Journal of Economic Perspectives—Volume 29, Number 3—Summer 2015—Pages 61–80

Promises and Perils of Pre-Analysis Plans[†]

Benjamin A. Olken

Imagine a nefarious researcher in economics who is only interested in finding a statistically significant result of an experiment. The researcher has 100 different variables he could examine, and the truth is that the experiment has no impact. By construction, the researcher should find an average of five of these variables statistically significantly different between the treatment group and the control group at the 5 percent level—after all, the exact definition of 5 percent significance implies that there will be a 5 percent false rejection rate of the null hypothesis that there is no difference between the groups. The nefarious researcher, who is interested only in showing that this experiment has an effect, chooses to report only the results on the five variables that pass the statistically significant threshold. If the researcher is interested in a particular sign of the result—that is, showing that this program "works" or "doesn't work"— on average half of these results will go in the direction the researcher wants. Thus, if a researcher can discard or not report all the variables that do not agree with his desired outcome, the researcher is virtually guaranteed a few positive and statistically significant results, even if in fact the experiment has no effect.

This is of course the well-known problem of "data-mining." If the researcher can choose which results to report, it is easy to see how results can be manipulated. Casey, Glennerster, and Miguel (2012), for example, demonstrate in a real-world economics context how researchers with opposite agendas could hypothetically string together two opposite but coherent sets of results by cherry-picking either positive or negative statistically significant results.

■ *Benjamin A. Olken is Professor of Economics, Massachusetts Institute of Technology, Cambridge, Massachusetts. His email is bolken@mit.edu.*

[†] To access the Appendix and disclosure statements, visit
http://dx.doi.org/10.1257/jep.29.3.61

doi=10.1257/jep.29.3.61

The parable of a nefarious researcher offers the most straightforward version of the data mining problem, but similar problems can arise in less-extreme forms. For example, real-world data are messy, and are often "cleaned" before analysis—for example, to remove data outliers like a person whose height is reported in the data as being 20 meters tall instead of 2.0 meters. However, in many cases the issue of whether to "clean" the data of certain observations will involve a judgment call, and the researcher will often know how including certain observations will tend to affect the final results. There are also many decisions to make about specifications: what regression form to use, what control variables to include, what transformations to make to the data, how to define variables, and so on (Leamer 1983). Even researchers who have the noblest of intentions may end up succumbing to the same sorts of biases when trying to figure out how, in the process of their analysis, to make sense of a complex set of results.

One potential solution to these issues is to pre-specify in a precise way the analysis to be run before examining the data. A researcher can specify variables, data cleaning procedures, regression specifications, and so on. If the regressions are pre-specified in advance and researchers are required to report all the results they pre-specify, data-mining becomes much less of a problem. In the "confirmatory" trials used for approval of pharmaceuticals by the Food and Drug Administration, pre-specified statistical analysis plans are required that explicitly spell out how data will be handled—and these analysis plans must be finalized and archived before researchers actually run regressions on the unblinded data (Food and Drug Administration 1998).

But pre-specifying analysis plans comes at a cost. A pre-analysis plan is relatively straightforward to write if there is a single, simple hypothesis, with a single, obvious outcome variable of interest. But in practice, most research is much more complicated than this simple ideal. In economics, the typical research paper is trying to elucidate or test various predictions from economic theory, rather than estimate a single parameter with a single hypothesis test. Most research papers test a large number of hypotheses. Hypotheses are often themselves conditional on the realizations of other, previous hypothesis tests: the precise statistical question a paper might tackle in Table 4 depends on the answer that was found in Table 3; the question posed in Table 5 depends on the answer in Table 4, and so on.

Pre-specifying the entire chain of logic for every possible realization of the data can quickly become an overwhelming task for even the most committed pre-specifier. And in practice, researchers often get ideas for new hypotheses from seeing realizations of the data that they did not expect to see. The most rigid adherents to pre-specification would discount any such results that were not rigorously specified in advance of the data. Usually, though, these later additions to the analysis would be allowed, but would be considered "exploratory"—that is, not of the same rigorous statistical standards as confirmatory trials.

In a world with unlimited resources and unlimited time to make decisions, one could imagine a sequence of studies on any single topic. Exploratory analysis would be used to generate hypotheses, and then in turn subsequent, separate pre-specified

confirmatory trials would be run to test those hypotheses more rigorously. Exploratory analysis from those trials could form the basis of future trials, and so on. In practice, though, there are time and particularly budgetary constraints—true everywhere, but particularly so in economics where the entire budget of the National Science Foundation for social and economic sciences—about $100 million in 2013 (National Science Foundation 2013)—pales in comparison with the billions spent annually on drug trials, where pre-specification is most rigorous. Such constraints mean that most of these follow-up confirmatory trials will never be done, and the "exploratory" analysis is all the community will have to go on. Thus, the question of how much to discount such exploratory analysis in assessing the results of studies—either for journal publications or as the basis of policy—is a substantive question of serious importance.

The purpose of this paper is to help think through the advantages and costs of rigorous pre-specification of statistical analysis plans in economics. I begin by laying out the basics of what a statistical analysis plan actually contains, so that those researchers unfamiliar with the issue can better understand how it is done. In so doing, I have drawn both on standards used in clinical trials, which are clearly specified by the Food and Drug Administration, as well as my own practical experience from writing these plans in economics contexts.

I then lay out some of the advantages of pre-specified analysis plans, both for the scientific community as a whole and also for the researcher. Even researchers with the noblest of intensions may end up succumbing to their biases when trying to figure out how to make sense of a complex set of results, and pre-analysis plans can also be a useful tool when research partners have strong vested interests. I also explore some of the limitations and costs of such plans. I then review a few pieces of evidence that suggest that, in many contexts, the benefits of using pre-specified analysis plans may not be as high as one might have expected initially. I suspect the possible explanations include that most researchers are not nefarious and that existing safeguards place limits on the ability of researchers to data mine. Such safeguards may include referee and editor preferences for robustness checks, open availability of data, the possibility of replication, and the structure imposed by economic theory.

Most of my examples will focus on the relatively narrow issue of pre-analysis for randomized controlled trials.[1] Such studies fit the idea of a pre-analysis plan well, because they are designed and in place for a time before the data become available. However, the issues and tradeoffs I will discuss potentially apply to other empirical research in economics, too. In principle, there is no reason, for example, that a researcher could not completely pre-specify a statistical analysis plan before

[1] The registration of trials is a separate, though related, issue. If researchers only report those trials that happen to have a particular result, then the sample of trials that readers see will be biased. One solution to this issue is to register trials before the results are known. The American Economic Association sponsors a registry for this purpose for social science trials (http://www.socialscienceregistry.org/). For clinical trials in medicine, the US National Institute of Health sponsors a similar registry (https://clinicaltrials.gov/), which to date includes over 170,000 studies.

downloading the US Current Population Survey, or any other pre-existing datasets. While such an approach would be possible with existing datasets (for example, Neumark 2001), there is no obvious before-and-after date when a pre-analysis plan would be formulated and then later carried out, so doing so becomes more complicated. While perhaps such an approach could be useful, as some have advocated (for instance, Miguel et al. 2014), it is not something I explicitly consider here.

What Is a Pre-Analysis Plan?

The Basics: What Features Should Statistical Pre-Analysis Plans Include?

Virtually all pre-analysis plans typically share a few common features, summarized in Table 1. In describing these features, I draw heavily on accepted practice in perhaps the most rigorous and heavily regulated setting where they are used: the "Statistical Principles for Clinical Trials" specified for full-scale confirmatory trials used by the US Food and Drug Administration (1998) to approve drugs and other medical products. I will also discuss how these approaches may need to be adapted to a social science context. The interested reader may also wish to consult Casey, Glennerster, and Miguel (2012), which also discuss related issues in framing pre-analysis plans in economics.

A primary outcome variable. Given that one of the key motivations of pre-specifying an analysis plan is to avoid temptations for data mining, a key decision that needs to be made is the primary outcome variable one plans to examine to judge the outcome of a project. The idea is to solve the multiple inference problem by designating in advance a single outcome metric to evaluate the study. In designating the primary outcome variable, one should be as precise as possible: not just the general topic one intends to study, but the precise variable definition one intends to use.

Designating a single primary outcome variable can turn out to be surprisingly hard. In medical clinical trials, conventions have evolved concerning how to evaluate many topics, thus allowing comparability across studies, but in social sciences, more choices are available to the researcher. For example, suppose you are designing a study to evaluate an after-school tutoring program for disadvantaged youth. Possible outcomes could include school dropout rates, attendance rates, test scores, juvenile delinquency, teen pregnancy, and others. The researcher must make a choice of which outcome to focus on. If the researcher does this right, he or she will have substantially increased the believability of the research. Of course, one must choose carefully: if the study designated test scores as its primary outcome variable, and found no impact on test scores, but instead found that the program improved school dropout rates by an economically meaningful and statistically significant amount, the logic of pre-analysis plans suggests that policymakers should be much less likely to rely on those results than if the researcher had designated dropout rates beforehand as the primary outcome variable.

If a researcher wants to designate more than one outcome variable as primary, there are two options. First, one can designate multiple co-primary outcome variables,

Table 1
Pre-Analysis Plan Checklist

Item	*Brief description*
Primary outcome variable	The key variable of interest for the study. If multiple variables are to be examined, one should know how the multiple hypothesis testing will be done.
Secondary outcome variable(s)	Additional variables of interest to be examined.
Variable definitions	Precise variable definitions that specify how the raw data will be transformed into the actual variables to be used for analysis.
Inclusion/Exclusion rules	Rules for including or excluding observations, and procedures for dealing with missing data.
Statistical model specification	Specification of the precise statistical model to be used, hypothesis tests to be run.
Covariates	List of any covariates to be included in analysis.
Subgroup analysis	Description of any heterogeneity analysis to be performed on the data.
Other issues	Other issues include data monitoring plans, stopping rules, and interim looks at the data.

but a researcher who chooses multiple hypotheses needs to adjust the statistical tests for each hypothesis to account for the multiple inference hypotheses. The simplest way to do this, known as a Bonferroni adjustment (Dunn 1961), simply divides the required p-value by the number of tests conducted: thus, if a study chooses three outcome variables and tests at the 5 percent significance level, one would require each test to have significance $0.05/3 = 0.0166$ before it would be viewed as statistically significant. There are other, more sophisticated ways to multiple-inference adjust that have less of an impact on statistical power, like the step-down approach (for example, Westfall and Young 1993). But the general principle is that each additional co-primary outcome comes at a meaningful cost in terms of statistical power.

Second, a researcher can aggregate the primary outcome variable into an index or composite variable. If variables have comparable scales, one can take a simple average. Otherwise, the most common approach in economics is to compute "average standardized effects," where one divides each variable by its standard deviation and then takes the average of these normalized variables (Kling, Liebman, and Katz 2007). The index approach can be more powerful than a joint hypothesis test because an index lines up the variables so that "better" results tend to be averaged together in the same direction, whereas a joint test is agnostic about the sign of different results. Alternatively, one can use principal components analysis, which looks at the covariance between the individual variables and weights them accordingly. These various techniques create a single hypothesis test, rather than multiple hypothesis tests, which improves power. The potential downside of an index approach is that, if one finds results, it is hard to know statistically precisely what is driving the results. Policymakers may find it difficult to act based on a change in an index number.

Secondary outcome variables. Many pre-analysis plans also specify secondary outcome variables, which are outcomes that may help shed light on the findings but would not themselves be "confirmatory." For example, if the Food and Drug Administration were considering whether to approve a drug, and a trial found meaningful results on a secondary outcome but not on a primary outcome, the drug would generally not be approved. In social science papers, secondary outcome variables often play a crucial role, because they illuminate the "mechanisms" or pathways that lie behind the results, which in turn helps guide the researcher in both enhancing understanding of the problem and in being able to say something sensible about external validity. Outside of regulatory contexts where secondary outcomes have a precise meaning (in particular, drug makers can be allowed to market a drug based on a proven secondary outcome if it is listed in advance and if they also found results on the primary outcome), researchers are in practice often somewhat laxer about multiple inference testing with secondary outcome variables. As I will discuss in more detail below, the pre-specification of secondary outcomes can become quite challenging in social science papers, because the set of secondary outcome variables to be examined depends on the results from primary outcome variables.

Variable definitions. A pre-analysis plan requires a *precise* variable definition. Continuing the earlier example, suppose that test scores are the primary outcome of interest. What test and test subjects are included? Will the outcome variable be the test score in levels or logs? Will it be in standard deviations, the percentile of the test score, a binary variable for passing the test, a binary variable for being above the 25th percentile, or the 50th percentile, and so on? Will the score be in levels or an improvement from a baseline? If there are multiple subjects, like math and reading, how will the scores be aggregated into a single outcome variable? Are there any rules for trimming or excluding outliers? A good rule of thumb is that if you gave the pre-analysis plan to two different programmers, and asked them each to prepare the data for the outcome variable, they should be both able to do so without asking any questions, and they should both be able to get the same answer.

Inclusion or exclusion rules. A precise set of rules lead to the "analysis set"—that is, the final set of data to be analyzed. As a general principle, of course, the analysis set should be as close as possible to the actual observations. However, if there are legitimate reasons to drop observations, they should be specified in advance in the analysis plan. Relatedly, one should discuss the plans for handling missing values and attrition, although a challenge is that one cannot always foresee the reasons one might want to exclude certain observations.[2]

[2] For example, in Kremer, Miguel, and Thornton's (2009) study of scholarships in Kenya, several schools withdrew from the study after a Teso-ethnicity school was hit by lightning and some Teso-ethnicity community members associated the lightning strike with the nongovernmental organization running the scholarship program. In some specifications, the authors restrict analysis to schools that did not withdraw due to this concern. A lightning strike seems like exactly the sort of legitimate reason one might want to exclude observations, but the possibility of lightning strikes and superstitious villagers would have been very hard to foresee in a pre-analysis plan.

Statistical model and covariates. An analysis plan should spell out the precise estimating equation and statistical model, including functional form and estimator (ordinary least squares, probit, logit, Poisson, instrumental variables, and so on). If fixed effects are going to be used, or comparisons to baseline values, or first differences of data, all this should be spelled out. The pre-analysis plan also states how standard errors will be treated (including any clustering, bootstrapping, or other techniques). If one is using nonstandard hypothesis tests, and in particular one-sided tests, it should be spelled out in advance.

Specification of the model should also be clear about which covariates should be included in regressions, because a typical study might collect tens or even hundreds of variables that could, potentially, be included as covariates. After all, researchers could potentially cherry-pick control variables to maximize statistical significance. Relatedly, it has become standard practice in most randomized controlled trials in economics to present a table showing that baseline covariates appear balanced across treatment and control groups. If the authors intend to present a balance test, it is also common sense to pre-specify in the analysis plan the variables that will be used to check covariate balance.

Subgroup analysis. Pre-specification of subgroup analysis matters because there are many possible ways of cutting the data into various subgroups—men versus women, old versus young, rural versus urban, and so on. Again, researchers could first do the analysis and then pick a subgroup with a statistically significant result, which is a frequent critique of some randomized trials in development economics (Deaton 2010). If heterogeneity analysis is likely to be important, pre-specification can be quite helpful to increase confidence in the results.

Other aspects. Other issues that are often considered in pre-analysis plans in the medical world include data monitoring plans, safety checks, stopping rules, and interim looks at the data. In particular, in medical trials one often checks the data in the middle of the trial to ensure that the outcome is not causing unexpected harm (in which case the trial might be stopped) and to learn whether results are so good that the trial can be declared a success early. A recent area of research has been to allow for adaptive trials, which are trials whose design evolves over time based on the data but according to pre-specified rules (Food and Drug Administration 2010). These issues can all be pre-specified in the analysis plan.

When Should You Write a Statistical Analysis Plan?

In the classic Food and Drug Administration model, the primary outcome and usually secondary outcomes would be specified in the formal trial protocol before the trial begins. However, the statistical portion of the pre-analysis plan can be finalized later—including issues such as covariates, regression specification, and handling of missing data—as long as it is written without ever unblinding the data, that is, without ever separating the data by treatment groups.[3]

[3] In fact, it is possible in rare circumstances to change the primary outcome variable of a trial once the trial has begun, if one can demonstrate that the original primary outcome variable no longer makes

Allowing researchers to design the statistical portion of their pre-analysis plans based on the blinded outcome data can be quite useful. In many cases in social science, the outcome variables that people study are sufficiently novel, and the data on relevant populations is sufficiently limited, that researchers have only limited information about the distribution of the variables when designing studies. For example, imagine that one of the variables in the study is the level of juvenile delinquency. Presumably, the researcher has some informed guess about the expected mean and standard deviation for this variable. But perhaps in this particular dataset, the standard deviation is much larger than usual. (Perhaps there was an unusual crime wave during the period of the study, or for some reason the study sample differs from the population.) Looking at the blinded data helps the researcher to discover if the outcome variables behave sensibly—that is, if they have reasonable relationships with each other and with the covariates—which helps to assure that the variables were measured well. If not, they can be excluded. Another use is to examine the blinded data to determine which covariates best predict the outcome variable, reducing standard errors by reducing the variance of the residuals. Especially when the outcome data or covariates are novel variables, it can be useful to examine the actual blinded data for this purpose.[4]

In my experience, it can be quite useful to write statistical programs, run them on the blinded data, and use the results to update the statistical analysis plan in the process. Indeed, one can generate a "fake" randomization—that is, one can run and rerun a randomization program with different starting seed values to generate the actual standard errors one would expect when running regressions. The exercise of writing the computer code and looking carefully at the data also forces the authors to make detailed choices about variable definitions and coding: for example, a researcher can make decisions about how to exclude outliers before knowing whether they are in treatment or control groups and how their exclusion will affect the results.

A trickier issue is the use of qualitative data, particularly for many social science trials, which are often not blinded (that is, both those administering the trial and the subjects know who is in the control group and who is in the treatment group). In this context, even if the statistical data is blinded, one may learn something from the qualitative findings of the trial. For example, one might observe that those in the after-school support program seem to be happy, so one might think to add subjective well-being measures to an analysis protocol. Even though

sense. For example, suppose that the primary outcome variable of a study was mortality, but the blinded data revealed that the overall mortality rate was much lower than expected and the trial was underpowered. It might be possible then to amend the trial protocol to change the outcome variable to be a combination of mortality and morbidity.

[4] Note that in doing so, it is often advisable to look at the complete, blinded data rather than looking at the control group and hiding the treatment group. There are a number of reasons for this. One reason is that in practice, researchers will have often seen summary statistics for the entire data: if one has, and also sees the control group, one can subtract to obtain the treatment group estimates. It is also easier to ensure that the data are not accidentally unblinded if the treatment and control assignments are kept entirely separate from the data one is using to construct the analysis plan.

this is based on qualitative observations, not the quantitative data, it has the same effect as looking at the unblinded data. For this reason, trial purists may prefer analysis plans to be finalized before the trial begins. The degree to which this makes sense depends on weighing the benefits from making sure this type of qualitative bias does not enter, against the costs in terms of missed advantages from trying out the analysis on blinded data, as discussed above.

What Do You Do With a Statistical Analysis Plan after You Write It?

Ideally, a pre-analysis plan should be added to a public archive. As discussed above, the American Economic Association operates a trial registry, where authors can also archive a statistical analysis plan, with specific timestamps marking exactly when it was registered. The timestamps can credibly convey to the reader that it was filed before all data was collected, for example, and that it was not modified later. The registry also allows authors to register a statistical analysis plan but not make it public until a later date (or not at all). In this way, authors who are concerned about others scooping their work could obtain the credibility benefits of pre-registration—that is, they could document to editors or referees that their analysis plan was pre-registered—while avoiding publicity about their work months or years before it is complete.

Benefits of Pre-Analysis Plans in Economics

The most obvious benefit from pre-specification is that a careful pre-analysis plan can address a substantial proportion of data-mining problems. For readers, referees, editors, and policymakers, knowing that analysis was pre-specified offers reassurance that the result is not a choice among many plausible alternatives, which can increase confidence in results.

However, a pre-analysis plan also offers some other useful benefits for researchers themselves, which are perhaps less obvious and therefore worth elaborating in further detail. The exercise of creating a pre-analysis plan can be useful for researchers to make sure that they think through, and collect, the data they need for the analysis. Beyond that, the act of commitment to an analysis plan per se offers some additional advantages.

First, pre-specified analysis plans allow researchers to take advantage of all the statistical power of well-designed statistical tests and to worry less about robustness to specifications. After seeing the results, it can be challenging for even well-intentioned researchers not to choose specifications that lead to more statistical significance—well-intentioned researchers might conclude, for example, that the specification that led to the smallest standard errors was the one that best fit the data, and it is hard to prefer intentionally a specification that makes ones' results look weaker. But given this, if specifications are not pre-specified, researchers will be required by referees and editors to report robustness results to a wide range of alternative specifications and will likely judge results by the average level of statistical significance across specifications rather than use the statistical significance from the preferred

specification. A pre-specified analysis plan could help discourage readers of the article—including journal referees—from expecting an endless series of robustness checks and accepting only those results that survive all possible permutations.

A second benefit to researchers, related to the first, is that pre-commitment can also potentially allow researchers to increase statistical power by using less-conventional statistical tests, if they really believe that such tests are appropriate in a given case, because they know that pre-committing to such a test means that they cannot be justly accused of cherry-picking the test after the fact. For example, convention typically dictates two-sided hypothesis tests, so that researchers can reject the null hypothesis of no effect of a program at the 5 percent level if the estimate is in the upper or lower 2.5 percent of the distribution under the null hypothesis. In practice, however, researchers are often interested in only knowing whether a program works or not, which could lead to a one-sided hypothesis test. Such a researcher might instead use a one-sided test rather than a two-sided test. Of course, a one-sided test has trade-offs, too. By committing to a one-sided test, researchers need to be prepared that even if they receive a very, very negative outcome—for example, an outcome in the bottom 0.005 percentile—they would interpret that outcome as no different from the null rather than report a negative and statistically significant result. The prospect of such a result is uncomfortable enough to cause most researchers to prefer two-sided tests.[5] Moreover, in addition to hypothesis tests, researchers often want to report confidence intervals. In a one-sided testing framework, the confidence interval has an infinite bound on one side, which may be less useful from a decision-making perspective. Clearly, there are often reasons for the conventional choices, like two-sided hypothesis tests, and researchers should proceed cautiously before pre-committing to alternatives.

A final major benefit to researchers is that pre-specification can be useful vis-à-vis their research partners. In practice, a substantial share of large-scale randomized controlled trials and other large-scale social science research is done in collaboration with partners who have some vested interest in a program's outcomes, like the government, a nongovernment organization, or a private sector program or firm. Even if sponsors don't have explicit rights of review of articles or research findings, researchers often develop close relationships with partners over time, which can potentially lead to awkward situations when results do not come out the way partners might have hoped. By creating an opportunity for researchers and partners to agree and commit before results are observed on how the program will be evaluated, pre-specification can provide researchers with protection from pressure from their partners to slant results in a more favorable light.

[5] An interesting hybrid alternative would be to pre-specify asymmetric tests: for example, to reject the null if the result was in the bottom 1 percent of the distribution or in the top 4 percent, or the bottom 0.5 and the top 4.5 percent, and so on. These asymmetric tests would gain much of the statistical power from one-sided tests, but still be set up statistically to reject the null in the presence of very large negative results. One could apply decision theory, where one specifies losses for each type of error, to determine the appropriate asymmetric approach to use. Although I have not seen this approach taken in economics, it seems like a potentially useful approach for researchers to consider.

Costs of Pre-Analysis Plans in Economics

When laid out in this way it seems hard to be against pre-analysis plans. After all, how can one argue against the idea that one should do hypothesis testing properly to get correct *p*-values, and that the research community should protect itself against data mining? However, restricting analysis to pre-specified hypotheses has some fairly important costs, which need to be weighed against the benefits.

One important challenge is that fully specifying papers in advance is close to impossible. Economics papers typically ask not just the result of a treatment, but also try to elucidate the mechanisms that underlie the treatment, such that the results quickly become too complex to map out in advance. For example, suppose that a paper has one main table and then ten follow-up tables of results, and each table can have three possible results—"positive," "zero," and "negative." In addition, suppose the question one would want to ask in each table depends on the outcome of the previous table. Pre-specifying the entire "analysis tree" would therefore involve writing out $3^{10} = 59{,}049$ possible regressions in advance. Even for the most dedicated pre-specifier, thinking through 59,049 possible regressions in advance would clearly be too taxing. It would also be inefficient—we would prefer that researchers spend their time and intellectual energy on those parts of the tree that are actually relevant rather than working down the branches of the tree that are meaningless.

Faced with this conundrum, researchers can take several possible tacks. First, they can try to pre-specify as much as possible. Many early pre-analysis plans in economics ended up voluminous, exceeding 50 pages, trying to pre-specify not just primary outcome variables but all of the secondary tests one would want to run conditional on results from the primary outcome variable. The result can be an unwieldy paper that reports all kinds of results that are not primarily of interest to the reader since they were relevant only conditional on realizations of the primary outcome variable that did not materialize.

More important, because researchers are spreading their thinking energy over the entire space of possible regressions they might want to run, they often do not focus on aspects of the space that end up being important, and they may miss out on important hypotheses. After all, scientific breakthroughs sometimes come from unexpected surprise results.[6] Limiting oneself only to the regressions that are pre-specified, and not including or severely discounting any additional analysis of the data inspired by surprise results, seems an inefficient way to learn the most from the data at hand.

[6] Limiting oneself strictly to pre-specified analysis at some point becomes absurd. Easterly (2012), for example, imagines what would have happened if Christopher Columbus had had to pre-specify an analysis plan for his 1492 voyage to the Indies: "(2) Columbus gets funding to test Going West approach [to reach the Indies]. (3) Rigorous Evaluation design filed in advance to test effect of Going West on Reach the Indies. (4) October 12, 1492. [Columbus discovers America.] (5) Rigorous Evaluation finds no effect of Going West approach on Reach the Indies. (6) Rigorous methodology means Evaluation not permitted to include any ex post outcomes of Going West not filed in advance in the design. (7) Going West approach declared ineffective, funding ends."

An alternative, more moderate approach is to focus the pre-analysis plan on a single primary outcome (or a narrow set of such outcomes), and then leave the remainder of the paper for exploring potential mechanisms as only exploratory and not pre-specified. Of course, this strategy would be explained as such in the pre-analysis plan and the article itself. In some sense, this is how pre-specification is supposed to proceed: a given trial is designed to test a single, well-specified hypothesis, and then the data is used in a variety of exploratory ways to come up with new hypotheses that are in turn then the subject of future pre-specified trials.

The problem with narrow pre-specification and extensive exploratory analysis is that, in practice, there are not enough resources to conduct repeated streams of separate trials simply to solve the pre-specification issue. Budgets for social science research are several orders of magnitude smaller than for medical research—and even in medicine, some journals would acknowledge that for many less-common areas, the exploratory results may be the only results that the scientific community will have.[7] The difference in magnitudes here is enormous: the registry for medical trials, http://clinicaltrials.gov, currently lists over 176,000 studies registered since the site was launched; by comparison, a reasonable estimate for the number of randomized controlled field experiments conducted in social science over a similar period is on the order of 1,000.[8]

This argument does not imply that researchers running any given trial would be better off by not pre-specifying analysis for that trial. But it does suggest that if journal editors were to restrict themselves to publishing studies based on the limited, pre-specified, confirmatory parts of analysis, and relegating exploratory analysis to second-tier status, a substantial amount of knowledge would be lost. We do not have near-infinite resources to run sequences of pre-specified trials iteratively, where each set of exploratory analysis from one trial was the subject of a subsequent, pre-specified confirmatory trial, and so it seems important to continue to allow researchers to publish, and the broader community to use, important results that were not necessarily pre-specified.

A related issue is that papers following rigorous pre-specified analysis plans may miss the nuance that categorizes social science research. Pre-analysis plans work particularly well for relatively simple papers: there was a trial of some particular

[7] For example, the total annual 2014 National Science Foundation budget for Social and Economic Sciences is $102 million. By comparison, the total National Institutes of Health budget in 2014 is approximately is $30.1 billion, and this does not include the billions spent by the private sector on clinical trials for pharmaceuticals and other medical products. While both of these budgets fund many activities that are not randomized trials, the difference in scale is remarkable.

[8] We do not have a precise number of trials in social science over this period. However, members of several of the largest organizations supporting such trials in development economics, the Abdul Latif Jameel Poverty Action Lab, Center for Effective Global Action, and Innovations for Poverty Action have each completed or are in the process of running about 500 trials since their respective founding in the early 2000s; since many of these trials are counted by several of these organizations, the total is likely closer to 750 or so. We do not have a formal count of other trials in economics outside these organizations (for example, trials run by the World Bank or organizations like MDRC are not included in these totals), but it seems safe to say that the total is on the order of a few thousand at most.

intervention, there is some key outcome metric to decide if it "works" or not, and the researcher compares that outcome across treatment and control groups. This framework naturally leads one to specify a primary outcome variable (the metric of whether the program "works" or not), and so on.

However, many empirical economics papers are instead seeking to test theoretical mechanisms to see whether they are borne out in practice. In many contexts, the point of the study is not just that this particular trial had this particular effect, but rather to show the existence in practice of a theoretically posited mechanism that may be of use elsewhere. Papers thus use a constellation of tests to elucidate economic mechanisms and test theories. While it may be possible to pre-specify complex papers, as discussed above, given the exponentially increasing challenges of pre-specifying complex analysis trees, pre-specification of analysis works best for simple setups, when there is a clear "primary outcome" or set of primary outcomes. One would not want the quest for pre-specification to come at the cost of writing only simple papers and losing the nuance that characterizes some of the best social science work.

A more prosaic but still important concern with requiring pre-analysis plans involves the intricacies of needing to monitor program implementation using unblinded data while at the same time finalizing the analysis plan based on blinded data—all with a limited staff. In principle, there are two distinct things one would like to do with the data while the trial is ongoing. First, as discussed above, one would like to look at the *blinded* data before finalizing the pre-analysis plan to improve the plan: for example, by checking means and standard deviations, doing data cleaning on the blinded data, or even just having more time to reflect on how to analyze the trial after the effort of launching the fieldwork has been completed. Second, one would also like to look at the *unblinded* data while the trial is ongoing to provide real-time feedback to implementing partners, ensure that implementation is going on correctly, and so on. For example, interim looks at the data can be used in a medical trial to see if a drug is causing an adverse reaction, to know if the trial should be stopped. One can similarly imagine that a business or government that is partnering with a social science researcher in a trial may require ongoing analysis of the trial to ensure that the experiment is not actively harming their business or program. Often follow-up trials need to be planned before a trial is complete, so interim looks at the data can be useful for that purpose as well.

In principle, in a really large research team, one could have two different sets of sub-teams, one looking at the blinded data during the trial and refining the analysis plan, and one looking at the unblinded data during the trial for management and safety purposes. In medicine, budgets are large enough that one can really have two completely different teams of people doing these tasks, with a firewall between them. For example, many medical trials have separate Data Monitoring Committees (DMCs) regularly inspecting unblinded data to verify safety and implementation—while keeping the unblinded data securely away from principal investigator's eyes. But many social science trials are on sufficiently tight budgets that having two separate teams of people for these two tasks isn't feasible. In that sense, requiring a

pre-analysis plan comes at a cost, since the researcher must forego one or the other of these during-trial activities.

A final, if perhaps less-persuasive, cost of pre-specification is that it prevents you from learning about your data as you analyze it. As all researchers who have worked with empirical data realize, a myriad of real-world issues arise: how should variables be defined, how to deal with outliers, and so on. In principle, perhaps, there is no reason that these issues cannot be sorted out on blinded data, and programs written in advance. In practice—much for the same reason that it is hard to think through every possible regression in advance—researchers frequently realize features of their data only during the process of analysis. For example, seeing surprisingly large standard errors on a regression may make authors realize that the distribution of a variable was more skewed or plagued with outliers than they had initially appreciated. Addressing these problems iteratively as they come up raises the possibility of data mining, but preventing researchers from dealing with these issues if they come up also may limit the amount we can learn from a given study.

These costs should not be necessarily viewed as dispositive, or arguing against pre-analysis plans in all cases. However, they do suggest that the degree to which requiring pre-analysis plans makes sense for the discipline depends on the extent to which the key problem—data mining—is actually a problem, an issue I explore in the next section.

Is There Much Need for Pre-analysis Plans in Practice?

How Bad Is the Problem?

The arguments in the previous section suggest that pre-specification of analysis has important benefits—preventing data mining and specification searching, limiting influence of partners, and so on. But imposing standards such that the only analysis that the scientific community trusts or that journal editors are willing to consider for publication is pre-specified also has costs. Authors may be limited in their ability to learn during the process of analysis, and as such will likely write papers of less-general interest focused only on those hypotheses that were pre-specified rather than on more potentially interesting findings discovered later.

The extent to which the community should reward pre-specification therefore depends, in practice, on how substantial the data mining concerns are. That is, many of the arguments in favor of pre-specification assume the worst about researchers: they are inherently biased and data mine as much as possible until they find results. But how common is the nefarious researcher in practice?

Several recent studies in social science suggest the problem is not as bad as the pessimists might believe. One recent study by Brodeur, Lé, Sangnier, and Zylberberg (forthcoming) tried to quantify the extent to which there is inflation of *p*-values through specification mining. The strategy was to look at the distribution of *p*-values in a wide range of studies, and to find out whether the *p*-values are bunched just below critical statistical significance values of 0.10 or 0.05. Brodeur

et al. examined all empirical regressions from the *American Economic Review*, *Journal of Political Economy*, and *Quarterly Journal of Economics*, between 2005 and 2011, examining over 50,000 regression results from 3,389 tables in 641 articles. They do find bunching of *p*-values, in a way that suggests that between 10 and 20 percent of all tests that show *p*-values in the conventional range of statistical significance between 0.05 to 0.0001 are in fact misallocated and instead should be in the range, not thought of as statistically significant, between 0.10 to 0.25. However, Brodeur et al. find no evidence of this problem arising in randomized trials, which suggests that at least as detected by their methodology, there is little bias problem in the context for which pre-analysis plans are most applicable.

Even to the extent Brodeur et al. (2013) do find excess bunching, their results imply that it may not be quantitatively as severe as one might have thought. Their results imply that out of 100 studies, instead of obtaining a nonsignificant result in 95 percent of studies where the null is in fact correct—as one would expect with a *p*-value of 0.05—we are in fact doing so for 92.75 percent of such studies.[9] Addressing this problem would be beneficial, but if it came at the cost of substantially excluding a variety of important and interesting findings that were discovered in after-the-fact analysis, it might not be worth the cost.

An alternative approach to searching for publication bias is to carry out the study again, in as similar a way as possible. Replication of large-scale field studies in economics is rare; in fact, given the costs of these studies and limited budgets, it probably makes sense in most cases to prioritize new experiments rather than funding replications of existing experiments. However, in social psychology where experiments can be conducted in the lab there have been some attempts to replicate main findings. A recent paper by Klein et al. (2014) reports an enormous effort (with more than 50 coauthors) to replicate 13 well-known psychology findings using labs around the world. Roughly speaking, they found that (depending on the standard applied) 10 or 11 of the 13 studies replicated well. Recall that even with correct 0.05 *p*-values, we would not at all be surprised if 1 out of the 13 (7.6 percent) failed to replicate. We would also not be surprised if some studies did not replicate given changes in subject pools, changes in experimenters, and so on. So on balance, while there appears to be evidence of a slight inflation of statistical significance, this replication-based approach suggests that in this context, major findings are holding up reasonably well.

Why Isn't the Problem Worse?

Economists and other social scientists may be closer to the world of correct *p*-values than to the world of the nefarious researcher who is cherry-picking results left and right. I suspect there are several reasons for this.

[9] To be concrete, let us suppose that 15 percent of tests should have *p*-values of 0.20 instead of 0.05. What would this mean for inference? It implies that the "correct" *p*-value, conditional on seeing a *p*-value of 0.05 and knowing that 15 percent of them should have *p*-values of 0.20, is $(0.85 \times 0.05) + (0.15 \times 0.20) = 0.0725$.

First, theory combined with experimental design provides some guidance that limits the degree to which researchers can engage in data mining. In many contexts, the primary outcome variable or variables for a given study will be fairly obvious. If you are studying an intervention to reduce teacher absence (as in Duflo, Hanna, and Ryan 2012), it is reasonably clear that you should show results on teacher absence and probably also results on student test scores. Any reasonable reader, referee, or journal editor would ask for such results if the authors did not report them. While there is some degree of manipulation researchers could do (for example, by reporting only math test scores and not language test scores), it is substantially limited by the expectations of readers concerning what outcomes the researchers should want to examine.

Second, authors are typically required to both show robustness and, for many journals, to make their data publicly available. Showing that main results are robust to a variety of specifications is statistically inefficient, because it means that papers are often judged by the average *p*-value across all specifications, rather than by a single, correctly specified *p*-value, but it has the advantage of making sure that authors are not systematically manipulating specifications to artificially improve their results. Making data available provides another check to make sure that researchers do not wildly mis-analyze their data. For example, the *American Economic Review*, and the other journals of the American Economic Association, along with *Econometrica,* the *Journal of Political Economy*, and the *Review of Economic Studies*, all require publication of data and programs for accepted articles.

Third, and perhaps most important, most academic researchers probably do not behave as the "nefarious" straw man I discussed in the beginning. To be sure, there are strong career and funding incentives, and everyone likes having strong and statistically significant results rather than statistically imprecise mush. But economics has no equivalent of the pharmaceutical trials where billions of dollars may depend on whether a single *p*-value is 0.049 or 0.051.

What Do Actual Papers Look Like?

To assess some of the challenges with pre-analysis plans in practice, I examined a set of recent papers that were using randomized controlled field trials. In particular, I looked at all such papers from the *American Economic Review*, *Quarterly Journal of Economics*, *Econometrica*, *Review of Economic Studies*, and *Journal of Political Economy* published from the start of 2013 until the middle of 2014: a total of 18 papers.[10] It is worth noting that none of these papers (as far as I could tell) had pre-analysis plans, which illustrates the degree to which pre-analysis plans are currently the exception, not the norm, in the economics profession.

For each of these papers, I examine the number of "primary" outcome variables and then the number of "conditional" tables of regressions, which potentially might have been specified in a different way if the primary outcome variables had

[10] The papers are listed in an online Appendix available with this paper at http://e-jep.org. I particularly thank John Firth for his help with this analysis.

realizations other than the ones that actually occurred. Since economists don't usually officially designate which outcomes are primary and which are secondary, and we cannot know for sure which tables would have been run conditional on the particular realization of outcomes and which would have been run regardless, this requires some judgment calls. Nevertheless, the exercise is useful to gauge some patterns and magnitudes.

First, these papers are complicated. The median paper has four treatment arms—three treatments groups and one control group—along with four main outcome variables. If we assume that each outcome variable could be positive, zero, or negative compared to the control group, that implies that each treatment arm has $3^4 = 81$ possible configurations of outcomes. Across three treatments, there are $81^3 = 531{,}441$ possible configurations of outcomes vis-à-vis controls. Second, it is common to look at secondary outcomes. The median paper in this group has 6.5 secondary outcomes, in addition to the primary outcomes. Third, I examine whether papers seem to be hovering near borderline statistical significance. If one was concerned that data mining was prevalent, one might expect most of the statistically significant *p*-values to be close to the 0.05 threshold.[11] However, these papers as a group are publishing statistically significant outcomes that are not close to the 0.05 threshold; they are much more statistically significant than that. Fourth, most of these papers use the robustness approach to convince readers that results are not spurious: specifically 10 of 18 papers show robustness tests to include controls of various types. Finally, 13 of the 18 papers examine subgroup heterogeneity.

This analysis suggests that complete pre-specification is not going to work without losing certain nuances that seem common in papers currently in top journals in economics. For example, supposing only one layer of conditionality, there are 531,441 possible combinations of primary outcome variables and results. Even if theory provides some guide for grouping these outcomes together, clearly the number of cases one would need to consider in writing a pre-analysis plan quickly becomes insuperable. Moreover, *p*-values are much more significant than 0.05, suggesting that fiddling around the margins is unlikely to be driving statistical significance in most of these studies. While the frequent use of heterogeneity analysis suggests that pre-specifying these issues may be important, overall these examples give some pause to the idea that requiring, or even strongly privileging, pre-specification for journal publication would on net improve the amount we learn from these trials.

[11] Specifically, for all statistically significant main outcomes (that is, all outcomes with *p*-values below 0.05), we calculate the *z*-statistic associated with it, and take the average. Across all significant outcomes in all papers, the average *z*-statistic is 3.18, which would correspond to a *p*-value of 0.0014. By comparison, if *p*-values were uniformly distributed between 0.00 and 0.05, one would expect an average *z*-statistic of 2.33, which would correspond to a *p*-value of 0.02; if there was substantial *p*-hacking, one might expect *p*-values closer to 0.05 and even lower average *z*-statistics. The reason it is not an average of 0.025 is because very low *p*-values have disproportionately high *z*-statistics, so the average *z*-statistic does not correspond to the average *p*-value.

Thoughts on the Way Forward

Economics papers tend to be complicated, and pre-specifying the entire chain of analysis is probably impossible for the median paper in economics. Forcing all papers to be fully pre-specified from start to end would likely result in simpler papers, which could potentially lose some of the nuance of current work. If economists were to exclude from publication or policy consideration all non-pre-specified, exploratory results in the name of increased transparency, we would be losing more than we would gain.

That said, in many contexts, pre-specification of one (or a few) key primary outcome variables, statistical specifications, and control variables offers a number of advantages. In cases where there is a partner with any kind of vested interest in the outcome, pre-specification of outcomes and analysis can be a huge advantage to all parties. Even when there is not a strong interested party, the rigor of researchers specifying a small number of primary outcomes in advance is a useful exercise that will help ensure that when data are analyzed, they know what to focus on. For the many decisions where there is no clear hard decision to make—what statistical model to use, what control variables to include, and so on—pre-specification frees the author from the need to report a large number of robustness checks and in so doing make their effective statistical power worse than it needs to be. Even if journals do not require pre-specification, individual researchers may choose to do so in order to enhance the credibility of their results, and mechanisms like the AEA registry that allow them to commit publicly to pre-registration can be useful to allow them to do so.

■ *I thank Abhijit Banerjee, Paul Catalano, Esther Duflo, Amy Finkelstein, Marc Fisher, Rachel Glennerster, Lisa LaVange, Heather Lanthorn, Edward Miguel, Brian Nosek, Sharon-Lise Normand, Robert O'Neil, Uri Simonsohn, Robert Temple, Marta Wosinska, and participants at the Berkeley Initiative for Transparency in Social Sciences conference for many helpful discussions on these topics; Gordon Hansen, Enrico Moretti, and Timothy Taylor for helpful editorial suggestions; and John Firth for comments and research assistance. I also thank my many coauthors with whom I have worked on preparing these analysis plans from scratch for our research. The views in this paper are those of the author alone and do not represent any of the individuals acknowledged here or their respective institutions.*

References

Ashraf, Nava, Erica Field, and Jean Lee. 2014. "Household Bargaining and Excess Fertility: An Experimental Study in Zambia." *American Economic Review* 104(7): 2210–37.

Avvisati, Francesco, Marc Gurgand, Nina Guyon, and Eric Maurin. 2014. "Getting Parents Involved: A Field Experiment in Deprived Schools." *Review of Economic Studies* 81(1): 57–83.

Balafoutas, Loukas, Adrian Beck, Rudolf Kerschbamer, and Matthias Sutter. 2013. "What Drives Taxi Drivers? A Field Experiment on Fraud in a Market for Credence Goods." *Review of Economic Studies* 80(3): 876–91.

Blattman, Christopher, Nathan Fiala, and Sebastian Martinez. 2014. "Generating Skilled Self-Employment in Developing Countries: Experimental Evidence from Uganda." *Quarterly Journal of Economics* 129(2): 697–752.

Bloom, Nicholas, Benn Eifert, Aprajit Mahajan, David McKenzie, and John Roberts. 2013. "Does Management Matter? Evidence from India." *Quarterly Journal of Economics* 128(1): 1–51.

Brodeur, Abel, Mathias Lé, Marc Sangnier, and Yanos Zylberberg. Forthcoming. "Star Wars: The Empirics Strike Back." *American Economic Journal: Applied Economics.*

Bursztyn, Leonardo, Florian Ederer, Bruno Ferman, and Noam Yuchtman. 2014. "Understanding Mechanisms Underlying Peer Effects: Evidence From a Field Experiment on Financial Decisions." *Econometrica* 82(4): 1273–1301.

Casey, Katherine, Rachel Glennerster, and Edward Miguel. 2012. "Reshaping Institutions: Evidence on Aid Impacts Using a Preanalysis Plan." *Quarterly Journal of Economics* 127(4): 1755–1812.

Crépon, Bruno, Esther Duflo, Marc Gurgand, Roland Rathelot, and Philippe Zamora. 2013. "Do Labor Market Policies Have Displacement Effects? Evidence from a Clustered Randomized Experiment." *Quarterly Journal of Economics* 128(2): 531–80.

Dal Bó, Ernesto, Frederico Finan, and Martín A. Rossi. 2013. "Strengthening State Capabilities: The Role of Financial Incentives in the Call to Public Service." *Quarterly Journal of Economics* 128(3): 1169–218.

Deaton, Angus. 2010. "Instruments, Randomization, and Learning about Development." *Journal of Economic Literature* 48(2): 424–55.

Duflo, Esther, Michael Greenstone, Rohini Pande, and Nicholas Ryan. 2013. "Truth-telling by Third-party Auditors and the Response of Polluting Firms: Experimental Evidence from India." *Quarterly Journal of Economics* 128(4): 1499–545.

Duflo, Esther, Rema Hanna, and Stephen P. Ryan. 2012. "Incentives Work: Getting Teachers to Come to School." *American Economic Review* 102(4): 1241–78.

Dunn, Olive J. 1961. "Multiple Comparisons among Means." *Journal of the American Statistical Association* 56(293): 52–64.

Dupas, Pascaline. 2014. "Short-Run Subsidies and Long-Run Adoption of New Health Products: Evidence from a Field Experiment." *Econometrica* 82(1): 197–228.

Dupas, Pascaline, and Jonathan Robinson. 2013. "Why Don't the Poor Save More? Evidence from Health Savings Experiments." *American Economic Review* 103(4): 1138–71.

Easterly, William. 2012. "If Christopher Columbus Had Been Funded by Gates." Blog post, NYU Development Research Institute, October. http://www.nyudri.org/2012/10/15/if-christopher-columbus-had-been-funded-by-gates/.

Eriksson, Stefan, and Dan-Olof Rooth. 2014. "Do Employers Use Unemployment as a Sorting Criterion When Hiring? Evidence from a Field Experiment." *American Economic Review* 104(3): 1014–39.

Feigenberg, Benjamin, Erica Field, and Rohini Pande. 2014. "The Economic Returns to Social Interaction: Experimental Evidence from Microfinance." *Review of Economic Studies* 80(4): 1459–83.

Field, Erica, Rohini Pande, John Papp, and Natalia Rigol. 2013. "Does the Classic Microfinance Model Discourage Entrepreneurship among the Poor? Experimental Evidence from India." *American Economic Review* 103(6): 2196–226.

Food and Drug Administration. 1998. "Guidance for Industry: E9 Statistical Principles for Clinical Trials." US Department of Health and Human Services. http://www.fda.gov/downloads/drugs/guidancecomplianceregulatoryinformation/guidances/ucm073137.pdf.

Food and Drug Administration. 2010. "Adaptive Design Clinical Trials for Drugs and Biologics." Draft Guidance. U.S. Department of Health and Human Services. http://www.fda.gov/downloads/Drugs/.../Guidances/ucm201790.pdf.

Jessoe, Katrina, and David Rapson. 2014. "Knowledge Is (Less) Power: Experimental Evidence from Residential Energy Use." *American Economic Review* 104(4): 1417–38.

Karlan, Dean, Robert Osei, Isaac Osei-Akoto, and Christopher Udry. 2014. "Agricultural Decisions after Relaxing Credit and Risk Constraints." *Quarterly Journal of Economics* 129(2): 597–652.

Klein, Richard A. et al. 2014. "Investigating Variation in Replicability: A "Many Labs" Replication Project." *Social Psychology* 45(3): 142–52.

Kling, Jeffrey R., Jeffrey B. Liebman, and Lawrence F. Katz. 2007. "Experimental Analysis of Neighborhood Effects." *Econometrica* 75(1): 83–119.

Kremer, Michael R., Edward Miguel, and Rebecca Thornton. 2009. "Incentives to Learn." *Review of Economics and Statistics* 91(3): 437–56.

Kroft, Kory, Fabian Lange, and Matthew J. Notowidigdo. 2013. "Duration Dependence and Labor Market Conditions: Evidence from a Field Experiment." *Quarterly Journal of Economics* 128(3): 1123–67.

Leamer, Edward E. 1983. "Let's Take the Con out of Econometrics." *American Economic Review* 73(1): 31–43.

Miguel, E., et. al. 2014. "Promoting Transparency in Social Science Research." *Science* 343(6166): 30–31.

National Science Foundation. 2013. "Social, Behavioral, and Economic Sciences." https://www.nsf.gov/about/budget/fy2013/pdf/10-SBE_fy2013.pdf. In *National Science Founation FY 2013 Budget Request to Congress.* https://www.nsf.gov/about/budget/fy2013/.

Neumark, David. 2001. "The Employment Effects of Minimum Wages: Evidence from a Prespecified Research Design." *Industrial Relations* 40(1): 121–44.

Tarozzi, Alessandro, Aprajit Mahajan, Brian Blackburn, Dan Kopf, Lakshmi Krishnan, and Joanne Yoong. 2014. "Micro-loans, Insecticide-Treated Bednets, and Malaria: Evidence from a Randomized Controlled Trial in Orissa, India." *American Economic Review* 104(7): 1909–41.

Westfall, Peter H., and S. Stanley Young. 1993. *Resampling-based Multiple Testing: Examples and Methods for* p*-value Adjustment.* Hoboken, NJ: John Wiley & Sons.