The Oft-Neglected Role of Protocol in the Design and Reporting of Experiments

Thomas J. Leeper
Northwestern University
leeper@u.northwestern.edu

December 6, 2010

An average treatment effect is relatively simple to estimate. The basic requirements of experimental research that make that estimation straightforward and unconfounded – manipulation of causal variables and control over all else – are often far more difficult to achieve than is often imagined or implied. In this article, I continue the discussion started by Lupia and Druckman in the inaugural issue of this newsletter by emphasizing how the maintenance and publication of a detailed experimental protocol is critical for the analysis and reporting of experiments.

Developing Experimental Protocol

Like any novel data collection, an experiment is complicated. It begins with research questions, theory-building, and hypothesis generation. Experiments are unique, however, in that these initial stages are followed by a somewhat murky process of developing stimulus (treatment) materials, disseminating those stimuli, and assessing outcome measures. The idea of "just doing an experiment" has probably occurred to many political scientists – including non-experimentalists – but the process of executing an experiment, and doing it well, is rarely a topic heavily emphasized in published descriptions of experiments. Executing an experiment well requires quality protocol, by which I mean the detailed justification and explanation of the experiment that will be implemented and the record of how that intended experiment may not have occurred, for what units, and why.

A well-thought-out experimental protocol is the experimenter's version of the observationalist's well-thought-out statistical model of causes, outcomes, and controls. Rosenbaum (2010) describes a "poorer observational study" where "if sufficiently many analyses are performed, something publishable will turn up sooner or later." This is not the way research – experimental or observational – should be conducted. Instead, he offers that "before beginning the actual experiment, a written protocol describes the design, exclusion criteria, primary and secondary outcomes, and proposed analyses" (7). Lupia (2010) suggests that researchers record "Record all steps that convert human energy and dollars into datapoints."

Both quotes provide useful general advice, but the experimental protocol should also includes many specific features about the intended experiment, including:
Theory and Hypotheses

- Details of what outcome variable is under examination (including the values of its potential outcomes) and the causal factors (alone or in combination) that are theorized to affect the outcome variable
- References to relevant prior designs (published or unpublished) that inform the design, implementation, and/or analysis of this experiment
- Exact listing of hypotheses with discussion of how each hypothesis is reflected in one or more features (manipulations or controls) of the intended design
- Discussion of how hypotheses were used to construct the design (rather than design to construct hypotheses), including how each hypothesis is testable (i.e., falsifiable) in the intended design, including any anticipated confounding between manipulated causes and observed or unobservable alternative causes

Instrumentation

- Details of how theoretical constructs are manipulated by the experimenter, including exact wording and presentation of stimulus/treatment materials
- Details of how covariates and outcome variables are measured and scaled and exact question wordings or coding schemes if subjective and/or self-report measures are used
- Explanation of how stimuli are predicted to affect only the intended causal construct[1]
- Details of pretesting (of stimuli and/or outcome measurement techniques) to validate whether stimuli manipulate the causal constructs as anticipated by the researcher, by manipulating only the intended cause, in the theorized direction, and with what intensity
- Intended mode(s) in which each unit will be exposed to each stimulus (e.g., in-person, via phone, internet, mail, television, radio, print, etc.) and the mode in which outcomes will be observed (e.g., in-person, laboratory, internet, phone, voting records, etc.)

Population, Sample, and Treatment Assignment

- Details of sample construction, including how a sampling frame was constructed of population units (if applicable) and how units from that frame were selected for or excluded from treatment assignment

[1]If experimenters are uncertain about how a given stimulus will alter the theoretical construct, pretesting should be conducted (as described). If a stimulus potentially affects more than one causal construct (and alternative stimuli are infeasible), nonequivalent outcome measures – that measure the desired construct and separately measure other causes that should be unaffected by stimuli – should be developed to demonstrate that the intended cause produced the observed outcome and a confounding cause was not at work.

- Details of randomization procedures, including how random numbers were generated (whether a uniform or some other probability distribution was used and from where those numbers were obtained) and assigned (across the entire sample, or within blocks, clusters, or a combination thereof)

Implementation

- Intended (and executed) schedule for when, where, and how treatments will be applied and outcome variables measured and by whom (the experimenter or an assistant, or a government agency, corporation, etc.); if multiple sessions or repeated exposure to stimuli are part of the design, the schedule should specify and justify the order of what stimuli are applied at each point in time

- Procedures for how units are made blind to their treatment assignment, how data are collected and stored regarding each unit prior to analysis, and how anyone implementing the study are blind to each unit's treatment assignment and value(s) of outcome variable(s) until analysis

- Details of manipulation checks and post-randomization covariate balance tests, including how and when manipulation checks will be administered and what balance tests will be used and what predetermined degree of imbalance will be considered problematic for experimental inference

- Procedures for how deviations from protocol will be recorded and utilized in data analysis, including errors made by experimenters, item and unit noncompliance or attrition, and other relevant notes

Analysis

- Definition of specific causal effects to be estimated for testing each hypothesis; if more than two treatment groups are implemented, this should include what treatment groups or pooled treatment groups are used to estimate those effects (i.e., specification of contrasts)

- Explanation of specific statistical analysis to be performed, including how covariates will be used in the analysis if at all (i.e., for regression, subclassification, subgroup analysis, etc.)

- Plan for how to analytically address noncompliance, attrition, missing data, or covariate imbalance (through dropping observations, multiple imputation, intent-to-treat analysis, etc.)

- Additional analyses or changes to this protocol should be recorded in a "lab book" (Lupia 2010) and a record should be kept of all analysis performed (in order to reduce the risk of Type I errors)

In the same way that a regression model includes putatively causal variables and "controls for" covariates and confounds, the experimental protocol establishes what factors vary

between experimental units and what factors are "controlled for" by uniform execution of the experiment across randomized units. The protocol determines what features (of subjects, of context, and of experimenters) are included (observed and analyzed), excluded (unobserved), or controlled for (observed or held constant across units) in the final analysis, even if that analysis only consists of a simple difference of means or difference of proportions test for the average causal effect. Thinking about protocol forces the experimenter to make and record decisions about what is controlled in the experiment decisions that might otherwise not be made, leading to unclear control or a loss of control due to uneven implementation of the intended design.

Deviations from Protocol

The protocol should include description and justification of all decisions regarding the *intended experiment* and *implemented experiment*: i.e., if cost, feasibility, sample size, time, ethics, or other constraints alter the intended procedure, a record of these changes in the protocol document (and the justification thereof) is the only way to fully communicate what choices were made by experimenters and why. Druckman (2010) noted that experimental research is characterized by several myths; the myth of experiments always working perfectly is probably a necessary addition to his list; all experiments are – at some level – broken experiments. The experimenter's concern needs to be placed on how to prevent, record, and compensate for the deviations from protocol that manifest in the implemented experiment. Compensating for deviations that have already occurred is more difficult than preventing them, but can only be done when deviations are anticipated and properly recorded.

Rather than pretend that deviations from protocol do not occur, researchers should keep the "lab book" called for by Lupia to record statistical procedures, but they need to also record the earlier unit-, treatment group-, and sample-level deviations from the intended experiment protocol before any statistical analysis is performed. These deviations (especially if they correlate with treatment group assignment) may undermine the elegance of experimental inference by confounding causal factors and altering other aspects of experimental control. At a minimum, if any variations in implementation emerge among units, unit-level variables should to be incorporated into the experimental dataset that record:

- the date, time, and location that stimuli were distributed or applied
- the date, time, and location that outcomes were measured
- location and/or mode of each stage of the experiment
- waves of implementation, which place units in different broader contexts even if procedures in each wave are identical
- exogenous contextual events that influenced some units but not others (due to timing of implementation, geographical locale, etc.)
- any changes in questionnaires, changes in stimuli, or changes in the contact between experimenter and experimental units that affect only some units

- different experimenters/interviewers/coders/assistants who administered stimuli, collected outcomes, or coded data[2]
- other research-relevant, unit-level or group-level deviations from protocol[3]

Reading a typical journal article, it would seem the types of deviations from protocol included in the above list are rarely seriously anticipated by researchers, considered in data analysis, or honestly reported in publication. Dropping observations and/or conducting intent-to-treat analysis are common answers to deviations from protocol (often for noncompliance, application of the incorrect treatment, or attrition), but these are only applicable in some cases and each may introduce bias into estimates of causal effects depending on the type and extent of deviations. Other answers certainly exist, but before we can address these problems, we need to hold ourselves to high and uniform standards of adopting, recording, and disseminating our experimental protocols (see Gerber, Doherty, and Dowling n.d.).

Reporting Protocol (and Deviations)

This level of full disclosure of experimental protocols is relatively rare.[4] While a journal article will often include a brief discussion of protocol, specific details of protocol (and deviations) are often omitted to save space or because those details are seemingly unimportant (especially in the eyes of non-experimenters). But, this is unfortunate and often reflects, at best, partial compliance with standards for reporting observational research. Druckman (2010) points out "these standards [for surveys] do not ask for the reporting of information critical to experimentation, such as randomization checks, manipulation checks, pre-test results" (10). Reporting on experiments requires more thorough description in order for consumers of that research to understand how the causal effect(s) might be constrained by particular context, unit characteristics, poorly implemented protocol, and so forth.[5]

Including the full protocol as an (online) appendix is better than providing insufficient detail in an abbreviated methods section. Reviewers would criticize a paper with multiple statistical models if only one model were fully described and would certainly take issue with analysis that did not clearly specify the variables included in the analysis and the rationale for controlling for or excluding particular factors. Presenting a detailed experimental protocol is the experimenter's means of justifying relatively simple statistical analyses. But an average treatment effect is not just a difference in expected outcomes between groups; it is a difference conditional on a large number of features of the experimental design and broader context

[2]Demographic variables for these individuals may need to be included if they might influence unit behavior on sensitive issues (such as race or gender for studies that examine those issues) or if data were gathered on each unit in systematically different (but unobserved or unobservable) ways.

[3]This could include units expressing anger or anxiety about the researcher's intent in a study of emotions in politics, units communicating with other units (which violates SUTVA), units that receive a treatment other than the one assigned or were exposed to more than one treatment, etc.

[4]Although social scientists rarely share their protocols, online sharing of protocols in the biological sciences is common at sites such as Nature Protocols (`http://www.nature.com/protocolexchange/`) or Springer-Protocols (`http://www.springerprotocols.com/`).

[5]Given that experimental research enables meta-analytic review, thinking as a meta-analyst can also help researchers determine what protocol information to report for a given study (see Lipsey 1994; Stock 1994).

that are only clear if they are considered, recorded, and explained by the experimenter in the protocol document.

Seemingly similar experiments that yield different effects estimates or substantive conclusions – think cold fusion! – may be due to dissimilarities in protocol and implementation that fail to be noted in lab books or reported in published results. Experimental researchers need to commit to incorporating deviations from protocol into their final datasets, to sharing of those data, and to publication of full protocols for the purposes of replication, meta-analysis, and scholarly critique. Reproductions of stimuli, exact questionnaires (or other outcome measurement tools), raw datasets, statistical syntax files, and detailed protocols are all necessary addenda to any published experiment.

Conclusion

The broad agenda of any experimental science is replication and complication – the progressive addition of moderating factors, search for mechanisms, and exploration of alternative causes and outcomes. Writing and reporting protocol is a critical best practice to establish strict experimental control, to minimize noncompliance and uneven implementation, to produce clear and accurate presentation of results, and to properly interpret and generalize experimental findings. Given the importance of the (often unwritten) information contained within (often unpublished) protocol to progressive scientific inquiry (including replication and critique), we should not be afraid of producing and sharing detailed records of our scientific endeavors. But, more importantly, we should not forget the importance of producing and enforcing protocol for preserving the elegance of experimental inference.

References

Druckman, James N. 2010. "Experimental Myths." *The Experimental Political Scientist* 1(1): 9-11.

Gerber, Alan S., David Doherty, and Conor Dowling. 2009. "Developing a Checklist for Reporting the Design and Results of Social Science Experiments." Presented at Experiments in Governance and Politics Network meeting, Yale University, April 24-25, 2009.

Lipsey, Mark W. 1994. "Identifying Potentially Interesting Variables and Analysis Opportunities." In Harris Cooper and Larry V. Hedges, eds., *The Handbook of Research Synthesis.* New York: Russell Sage Foundation.

Lupia, Arthur. 2010. "Procedural Transparency, Experiments and the Credibility of Political Science." *The Experimental Political Scientist* 1(1): 5-9.

Rosenbaum, Paul R. 2010. *Design of Observational Studies.* New York: Springer.

Stock, William A. 1994. "Systematic Coding for Research Synthesis." In Harris Cooper and Larry V. Hedges, eds., *The Handbook of Research Synthesis.* New York: Russell Sage Foundation.