Standard Operating Procedures:
A Safety Net for Pre-Analysis Plans

Winston Lin[*] and Donald P. Green[†]

August 13, 2015

Abstract: Across the social sciences, growing concerns about research transparency have led to calls for pre-analysis plans, documents that lay out in advance how researchers intend to analyze the data they are about to gather. Such plans help readers to distinguish between exploratory and confirmatory analysis, thereby improving the credibility of the reported results. Pre-analysis plans, however, impose costs on researchers. They are time-consuming to write, especially if researchers attempt to describe in detail how they would handle the many contingencies that may arise in the course of data collection. In this article, we make the case for "standard operating procedures," default practices that researchers can fall back on in the event that their pre-analysis plan fails to address these contingencies. We offer an example of a documented set of standard operating procedures that may be adapted by other researchers seeking to place a safety net beneath their pre-analysis plans.

Concerns about data fishing and publication bias have sparked a growing movement to promote transparency in social science research (Miguel et al. 2014; Nosek et al. 2015). One recent innovation is the public archiving of pre-analysis plans (PAPs) that specify details of the analysis (e.g., statistical methods, sample exclusions, outcome measures, covariates, and subgroup definitions) before the researchers see unblinded outcome data.[1] Deviations from the plans are not prohibited, but "when such deviations arise they [should] be highlighted and the effects on results reported" (Humphreys, Sanchez de la Sierra, and van der Windt 2013, 13).

In principle, PAPs have three main advantages. First, pre-specification limits the extent to which researchers can make decisions that consciously or unconsciously tilt a study toward a desired result (Rubin 2007; Tukey 1993). Second, the validity of frequentist statistical inference (standard errors, confidence intervals, p-values, and significance tests) hinges on the assumption that the analysis follows a pre-specified strategy (Simmons, Nelson, and Simonsohn 2011; Tukey 1993). Third, publicly archived PAPs enable readers to see which analyses were pre-specified and to take that into account when assessing the credibility of results (Casey, Glennerster, and Miguel 2012; Chan et al. 2013; Freedman 2008, 2010; Humphreys, Sanchez de la Sierra, and van der Windt 2013; Monogan 2013, 2015; Tukey 1993).

In practice, researchers have found that PAPs have important benefits but can be challenging to write. On the one hand, writing a PAP can help researchers clarify their own thinking about research design and data collection before it is too late, and getting sponsors or partners on board can protect against pressures for ex post changes in the analysis (Casey, Glennerster, and Miguel 2012; Humphreys, Sanchez de la Sierra, and van der Windt 2013; McKenzie 2012; Olken 2015). On the other hand, detailed PAPs are time-consuming to compose, and PAPs can easily fail to cover strokes of good or bad fortune, such as

[*] Postdoctoral Research Scholar, Dept. of Political Science, Columbia University (winston.lin@columbia.edu).
[†] Professor of Political Science, Columbia University (dpg2110@columbia.edu).

new data sources becoming available (Humphreys, Sanchez de la Sierra, and van der Windt 2013) or a school being hit by lightning (Olken 2015). As Humphreys, Sanchez de la Sierra, and van der Windt (2013, p. 11) write:

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

Keeping in mind Voltaire's aphorism that the best is the enemy of the good (O'Donoghue and Rabin 2001), our proposal in this article is to build off-the-shelf sets of *good* practices for *some* issues. We and a colleague have prepared such a document (Lin, Green, and Coppock 2015) for our research group, focusing on issues we have encountered in analyzing data from randomized experiments. The idea is to have a series of defaults, which we call "standard operating procedures" (SOP), to guide decisions that have not been made explicit in a PAP. The SOP document can support and flesh out PAPs, making them easier to write. The SOP does not replace PAPs, nor does it override the explicit decisions in PAPs. Rather, an SOP lightens the burden of preparing a PAP, especially when experimental opportunities arise suddenly and require the researcher to make plans under a tight deadline. This article briefly summarizes some of the potential benefits of SOPs and offers an example from our lab that others are welcome to adapt to suit their own research needs.

Benefits of SOPs

Here are some example scenarios where an SOP document can provide guidance in the event that the PAP has not explicitly addressed the issue. Each scenario raises a question that a PAP could easily fail to anticipate.

- A project sponsor reveals to you that if a particular unit had not been assigned to treatment, the sponsor would have canceled the experiment. Thus, although treatment assignment was randomized, not every randomization would have yielded a reportable study. Should you still report the results, and if so, how should you analyze the data?

- After treatment has begun, you learn that some subjects were randomly assigned more than once. (For example, when applicants for a social program are randomly assigned as their applications are processed, randomization may go on for months or years, and in unusual cases, a persistent applicant who was originally assigned to the control group may later succeed in getting assigned to treatment.) How should their data be analyzed?

- You are conducting a randomized experiment to study the persuasive effects of a telephone canvassing effort, and have specified in the PAP that you will use an instrumental variables method (Angrist, Imbens, and Rubin 1996) to estimate the average effect of contact on those who were contacted. In the following situations, should the subject be coded as "contacted"?

    - The subject hung up right after the canvasser's initial greeting.
    - The canvasser never spoke to the subject but left a message with a housemate.
    - No one answered the phone, but the canvasser called from a number with a recognizable caller ID that identified the campaign.

SOPs can codify a research group's standing decisions on such issues, as well as others that are more routinely encountered, such as whether to report a one-tailed or a two-tailed test and how to handle missing covariate values. By specifying these decisions in the SOP, researchers eliminate the need to state them again and again when writing PAPs. Just as important, the SOP protects the researcher who might otherwise neglect to specify the procedure in a PAP. The SOP makes recurrent practices explicit and documents them *ex ante* so that researchers do not have to contend *ex post* that they were implicit.

Developing and updating SOPs

Developing an SOP takes some up-front work, but we think the investment can be more helpful than onerous. To save time, one research group can borrow another group's existing SOP and modify it to fit their own needs and preferences. Different groups can collaborate on SOPs and learn from each other.

SOPs can be amended to reflect methodological innovations and lessons from experience. However, readers need some assurance that changes to SOPs are not just another form of data fishing. We suggest that each PAP either include the SOP as an appendix, or reference a specific SOP document that is archived and time-stamped in the same registry as the PAP. If an analysis follows the pre-registered PAP and SOP, it is clearly pre-specified. If it is guided by later amendments to the SOP, it falls into what Humphreys et al. (2013, p. 18) call "a gray zone in which analysis may stay true to the intent of the registered design but the defense of the details of implementation must be provided *ex post* rather than *ex ante*." Pre-specified, gray-zone, and exploratory analyses can all be valuable, but readers need to know which is which.

Of course, any SOP document will have gaps. When situations arise that are covered neither by the PAP nor by the SOP, we would still like to avoid letting our decisions be influenced by their likely effects on results. One possible strategy is to consult a "jury" of colleagues who cannot see the unblinded outcome data or any information that might suggest whether a particular decision would make the estimated effects bigger or smaller. To make efficient use of jurors' time and expertise, such a jury might be invited to make binding decisions on a series of questions that were not anticipated by the PAP or SOP. The reasoning behind these decisions should be documented, and, if appropriate, the SOP should be amended to cover similar situations in the future. In time, as experience and "common law" decisions accumulate, SOPs and the decision rules they embody will gradually become more comprehensive in scope.

Overview of our current SOP

Our SOP can viewed on GitHub, a Web-based repository platform that allows us to publicly archive previous versions with tracked changes and allows users to post requests for additional issues to be addressed. The document can be downloaded at https://github.com/acoppock/Green-Lab-SOP/raw/master/Green_Lab_SOP.pdf without a GitHub account.

The principal motivation for the SOP is to support PAPs in pre-specifying analyses and credibly protecting against data fishing. Thus, the SOP focuses on data analysis, not experimental design, and it specifies our fallback plans for various analytic issues in case these were not addressed in the PAP. It is a document of default practices, not a manual of recommended practices. Our PAPs may deviate from the SOP when we believe a different approach is more appropriate for a particular study. Each PAP will include a statement that for any decisions not explicitly specified in the PAP, we plan to follow the SOP.

The SOP is a living document and will be expanded over time. Currently, it addresses several general topics (such as attrition, noncompliance, and use of covariates), as well as some nonstandard situations we have encountered (such as learning that some subjects were randomly assigned more than once) and some issues specific to voter turnout experiments (such as the coding of contact in canvassing experiments). It does not attempt to cover all issues that may be important in analyzing experimental data.

For example, so far it does not address the multiple comparisons problem—not because we think this issue is unimportant, but because we do not have an off-the-shelf recommendation for handling it. The issue becomes more important as the number of outcome measures, treatments, or subgroups analyzed grows. Other researchers may find it useful to codify their multiple-comparisons practices in SOPs, especially if they typically analyze many outcome measures in a single study. We look forward to learning from their approaches and may address the issue in PAPs for specific projects and, if appropriate, a future version of our SOP.

We do not regard all of the defaults in our SOP as clearly superior to the alternatives. For example, in the section on covariate adjustment, we recommend that covariates be pre-specified "on the basis of their expected ability to help predict outcomes," give rules of thumb for the maximum number of covariates, and suggest how a jury can be used in exceptional cases (e.g., when a new source of baseline data becomes available after random assignment). We considered the alternative of adopting automated model selection methods, but would like to see more evidence that (1) valid confidence intervals can be constructed when such methods are used and (2) the benefits of such methods (possible improvements in precision) outweigh the costs (increased computing time, possible loss of transparency to non-expert readers). This is just one example of a topic where, as the literature evolves and experience accumulates, our SOP may evolve as well.

Our SOP intentionally uses some arbitrary thresholds. For example, in several places in the sections on noncompliance and attrition, we specify statistical tests to compare baseline characteristics across treatment arms and write that "p-values below 0.05" will be considered evidence of noncomparability or asymmetric attrition, triggering changes in the analysis strategy. It may be wiser to pre-specify a rule based on substantive rather than statistical significance, and we may do so in PAPs, making use of subject-matter knowledge or simulations based on relevant data. However, the purpose of the SOP is to provide a fallback that constrains the analyst's discretion if the PAP does not address the issue, and a specific but arbitrary threshold serves this purpose more effectively than vague but judicious guidance.

In sharing our SOP, we are not seeking to persuade other research groups to adopt the same default practices we have chosen. In fact, we welcome debate and discussion about these practices and more opportunities to learn from other researchers' choices. We believe that by building, borrowing, and discussing SOPs, researchers can share useful ideas about methodological issues and bolster the contributions of PAPs toward improving transparency.

Acknowledgments

We thank Susanne Baltes, Jake Bowers, Alex Coppock, Al Fang, Macartan Humphreys, Hal Lin, Anselm Rink, Michael Schwam-Baird, Uri Simonsohn, Amber Spry, Dane Thorley, Anna Wilke, and Jose Zubizarreta for helpful comments and discussions.

[1] One of us (Lin) worked at program evaluation firms that were already pre-specifying analyses of social experiments in the early 1990s. However, the public registration and archiving of time-stamped PAPs in the social sciences is a recent development. For valuable discussions, see Casey, Glennerster, and Miguel (2012), McKenzie (2012), Monogan (2015), Nyhan (2015), and the symposia in *Political Analysis* (Winter 2013) and *Journal of Economic Perspectives* (Summer 2015).

References

Angrist, Joshua D., Guido W. Imbens, and Donald B. Rubin (1996). Identification of causal effects using instrumental variables. *Journal of the American Statistical Association* 91: 444–455.

Casey, Katherine, Rachel Glennerster, and Edward Miguel (2012). Reshaping institutions: Evidence on aid impacts using a preanalysis plan. *Quarterly Journal of Economics* 127: 1755–1812.

Chan, An-Wen, Jennifer M. Tetzlaff, Peter C. Gøtzsche, Douglas G. Altman, Howard Mann, Jesse A. Berlin, Kay Dickersin, Asbjørn Hróbjartsson, Kenneth F. Schulz, Wendy R. Parulekar, Karmela Krleža-Jeric, Andreas Laupacis, and David Moher (2013). SPIRIT 2013 explanation and elaboration: Guidance for protocols of clinical trials. *BMJ* 346: e7586.

Freedman, David A. (2008). Editorial: Oasis or mirage? *Chance* 21(1): 59–61.

Freedman, David A. (2010). Survival analysis: An epidemiological hazard? In *Statistical Models and Causal Inference: A Dialogue with the Social Sciences,* eds. David Collier, Jasjeet S. Sekhon, and Philip B. Stark. New York: Cambridge University Press, 169–192.

Humphreys, Macartan, Raul Sanchez de la Sierra, and Peter van der Windt (2013). Fishing, commitment, and communication: A proposal for comprehensive nonbinding research registration. *Political Analysis* 21: 1–20.

Lin, Winston, Donald P. Green, and Alexander Coppock (2015). Standard operating procedures for Don Green's lab at Columbia. Available at https://github.com/acoppock/Green-Lab-SOP.

McKenzie, David (2012). A pre-analysis plan checklist. Blog post, available at http://blogs.worldbank.org/impactevaluations/a-pre-analysis-plan-checklist.

Miguel, E., C. Camerer, K. Casey, J. Cohen, K. M. Esterling, A. Gerber, R. Glennerster, D. P. Green, M. Humphreys, G. Imbens, D. Laitin, T. Madon, L. Nelson, B. A. Nosek, M. Petersen, R. Sedlmayr, J. P. Simmons, U. Simonsohn, and M. van der Laan (2014). Promoting transparency in social science research. *Science* 343: 30–31.

Monogan, James E., III (2013). A case for registering studies of political outcomes: An application in the 2010 House elections. *Political Analysis* 21: 21–37.

Monogan, James E., III (2015). Research preregistration in political science: The case, counterarguments, and a response to critiques. *PS: Political Science and Politics* 48(03): 420–424.

Nosek, B. A., G. Alter, G. C. Banks, D. Borsboom, S. D. Bowman, S. J. Breckler, S. Buck, C. D. Chambers, G. Chin, G. Christensen, M. Contestabile, A. Dafoe, E. Eich, J. Freese, R. Glennerster, D. Goroff, D. P. Green, B. Hesse, M. Humphreys, J. Ishiyama, D. Karlan, A. Kraut, A. Lupia, P. Mabry,

T. Madon, N. Malhotra, E. Mayo-Wilson, M. McNutt, E. Miguel, E. L. Paluck, U. Simonsohn, C. Soderberg, B. A. Spellman, J. Turitto, G. VandenBos, S. Vazire, E. J. Wagenmakers, R. Wilson, and T. Yarkoni (2015). Promoting an open research culture. *Science* 348: 1422–1425.

Nyhan, Brendan (2015). Increasing the credibility of political science research: A proposal for journal reforms. *PS: Political Science and Politics* 48(S1): 78–83.

O'Donoghue, Ted and Matthew Rabin (2001). Choice and procrastination. *Quarterly Journal of Economics* 116: 121–160.

Olken, Benjamin A. (2015). Promises and perils of pre-analysis plans. *Journal of Economic Perspectives* 29(3): 61–80.

Rubin, Donald B. (2007). The design *versus* the analysis of observational studies for causal effects: Parallels with the design of randomized trials. *Statistics in Medicine* 26: 20–36.

Simmons, Joseph P., Leif D. Nelson, and Uri Simonsohn (2011). False-positive psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychological Science* 22: 1359–1366.

Tukey, J. W. (1993). Tightening the clinical trial. *Controlled Clinical Trials* 14: 266–285.