NBER WORKING PAPER SERIES

USING SPLIT SAMPLES TO IMPROVE INFERENCE ABOUT CAUSAL EFFECTS

Marcel Fafchamps
Julien Labonne

Working Paper 21842
http://www.nber.org/papers/w21842

NATIONAL BUREAU OF ECONOMIC RESEARCH
1050 Massachusetts Avenue
Cambridge, MA 02138
January 2016

We thank Rob Garlick for fruitful discussions while working on this paper and Kate Vyborny for comments. All remaining errors are ours. The views expressed herein are those of the authors and do not necessarily reflect the views of the National Bureau of Economic Research.

Using Split Samples to Improve Inference about Causal Effects
Marcel Fafchamps and Julien Labonne
NBER Working Paper No. 21842
January 2016
JEL No. C12,C18

ABSTRACT

We discuss a method aimed at reducing the risk that spurious results are published. Researchers send their datasets to an independent third party who randomly generates training and testing samples. Researchers perform their analysis on the former and once the paper is accepted for publication the method is applied to the latter and it is those results that are published. Simulations indicate that, under empirically relevant settings, the proposed method significantly reduces type I error and delivers adequate power. The method – that can be combined with pre-analysis plans – reduces the risk that relevant hypotheses are left untested.

Marcel Fafchamps
Freeman Spogli Institute
Stanford University
616 Serra Street
Stanford, CA 94305
and NBER
fafchamp@stanford.edu

Julien Labonne
Yale-NUS College
College Avenue West
Singapore
julien.labonne@yale-nus.edu.sg

A online appendix is available at http://www.nber.org/data-appendix/w21842

1 Introduction

The gap between econometric theory and practice makes it challenging to assess the reliability of empirical findings in economics and political science (Leamer, 1974, 1978, 1983; Lovell, 1983; Glaeser, 2006). This is due to a combination of researchers' degree of freedom and publication bias. As a result, the probability of Type I error in published research is believed to be larger than the commonly accepted five percent. For example, Gerber and Malhotra (2008) and Brodeur et al. (forthcoming) report that there is a bunching of p-values just below the 0.05 threshold in top economics and political science journals. This is consistent with researchers and editor unconsciously or consciously selecting outcome variables, regression methods, estimation samples, and control variables to deliver significant results.[1]

A number of reforms of the reviewing process have been proposed to decrease the risk that spurious findings are published and cited (Green, Humphreys and Smith, 2013). The common objective is to encourage researchers to transparently select which statistical tests to implement *before* accessing the data on which they will be run. A prominent example is the introduction of pre-analysis plans (PAPs).[2] Such plans are written – and possibly shared with the research community – before any analysis is carried out. This reduces the risk that researchers select hypotheses that can be rejected with the available data (Casey, Glennerster and Miguel, 2012; Olken, 2015). Some researchers are more skeptical and argue that the profession should encourage replications instead (Coffman and Niederle, 2015). The profession has also insisted on the need to correct for multiple comparisons. While filing a PAP may offer some protection in lab and field experiments, it offers no protection in the case of analysis of observational data.

In this paper we discuss a related – and possibly complementary – method that can

[1] This behavior isn't confined to economics and political science, however. As reported by Shea (2013), Dirk Smeesters, a social psychologist at Erasmus University Rotterdam, whose own university publicly announced that it had "*no confidence in the scientific integrity*" of three of his articles, stated that "*many authors knowingly omit data to achieve significance, without stating this*".

[2] Building on J-PAL Hypothesis Registry, The American Economic Association has recently set-up a RCT registry. To-date the American Political Science Association hasn't followed suit but the E-GAP (Experiments in Governance and Politics) network allows researchers to register both experimental and observational studies.

be applied to both new and existing datasets. The process involves sending the data to an independent third party who randomly generates two non-overlapping subsets of the data. Researchers only have access to one subset – called the training dataset – while the third party keeps the second one – called the testing dataset. Researchers are free to analyze the training dataset and can adapt ideas from seminar audiences, editors and referees and incorporate them into the analysis. Once the paper has been accepted for publication it is akin to a detailed pre-analysis plans that fully specifies the regressions to be estimated on the testing sample. The analysis is implemented, unchanged, on the testing dataset, and these results are the ones that are published.[3] An important feature of the method is that only a subset of hypotheses tested on the training sample - potentially those that are rejected on that sample - are carried over to the testing sample.[4] As we correct for multiple testing, this compensates for the loss of power due to smaller sample sizes. Importantly, this method can be combined with a PAP as a researcher could register a number of hypotheses in a PAP and carry out more exploratory analyses with the proposed method. In addition, the method can be applied by researchers using existing data. In those situations a PAP might not be as credible.

The proposed method offers three methodological benefits: increased ability to learn from the data and to test hypotheses that researches did not think about when they started their project – an issue that is not addressed by PAPs; reduced type I errors, as in PAPs; and reduced the risk of publication bias. The main potential cost of split samples is loss of power. First, the method allows researchers to test hypotheses that they did not think about before starting their analysis: researchers can refine their research plans based on initial findings, interactions with seminar audiences, and re-

[3] In the recent field of genoeconomics, researchers often attempt to replicate their findings by testing whether the identified genes are correlated with the outcomes of interest on another sample (Benjamin et al., 2011). Other researchers have used a related method and construct a genetic score on a subsample and check its predictive accuracy on the remaining sample (Benjamin et al., 2012). In both cases, no third-party is involved and researchers have control over the choice of the other sample and over the split between the training and the testing sample.

[4] In practice, we do not recommend a purely mechanical approach with *all* hypotheses rejected on the training set tested on the testing set. Rather, researchers are expected to select a subset of those hypotheses. This should be guided by both economic theory and additional robustness tests.

quests from referees and editors.[5] Second, it reduces the risk of Type I error because researchers fully specify the regressions they want to estimate before having access to the dataset on which hypotheses will be tested. This reduces the risk of focusing on specifications where a spurious null happens to be rejected. Third, the method reduces the risk of publication bias because journal editors decide whether to publish a paper before seeing the final results.[6] This last benefit could also be achieved by PAPs if journal editors were willing to accept a paper on the basis of the quality of PAP design alone.

To capture these features in a simple way, we imagine a situation in which the researcher wishes to test multiple hypotheses, without strong a priori information on which hypotheses are most relevant. In such a situation, it is common for researchers to adjust for multiple testing. We present results from simulations quantifying the trade-off between reduced type I error and loss of power, compared to a situation where the researcher test hypotheses on the full sample. In both cases we adjust for multiple testing. We show that the loss of power from using split samples instead of the full sample is lowest when the total number of tests is large – that is, when the researcher most wishes to learn from the data. In this case, multiple comparison adjustments can induce a large reduction in power when using the full sample. The split sample approach allows the researcher to curtail the number of tests carried on the testing sample, and this compensates for the loss of power due to smaller sample size. This is because researchers decide on the few hypotheses to test based on initial work with the data which limits the loss of power associated with multiple testing adjustments. We also provide guidance on the optimal way of splitting the full sample into training and testing subsamples.

Results presented in the paper indicate that in a large number of relevant empirical settings, the loss of power associated with the split sample is manageable. Economi-

[5]One could argue that additional hypotheses can be addressed in future research. But given the cost of collecting additional data and the long publication lag in economics, this would unnecessarily delay the availability of evidence.

[6]Franco, Malhotra and Simonovits (2014) take advantage of an NSF-sponsored program to quantify publication bias. They show that strong results are much more likely to be published. This effect is partially explained by the fact that researchers do not write up null findings (Franco, Malhotra and Simonovits, 2014), and partly by the fact that editors and referees are reluctant to publish null results.

cally significant effect size (above 0.2 standard deviation) can be detected with power comfortably above 80 percent as soon as sample size is above 3,000. For a smaller effect size (*e.g.,* 0.1 standard deviation), a sample of 10,000 observations or more is required. In addition, we provide evidence that the split sample approach is more likely than a PAP approach to identify a null hypothesis that should be rejected. When using a PAP, researchers often keep the number of tested hypotheses small to counteract the loss of power due to multiple testing. Our proposed method allows researchers to test a large number of null hypotheses with only a small loss in power.

Results further suggest that the method increases the likelihood that relevant hypotheses are tested. Indeed, due to the expected loss in power associated with multiple testing adjustments researchers often limit the number of hypotheses included a PAP. In those situations, researchers are unable to learn from observations made during data collection and field experimentation. For effect sizes of .3 , we show that as long as there is a small likelihood that the relevant hypothesis isn't included in the PAP, the split sample approach will deliver more power as soon as sample size is above 2,000.

We argue that the method is especially relevant as the profession is entering the *age of big data* (Einav and Levin, 2014). Researchers now have access to large datasets from both the public and private sectors and are increasingly able to run experiments on a large number of subjects. Those datasets often contain a large number of potential outcome and control variables which creates great opportunities for exploring previously untestable hypotheses. It appears important to develop methods that deliver credible results (Athey and Imbens, 2015; Belloni, Chernozhukov and Hansen, 2014).

It is important to note that there are other ways through which spurious results can be published, but dealing with them is beyond the scope of this paper. The method would still deliver biased estimates if researchers use unreliable data, or faulty code and software. For example, Bell and Miller (forthcoming) could replicate Rauchhaus (2009)'s findings in STATA but not in R, which they attribute to a problem in STATA. More perniciously, some researchers have been caught fabricating data. In line with current practice, we argue that the best way to deal with those issues is to ask researchers to make their code and data publicly available after publication. This would

increase the likelihood that potential mistakes are quickly identified.

The remainder of the paper is organised as follows. In Section 2, we present a canonical setup often encountered in empirical work. The proposed method is described in Section 3. Results on power and the family-wise error rates are discussed in Section 4. Section 5 highlights three additional benefits: the method allows researchers to learn from the data, controls referee degrees of freedom and helps editors decide whether to publish null results. Section 6 concludes.

2 The Problem

In this section, we discuss current empirical practices and why they might lead to the publication of spurious findings. We also describe how researchers currently attempt to deal with those issues.

2.1 Canonical set-up

We consider the following canonical setup. Researcher A is interested in estimating the effect of an exogenous treatment T (with $T = 1$ for half of the observations and $T = 0$ otherwise). She has access to a sample S of size N that includes a set of m potential outcome variables $(y^k)_{k=1,...,m}$. The m outcome variables can either capture different concepts, related concepts, or different ways of measuring the same concept. For example, the researcher may have access to firm data on firms' hiring practices, number of employees, value-added, profits, etc. Unsure of which aspects of firm performance is affected by treatment, the researcher runs regressions of the form:

$$y^k = a + b_k T + u \tag{1}$$

The researcher then runs a series of tests $H_0^k : b_k = 0$. Some of these null hypotheses are true, some are non-true. The researcher faces a multiple comparison problem: without adequate adjustments, the probability that a true null hypothesis is rejected is higher than the level α at which each individual test is carried out. The set-up, adapted from

Benjamini and Hochberg (1995), is summarised in Table 1.

Table 1: Set-up

	Declared Non-significant	Declared Significant	Total
True null hypotheses	U	V	m_0
Non-true null hypotheses	T	S	$m - m_0$
Total	m-R	R	m

The researcher is concerned about Type I errors and wants to find ways to control the Family Wise Error Rate (FWER).

Definition 1 *The Family Wise Error Rate is the probability of rejecting at least one true null hypothesis. In the notation of Table 1, it is equal to* $Pr(V > 0)$.

The most basic way to keep the FWER in check is to make Bonferroni adjustments: instead of rejecting H_0 if the p-value is smaller than α, reject if it is smaller than α/m. Let R_k be a variable indicating whether hypothesis k was rejected. It is straightforward to show that the adjustment controls the FWER:

$$P(V > 0) \leq P(R > 0) = P(\bigcup_{k=1}^{m} R_k) \leq \sum_{k=1}^{m} P(R_k) = m * \frac{\alpha}{m} = \alpha$$

The adjustment is only valid if all null hypotheses are true $(m = m_0)$ and all tests are independent. It is well known that this correction tends to be very conservative and can lead to serious loss of power. In addition, the method is only valid if the researcher can keep track of all tests she performed. If for example, the researcher ran $m^{'}$ tests and attempt to control the FWER as if only m tests had been carried out (with $m < m^{'}$), the reported FWER will understimate the actual FWER.

Let α be the significance level used to test H_0^k and let δ_k be the standardized effect size for the $m - m_0$ non-true null hypotheses. In this convenient set-up we can use standard power calculations formula (see McConnell and Vera-Hernández (2015)). Power, denoted as $1 - \beta$, is the probability of rejecting the null hypothesis when the

alternative is correct. Under our assumptions, it is given by:

$$1 - \beta_k = \Phi(\delta_k \sqrt{\frac{N}{4}} - Z_{1-\frac{\alpha}{2}}) \tag{2}$$

where Φ is the cumulative distribution function for the standard normal distribution. The detailed calculations are available in the Appendix. If the researcher carries out Bonferonni corrections, power becomes:

$$1 - \beta_k^{Bonf} = \Phi(\delta_k \sqrt{\frac{N}{4}} - Z_{1-\frac{\alpha}{2m}}) \tag{3}$$

Comparing the two formulas directly shows that Bonferonni corrections lead to a loss of statistical power. This loss is increasing in m, the number of tests that are carried out.

Since the probability of rejecting each true null hypotheses is α, the probability of rejecting at least one is given by:

$$FWER = 1 - (1 - \alpha)^{m_0} \tag{4}$$

where m_0 is the (unknown) number of true null hypotheses. It is important to note that the FWER is not a function of sample size or effect size. If the researcher carries out Bonferonni corrections, the FWER becomes:

$$FWER^{Bonf} = 1 - (1 - \frac{\alpha}{m})^{m_0} \tag{5}$$

Researchers are now using alternative p-values adjustments to correct for multiple testing (*e.g.* the methods proposed by Benjamini and Hochberg (1995) and Benjamini, Krieger and Yekutieli (2006)) and we will consider those approaches in the simulations.

2.2 Pre-Analysis Plan

Before having access to the data, the researcher can prepare and register a pre-analysis plan (Coffman and Niederle, 2015; Olken, 2015). Such a plan lists the hypotheses to

be tested and describes how they will be tested, including which variables to include, how they will be included, and how researchers intend to deal with the multiple comparison problems.

This approach is appealing but it has some drawbacks. First, following a PAP to the letter does not allow researchers to learn from the data, and this can slow down the pace of new discoveries. Indeed, PAPs can only cover hypotheses that the researcher could think of before carrying out their experiment. There often are other testable hypotheses that the researcher did not think of beforehand. A number of social scientists have recently argued that some of their most important findings were the direct result of time spent with the data (Laitin, 2013; Gelman, 2014). For example, Simonsohn (cited by Laitin (2013)) argues that: "*I also think of science as a process of discovery . . . Every paper I have [written] has some really interesting robustness, extensions, follow-ups that I would have never thought about at the beginning*." Similarly, Gelman (2014) states that "*Many of my most important applied results were interactions that my colleagues and I noticed only after spending a lot of time with our data*."

Second, unless pre-analysis plans fully specify the regressions to be estimated, it still leaves some room for data mining. As a result, Humphreys, Sanchez de la Sierra and van der Windt (2013) argue that researchers should write a mock report with fake data. This forces researchers to make all decisions regarding the analysis (including micro-decisions such as the precise way of defining all variables) before having access to the dataset on which the regressions will be estimated. The methodology is then applied to the real data.

Third, it is difficult to credibly implement PAPs in observational studies because it is difficult to guarantee that the researcher has not run the regressions before registering the PAP. This concern is especially acute in situations where the data have already been used by other researchers. PAPs are better suited for analysis of experimental data.

Fourth, unless editors are willing to unconditionally accept papers based on a detailed pre-analysis plan, there is always room for what Pepinsky (2013) refers to as *referee degree of freedom*, i.e., the referees (and editor) may require the researcher to con-

duct analysis that was not in the PAP.

Fifth, PAP forces researchers to divulge their research design with other, possibly competing researchers at an early stage of the research process. Given the long publication lags in economics, this opens the door to abuse.[7]

Finally, as long as the decision to publish results is based on whether or not some null hypothesis is rejected, there remains a risk that, even if all research follows a PAP, many published findings are spurious. To illustrate, imagine m researchers, each with access to data on treatment T and one of the outcome variables $(y^k)_{k=1,...,m}$. Each of these m researchers registers a PAP to estimate the effect of T on a single y^k. All tests for which the null is rejected are then published. Ioannidis (2005) argues that since there are many more true null hypotheses than false ones, as long as m is sufficiently large there will be more cases of Type I error than of cases where the null is correctly rejected.

3 The Method

We now describe the split sample approach in details.[8] As above, we assume that researcher A is interested in estimating the effect of T on a list of possible outcomes $(y^k)_{k=1,...,m}$. There is some uncertainty regarding which particular hypotheses to test and how to best test them. The research project proceeds as follows:

- Step 1: Guided by theory and existing evidence, researcher A puts together a sample S including a number of variables that broadly captures the general set of hypotheses that she wants to test. The researcher also includes variables used to test for potential heterogeneous effects.

- Step 2: The data is then sent to a third-party B who randomly generates two non-overlapping subsets. If the researcher is interested in studying particular sub-

[7] A number of researchers have opted to gate their PAP to address those concerns.

[8] Our method differ from earlier efforts to use split-sample in applied econometrics. Researchers focused on pre-testing bias; more specifically of how the potential bias arising from dropping regressors based on the associated t-statistics in both OLS and IV estimation (e.g. Angrist and Krueger (1995)). Researchers were concerned about the determinants of a single outcome variables. We are concerned about how one treatment variable affects a large number of potential outcome variables.

groups the sample should be stratified accordingly. The first sub-sample (*training sample*) is sent back to A. The third-party keeps the second one (*testing sample*). All relevant IDs are scrambled during the process so that A is unable to 'reverse engineer' the randomization.

- Step 3: A runs regressions, presents the results at seminars and conferences, and refines the methodology based on feedback received.

- Step 4: The paper is submitted to a journal, referees make their comments and A amends her analysis in response, possibly several times.

  The discovery process described by steps 3 and 4 identifies a final subset J of the m outcome variables such that each of these outcome variables is significant at the α level in the training set, conditional on a choice of estimator, control variables, and standard error correction. We call this the final methodology for analysis. According to our simulations discussed in the next Section, in most contexts $J \ll m$ which compensates somewhat for the loss of power due to lower sample size when correcting for multiple testing.

- Step 5: The editor accepts the paper conditional on the agreed upon final methodology for analysis. A then secures the testing sample from B and applies the agreed upon methodology to it. The published version of the paper only includes the results obtained from the testing sample.

  We argue that editors might be less reluctant to accept a paper based on results from the training set than a PAP design. Indeed, it contains more information about the results and, in the case of a RCT, about the quality of its implementation. The strength of the main results and associated robustness checks on the training sample provide some information as to whether they will hold on the testing sample. In addition, one can make a case that precisely estimated zeroes should be published (as opposed to underpowered studies) and results from the training set provide useful information on the study's statistical power. We discuss this in more details in Section 5.3.

Importantly, even in cases where the editor requires to see the results on the testing sample before accepting the paper, the authors could register a PAP containing all relevant details before running the regressions on the testing sample. If the editor declines to publish the paper after seeing the results, this would allow authors to have a record of a pre-registered design when they submit the paper to the next journal.

We think of Steps 3 and 4 as a way for researchers to refine their research plan. The methodology that is accepted in Step 5 is akin to a detailed pre-analysis plans that fully specifies the regressions to be estimated on the testing sample. As researchers can adapt ideas from seminar audiences, editors and referees and incorporate them into the analysis, there is room in the analysis plan to incorporate interesting hypotheses that A would not have tested otherwise.

More formally, the process looks as follows. Researcher A puts together a sample S that includes N observations and a set of $m+1$ variables: T_i and $(y^k)_{k=1,\dots,m}$. A third party B then randomly splits the dataset into two sub samples S_1 and S_2 such that: $S = S_1 \cup S_2$ and $S_1 \cap S_2 = \emptyset$. At first, the researcher does not have access to S_2. The researcher starts with a set of specific hypotheses to test. Feedback from other researchers is then used to help A finalize a list of hypotheses to test. This list can be represented most generally as a series of J triplets consisting of: (1) a set of outcome variables $(z^j)_{j=1,\dots,J}$ which we allow to be transformations of the original data $(y^k)_{k=1,\dots,m}$ such that $\forall j$ we have $z_i^j = f(y_i^1, \dots, y_i^m)$;[9] (2) an estimation method (e.g., estimator, control variables); and (3) a set of rules that define the estimation sample (e.g., excluded outliers). Once the J triplets are agreed with an editor, the associated regressions are estimated on S_2 and this is the set of results that are published. The method's key feature is that, given that the training and testing samples are independent, the probability of type I error in the two samples are independent.

[9] In the simplest case, the $(z^j)_{j=1,\dots,J}$ are simply the subset of the $(y^k)_{k=1,\dots,m}$ variables that are significant at the α level.

4 The Split Sample Approach, Power and Family-Wise Error Rate

4.1 Set-up

We now illustrate the method for the canonical setup described above. We compute power and FWER under the full sample approach and the proposed split sample approach. In both cases, we present results both with and without Bonferonni adjustments. We show the sensitivity of power and FWER to variation in the following parameters: the sample size (N); the standardized effect size (δ); the number of tested hypotheses (m); the number of tested null hypotheses that are true (m_0); and the share of the total sample that is allocated to the training set (s).

Throughout we assume that the researcher starts with m possible null hypotheses. Of these, a subset J are found to be significant at the α level in the training set and interesting. This subset determines the list of tests estimated on the testing set. To illustrate, let $m = 20$ and imagine that, in the training sample, treatment is significant at the $\alpha = 5\%$ level for seven of these 20 outcome variables. Then we only regress treatment on these seven outcome variables in the testing sample. It is this shrinking of the set of hypotheses that delivers power while keeping FWER low, as we now demonstrate.

Split sample without Bonferonni correction We start by showing how the formula introduced in Section 2 can be adjusted to compute power and the FWER with the split sample methodology. For a null hypothesis to be considered to be rejected, it is necessary that it be rejected first on the training sample, and then again on the testing sample. As a result, power is given by:

$$1 - \beta_k^{Split} = \Phi\left(\delta_k \sqrt{\frac{sN}{4}} - Z_{1-\frac{\alpha}{2}}\right) \Phi\left(\delta_k \sqrt{\frac{(1-s)N}{4}} - Z_{1-\frac{\alpha}{2}}\right) \tag{6}$$

Split sample with Bonferonni correction With Bonferonni correction, the calculations are as above except that we need to account for the number of tests carried out

on the testing sample. Power is the expected value of the following random variable:

$$1 - \beta_k^{Split/Bonf} = \Phi\left(\delta_k \sqrt{\frac{sN}{4}} - Z_{1-\frac{\alpha}{2}}\right) \Phi\left(\delta_k \sqrt{\frac{(1-s)N}{4}} - Z_{1-\frac{\alpha}{2B}}\right) \tag{7}$$

where B is the number of tests carried out on the testing sample. It distributed according to:

$$B(m_0, \alpha) + B\left(m - m_0, \Phi\left(\delta_k \sqrt{\frac{sN}{4}} - Z_{1-\frac{\alpha}{2}}\right)\right) \tag{8}$$

where $B(n, p)$ is a binomial distribution with n trials and p probability of success in each trial. The number of tests conducted on the testing sample is the sum of two terms: the number of true null hypotheses that are incorrectly rejected on the training sample; and the number of non-true null hypotheses that are correctly rejected on the training sample. To obtain an approximation of expected power, we take 10,000 draws of the distribution B using (8), compute power (7) for each iteration, and then take the average over all 10,000 iterations.

4.2 Results

We now present the results from applying the above formulas and simulation method to various parameter values. To capture the idea that there are many more true null hypotheses than false ones, we organize the simulations around the assumption that, out of 100 possible null hypotheses, only one is non-true, i.e., should be rejected. Hence, unless stated otherwise, the results presented below are based on $m = 100$ and $m_0 = 99$. Given these parameter values, the majority of the results found significant are spurious. For instance, if $\alpha = 5\%$, there will on average be five false rejections and, provided that power is high enough, one true rejection in the training sample. For now we use a 50-50 split between the training and testing samples, i.e., we set $s = 0.5$.

We organize our simulations around four stylized testing scenarios: (1) testing all 100 null hypotheses on the full sample without correction; (2) testing all 100 null hypotheses on the full sample with Bonferonni corrections; (3) testing all 100 null hy-

potheses on the training sample, and only testing on the training sample those null hypotheses that were significant in the training sample; and (4) proceeding as in (3) but adding Bonferroni corrections to the testing sample results.

We start by investigating the effect of sample size on the power to detect a true effect of size 0.2. In other words, we compute the likelihood of rejecting the null hypothesis when this hypothesis is false and the true effect is 0.2. Figure 1 plots power under the four scenarios for sample sizes varying between 500 to 10,000. Even with Bonferonni corrections, power under the split sample approach is well above 0.8 for the kind of sample sizes of 3,000 or more that are commonly encountered in empirical work. As expected, the Bonferonni corrections lead to a loss in power. But this loss of lower is less with the split sample approach than with the full sample approach. This makes sense because the split sample approach reduces the number of tests that are carried out on the testing sample.

Figures 2 and 3 plot similar results for different effect sizes of 0.1 and 0.3. Larger, but still relatively common, sample sizes are required to have power above 0.8 with smaller expected effect sizes (Figure 2). For example, with a small expected effect size of 0.1, raising power above 0.8 under the split sample approach requires sample sizes of 10,000 or more. When the expected effect size is 0.3, power under the split sample approach reaches 0.8 as soon as sample size is above 1,500.

So far we have set $m = 100$ and $m_0 = 99$. Next, we simulate what happens to power when we vary the total number of hypotheses that are being tested (m) and the number of non-true hypotheses (m_0). The effect size that we are trying to detect is 0.2, as in Figure 1. Figure 4 shows our simulation results for scenario (4) – the split sample approach with Bonferroni correction applied to the testing sample results. Results show that power is a decreasing function of m and m_0. This is because the Bonferroni correction becomes more stringent as m or m_0 increase.

Having shown that the split sample approach need not have a prohibitive cost in terms of loss of power, we now turn to its advantages in terms of minimizing the risk of false rejection. In Table 2 we compare the FWER under our four scenarios. Recall that the FWER is the probability of rejecting at least one true null hypothesis. We start

by observing that, for $m = 100$ and $m_0 = 99$, the FWER is close to 1 when we test all 100 null hypotheses on the full sample without Bonferroni corrections. Even without Bonferroni correction, moving from the full sample approach to the split sample approach result in a massive reduction in the FWER from 0.994 to 0.219. This improvement is due solely to the reduction in the number of hypothesis tests that are carried out on the testing sample. If we add Bonferroni corrections, the FWER falls below 5% with or without split sample. The formal similarity between the two approaches is misleading, however. For the FWER to be truly below 5% in the full sample approach, the researcher must credibly track and report all the tests they run. We argue that this is unlikely to be the case in most empirical applications (Gelman, 2013). In contrast, the split sample approach does not suffer from this type of under-reporting bias.

We also investigate whether it is optimal to split the sample 50-50 between training and testing sets. We continue to focus on scenario (4) – sample split with Bonferroni corrections – and we simulate power under alternative sample splitting rules, i.e., 30/70 and 70/30. The results, displayed in Figure A.2, indicate that, across all considered sample sizes, a 50/50 split delivers the best power.

Next, we investigate whether power in the split sample approach with Bonferroni correction depends on the threshold level of significance used to select hypotheses in the training sample. So far we have assumed that this threshold is the same in the training and testing samples, i.e., $\alpha = 0.05$. We now compare this situation to using a threshold of 0.1 when selecting hypotheses on the training sample. Three effect sizes are considered: $0.1, 0.2$ and 0.3. We find that, for all three effect sizes, power appears to be marginally larger with a 0.05 threshold than a 0.1 threshold. This is because applying a less restrictive threshold to the training sample increases the number of true null hypotheses that are rejected, and thus the number of hypotheses that are tested on the testing sample. A larger number of hypotheses means that a stronger Bonferroni correction is required on the testing sample, and this is what drives the loss of power.

4.3 Extensions

Clustered samples. Up to now we have assumed that researchers have access to an unclustered sample (or that inter-cluster correlation is sufficiently low to be ignored). In a number of settings this assumption is likely to be violated and we now report results from simulations with clustered sample. In a sample with c clusters and an intra-cluster correlation coefficient of ρ power is given by:

$$1 - \beta_k^{Clustered} = \Phi(\delta_k \sqrt{\frac{N}{4 * (1 + (c-1)\rho)}} - Z_{1-\frac{\alpha}{2}}) \tag{9}$$

We can easily adjust the formula to obtain power both for the full sample approach and the split sample approach with Bonferonni corrections. As before, we run 10,000 simulations. We compute power for sample sizes varying from 500 to 10,000 with 20 observations per clusters. We assume that ρ is either .05 or .1. Results are available in Figure A.3. As expected power is lower than what it is with an unclustered sample. For example, with $\rho = .05$ power is above .8 with the split sample approach for samples of 5,000 observations and more. If $\rho = 1$, sample sizes of about 8,000 are required for power to be above .8.

Alternative p-values adjustments. Up to now we have assumed that researchers are interested in controlling the FWER and thus rely on Bonferonni corrections. In some contexts researchers are interested in controlling the False Discovery Rate (FDR) instead.

Definition 2 *The False Discovery Rate is the expected proportion of errors among the rejected hypotheses. In the notation of Table 1, it is equal to $E(Q)$; where $Q = \frac{V}{R}$ if $R > 0$ and $Q = 0$ if $R = 0$.*

The concept, introduced Benjamini and Hochberg (1995), captures the idea that, in a number of relevant cases, it is acceptable to reject true null hypotheses as long as such rejections constitute a small share of total rejections. The intuition is that the decision-maker would reach the same conclusion regardless of whether or not those true null hypotheses are rejected.

Benjamini and Hochberg (1995) proposed a method to control the FDR. The BH method proceeds as follows:

1. Carry out the m tests and get the associated p-values $p_1, \ldots, p_m$
2. Rank the p-values from smallest to largest $p_{(1)}, \ldots, p_{(m)}$
3. Get $k = Max\{i | p_{(i)} \leq \frac{i}{m} q\}$. q is the level at which the researcher would like to control the FDR.
4. Reject all $H_{(i)}$ for $i \leq k$

Benjamini and Yekutieli (2001) show that the method is conservative as it controls the FDR at level $\frac{m_0}{m} q$. The proof relies on the fact that while for true null hypotheses the p-values are uniformly distributed over $[0, 1]$, they tend to be bunched towards 0 for non-true null hypotheses. As a result, when observing two p-values the hypothesis associated with the smallest one is more likely to be non-true. Simulations presented in Benjamini and Hochberg (1995) indicate that power is significantly larger than for methods that control the FWER. Benjamini, Krieger and Yekutieli (2006) extend the method to a two-stage procedure where the first stage is used to get an estimate of m_0. The sharpened q-values are obtained as follows:

1. Apply the BH procedure at level $q' = q/(1+q)$. Let c be the number of hypotheses rejected. If $c = 0$, stop; otherwise, continue to step 2.
2. Let $\hat{m}_0 = M - c$
3. Apply the BH procedure at level $q^* = q' m / \hat{m}_0$

We run simulations computing q-values for $m = 100$, $m_0 = 90$, $\delta = .2$ and sample sizes varying from 500 to 5,000 in 100 increments. We assume that half of the observations are allocated to the training set. For both the full sample and the split sample approach we compute power as the share of the 10,000 iterations for which the q-value is below .05. Results are available in Figure A.4. As expected under both the full sample and the split sample approaches, power is higher when using sharpened q-values

than when using Bonferonni corrections. In addition, power under the split sample approach is now above .8 as soon as sample sizes are larger than 2,000 observations.

5 Additional Benefits from the Split Sample Approach

5.1 Researchers' Ability to Learn

The split sample approach has one important additional benefit: it allows the researcher to test a large number of hypotheses with little loss in power. When using a PAP with full sample analysis, the researcher is often induced to select a short list of tested hypotheses in order avoid the loss of power due to Bonferroni corrections. This short list typically includes hypotheses that the researcher a priori believes are most likely to be rejected. This means that many hypotheses (e.g., outcome variables) are excluded from the PAP, thereby preventing the researcher from learning from observations made during data collection and field experimentation. Because our method reduces the loss of power due to multiple testing, it allows researchers to learn from the data and to test hypotheses that they did not think about when they started the project.

To illustrate, let's imagine that the researcher has a dataset with 100 potential outcome variables but decided to only include 10 of them in the PAP. We keep other assumptions unchanged. In particular, we continue to assume that the null hypothesis should only be rejected for one of the 100 potential outcome variables. The question is whether this hypothesis is included in the shortlist or not. If it is, the shortlist approach yields correct inference. But if it is left out, the researcher might wrongly declare that the treatment has no effect. We now show that under a variety of settings the split sample approach reduces that risk.

Let ψ be the likelihood that the one hypothesis to be rejected is included in the shortlist of 10 tests. Once Bonferonni adjustments are taken into account, power under the full sample is given by:

$$Power_{PAP} = \psi * \Phi(\delta_k \sqrt{\frac{N}{4}} - Z_{1-\frac{\alpha}{2*10}}) \tag{10}$$

Power under the split sample approach with Bonferonni corrections is given as before by equation (7). Using these two formulas, we can compute the value ψ^* at which the two methods yield similar power. In Figure 5, we plot the value of ψ^* for various effect sizes (.1, .2 and .3). For all values of ψ below the curve, the split sample approach delivers more power. In a large number of cases, ψ needs to be close to one for the full sample approach with a PAP to be superior (or equivalent) to the split sample approach. For example, for effect sizes of .3 as soon as sample size is above 2,000, ψ needs to be one for the two approaches to yield similar results. Even with an effect size of .1 and a sample size of 7,000, ψ needs to be above .6 for the full sample approach with a PAP to dominate. This set of results thus confirms that the split sample approach increases researchers' ability to learn from the data.

5.2 Controlling Referee Degrees of Freedom

In economics and political science referees and editors take a very active role in defining the paper's methodology. They often suggest alternative estimation strategies, alternative outcome variables and alternative sub-group analyses. In particular they are often interested in seeing evidence of mechanisms (*e.g.*, does the variable of interest affect a related concept?). By design, those additional analyses can't be incorporated in a PAP and so they are subject to potential criticism of data mining due to what Pepinsky (2013) refers to as *referee degree of freedom*.

Once researchers have been invited to revise and resubmit a paper, they will revise the paper and attempt to find a variable Y that can be considered to proxy for the related concept introduced by the referees and for which they can reject the null. The issue is whether the additional hypothesis that the referee wants the researchers to explore is true (in the sense that it should not be rejected). Let's call ϕ that likelihood and γ the likelihood that the researchers find a Y for which they can reject the null. Under the full sample approach, the probability that an error is made is then $\phi * \gamma$. Under the split sample approach, the probability that an error is made is $0.05 * \gamma$. So as soon as ϕ is larger than 0.05 the split sample approach leads to fewer type I errors than a PAP once the editorial process has been taken into account.

5.3 Helping Editors Decide Whether to Publish Null Results

The split sample protocol proposed in this paper can assist editors decide which results to publish – and in particular whether to publish so-called 'null results'.

There are essentially two types of null results: situations in which the null hypothesis cannot be rejected, but the alternative hypothesis cannot be rejected either; and situations in which the alternative hypothesis can be rejected, in some sense clarified below. The first situation arises when the standard error of the estimated coefficient b_k is large (relative to $\widehat{b}_k$). The second situation arises when the estimated b_k is close to the null hypothesis and its standard error is small. In the first situation, the data are uninformative about the hypothesis of interest and inference is inconclusive. In the second situation, the data are informative about the range of likely values of b_k and inference is conclusive. We discuss how the split sample approach can be relied on to improve publication decisions.

Let us begin with the situation in which the estimated b_k is close to zero and its standard error is small. Formally, let the researcher set a relevant effect size d_k such that, if $|b_k| \leq d_k$, the magnitude of the effect of treatment on the outcome variable is regarded as economically negligible – i.e., a 'true zero'. This hypothesis ($H_0 : |b_k| > d_k$) can be tested by verifying whether the α-sized confidence interval of $\widehat{b}_k$ (*i.e.*, $[\widehat{b}_k \pm z_{1-\alpha/2} * \widehat{se}_k]$) lies entirely within the interval $[-d_k, d_k]$. Indeed, in this case researchers can reject both $H_0^- : b_k = -d_k$ (or any value less than $-d_k$) and $H_0^+ : b_k = d_k$ (or any value above d_k) – and hence can conclude that the true b_k is either zero or close enough to zero to be ignored.

The split sample approach can be applied to this alternative hypothesis as well. The training sample is used to identify the k relevant outcome variables for which $H_0 : |b_k| > d_k$ is rejected. The testing sample is then used to confirm that this inference is not a false positive. The same logic applies as for testing $b_k = 0$, except that here the object of inference is to identify k outcome variables on which treatment has no (economically relevant) effect.

We now discuss what happens if either type of result obtained on the training sample is not confirmed on the testing sample. Should the editor accept the results for

publication? The tendency of editors in economics is to reject papers with results that are not significantly different from zero. The implicit justification is that non-significant results are due to insufficient power – and hence are uninformative. The downside of this approach is publication bias: because studies that find no effect are not published (i.e., end up in the 'file drawer') – or are only published in less well-ranked journals – the profession overweighs studies that document an effect and, so doing, draws biased inference from meta analysis (Ferguson and Brannick, 2012; Franco, Malhotra and Simonovits, 2014; Simonsohn and Nelson, forthcoming).

The split sample approach that we propose protects against publication bias in several ways. First, by only reporting inference results on outcome variables that are significant in the testing sample, the method offers protection against publishing underpowered results: as we have documented earlier, a large sample is required to confirm a significant effect on the testing sample. Second, sufficient power is required to obtain false positives on the training sample. The training sample hurdle thus offers some protection against publishing insignificant results due to lack of power. More importantly, the training sample provides a consistent estimator of the variance of the error term u for each outcome variable k. This allows the researchers to calculate power on the testing sample, and thus enables the editor to judge whether power is sufficient to detect an economic relevant effect on the testing sample. Finally, by committing to publish the testing sample results before observing these results – but after calculating power on the testing sample – the editor adopts a strategy that will lead to high powered *null* results being published while, at the same time, documenting occurences of false positives. In line with Vivalt (2015) and the studies cited earlier, we believe that both of these features will, over time, help improve inference achieved over multiple studies.

6 Conclusion

In this paper we contribute to the nascent literature on ways to increase the likelihood that published findings are true. We investigate the effectiveness of a method that

can be applied to both new and existing datasets. The method relies on a third-party randomly splitting the data in two non-overlapping subsets. Researchers use the first half to refine their research plan, present their findings during seminars and conference and submit them to journals. Once the paper is accepted, the precise research plan is then implemented on the second half and this is the set of results that are published.

We find that for a large number of empirically-relevant settings, the loss in statistical power associated with the split sample approach is manageable and we strongly encourage researchers to adopt the approach. This is especially true for quasi-experiments and observational relying on large datasets. For experiments, If researchers have strong prior that some hypotheses are true, they could set up a PAP for this subset of hypotheses. They could then use the split sample approach to test other, more exploratory hypotheses.

We believe that either journals or a professional association should set up and maintain an online platform where researchers can upload their dataset and have someone carry out the split sample. Importantly, the method can still be implemented by researchers working with proprietary data, e.g., researchers can send their anonymized dataset with garbled variable names to the third party.

References

Angrist, Joshua D. and Alan B. Krueger. 1995. "Split-Sample Instrumental Variables Estimates of the Return to Schooling." *Journal of Business & Economic Statistics* 13(2):225–235.

Athey, Susan and Guido Imbens. 2015. "Machine Learning Methods for Estimating Heterogeneous Causal Effects." *mimeo, Stanford University* .

Bell, Mark and Nicholas Miller. forthcoming. "Questioning the Effect of Nuclear Weapons on Conflict." *Journal of Conflct Resolution* .

Belloni, Alexandre, Victor Chernozhukov and Christian Hansen. 2014. "High-Dimensional Methods and Inference on Structural and Treatment Effects." *Journal of Economic Perspectives* 28(2):29–50.

Benjamin, Daniel J., David Cesarini, Christopher F. Chabris, Edward L. Glaeser, David I. Laibson, Vilmundur Guonason, Tamara B. Harris, Lenore J. Launer, Shaun Purcell, Albert Vernon Smith, Magnus Johannesson, Patrik K. E. Magnusson, Jonathan P. Beauchamp, Nicholas A. Christakis, Craig S. Atwood, Benjamin Hebert, Jeremy Freese, Robert M. Hauser, Taissa S. Hauser, Alexander Grankvist, Christina M. Hultman and Paul Lichtenstein. 2011. "The Promises and Pitfalls of Genoeconomics." *Annual Review of Economics* 4(1):627–662.
URL: *http://dx.doi.org/10.1146/annurev-economics-080511-110939*

Benjamin, Daniel J., David Cesarini, Matthijs J. H. M. van der Loos, Christopher T. Dawes, Philipp D. Koellinger, Patrik K. E. Magnusson, Christopher F. Chabris, Dalton Conley, David Laibson, Magnus Johannesson and Peter M. Visscher. 2012. "The genetic architecture of economic and political preferences." *Proceedings of the National Academy of Sciences* 109(21):8026–8031.
URL: *http://www.pnas.org/content/109/21/8026.abstract*

Benjamini, Yoav, Abba M. Krieger and Daniel Yekutieli. 2006. "Adaptive linear step-up procedures that control the false discovery rate." *Biometrika* 93(3):491–507.
URL: *http://biomet.oxfordjournals.org/content/93/3/491.abstract*

Benjamini, Yoav and Daniel Yekutieli. 2001. "The Control of the False Discovery Rate in Multiple Testing Under Dependency." *The Annals of Statistics* 29:1165–1188.

Benjamini, Yoav and Yosef Hochberg. 1995. "Controlling the False Discovery Rate: A Pactrical and Powerful Approach to Multiple Testing." *Journal of the Royal Statistical Society. Series B (Methodological)* 57(1):289–300.

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

Casey, Katherine, Rachel Glennerster and Edward Miguel. 2012. "Reshaping Institutions: Evidence on Aid Impacts Using a Pre-Analysis Plan." *Quarterly Journal of*

Economics 127(4):1755–1812.
URL: *http://qje.oxfordjournals.org/content/early/2012/09/09/qje.qje027.abstract*

Coffman, Lucas C. and Muriel Niederle. 2015. "Pre-Analysis Plans are not the Solution Replications Might Be." *Journal of Economic Perspectives* 29(3):81–98.

Einav, Liran and Jonathan Levin. 2014. "Economics in the age of big data." *Science* 346(6210).

Ferguson, Christopher and Michael Brannick. 2012. "Publication bias in psychological science: Prevalence, methods for identifying and controlling, and implications for the use of meta-analyses." *Psychological Methods* 17(1).

Franco, Annie, Neil Malhotra and Gabor Simonovits. 2014. "Publication bias in the social sciences: Unlocking the file drawer." *Science* .
URL: *http://www.sciencemag.org/content/early/2014/08/27/science.1255484.abstract*

Gelman, Andrew. 2013. "False memories and statistical analysis.".
URL: *http://andrewgelman.com/2013/09/09/false-memories-and-statistical-analysis/*

Gelman, Andrew. 2014. "Preregistration: what's in it for you?".
URL: *http://andrewgelman.com/2014/03/10/preregistration-whats/*

Gerber, Alan and Neil Malhotra. 2008. "Do statistical reporting standards affect what is published? Publication bias in two leading political science journals." *Quaterly Journal of Political Science* 3:313–326.

Glaeser, E. 2006. "Researcher Incentives and Empirical Methods." *Harvard Institute of Economic Research, Discussion Paper Number 2122* .

Green, Don, Macartan Humphreys and Jenny Smith. 2013. "Read it, understand it, believe it, use it: Principles and proposals for a more credible research publication." *Columbia University, mimeo* .

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):1–20.
URL: *http://pan.oxfordjournals.org/content/21/1/1.abstract*

Ioannidis, John. 2005. "Why Most Published Research Findings Are False." *PLOS Medicine* 2(8).

Laitin, David D. 2013. "Fisheries Management." *Political Analysis* 21:42–47.

Leamer, Edward. 1974. "False Models and Post-Data Model Construction." *Journal of the American Statistical Association* 69(345):pp. 122–131.
URL: *http://www.jstor.org/stable/2285510*

Leamer, Edward. 1978. *Specification Searches. Ad Hoc Inference with Nonexperimental Data*. New York, NY: Wiley.

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

Lovell, M. 1983. "Data Mining." *Review of Economic and Statistics* 65(1):1–12.

McConnell, Brendon and Marcos Vera-Hernández. 2015. "Going beyond simple sample size calculations: a practitioner's guide." *IFS Working Paper W15/17* .

Olken, Benjamin. 2015. "Pre-Analysis Plans in Economics." *Journal of Economic Perspectives* 29(3):61–80.

Pepinsky, Tom. 2013. "The Perilous Peer Review Process.".
URL: *http://tompepinsky.com/2013/09/16/the-perilous-peer-review-process/*

Rauchhaus, Robert. 2009. "Evaluating the Nuclear Peace Hypothesis A Quantitative Approach." *Journal of Conflict Resolution* 53(2):258–277.

Shea, Christopher. 2013. "The Data Vigilante." *The Atlantic* .

Simonsohn, Uri and Leif D. Nelson. forthcoming. "P-Curve: A Key to the File Drawer." *Journal of Experimental Psychology: General* .

Vivalt, Eva. 2015. "How Much Can We Generalize from Impact Evaluations? Are They Worthwhile?" *mimeo, Australian National University* .

Wittes, Janet. 2002. "Sample Size Calculations for Randomized Controlled Trials." *Epidemiologic Reviews* 24(1):39–53.

Table 2: Family Wise Error Rate

m	m_0	Full Sample		Split Sample	
		Bonferonni Corrections:			
		No	Yes	No	Yes
10	9	0.370	0.044	0.022	0.018
100	90	0.990	0.044	0.202	0.016
100	99	0.994	0.048	0.219	0.048
1,000	900	1	0.044	0.895	0.016
1,000	990	1	0.048	0.916	0.041

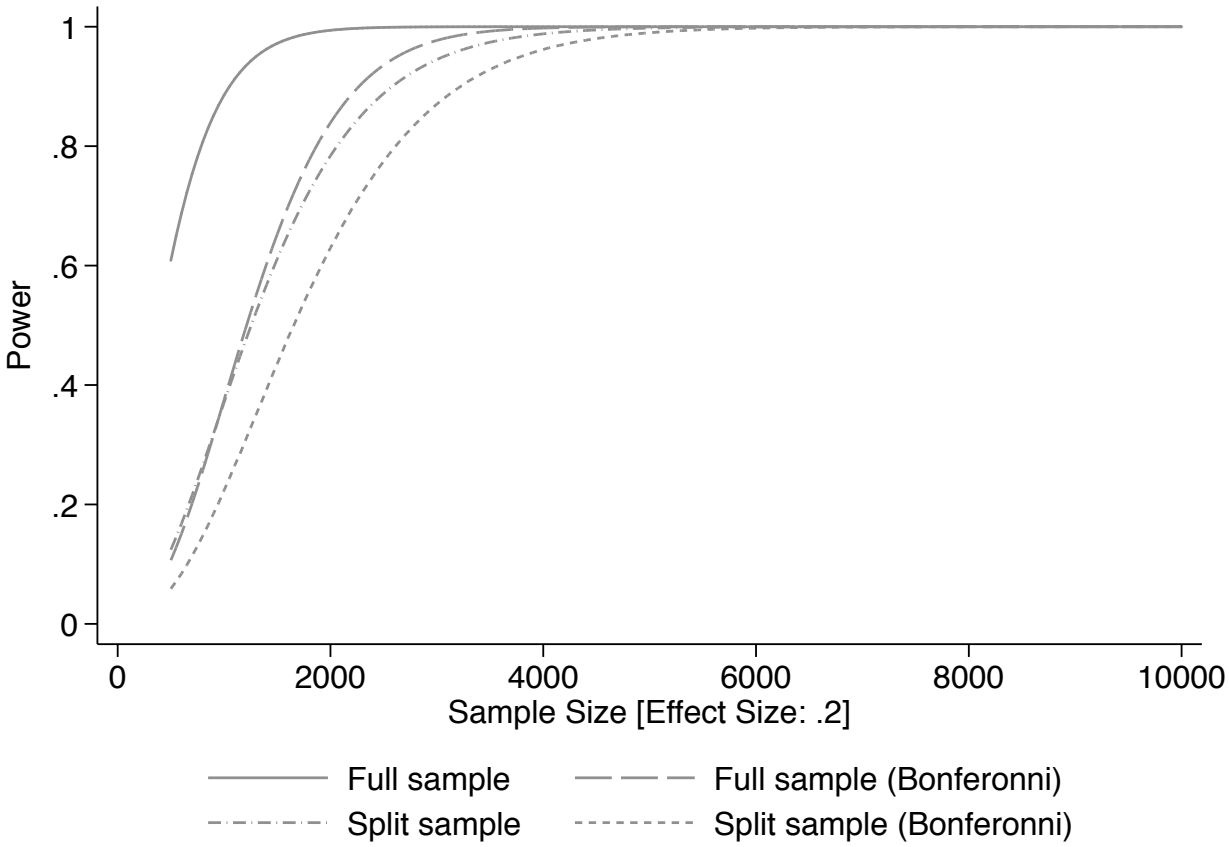


Figure 1: Comparing Power : Full Sample vs. Split Sample [Effect size = .2]

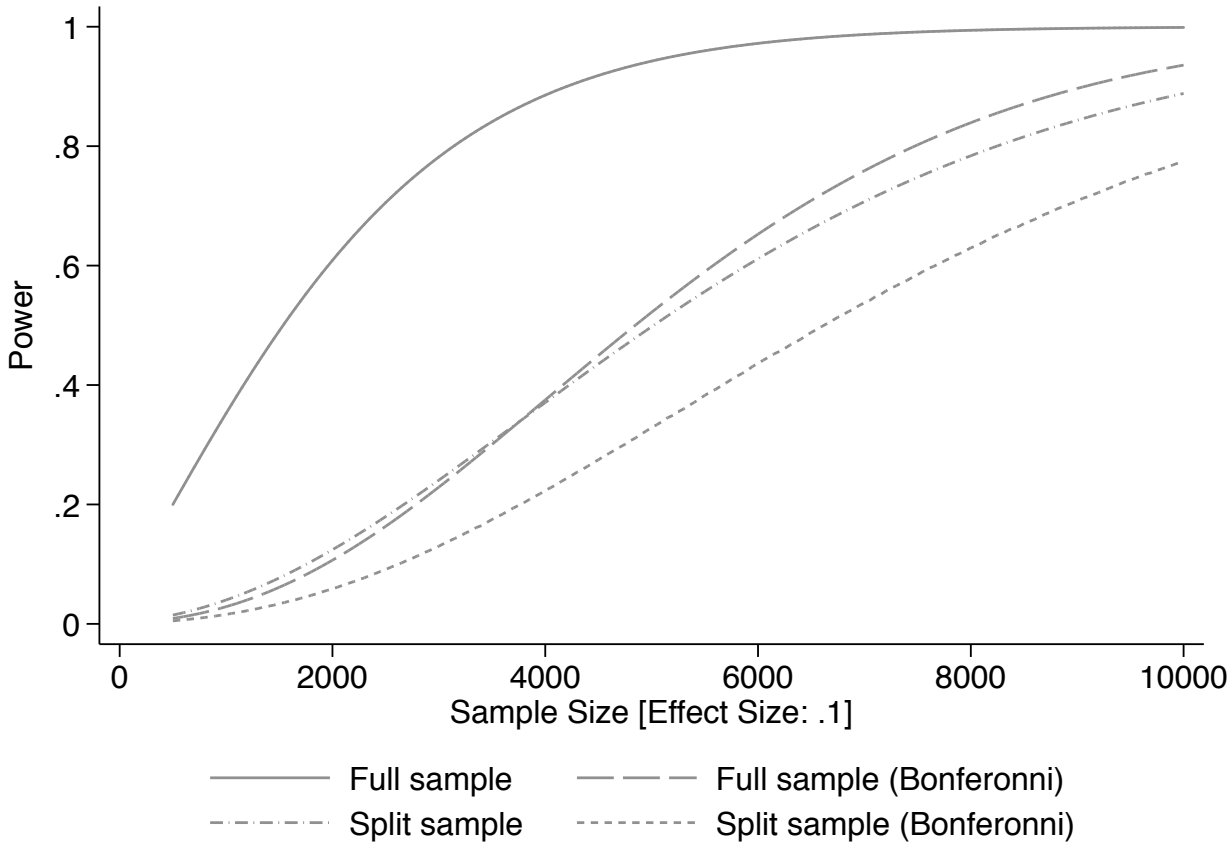


Figure 2: Comparing Power : Full Sample vs. Split Sample [Effect size = .1]

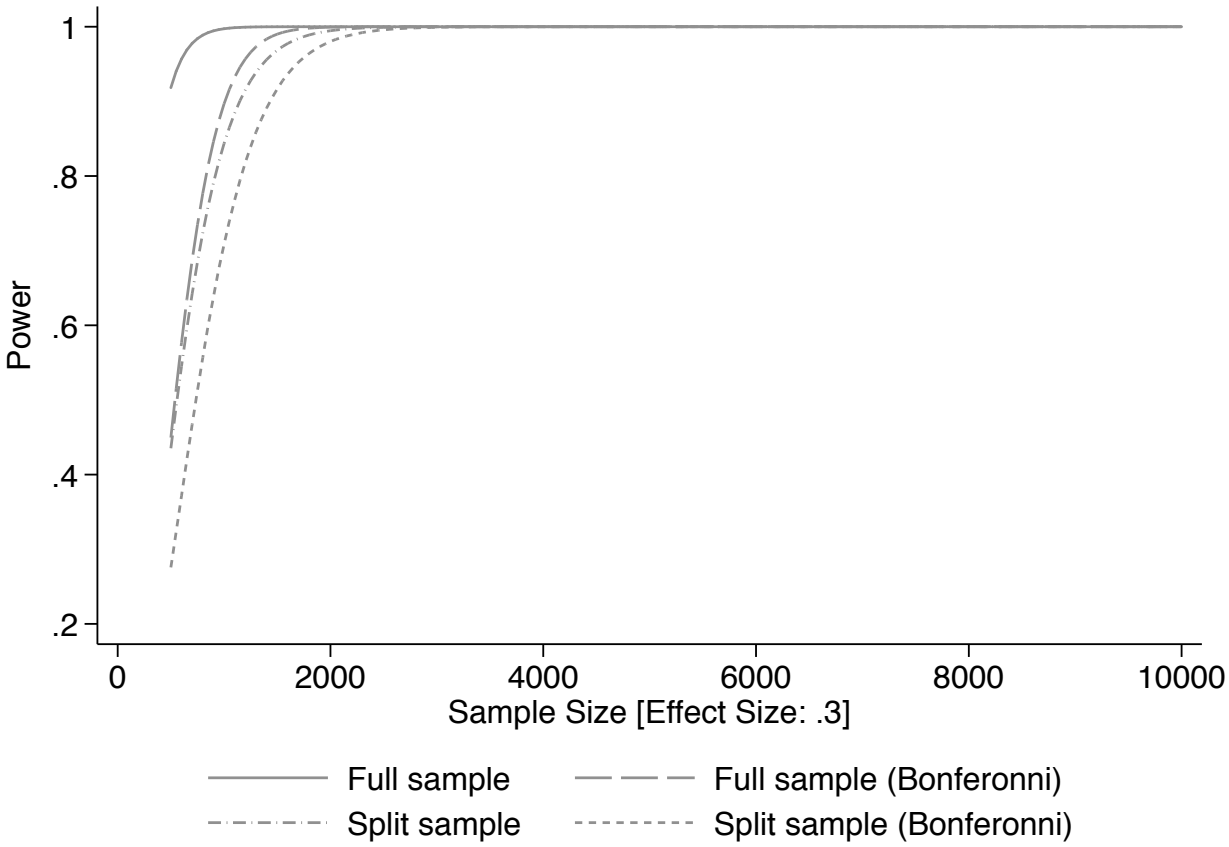


Figure 3: Comparing Power : Full Sample vs. Split Sample [Effect size = .3]

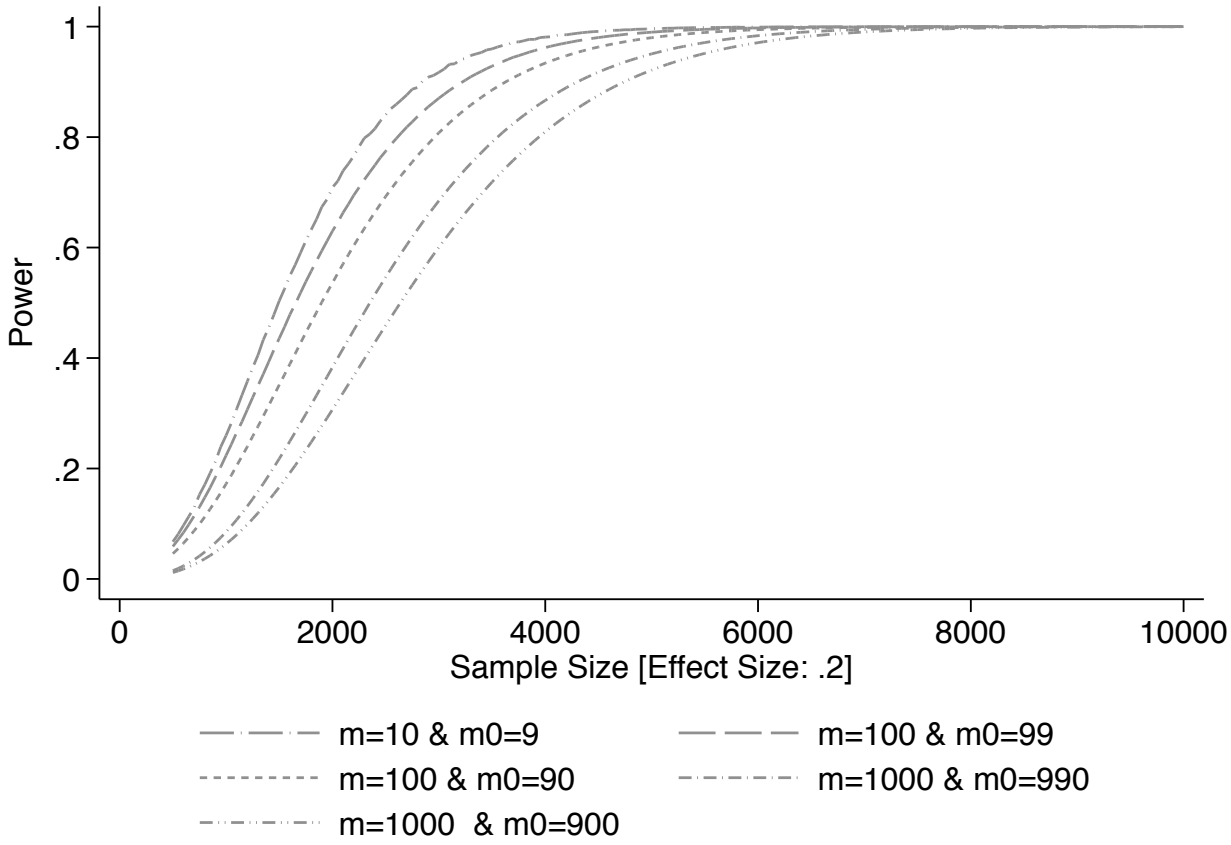
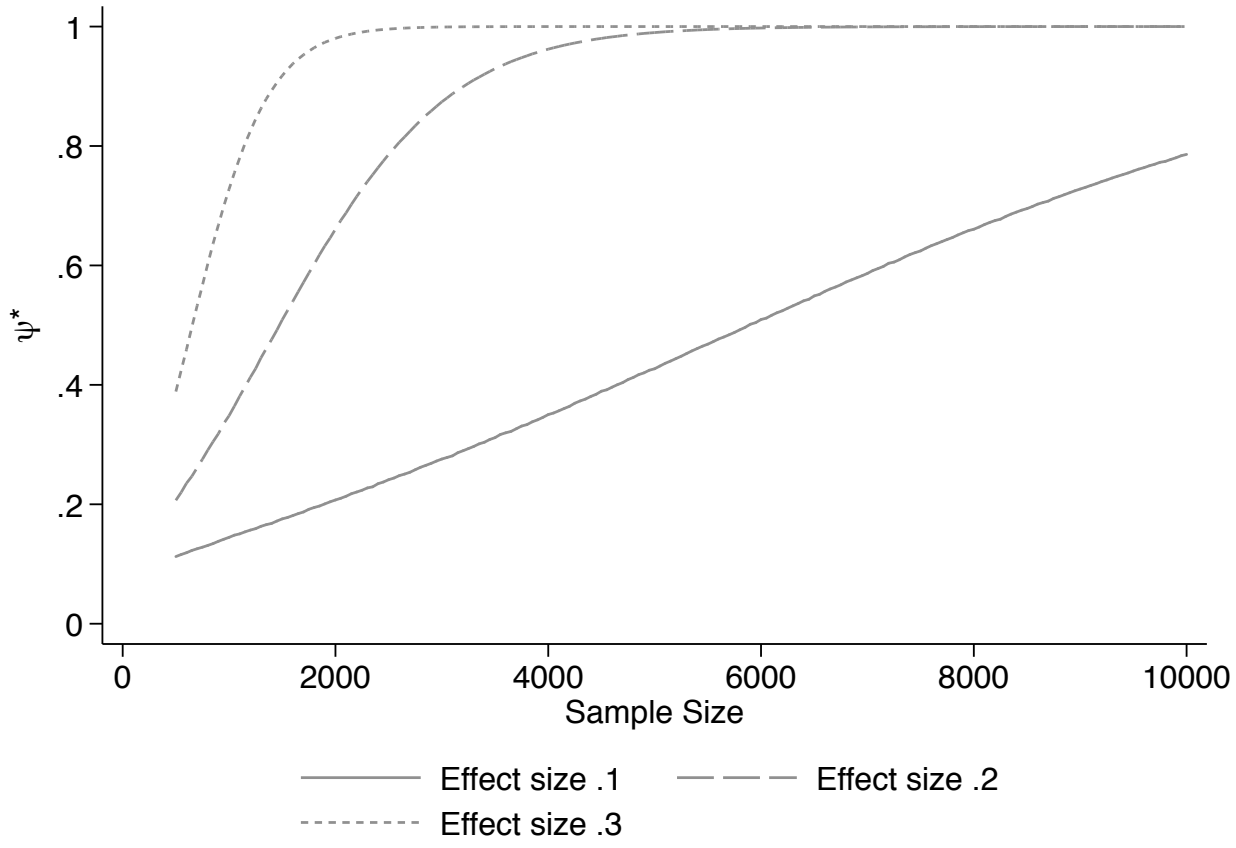


Figure 4: Power Under the Sample Split Approach with Bonferonni Corrections: Number of variables

Figure 5: Value of ψ at which the full sample approach with a PAP and the split sample approach yields the same power.



Note: ψ is the likelihood that the non-true hypothesis is in the set of tests included in the PAP.