the results from all studies. Under the latter model, the collection of studies would ideally be reviewed on a results-blind basis. This approach can substantially limit the publication biases that might otherwise occur, in which studies showing large effects are published whereas those suggesting null effects languish as working papers (Dunning & Hyde 2014).

In summary, this new model for coordinating and funding research seeks to overcome several of the incentive problems that appear to undergird the challenges discussed in this article: publication biases, replication failures, and ultimately obstacles to cumulative learning. It is important to emphasize the importance of humility in this regard. For experimental and observational research alike, cumulation is an old and vexing problem. In our studies, designs, interventions, and outcome measures are harmonized to the extent possible; in each study it is possible, for instance, to assess whether voters are provided with "good news" or "bad news" relative to their measured prior beliefs, and to assess the consequences for individual voting decisions that are measured in similar ways across studies. However, the nature of the information varies substantially across studies—as do the research contexts. It is exceedingly difficult to identify the effects of mechanisms, and to hypothesize about how research context may make particular mechanisms operative, even with very extensive variation in treatments [what Green & Gerber (2012) call the "implicit mediation" approach to identifying mechanisms]. Each of the field experiments included in this initiative features some variation in treatment across the first and second treatment arms, yet this variation is obviously limited. This pilot of the Metaketa model is in many ways a "proof of concept" that may offer beneficial lessons for similar research conducted in the future. Several important questions remain, however—for instance, about the point in a particular research literature at which such an investment (both in terms of money and researcher time) is merited.

CONCLUSION

By replicating experiments, sometimes in new contexts, researchers may assess both the reliability and the portability of previous research results. This article has discussed several challenges to the cumulation of knowledge, as well as several key practices that may foment cumulative learning. The recommendations discussed include simple analysis, transparent reporting, third-party replication, and coordinated research. That some of these practices are not more widespread can arguably be traced to the tendency to privilege innovation and devalue follow-up studies, particularly those that do not lead to surprising results. Innovation has an important place for knowledge production, but so do verification and validation. Without overcoming these challenges, the payoff to learning from experiments may be substantially lessened.

The solutions described in this article offer possible correctives, yet there are also many difficulties. Research prespecification is increasingly utilized by experimental researchers, yet the specificity and utility of preanalysis plans vary widely. Results-blind review seems to be the best corrective to publication bias, yet it remains to be seen how well reviewers can evaluate projects only on the strength of the question, theory, and design—absent the findings. Models for coordinated research may shift incentives in favor of greater and more harmonized replication, yet whether those models are themselves sustainable is also an open question. One of the ironies of recent initiatives to coordinate and replicate research is that they are innovative—and therefore may command more attention and funding than will follow-up efforts. Another interesting

question is whether the descriptive data produced by experimental results will have a long intellectual half-life. Survey data or other data generated through many observational studies can have utility that long outstrips their intended purpose, and whether the data generated through initiatives such as Metaketa, described here, will have that salutary side effect remains to be seen.

There are also reasons for optimism. As a consequence of the recent emphasis on design-based inference, the quality of empirical work appears to have improved in the direction of simpler and more credible data analysis. Even apparent setbacks—such as the discovery of fraud in a leading experimental study, as described in the introduction—may indicate progress. Without data-archiving requirements at leading journals, and without at least some incentives to engage in replication in the first place, the fraud would likely not have been discovered. The challenges described in this article are difficult to surmount—and indeed are important obstacles for many kinds of research, not just for experiments—yet the new practices I describe may offer substantial benefits for experimental and observational research alike. Thus, although the question of what experiments can achieve in political science remains open, supplementary practices such as prespecification, replication, and coordination may help experimental research attain what randomization alone cannot achieve.

APPENDIX

In this appendix, I present R code for the simulation mentioned in the "Simple Analysis" section of the text.[20] The simulation demonstrates that under the box model described in that section, differences of proportions (such as differences between the percentages in **Table 1**) are unbiased estimators for average treatment effects. Certain estimators based on the logistic regression model are numerically equivalent to these differences of proportions and are therefore also unbiased for those estimands. The simulation also shows that the conservative standard errors described in the text are very accurate.

First, I generate potential outcomes for 500 mothers to produce averages that approximately match the percentages in **Table 1.** I "freeze" these potential outcomes for use in repeated randomization in the simulation.

```
set.seed(54321)
po <- as.data.frame(cbind(as.numeric(1:500<=.105*500),
                          as.numeric(1:500<=.09*500),
                          as.numeric(1:500<=.148*500),
                          as.numeric(1:500<=.197*500)))
names(po) <- c("trueA", "trueB", "trueC", "trueD")
true <- apply(po, 2, mean)
```

Now, I build a function that randomly assigns units to one of the four experimental arms in the factorial design; calculates differences of proportions between each treatment group and the control group; and estimates the standard errors of the differences using the conservative formula mentioned in the text. Finally, it fits a logistic regression of the observed outcome on the treatment vectors and calculates standard errors using Fisher information.

[20] The code is produced in an R markdown file in RStudio Version 0.99.887, running R Version 3.3.0 on Mac OS X 10.11.4.

```
factorial.result <- function(po){

  # define n, number of arms and the size of each arm from matrix of potential outcomes
  n <- nrow(po); num_arms  <-  ncol(po); groupsize <- n/num_arms

  # Assign units to one of the four arms and get observed outcomes
  treat <- sample(rep(1:num_arms, each=groupsize), n, replace=F)
  y_observed <- (treat==4)*po$trueA + (treat==1)*po$trueB +
    (treat==2)*po$trueC + (treat==3)*po$trueD
  data  <- as.data.frame(cbind(y_observed, treat))

  # Get observed values and averages as a function of treatment assignment
  est_cells <- table(data$y_observed, data$treat)[2, c(4, 1:3)]/groupsize
  names(est_cells) <- c("est. A","est. B", "est. C", "est. D")

  # Calculate estimated effects (differences of proportions)
  diffmeans_est_effects <- est_cells[2:4] - est_cells[1]
  names(diffmeans_est_effects) <- c("diff B-A", "diff C-A", "diff D-A")

  # Calculate SEs for the difference-of-proportions estimators
  se_effects <- NA
  for (i in 1:3){ se_effects[i] <- sqrt(var(data$y_observed[data$treat==i])/
                                        groupsize +
                                        var(data$y_observed[data$treat==4])/
                                        groupsize) }

  # Now, construct right-hand-side indicator variables and fit the logistic regression:
  case <- as.numeric(data$treat==1|data$treat==3)
  cash <- as.numeric(data$treat==2|data$treat==3)
  casecash <- as.numeric(data$treat==3)
  fit <- glm(data$y_observed ~ case + cash + casecash,
             family=binomial(link=logit))

  # Calculate predicted probabilities that Y=1 under each possible assignment.
  newdata1 <- as.data.frame(rbind(c(0,0,0), c(1,0,0), c(0,1,0), c(1,1,1)))
  predicted <- as.data.frame(predict.glm(fit, newdata = newdata1, se.fit=TRUE,
                                         type = "response"))
  probs <- predicted[,1]
  names(probs) <- c("alpha","alpha+beta1","alpha+beta2","alpha+beta1+beta2+beta3")

  # Calculate estimated effects from the regression (differences in Prob Y=1)
  logistic_effects <- probs[c(2,3,4)] - probs[1]
  names(logistic_effects)<-c("logistic_B-A","logistic_C-A","logistic_D-A")

  # standard errors for the regression based on predicted probabilities
  se_log_effects <- NA
  se_log_effects[1]<-sqrt(predicted[2,2]^2+predicted[1,2]^2) #se_log_BA
  se_log_effects[2]<-sqrt(predicted[3,2]^2+predicted[1,2]^2) #se_log_CA
  se_log_effects[3]<-sqrt(predicted[4,2]^2+predicted[1,2]^2) #se_log_DA

  # now return all the important output
  return(c(est_cells, probs, diffmeans_est_effects, logistic_effects,
           se_effects, se_log_effects))
}
```

I now use this function to conduct a simulation in which 125 of 500 units are assigned to each of four treatment arms in each replicate. There are 10,000 replicates.

```
reps <- replicate(10000, factorial.result(po))

# take the means across the 10,000 assignments
average_est <- round(rowMeans(reps), digits=3)
```

What does the simulation demonstrate? First, because averages of random samples are unbiased estimators for averages in the population from which they are drawn, we can see that (1) in each treatment condition, the average proportion of units with $Y_i = 1$, across the 10,000 replicates, equals (2) the corresponding average potential outcomes—i.e., the truth.

```
## est. A est. B est. C est. D
##  0.104  0.090  0.148  0.196


## trueA trueB trueC trueD
## 0.104 0.090 0.148 0.196
```

Moreover, because predicted probabilities under the logistic regression fit are numerically equivalent to those proportions, (3) the average estimates of $\Lambda(\alpha)$, $\Lambda(\alpha + \beta 1)$, $\Lambda(\alpha + \beta 2)$, and $\Lambda(\alpha + \beta_1 + \beta_2 + \beta_3)$ coincide with (1)—and therefore with the truth.[21]

```
##                   alpha             alpha+beta1             alpha+beta2
##                   0.104                   0.090                   0.148
## alpha+beta1+beta2+beta3
##                   0.196
```

In consequence, both the differences of proportions and the differences of predicted probabilities are unbiased for differences in average potential outcomes—that is, for the average treatment effects:

```
## diff B-A diff C-A diff D-A
##   -0.014    0.044    0.092


## logistic_B-A logistic_C-A logistic_D-A
##       -0.014        0.044        0.092


## true_B-A true_C-A true_D-A
##   -0.014    0.044    0.092
```

What about the standard errors? We can compare (*i*) the standard deviation of the simulated distribution of difference-of-means estimators, across the 10,000 replicates, to (*ii*) the average of the nominal "conservative" standard errors described in the text and (*iii*) the average of the model-based standard errors, where the latter two averages are also taken across the replicates:

```
print(apply(reps, 1, sd)[9:11],digits=2)


## diff B-A diff C-A diff D-A
##    0.037    0.041    0.043


## se B-A se C-A se D-A
##  0.037  0.042  0.045


## se_log_BA se_log_CA se_log_DA
##     0.037     0.042     0.045
```

[21] The predicted probabilities are $\Lambda(\widehat{\alpha})$, $\Lambda(\widehat{\alpha} + \widehat{\beta}_1)$, $\Lambda(\widehat{\alpha} + \widehat{\beta}_2)$, and $\Lambda(\widehat{\alpha} + \widehat{\beta}_1 + \widehat{\beta}_2 + \widehat{\beta}_3)$, respectively; as a shorthand, these are indicated by `alpha`, `alpha + beta1`, etc. in the code.

As the simulation shows, the conservative standard errors are very close to the standard deviation of the simulated sampling distribution. Here, the model-based standard errors are fine, too, though they must be based on the predicted probabilities (as in the function factorial.result).

DISCLOSURE STATEMENT

The author is chair of the inaugural Evidence in Governance and Politics (EGAP) Metaketa selection committee and directs the Center on the Politics of Development (CPD) at the University of California, Berkeley, which administers grant funding associated with the Metaketa initiative. He has not received financial compensation for these roles.

ACKNOWLEDGMENTS

I am grateful to Guy Grossman, Macartan Humphreys, Susan Hyde, Mathias Poertner, and Guadalupe Tuñón for excellent comments and suggestions.

LITERATURE CITED

Angrist JD, Pischke J-S. 2009. *Mostly Harmless Econometrics: An Empiricist's Companion*. Princeton, NJ: Princeton Univ. Press

Angrist JD, Pischke J-S. 2014. *Mastering 'Metrics: The Path from Cause to Effect*. Princeton, NJ: Princeton Univ. Press

Angrist JD, Pischke J-S. 2015. Mastering metrics: teaching econometrics. *VOX: CEPR's Policy Portal*. May 21. **http://www.voxeu.org/article/mastering-metrics-teaching-econometrics**

Banerjee AV, Banerji R, Duflo E, Glennerster R, Khemani S. 2008. *Pitfalls of participatory programs: evidence from a randomized evaluation in education in India*. NBER Work. Pap. 14311

Banerjee AV, Duflo E. 2009. The experimental approach to development economics. *Annu. Rev. Econ.* 1:151–78

Benjamini Y, Hochsberg Y. 1995. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. R. Stat. Soc. Ser. B* 57(1):289–300

Björkman M, Svensson J. 2009. Power to the people: evidence from a randomized field experiment on community-based monitoring in Uganda. *Q. J. Econ.* 124(2):735–69

Brady HE, Collier D, eds. 2010. *Rethinking Social Inquiry: Diverse Tools, Shared Standards*. Lanham, Maryland: Rowman & Littlefield. 2nd ed.

Broockman D, Kalla J, Aronow P. 2015. *Irregularities in LaCour 2014*. Work. pap., Stanford Univ. **http://stanford.edu/~dbroock/broockman_kalla_aronow_lg_irregularities.pdf**

Cartwright N, Hardie J. 2012. *Evidence-Based Policy: A Practical Guide to Doing It Better*. New York: Oxford Univ. Press

Clemens M. 2015. *The meaning of failed replications: a review and proposal*. Cent. Glob. Dev. Work. Pap. 399

Collier D, Brady HE, Dunning T. 2015. *The set-theoretic comparative method (STCM): fundamental problems and better options*. Work. pap., Univ. Calif., Berkeley

De Rooij EA, Green DP, Gerber AS. 2009. Field experiments on political behavior and collective action. *Annu. Rev. Polit. Sci.* 12:389–95

Dunning T. 2012. *Natural Experiments in the Social Sciences: A Design-Based Approach*. New York: Cambridge Univ. Press

Dunning T, Hyde S. 2014. Replicate it! A proposal to improve the study of political accountability. *Monkey Cage blog. Washington Post*, May 16. **http://www.washingtonpost.com/blogs/monkey-cage/wp/2014/05/16/replicate-it-a-proposal-to-improve-the-study-of-political-accountability**

Dunning T, Monestier F, Piñeiro R, Rosenblatt F, Tuñón G. 2015. *Positive versus negative incentives for compliance: evaluating a randomized tax holiday in Uruguay*. Work. pap., Univ. Calif., Berkeley

Fang A, Gordon G, Humphreys M. 2015. *Does registration reduce publication bias? No evidence from medical sciences*. Work. pap., Columbia Univ.
Freedman DA. 2009a. *Statistical Models: Theory and Practice*. New York: Cambridge Univ. Press. 2nd ed.
Freedman DA. 2009b. Randomization does not justify logistic regression. In *Statistical Models and Causal Inference: A Dialogue with the Social Sciences*, ed. DA Freedman, D Collier, JS Sekhon, PB Stark, pp. 219–42. New York: Cambridge Univ. Press
Freedman DA, Pisani R, Purves R. 2007. *Statistics*. New York: W.W. Norton. 4th ed.
Gerber AS, Green DP, Nickerson DW. 2001. Testing for publication bias in political science. *Polit. Anal.* 9:385–92
Gerber AS, Malhotra N. 2008. Do statistical reporting standards affect what is published? Publication bias in two leading political science journals. *Q. J. Polit. Sci.* 3(3):313–26
Gherghina S, Katsanidou A. 2013. Data availability in political science journals. *Eur. Polit. Sci.* 12:333–49
Green D, Gerber A. 2012. *Field Experiments: Design, Analysis, Interpretation*. New York: W.W. Norton
Hamermesh DS. 2007. *Replication in economics*. NBER Work. Pap. No. 13026
Humphreys M, de la Sierra RS, van der Windt P. 2011. *Social and economic impacts of TUUNGUANE 1: mock report*. **http://cu-csds.org/wp-content/uploads/2011/03/20110304-MOCK-REPORT-SHARED-FOR-REGISTRATION.pdf**
Humphreys M, de la Sierra RS, van der Windt P. 2013. Fishing, commitment, and communication. *Polit. Anal.* 21(1):1–20
Humphreys M, Weinstein JM. 2009. Field experiments and the political economy of development. *Annu. Rev. Polit. Sci.* 12:367–78
Hutchings VL, Jardina AE. 2009. Experiments on racial priming in political campaigns. *Annu. Rev. Polit. Sci.* 12:397–402
Hyde S. 2015. Experiments in international relations: lab, survey, and field. *Annu. Rev. Polit. Sci.* 18:403–24
LaCour MJ, Green DP. 2014. When contact changes minds: an experiment on transmission of support for gay equality. *Science* 346(6215):1366–69
Laitin D. 2013. Fisheries management. *Polit. Anal.* 21(1):42–47
Lieberman ES, Posner DN, Tsai LL. 2013. Does information lead to more active citizenship? Evidence from an education intervention in rural Kenya. *World Dev.* 60:69–83
Lin W. 2013. Agnostic notes on regression adjustment to experimental data: reexamining Freedman's critique. *Ann. Appl. Stat.* 7(1):295–318
Malhotra N. 2014. *Publication bias in political science: using TESS experiments to unlock the file drawer*. Presented at West Coast Exp. Conf., Claremont Graduate Univ., May 9
Mauldon J, Malvin J, Stiles J, Nicosia N, Seto E. 2000. *Impact of California's Cal-Learn demonstration project: final report*. UC Data, Univ. Calif. Berkeley
McDermott R. 2002. Experimental methods in political science. *Annu. Rev. Polit. Sci.* 5:31–61
McCullough BD, McGeary KA, Harrison TD. 2006. Lessons from the JMCB archive. *J. Mon. Credit Banking* 38(4):1093–107
Miguel E, Camerer C, Casey K, Cohen J, Esterling KM, et al. 2014. Promoting transparency in social science research. *Science* 343:30–31
Monogan JE III. 2013. A case for registering studies of political outcomes: an application in the 2010 House elections. *Polit. Anal.* 21(1):21–37
Monogan JE III. 2015. Research preregistration in political science: the case, counterarguments, and a response to critiques. *PS Polit. Sci. Polit.* 48(3):425–29
Olken B. 2007. Monitoring corruption: evidence from a field experiment in Indonesia. *J. Polit. Econ.* 115(2):200–49
Palfrey TR. 2009. Laboratory experiments in political economy. *Annu. Rev. Polit. Sci.* 12:379–88
Ragin CC. 1987. *The Comparative Method: Moving Beyond Qualitative and Quantitative Strategies*. Berkeley: Univ. Calif. Press
Ragin CC. 2000. *Fuzzy-Set Social Science*. Chicago: Univ. Chicago Press

Rosenbaum PR. 2002. *Observational Studies*. Springer Ser. Stat. New York: Springer-Verlag. 2nd ed.
Simonsohn U, Neilson LD, Simmons JP. 2014. P-curve: a key to the file drawer. *J. Exp. Psychol.* 143(2):534–47
Singal J. 2015. Michael LaCour probably fabricated a document about research integrity. *New York*, June 1. **http://nymag.com/scienceofus/2015/06/lacour-probably-fabricated-an-integrity-document.html**
Tanner S. 2014. QCA and causal inference: a poor match for public policy research. *Qual. Multi-Method Res.* 12(1):15–24

Annual Review of Political Science
Volume 19, 2016

Contents

Indexes

Errata

An online log of corrections to *Annual Review of Political Science* articles may be found at http://www.annualreviews.org/errata/polisci