

Research Transparency and Data Archiving for Experiments

Rose McDermott, *Brown University*

Although still more common in medical studies and some other areas of social science such as psychology and behavioral economics, experimental work has become an increasingly important methodology in political science. Experimental work differs from other kinds of research because it systematically administers a specific treatment to part of a population while withholding that manipulation from the rest of a subject pool. The best studies strive to keep all other aspects of the experiment similar, so that any emergent difference between the treatment and control group that emerge provide unparalleled traction in determining causal inference. Many other valuable forms of social research use observation of the natural world, rather than depending on intervention to advance understanding. Because experimentalists can create the environment or process they want to study, this strategy of intervention and manipulation constitutes the main distinction between experimental work and other forms of social observation.

In spite of this critical separation, experimentalists confront some methodological challenges and opportunities that both mimic and diverge from those scholars who are engaged in other forms of qualitative and quantitative work that depends on observation of the natural world. Like those scholars who conduct quantitative research, experimentalists often work with large data sets involving multiple pieces of independent observations, most of which are analyzed statistically to determine results and interpreted in ways familiar to quantitative researchers. However, like qualitative researchers, experimentalists often work with sensitive populations with concerns about protecting individual identities. This work requires that subjects' safety and confidentiality are protected above all other values. In addition, also like qualitative researchers, interviews that take place as part of debriefing may prove informative and useful in ways that require a different kind of data archiving than standard number files provide. Finally, unlike either quantitative or qualitative work, experiments also involve particular experimental protocols, treatment assignments, manipulations, and even Consolidated Standards of Reporting Trials (CONSORT) files that show subject mortality that require unique characterizations. These files differ from other types of work and may require unique standards to achieve research transparency and proper archiving. CONSORT was created to improve the transparency in reporting on randomized clinical trials. The CONSORT statement encourages reporting of a 25-item checklist and flow diagram. This standard reporting strategy provides complete and transparent reporting of all aspects of design,

analysis, and interpretation of experimental investigations. The checklist includes title, abstract, introduction, methods, results, conclusion, and supplementary information. The flow chart shows how subjects move through the four stages of clinical trial from enrollment, assignment to condition, follow-up, and analysis.

This article in this symposium outlines some of the current standards and developments designed to achieve increased transparency and archiving of experimental work. The following section discusses some areas of consensus in reaching this goal as well as some challenges that confront scholars who wish to pursue experimental work, particularly in the context of field experiments. The second section outlines some potential strategies and next steps that may be useful to maximize research transparency and data archiving in concert with the goals pursued by other research traditions in political science.

RECENT DEVELOPMENTS

This brief overview examines recent attempts to increase transparency and archiving in experimental work, efforts particularly evident in at least two distinct areas. First, the new *Journal of Experimental Political Science (JEPS)*, initiated under the auspices of the APSA's Experimental Methods Section, whose initial editors will be Rebecca Morton and Joshua Tucker at New York University, has established clear standards for submission, review, and the conduct of research. Second, similar specific standards have been pursued and endorsed by Experiments in Governance and Politics (EGAP), which has tasked itself with "supporting experimental research on the political economy of development." This project encompasses many prominent scholars working in both the laboratory and field experimental areas and has been spearheaded by Jeremy Weinstein at Stanford University and Macartan Humphreys at Columbia University, among others. Although many other efforts exist, these two have been the most systematic attempts to assemble scholars working in experimental traditions and secure commitment to follow particular procedures designed, in part, to achieve transparency, accountability, and replicability.

A third effort has recently emerged: the Berkeley Initiative for Transparency in the Social Sciences (BITSS). Although started by a group of scholars primarily interested in development studies, BITSS encourages research transparency across a wide array of social science disciplines, including political science. In particular, it promotes study registries, data sharing, and replication through learning, discussing, and disseminating best practices. While many members of its leadership are experimentalists, BITSS does not limit itself

to experimental processes and procedures, nor does it restrict the content to development work as EGAP largely does. Many of its suggestions and strategies are applicable to advancing research transparency in experimental work, but BITSS's dictums supporting transparency and replication extend beyond interventional research into observational methods as well.

The new *JEPS* established a set of instructions for both contributors and reviewers that directly speak to many of the issues raised in this symposium. In addition, it directly endorses the reporting standards for experimental work developed by the Experimental Methods Section's Standards Committee. The Standards Committee that developed this document was headed by Alan Gerber and included Kevin Arceneux, Cheryl Boudreau, Conor Dowling, Sunshine Hillygus, and Tom Palfrey. These standards addressed many aspects of experimental design that affect experimental treatment across both laboratory and field settings. Specifically, these standards ask authors, first, to clearly state their hypotheses. Next, authors are asked to explicitly state which subjects were included or excluded from consideration, how and where they were recruited, and to provide the dates when the study was conducted. If a survey was used, authors are asked to supply response rate. Although not stated, best practices would expect that the survey instrument should also be provided, even if only as part of an online supplementary index. This subject information is crucial for achieving appropriate levels of research transparency, more so in experimental work because experimental work proceeds largely through aggregation that occurs as both subject population and context are expanded or shifted. Proper replicability procedures demand that subsequent researchers are aware of the previous populations that have been investigated, and the context under which they have been examined.

Then, the standards request that authors provide statements of what are called their "allocation method." This refers to information regarding whether and how processes of randomization were used in the experiment; experimenters are asked to provide information about how this was accomplished and evidence that it was achieved. Scholars are also asked to provide information about whether subjects, administrators, and analysts were blind to the conditions of the subjects across treatments. In addition to standard requirements about detailing the conditions of treatment and control, providing the instruments of measurement and assessment, following careful standards of analysis, noting institutional review board (IRB) approval and whether or not deception was used, the most unusual and potentially controversial standard asks authors to provide a CONSORT flow chart detailing how many subjects were lost across the course of the study by treatment condition. Where low noncompliance exists, authors are instructed that they can omit the diagram and replace it with a statement in the text. The goal here is to make clear how and why certain subjects may have dropped out of one condition more than another, possibly indicating a systematic difference in who is affected by the treatment and why that might otherwise be lost if only completed subjects were analyzed and presented in the final results. Although this requirement is often considered in medical experiments—where, for exam-

ple, drug side effects may cause more patients in one condition than another to drop out of a study and this information may be crucial for issues of patient compliance—this standard is not typical in either the psychological or economics experimental literature. The political science standards discussed here go one-step further to require authors to report statistics for intent-to-treat, which is another technique for ensuring that results do not reflect biased findings by failing to incorporate those lost to analysis at earlier phases of the experiment.

In its instructions to contributors and reviewers, *JEPS* goes beyond the standards for experimental work provided by the Standards Committee. Because the journal requests shorter articles, *JEPS* notes that some of the material required by the standards may be uploaded into online supplementary material so that it does not count against manuscript word count but will still allow other scholars to find the information. In addition, they require not only evidence of IRB authorization but also disclosure of potential conflict of interest; of course, this is very important in any cases where scholars also have a financial interest in companies that run surveys or experiments for profit. Perhaps most innovatively, a review history of the manuscript, which details which journals the article has previously been submitted to and their responses to requests for revision, is required. This may allow work, which was rejected for lack of wider audience interest or lack of experimental sophistication on the part of reviewers or editors, to receive more expedited review. Finally, the journal does require, for replication purposes, that all data relevant to an experiment be submitted. The instructions proceed as follows:

For experiments these files should include original experimental instructions or other experimental instruments used in the experiments such as surveys, videos, computer programs, etc., and the raw data from the experiment. For empirical papers, both using experimental or observational data, the final data set(s) and programs used to run the final models, plus a description of how previous intermediate data sets and programs were used to create the final data set(s) must be provided . . . Authors must provide a Readme PDF file listing all included files and documenting the purpose and format of each file provided, as well as instructing a user on how replication can be conducted. If a request for an exemption based on proprietary data is made, authors should inform the editors if the data can be accessed or obtained in some other way by independent researchers for purposes of replication. Authors are also asked to provide information on how the proprietary data can be obtained by others in their Readme PDF file. A copy of the programs used to create the final results is still required.

Similar instructions are provided to reviewers, noting most unusually that the journal encourages the submission of replication studies and null findings, thus explicitly encouraging authors to submit work designed to replicate studies conducted by other investigators, or studies whose lack of findings can save others from wasting time undertaking work others have already found to be unsubstantiated.

An additional major effort designed to achieve consistency, replicability, transparency, and accountability in experimental work has been undertaken by those scholars involved

in EGAP. Most of these studies, because of the content they examine, tend to take place in field contexts. In addition, although the explicit subject goals of this group are designed to investigate issues specifically related to political economy, governance, and development, many of the methodological issues they confront and address do not differ substantively from those facing any experimentalist.

Two main aspects are central to their campaign to enhance transparency and archiving. First, like *JEPS*, EGAP has developed a set of standards adopted by the unanimous vote of the membership. Endorsing these standards is a condition of membership for incoming new members. This statement is relatively simple and straightforward and encompasses human subject protection, transparency, rights surrounding review, and publications of data and findings and remuneration, which, while discouraged at the very least, must be disclosed. This last item appears similar to the conflict of interest statement requested by the editors of *JEPS*.

The second aspect revolves around various strategies designed to institutionalize procedures to ensure transpar-

enter into the overall literature, but many scholars may continue to reinvent the wheel, not knowing that previous work has shown that certain speculated relationships fail to exist. Another kind of registration system, which would require full-scale review of an overall research design, might require journal editors to make publication decisions based on the design prior to the collection of data so as to prevent such bias at the back end. Under this scheme, when accepted, journals would be required to publish papers regardless of the outcome of results. It is easy to see why journals may not want to comply with a strategy that ties its hands in this way, because there may be other reasons, including something as simple as bad writing, which may incline the journal to eschew publication at the final stages for reasons not evident in the presentation of the design.

The public nature of these prior commitments would severely reduce the incentive, and dramatically escalate the humiliation, associated with violating the original research goals set out by the experimentalists. One can imagine other kinds of registration strategies designed to “name and shame”

Similar instructions are provided to reviewers, noting most unusually that the journal encourages the submission of replication studies and null findings, thus explicitly encouraging authors to submit work designed to replicate studies conducted by other investigators, or studies whose lack of findings can save others from wasting time undertaking work others have already found to be unsubstantiated.

ency. Many of these involve various kinds of registration opportunities. For example, scholars are encouraged to register their preanalysis plans to reduce the likelihood of “data fishing” or what in the old days of social psychology used to be called “dust bowl empiricism,” which can become a serious problem especially with the tools available with current computing power. This registration tool, which already has been substantially used, allows scholars to state which aspects of their data they will analyze in which way, detailing both hypotheses and methods of analysis; these records are then freely available on the EGAP website for other scholars to see.

The typical posting on the EGAP website lists the particular hypotheses that experimentalists plan to test. However, a continuum of registration demands or designs could be included. Simple registration seeks to prevent the problem of scholars cherry-picking those aspects of their data that show the best or strongest results, or confirm a particular theoretical or ideological position. This registration may help keep authors honest about what they plan to investigate, although it may also unnecessarily restrict creativity by preventing the credible examination of true surprises that can emerge in the context of any data collection. Moreover, this simple registration strategy does nothing to address the problem of publication bias among journals that remain stubbornly resistant to publishing null results in particular. Without adequate representation of the full range of outcomes, not only does bias

norm violators providing successful avenues by which scholars can subtly, but powerfully, strengthen best practice norms in experimental methods.

One of the most interesting additional initiatives is the Transparency and Accountability Initiative funded by a host of high-profile private organizations as well as nongovernmental organizations, including the Ford Foundation, Open Societies, and the Hewlett Foundation. Although focused primarily on achieving these goals in the area of international development, this initiative appears designed to provide mechanisms which can allow citizens to hold their governments accountable through a wide variety of educational, technological and policy innovations. The link to this initiative can be found here: <http://www.transparency-initiative.org/about>.

CONVERGENCE AND CONTENTION IN BEST PRACTICES

Points and patterns of consistency appear to be clearly emerging in experimental research designed to enhance research transparency and data archiving. First, most notably, experimentalist largely consider both of these things to be not only good things, but necessary for their own research to proceed apace. Specifically, most experimentalists, perhaps more than those working in other research traditions, know that experimental work proceeds through a process of aggregation and replication, whereby findings from previous work are extended to new populations or within different contexts. For this work to be done well, it must be done carefully, to determine the

limits of particular phenomena and to understand the nature of particular contingencies on expected results. In other words, endorsing and enhancing these practices within the community of experimentalists improves everybody's work, and efforts that reinforce individual incentives are often easiest to encourage and expand.

Aside from issues related to transparency and archiving, experimentalists also seem to strongly endorse issues related to achieve accountability. This is most notable in the items related to requiring IRB approval for human subjects, but also in the statements revolving around conflict of interest.

Although political science did not traditionally require that data sets be mounted with publication, scholars who wanted to replicate studies could typically request such data from the authors, and authors might note that such information was available on request in a publication. But as standards across disciplines converged toward posting data with publication, political science journals are increasingly moving in that direction as well. Innovations including advance registration, such as that offered by EGAP, provides even higher standards to which scholars can hold themselves accountable even prior to analysis, write-up, or publication.

Second, experimentalists across the board, whether based in the lab or field settings, clearly endorse the protection of human subjects. This extends beyond the cynical enlightened self-interest that recognizes that abused subjects talk to others and can make future experiments more difficult at best, and rain down lawsuits at worst. However, even in places where IRB approval is not yet the norm, such as many institutions in Europe, scholars recognize that well-treated subjects are not only more cooperative but also supply more accurate information, not only in their experimental response, but in the often crucial insights they can provide in proper debriefing procedures.

This human rights issue, however, does raise concerns related to subject confidentiality. Even when every reasonable effort is made to protect subjects' identity, the consequences of exposure may feel great to some subjects, particularly when studies are conducted in war-torn or contentious regions, or across conflictual groups, as often occurs in examinations of inter-ethnic discrimination or civil war. When subjects feel that exposure can be easily gleaned from the sensitive nature of the questions or the idiosyncratic nature of truthful responses, subjects may be understandably reluctant to participate, or to give accurate responses. More important, investigators who include such people really may be placing them at risk, and thus the obligation to protect under such circumstances becomes particularly acute. Investigators who are genuinely concerned about negative consequences devolving to any of their subjects should not include such individuals in their studies, even if significant costs redound to the study. Exclusion under such conditions remains the only ethical path. However, determining when such conditions may arise or be in place may not always remain obvious and the subject's perception must always take precedence over the judgment of the investigator.

I learned a searing lesson in the perception of identity that has stayed with me ever since when I conducted my war games at Harvard. I was taking a variety of measures, including saliva for hormonal analysis, and a copy of their handprint to mea-

sure finger length ratio. I wanted subjects to have an id number that was not their name but that they would remember over several months because of the panel nature of the study. So I used the standard used in VA studies that involved the last four digits of a person's social security number, which are uncommon enough to make replication in a small set rare, but not so unique as to be identifying. On the second day of the study, a young African American woman came in and I started to explain the protocol to her and she physically pulled back and said, "Wait. You want my DNA, my fingerprints, my Social Security number for a study funded by the Department of Defense and you're telling me this is anonymous and confidential? And why am I supposed to believe you?" I was stunned, but I instantly saw how the experience looked completely different than my intent when seen through her eyes. More for my sake than hers, I asked her, without requesting any data, what I could do that would make her feel comfortable. She said she was not sure. I asked if she would feel better if she could pick her own id number. She nodded. She picked a number I still remember for its simplicity a decade later but the point was not that it could not be guessed; the point was that she picked it, not me. I then explained about the copy of the hand and she looked at me and said, so, if you blacked out my fingerprints, you would still want it?" I said yes. She copied her hand. I blacked it out, but I went one step further. I measured what I needed in front of her and took the number and then destroyed the copy while she watched. The information did not change, and the DNA I could extract from the saliva (but did not) was what it is: a totally unique identifier that could never be anything less because of its nature. But I had a completely different understanding of the nature of subject identity and the sensitivity and responsibility involved in protecting individuals not from what I would do, but from what they feared I could do.

However, issues related to protecting the identity of experimental subjects does remain distinct from the graver risks that may accompany the kind of in-depth interview work typically conducted by qualitative researchers. In experimental cases, the easiest way around subject identification, barring the use of biological data, is to never collect subject names; simply assign id numbers that tie relevant linked data together. Anyone who may want to know the identity of participants will never be able to ascertain this information because it was never collected. This becomes an issue, for example, when universities tried to use such information to find students who were in violation of immigration laws to pursue orders of deportation against them. If names that link status to a particular individual are never taken, such protection is ensured even if suspicions arise. With qualitative researchers using interview data, information may reveal the identity of a subject even barring the collection of a name because of the specificity of the information provided; this poses greater risks for the participants and greater challenges for the researcher. This topic is dealt with more in the Colin Elman and Diana Kapiszewski contribution on qualitative research.

This dictum to never collect subject names may run afoul of traditions in both survey research and economics that require compensation of subjects in a way that requires subject

identification. The typical way that this is addressed is to keep two distinct logs that are not merged, one which contains the data, and the second which contains the names of subjects along with contact information for purposes of remuneration or reimbursement. However, constructing a wall between names and data is not always feasible or even successful, often for reasons as simple as the order of entry between the two being matched one to one by research assistants who may not be familiar with the importance of confidentiality. The more unassailable way to address these concerns is to remunerate subjects on sight, either with cash or gift cards chosen from a menu of options. If this recompense needs to be provided through the Internet, it can be done through the generation of randomly assigned codes that can then be redeemed for particular rewards or benefits.

The unsettled challenges that remain at the level of large-scale norms seem to relate to the proprietary nature of data, which is an issue not exclusive to experimentalists. First, any scholar who expends tremendous time and effort designing and conducting a study and collecting data may not want to give it all away before they have had a chance to fully explore all of their potential findings. In this case, by parsing data into pieces and publishing and posting findings from particular pieces and parts of data this can be partially avoided. This strategy is not always viable, especially if many parts of the data are linked theoretically or empirically. Under these conditions, incentives can pull in opposite directions when scholars want to publish early but also to protect their data. In such cases, the researcher may have to decide how to approach these constraints on a case by case basis. Moreover, such a strategy can also run contrary to the expectations or demands imposed by the various registration strategies discussed previously. Although the timing of data release remains distinct from actual research design strategy, and thus need not be delineated in advance in a registry, scholars need to think seriously during the design phase of their work about how they might need or want to parse up their findings during the write-up and publication phase. While it has become increasingly common to break studies into ever smaller parts in search of the ever-larger quantities of publication demanded by promotion and tenure committees, work should be divided according to its conceptual or theoretical specifications rather than its strategic value.

The second case is more common and relates to the embargo standards imposed by Time Sharing Experiments in Political Science (TESS) funded by the National Science Foundation (NSF). When scholars have their proposals accepted and run by TESS, they get the data first, but after a year of embargo, it becomes publically available as mandated by all taxpayer funded work. This means that if investigators do not complete their work in this time, other scholars seeking data can use it. Although often such data goes unused by both investigator and observers, as Diana Mutz's book *Population Based Survey Experiments* (2011) illustrates so well, the time-limited nature of the data embargo does pose a risk to experimentalists who may lose the ability to publish first on their own data.

FUTURE CHALLENGES

The challenges that seem to confront experimentalists pursuing best practices and high standards for research transparency, accountability, replicability, and data archiving overlap with the challenges facing both qualitatively and quantitatively oriented scholars. And here overlap appears to be the key word. One of the risks with various groups pursuing the same agenda in different ways is that the norms that develop may become haphazard or too narrow in orientation. Specifically, for norms to become widely accepted, they must have wide adherence, and when various groups each develop standards and practices independent of one another, but seek to impose their particular branding on their contributors, reviewers, or participants, regulation may become burdensome rather than protective, especially if such rules and procedures have significant areas of disagreement or neglect. Such territoriality may work in opposition to the larger goals, as we have learned often happens in domestic and international politics as well.

Only the most disciplined scholars can achieve true freedom. Creativity does not result from luck or serendipity. Rather, it emerges when a prepared mind encounters unexpected processes in the midst of recognized patterns and structures. Just as it takes a dancer years and years to develop the physical prowess, muscle strength, and skill to express truly original movement, it requires the most tedious discipline and practice as a scientist to develop the experience and talent required to know when deviations from the standard will lead to total failure and when it just might instigate the spark of discovery known as genius.

Only the most disciplined scholars can achieve true freedom. Creativity does not result from luck or serendipity. Rather, it emerges when a prepared mind encounters unexpected processes in the midst of recognized patterns and structures. When best practices become habit, time and energy need no longer be spent on organization and logistics but rather can be allocated to the recognition or generation of such patterns and dynamic processes. Just as it takes a dancer years and years to develop the physical prowess, muscle strength, and skill to express truly original movement, it requires the most tedious discipline and practice as a scientist to develop the experience and talent required to know when deviations from the standard will lead to total failure and when it just might instigate the spark of discovery known as genius. Best practices and norms of transparency and accountability may need to be tailored to specific sub types of particular research methodologies. However, the broader goals need to be shared by journal and press editors, organized sections, and the wider political science community if they are to be adopted as functional and effective norms. Achieving consistency may be a campaign beset by obstacles, but accomplishing the successful adoption of widespread norms of research transparency, data archiving, accountability, and replicability is a goal worth striving for because it not only serves us as academics, helping us conduct better work and receive more credibility from the larger research community, but it also should allow us to communicate our results with more confidence, accessibility, and assurance to our students and the larger public. ■