# Channelling Fisher: Randomization Tests and the Statistical Insignificance of Seemingly Significant Experimental Results*

Alwyn Young
London School of Economics

This draft: September 2016

## Abstract

I follow R.A. Fisher's The Design of Experiments, using randomization statistical inference to test the null hypothesis of no treatment effect in a comprehensive sample of 53 experimental papers drawn from the journals of the American Economic Association. Randomization tests of the significance of treatment coefficients find that 10 to 20 percent of conventionally significant coefficients are not significant at the same level. In joint tests for equations with multiple treatment measures, 30 to 40 percent of equations with an individually significant coefficient cannot reject the null of no treatment effect. An omnibus randomization test of overall experimental significance that incorporates all of the regressions in each paper finds that only 40 to 50 percent of experimental papers are able to reject the null of no treatment effect anywhere. Bootstrap methods support and confirm these results.

*I am grateful to Alan Manning, Ben Olken, Steve Pischke and Eric Verhoogen for helpful comments, to Ho Veng-Si for numerous conversations, and to the following scholars (and by extension their co-authors) who, displaying the highest standards of academic integrity and openness, generously answered questions about their randomization methods and data files: Lori Beaman, James Berry, Yan Chen, Maurice Doyon, Pascaline Dupas, Hanming Fang, Xavier Giné, Jessica Goldberg, Dean Karlan, Victor Lavy, Sherry Xin Li, Leigh L. Linden, George Loewenstein, Erzo F.P. Luttmer, Karen Macours, Jeremy Magruder, Michel André Maréchal, Susanne Neckerman, Nikos Nikiforakis, Rohini Pande, Michael Keith Price, Jonathan Robinson, Dan-Olof Rooth, Jeremy Tobacman, Christian Vossler, Roberto A. Weber, and Homa Zarghamee.

## I: Introduction

In contemporary economics, randomized experiments are seen as solving the problem of endogeneity, allowing for the identification and estimation of causal effects. Randomization, however, has an additional strength: it allows for the construction of exact test statistics, i.e. test statistics whose distribution does not depend upon asymptotic theorems or distributional assumptions and is known in each and every sample. Randomized experiments rarely make use of such methods, relying instead upon conventional econometrics and its asymptotic theorems. In this paper I apply randomization tests to randomized experiments, using them to construct counterparts to conventional tests of significance within regressions and, more ambitiously, an exact omnibus test of overall significance that combines all of the regressions in a paper in a manner that is, practically speaking, infeasible in conventional econometrics. At the coefficient level, randomization tests reduce the number of significant coefficients by 10 to 20 percent. Joint tests of statistical significance in multi-treatment equations reduce the number of regression specifications with statistically significant treatment effects by 30 to 40 percent, while the omnibus test finds that, when all treatment outcome equations are combined, only 40 to 50 percent of papers can reject the null of no treatment effect. These results relate, purely, to statistical inference, as I do not modify published regressions in any way. I confirm them with bootstrap statistical inference, and use bootstrap sampling to establish the size biases of conventional methods and the fact that the power of randomization tests is virtually identical to that of conventional methods when these are exact and the assumptions on the independence of residuals are similar.

Two factors lie behind the discrepancy between the results reported in journals and those produced in this paper. First, published papers fail to consider the multiplicity of tests implicit in the many treatment coefficients within regressions and the many regressions presented in each paper. About half of the regressions presented in experimental papers contain multiple treatment regressors, representing indicators for different treatment regimes or interactions of treatment with participant characteristics. When these regressions contain a .01 level significant coefficient, there are on average 5.9 treatment measures, of which only 1.7 are significant. I find treatment measures within regressions are generally mutually orthogonal, so the finding of a significant coefficient in a regression should be viewed as the outcome of multiple independent rolls of 20-sided or 100-sided dice. However, only 31 of 1009 regressions with multiple treatment measures report a conventional F- or Wald-test of the joint significance of all treatment variables within the regression.[1]

While treatment coefficients within regressions are largely orthogonal, treatment coefficients across regressions, particularly significant regressions, are highly correlated. The typical paper reports 10 regressions with a treatment coefficient that is significant at the .01 level, and 27 regressions with no treatment coefficient that is significant at this level.[2] I find that the randomized and bootstrapped distribution of the coefficients and p-values of significant regressions are highly correlated across equations, while the insignificant regressions are much more independent. Thus, the typical paper presents many independent tests that show no treatment effect and a small set of correlated tests that show a treatment effect. When combined, this information suggests that most experiments have no significant effects. I should note that this result is unchanged when I restrict attention only to regressions with dependent variables that produce a significant treatment coefficient in at least one regression. Thus, it is not a consequence of combining the results of regressions of variables that are never significantly correlated with treatment with those concerning variables that are consistently correlated with treatment. Dependent variables that are found to be significantly related to treatment in a subset of highly correlated specifications are not significantly related to treatment in many other, statistically independent, specifications.

The second factor explaining the lower significance levels found in this paper is the fact that published papers make heavy use of statistical techniques that rely upon asymptotic theorems that are largely invalidated and rendered systematically biased in favour of rejection by their regression design. Chief amongst these methods are the robust and clustered estimates of variance, which are designed to deal with unspecified heteroskedasticity and correlation across

[1] These occur in two papers. In an additional 7 regressions in two other papers the authors make an attempt to test the joint significance of multiple treatment measures, but accidentally leave out some treatment measures. In another paper the authors test whether a linear combination of all treatment effects in 28 regressions equals zero, which is not a test of the null of no treatment effect, but is closer. F-tests of the equality of treatment effects across treatment regimes (excluding control) or in non-outcome regressions (e.g. tests of randomization balance) are more common.

[2] Naturally, I only include treatment outcome regressions in these calculations and exclude regressions related to randomization balance (participant characteristics) or attrition, which, by demonstrating the orthogonality of treatment with these measures, confirm the internal validity of the randomized experiment.

observations. The theorems that underlie these and other asymptotic methods depend upon maximal leverage in the regression going to zero, but in the typical regression design it is actually much closer to its upper limit of 1. High leverage allows for a greater spread in the bias of covariance estimates and an increase in their variance, producing an unaccounted for thickening of the tails of test distributions, which leads to rejection rates greater than nominal size. The failure and potential bias of asymptotic methods is, perhaps, most immediately recognized by noting that no less than one fifth of the equation-level coefficient covariance matrices in my sample are singular, implying that their covariance estimate of some linear combination of coefficients is zero, i.e. a downward bias of 100 percent. Using the bootstrap I show that the conventional test statistics of my experimental papers, when corrected for the actual thickness of the tails of their distributions, produce significant results at rates that are close to those of randomization tests.

Conventional econometrics, in effect, cannot meet the demands placed on it by the regressions of published papers. Maximal leverage is high in the typical paper because the authors condition on a number of participant observables, either to improve the precision with which treatment effects are estimated or convince sceptical referees and readers that their results are robust. These efforts, however, undermine the asymptotic theorems the authors rely on, producing test statistics that are biased in favour of rejecting the null hypothesis of no treatment effect when it is true. Randomization inference, however, remains exact regardless of the regression specification. Moreover, randomization inference allows the construction of omnibus Wald tests that easily combine all of the equations and coefficient estimates in a paper. In finite samples such tests are a bridge too far for conventional econometrics, producing hopelessly singular covariance estimates and biased test statistics when they are attempted. Thus, randomization inference plays a key role in establishing the validity of both themes in this paper, the bias of conventional methods and the importance of aggregating the multiplicity of tests implicitly presented in papers.

In a paper of this sort it is important to follow transparent rules rather than opaque discretion. To this end, I set up a set of criteria for inclusion in the sample (presented later) and test all treatment coefficients that can be analysed with randomization inference. Authors might argue that this dilutes power, mixing in trivial details with key treatment procedures in joint tests of significance. To address this, I use additional rules to implement alternative procedures. I separately analyse regressions where treatment measures divide the sample into mutually exclusive groups and where such regressions account for at least 1/3 of reported regressions. Because of the costs of implementation and their prominence in presentation, it is hard to argue that treatment measures that are applied to mutually exclusive groups were of no importance to the authors. I find, however, that the reduction in significance brought about by joint testing is similar to that found in the full sample. Authors might argue that only certain outcome variables were of interest to them. I address this by restricting the sample in the omnibus test to dependent variables that are associated with significant treatment effects somewhere in the paper, with little change in overall results. I supplement joint testing procedures, which maximize power for a diffuse alternative, with multiple testing procedures, which seek to maximize power on the axes (i.e. allowing some treatments to have effects and others not). The results, in terms of the number of equations and papers where significant treatment effects are found, are again quite similar. The typical experimental paper honestly and forthrightly reports a vast number of tests, within equations and across equations, most of which yield no individually significant effects. Any systematic procedure that accounts for all of this multiple testing must, inevitably, discount some of the individually significant results and conclude that there is much less evidence in favour of significant treatment effects than is nominally presented.

The bootstrap plays an important supporting role in this paper. Randomization tests are exact because they are based upon Fisherian thought experiments regarding experimental outcomes for a fixed experimental sample. Readers raised on Neyman's population sampling approach to statistical inference might find more credibility in the population resampling of the bootstrap. The bootstrap not only confirms the randomization results on statistical significance, but also allows, through its population resampling, an exploration of size and power. Bootstrap samples drawn from the experimental population itself provide a data generating process that mimics the characteristics of the experimental data, i.e. any heterogeneity in treatment effects or heteroskedasticity and correlation in errors. When test statistics are centered on the experimental population mean, these samples allow an analysis of size, and show that all of the covariance estimation techniques used by authors are on average strongly biased in favour of rejecting the null. When these bootstrap samples are used to test the null of no effects, they illustrate power when the alternative of average treatment effects in the amounts actually found in the experiments is true. In baseline simulations, the conventional methods used by authors appear more powerful than randomization tests for two reasons. First, conventional techniques are biased in favour of rejection, a weakness when the null is true, but a positive feature when it is false. Second, many authors cluster at a level below treatment groupings, maintaining that there is no cross-correlation in errors amongst individuals living together in regions or participating jointly in laboratory sessions. When the clustering techniques used by authors are adjusted to treatment levels, or randomization inference is adjusted to incorporate the independence assumptions made by authors, the two techniques have virtually identical power. Such adjustments, however, do little to change the significance rates in the analysis of the papers themselves. In sum, the bootstrap shows that size, and not power, explains the difference between conventional and randomization results.

The power of randomization tests is the subject of some confusion in casual discourse. First, they have a general reputation for low power, owing to the fact that they are often used in non-parametric tests, such as the Kolmogorov-Smirnov test of equality of distributions, which impose little structure on the problem. Second, the fact that their theoretical motivation is based upon testing sharp hypotheses, e.g. the null that treatment has zero effects for each and every participant, while the typical econometric null is one of zero, but potentially heterogeneous, average treatment effects, leads to the belief that either (a) as the sharp hypothesis is more restrictive, power must be even greater than it would be if one were testing for average treatment effects; (b) the sharp hypothesis is so restrictive it tells us nothing about average treatment effects. This discourse confuses motivations of size with determinants of power. Much of applied econometrics consists of calculating conditional means. When I implement randomization tests in this paper, I recalculate these conditional means, again and again, for different potential distributions of treatment. While the sharp null motivates the test, in its implementation it amounts to no more, nor less, than calculating how conditional averages vary across potential distributions of treatment. There is no intuition as to why the power of this test should be any different than that of conventional conditional mean calculations, and in fact it is not. As already noted, in the context of bootstrap samples that mimic the data generating process, with all its heterogeneity, in the experiments themselves, I find that randomization and conventional techniques have virtually identical power when the latter are exact and both methods make the same assumptions about the independence of residuals.

Notwithstanding its results, this paper confirms the value of randomized experiments. The methods used by authors of experimental papers are standard in the profession and present throughout its journals. Randomized statistical inference provides a solution to the problems identified in this paper, avoiding a dependence on asymptotic theorems that produce inaccurate and biased finite sample statistical inference and allowing the simple calculation of omnibus tests that incorporate all of the regressions and tests run in an analysis. While, to date, it rarely appears in experimental papers, which generally rely upon traditional econometric methods,[3] it can easily be incorporated into their analysis. As proven by Lehmann (1959), only a permutation test, and none other, can provide a finite sample exact test of a mean difference between two populations that does not depend upon knowledge of the characteristics of the disturbances.[4] Thus, randomized experiments are ideally placed to solve both the problem of identification and the problem of accurate statistical inference, making them doubly reliable as an investigative tool.

The paper proceeds as follows: Section II explains that the 53 paper sample is as comprehensive and non-discriminatory as possible, using virtually every paper published in the American Economic Review, American Economic Journal: Applied Economics and American Economic Journal: Microeconomics revealed by a search on the American Economic Association (AEA) website that satisfies a set of criteria derived from the needs and objectives of the analysis (i.e. public use data, do-files, data on participant characteristics that are used to condition regressions, and regressions that use conventional statistical inference but can be analysed using

[3] Of the 54 experimental papers that otherwise meet the criteria for inclusion in my sample (discussed below), only one uses randomization statistical inference throughout (and hence is not included in the final sample), while one other uses randomization inference to analyse results in some regressions and one more indicates that they confirmed the significance of results with randomization tests. Wilcoxon rank sum tests are reported in four other papers. These non-parametric tests are not precisely randomization tests, although Stata describes them as having randomization based Fisher exact distributions. To calculate the distribution of a test statistic based upon randomization inference, one must replicate the randomization process. Stata's Wilcoxon test reshuffles treatment at the observation level, but these tests are used in papers which applied treatment in groups or stratified treatment. Hence, the distributions used to evaluate the test statistic are not the distributions that could have been produced under the randomization null (unless one wishes to add additional assumptions about the unimportance of the treatment groupings).

[4] Naturally, this is only for a sharp mean difference, i.e. one where the values for each member of the population are specified. For average differences with unspecified heterogeneity, the randomization test is not exact, but then again, neither is any other method, as the error terms become heteroskedastic in an unknown fashion.

randomization techniques). About 70 percent of the 1954 regressions are ordinary least squares (OLS)[5] and about 70 percent use the clustered or robust estimate of covariance.

Section III provides a thumbnail review of the theory that underlies later empirical results. I show that an asymptotic maximum leverage of zero plays a role in many asymptotic theorems and that much of the sample is far from this ideal, with an average maximum leverage of .491 and .616 in robust and clustered OLS regressions, respectively. I argue that maximal leverage determines the bias and variance of the variance estimates of the robust and clustered covariance matrices. The theory underlying randomization and bootstrap statistical inference is reviewed and several alternative measures, with different theoretical properties, are presented. I note particular problems with the way authors implement the bootstrap, using non-pivotal statistics that depend upon a variance estimate whose sampling variation is not properly accounted for, producing spuriously high rejection rates.

Section IV presents the main empirical results. I begin by reviewing the results on statistical significance in joint and multiple tests and the within and across equation coefficient correlation discussed above. I then use bootstrap samples from the experiments themselves to explore the size and power properties of conventional and randomization tests. I show that the size distortions of conventional tests, with an average rejection rate of .02 at the .01 level, but ranging as high as .886 for particular coefficients, are determined by the bias and variance of the variance estimate which, in the case of the robust and clustered covariance matrices, are related to maximal leverage. For the type of average treatment effects present in the experimental population, the average power of conventional techniques in coefficient tests at the .01 level is .193, while that of different randomization tests is .156 and .168. This is principally because many authors cluster at below treatment levels. When conventional tests are adjusted to cluster at treatment levels, or randomization inference makes use of the independence assumptions authors make, the power differences between the two methods largely vanish. Adjusting for the size bias of authors' methods, one randomization method actually has as much as .052 ln proportional greater power at the .01 level than conventional methods, whereas without such adjustment its ln proportional power is still only -.025 less than biased conventional methods. Motivated by these

[5] Throughout the paper I use the term regression broadly, allowing it to denote any statistical procedure that yields coefficient and standard error estimates.

investigations, I apply similar adjustments to the analysis of the papers themselves, and find they have little effect on the relative number of significant results found in conventional and randomization inference, confirming that power is not the main issue. Section V concludes.

This paper follows R.A. Fisher, who in The Design of Experiments (1935) introduced the dual concepts of randomization tests and null hypotheses, arguing that permutations of possible treatments provided a "reasoned basis" of testing the null hypothesis of no effect without resort to distributional assumptions such as normality. Fisher's argument can be brought 80 years up to date simply by noting that it avoids dependence on asymptotic theorems as well. Randomized allocation of treatment has played a role in medical trials and social research for decades,[6] but the growth of randomized experiments in economics in recent years is largely due to Kremer and Miguel (2004), whose seminal field experiment sparked an enormous literature in development and other areas of economics. Duflo, Glennerster and Kremer (2008) provide a useful overview of methods. The growing dominance of randomized experiments in development research has inevitably led to a debate about its merits, with, as examples, Deaton (2010) providing a thought-provoking critique arguing that randomized experiments face conventional econometric problems and Imbens (2010) making the case for the importance of identification and the accuracy of randomization inference. This paper affirms both viewpoints, showing just how seriously biases in conventional econometric methods can undermine inference in randomized experiments, while arguing that randomization inference, available only to these papers, provides a natural solution to such problems.

The tendency of White's (1980) robust covariance matrix to underestimate the sampling variance of coefficients and produce rejection rates higher than nominal size was quickly recognized by MacKinnon and White (1985). The natural extension of White's single observation method to correlated group data, the clustered covariance matrix, has also been found to produce excessively high rejection rates in simulations by Bertrand, Duflo and Mullainathan (2004) and Donald and Lang (2007). The bootstrap samples of this paper affirm these results in a broad practical setting. Chesher and Jewitt (1987) identified the link between maximum leverage and bias bounds for robust covariance matrices, while Chesher (1989) extended the analysis by

[6] For a description of some of the early social experiments, and the problems they faced, see Burtless (1995) and Heckman and Smith (1995).

showing how maximal leverage determines bounds on the variance of these matrices and, hence, the thickness of the tails of the test statistic distributions. In this paper I provide systematic evidence that leverage, and not sample size, determines the bias and variance of variance estimates, and that these in turn determine the size bias of conventional t-tests. Maximum leverage, and not sample size, is the best indicator of potential problems in the use of the robust and clustered covariance matrices.

The addition of multiple treatment measures and interactions to estimating equations is a form of specification search. The need to find some way to evaluate, in its entirety, the information generated by specification searches was first raised by Leamer (1978), who addressed the problem using Bayesian methods. This paper follows Leamer in recognizing that specification search is in many respects a natural part of scientific inquiry and should be neither condemned nor ignored completely, but instead incorporated in some fashion into our evaluation of evidence. I use joint and multiple testing procedures of treatment coefficients to combine the information implicit in multiple tests within equations and across equations. The omnibus test, in particular, combines all the treatment information in all of the regressions run in a paper. In this, the integrity of authors in the presentation of the many specifications they ran allows, through the explicit consideration of the covariance of all the coefficients and equations, a null hypothesis test that fully incorporates all of the information generated by specification search.

All of the results of this research are anonymized. Thus, no information can be provided, in the paper, public use files or private discussion, regarding the significance or insignificance of the results of particular papers. The public use data files of the AEA provide the starting point for many potential studies of professional methods, but they are often incomplete as authors cannot fully anticipate the needs of potential users. Hence, studies of this sort must rely upon the openness and cooperation of current and future authors. For the sake of transparency, I provide code and notes (in preparation) that show how each paper was analysed, but the reader eager to know how a particular paper fared will have to execute this code themselves. Public use data files (in preparation) provide the results and principal characteristics of each regression in an anonymized fashion, allowing researchers to reproduce the tables in this paper and use the randomization and bootstrap data in further analysis.

## II. The Sample

My sample is based upon a search on www.aeaweb.org using the keywords "random" and "experiment" restricted to the American Economic Review, American Economic Journal: Applied Economics and American Economic Journal: Microeconomics which, at the time of its last implementation, yielded papers up through the March 2014 issue of the AER. I then dropped papers that:

(a) did not provide public use data files and Stata do-file code[7];
(b) were not randomized experiments;
(c) did not have data on participant characteristics;
(d) already used randomization inference throughout;
(e) had no regressions that could be analyzed using randomization inference.

Public use data files are necessary to perform any analysis, and I had prior experience with Stata and hence could interpret do-files for this programme at relatively low cost. Stata appears to be by far the most popular regression programme in this literature.

My definition of a randomized experiment excluded natural experiments (e.g. based upon an administrative legal change), but included laboratory experiments (i.e. experiments taking place in universities or research centres or recruiting their subjects from such populations).[8] The sessional treatment of laboratory experiments is not generally explicitly randomized, but when queried laboratory experimenters indicated that they believed treatment was implicitly randomized through the random arrival of participants to different sessions. I noted that field experiment terminology has gradually crept into laboratory experiments, with a recent paper using the phrase "random-assignment" no less than 10 times to describe the random arrival of students to different sessions, and hence decided to include all laboratory experiments that met the other criteria.[9] Laboratory experiments account for 15 of the 53 papers but only 193 of the 1954 regressions.

[7] Conditional on a Stata do-file, a non-Stata format data file (e.g. in a spreadsheet or text file) was accepted.

[8] A grey area is experiments that take place in field “laboratories”. If the experimental population is recruited from universities, I term these lab experiments (two papers), despite their location off campus.

[9] A couple of lab papers tried to randomize explicitly, by assigning students to sessions, but found that they had to adjust assignment based upon the wishes of participants. Thus, these papers are effectively randomizing implicitly based upon students’ selection of sessions, and I treat them as such in my analysis.

The requirement that the experiment contain data on participant characteristics was designed to filter out a sample that would use mainstream multivariate regression techniques with estimated coefficients and standard errors. This removed a number of laboratory experiments, which tend to not have data on participant characteristics, use atypical econometric methods and whose passive randomization, if they dominated the sample, might raise concerns. Conditional on a paper having public use data on participant characteristics, however, I included all regressions in a non-discriminatory fashion, including uncommon methods such as t-tests with unequal variances and tests of differences of proportions, as long as they produce a coefficient/parameter estimate and standard error. Subject to the other criteria, only one paper used randomization inference throughout, and was dropped. One other paper used randomization inference for some of its regressions, and this paper and its non-randomization regressions were retained in the sample.

Not every regression presented in papers based on randomized experiments can be analyzed using randomization inference. For randomization inference to be possible the regression must contain a common outcome observed under different treatment conditions. This is often not the case. If participants are randomly given different roles and the potential action sets differ for the two roles (e.g. in the dictator-recipient game), then there is no common outcome between the two groups that can be examined. In other cases, participants under different treatment regimes do have common outcomes, but authors do not evaluate these in a combined regression. Consider for example an experiment with two treatments, denoted by T equal to 0 or 1, and the participant characteristic "age". Under the null of no treatment effect, the regression

$$(1)\ y = \alpha + \beta_T T + \beta_{age} age + \beta_{T*age} T*age + \varepsilon$$

can be analysed by re-randomizing treatment T across participants, repeatedly estimating the coefficients $\beta_T$ and $\beta_{T*age}$, and comparing their distribution to the experimentally estimated coefficients. In many cases, however, authors present this regression as a paired set of "side-by-side" regressions of the form $y = \alpha + \beta_{age} age + \varepsilon$ for the two treatment regimes. These regressions are compared and discussed, but there is no formal statistical procedure given for testing the significance of coefficient differences across regressions. Within each regression there is no coefficient associated with treatment, and hence no way to implement randomization inference. One could, of course, develop appropriate conventional and randomization tests by stacking the regressions into the form given by (1), but this implicitly involves an interpretation of the authors' intent in presenting the side-by-side regressions, which could lead to disputes.[10] I make it a point to always, without exception, adhere to the precise regression presented in tables.

Within papers, regressions were selected if, following (e) above, they allow for randomization inference and:

> (f) appear in a table and either involve a coefficient estimate and standard error or a p-value;
> (g) pertain to treatment effects and not to an analysis of randomization balance, sample attrition, non-experimental cohorts, or first-stage regressions that do not involve treatment outcomes analysed elsewhere in the paper;

while tests were done on the null that:

> (h) randomized treatment has no effect, but participant characteristics or other non-randomized treatment conditions might have an influence.

In many tables means are presented, without standard errors or p-values, i.e. without any attempt at statistical inference. I do not consider these regressions. Alternative specifications for regressions presented in tables are often discussed in surrounding text, but catching all such references, and ensuring that I interpret the specification correctly is extremely difficult (see the discussion of do-file inaccuracies below). Consequently, I limited myself to specifications presented in tables. If coefficients appear across multiple columns, but pertain to a single statistical procedure, they are treated as one regression. Papers often include tables devoted to an analysis of randomization balance or sample attrition, with the intent of showing that treatment was uncorrelated with either. I do not include any of these in my analysis. This is of course particularly relevant to the omnibus test of overall experimental significance. To include regressions specifically designed to show that randomization successfully led to orthogonality between treatment and participant characteristics and attrition in the omnibus test of experimental significance would be decidedly inappropriate. Similarly, I drop regressions projecting the

[10] Stacking the regressions often raises additional issues. For example, there might be more clusters than regressors in each equation, but fewer clusters than regressors in the combined equation. Individually, the covariance matrix of each side-by-side regression is non-singular, but if one stacks the regressions one ends up with a highly singular covariance matrix. This issue (i.e. more regressors than clusters) is present in many papers which use the clustered covariance matrix. One could argue that it implicitly exists in this side-by-side example as well, but only if one assumes that the stacked regression was the authors' actual intent.

behaviour of non-treatment cohorts on treatment measures, which are typically used by authors to, again, reinforce the internal validity of the experiment. In difference in difference equations, I only test the treatment coefficients associated with differences during the treatment period. I also drop 14 first-stage regressions (in iv presentations) that relate to dependent variables that are *not* analysed as treatment outcomes elsewhere in the paper. As discussed in the next section, these pertain to cases, such as take-up of an offered opportunity, where the influence of treatment cannot, by construction, be in doubt (e.g. one cannot take up an opportunity unless one is offered the chance to do so).

I, universally, test the null of no randomized treatment effect, while allowing non-randomized elements to influence behaviour. For example, a paper might contain a regression of the form

$$(2)\quad y = \alpha + \beta_T T + \beta_{T_0 age} T_0 * age + \beta_{T_1 age} T_1 * age + \varepsilon$$

where T is a 0/1 measure of treatment and $T_0$ and $T_1$ are dummies for the different treatment regimes. The null of no treatment effect is given by re-expressing the regression as (1) earlier above and testing $\beta_T = \beta_{T*age}$=0, while allowing α and $\beta_{age}$ to take on any value.[11] In more complicated situations the paper might contain randomized overall treatments (e.g. the environmental information provided to participants) combined with other experimental conditions which were not randomized (e.g. whether the participant is offered a convex or linear payoff in each round). As long as the action space is the same under the different randomized treatments, I am able to test the null of no randomized treatment effect by re-randomizing this aspect across participants, while keeping the non-randomized elements constant.[12] Such cases are quite rare, however, appearing in only two or three papers. In most cases all experimental terms appearing in the regression were clearly randomized and all remaining regressors are clear non-experimental participant characteristics.

Having established (a)-(h) as my explicit sample selection guidelines, to avoid any implicit (and unknown) sample selection I did not allow myself the luxury of dropping papers or

[11] In these cases I am "changing" the regression specification, but the change is nominal. I must also confess that in the case of one paper the set of treatments and coefficients was so restrictive that I could not see what the null of no treatment effect was (or if it was even allowed), and so dropped that paper from my analysis.

[12] Thus, in the example just given, I test the null that the informational conditions had no effect, while allowing the payment scheme to have an effect.

regressions as it suited me. This led to uneven levels of effort across papers. The randomization, bootstrap and size analysis for some papers could be performed in less than an hour; for others, because of sample sizes and procedures, it took more than a year of dedicated workstation computing power. The do-files for many papers are remarkably clear and produce, exactly, the regressions reported in the papers. Other do-files produce regressions that are utterly different from those reported in the published paper, while yet others involve extraordinarily convoluted code (aimlessly loading, formatting, dropping, reloading and reformatting data again and again) that could never be implemented 10000 times (in randomization). In between, there are gradations of error and complexity. Rather than allowing myself to choose which papers were "too hard" to work through, I adopted the procedure of using the do-files, good or bad, as a guideline to developing shortened code and data files that would produce, almost exactly,[13] the regressions and standard errors reported in the tables of the paper. There are only a handful of regressions, across three papers, that I could not reproduce and include in my sample.[14]

Regressions as they appear in the published tables of journals in many cases do not follow the explanations in the papers. To give a few examples:

(a) a table indicates date fixed effects or location fixed effects were added to the regression, when what is actually added is the numerical code for the date or location.

(b) regressions are stacked, but not all independent variables are duplicated in the stacked regression.

(c) clustering is done on variables other than those mentioned, these variables changing from table to table.

(d) unmentioned treatment and non-treatment variables are added or removed between columns of a table.

(e) cluster fixed effects are added in a regression where aspects of treatment are applied at the cluster level, so those treatment coefficients are identified by two observations which miscoded treatment for a cluster (I drop those treatment measures from the analysis).

[13] That is, differing at most in rounding error on some coefficients or standard errors or in the value of only one isolated coefficient or another. Often, in my examination, I found that coefficients had been mistakenly placed in incorrect columns or rows. I also found that authors that took the AEA's instructions to provide code that produced tables too literally, i.e. by having the do-file try to extract the coefficients and put them in a table, generated the greatest number of errors. Code is generally much more accurate when it simply produces a screen output that the user can interpret.

[14] One additional paper had only one treatment regression, which I could not come anywhere near reproducing. It is dropped from my sample.

In addition, as noted earlier, covariance matrices are very often singular, and in many cases Stata notes this explicitly, either by telling the user that the estimation procedure did not converge or that the covariance matrix is remarkably singular. Initiating a dialogue with authors about these issues, as well as the many cases where the do-file code does not produce the regressions in the paper, would have generated needless conflict, created a moving specification target, and added yet more time to the three years spent in preparing the estimates of this paper. The programming errors inflicted on authors by their research assistants are enough to drive a perfectionist to distraction, but have no relevance for this paper, which concerns itself with statistical inference and not the appropriateness of regression specifications. I mention the above examples to forestall criticism that the regressions I analyse are not those described in the papers. This paper analyses statistical inference in regressions as they appear in tables in the journals of the profession, recognizing that in some cases these regressions may not reflect the intent of the authors.

To permute the randomization outcomes of a paper, one needs information on stratification (if any was used) and the code and methods that produced complicated treatment measures distributed across different data files. Stratification variables are often not given in public use files nor adequately or, upon careful examination, correctly described in the paper. Code producing treatment measures is often unavailable, and it is often impossible to link data files, as the same sampling units are referenced with different codes or without codes at all. I have called on a large number of authors who have generously answered questions and provided code and data files to identify randomization strata, create treatment measures and link data files. Knowing no more than that I was working on a paper on experiments, these authors have displayed an extraordinary degree of scientific openness and integrity. Only two papers, and an additional segment from another paper, were dropped from my sample because authors could not provide the information on randomization strata and units necessary to re-randomize treatment outcomes.

Table I below summarizes the characteristics of my final sample, after reduction based upon the criteria described above. I examine 53 papers, 15 of which are laboratory experiments and 38 of which are field experiments. A common characteristic of laboratory experiments, which recruit their subjects from a narrow academic population, is that treatment is almost always

Table I: Characteristics of the Sample

| 53 papers | | 1954 regressions | |
|---|---|---|---|
| location | journal | Type | covariance |
| 38 field | 29 AER | 1371 ordinary least squares | 447 default |
| 15 lab | 20 AEJ: Applied Economics | 322 maximum likelihood | 979 clustered |
| | 4 AEJ: Microeconomics | 67 generalized least squares | 311 robust |
| | | 194 other | 126 bootstrap |
| | | | 91 other |

Notes: AER = American Economic Review; AEJ = American Economic Journal.

administered at the sessional level and implicitly randomized, as noted earlier, through the random arrival of subjects to sessions. 29 of the papers in my final sample appeared in the American Economic Review, 20 in the American Economic Journal: Applied Economics, and only 4 in the American Economic Journal: Microeconomics. Turning to the 1954 regressions, 70 percent of these are ordinary least squares regressions and an additional 16 percent are maximum likelihood estimates (mostly discrete choice models). Generalized least squares, in the form of weighted regressions based upon a pre-existing estimate of heteroskedasticity or random effects models, make up another 3 percent of the sample. The final residual category, "other", accounts for 10 percent of regressions and includes handfuls of instrumental variables regressions, population weighted regressions, quantile regressions, seemingly unrelated estimates, tests of difference of proportions, t-tests with unequal variances, two-step Heckman models, two-step ordinary least squares regression estimates, and weighted average treatment effects.[15]

A little under a quarter of the regressions in my sample make use of Stata's default covariance matrix calculation. Half of all regressions, however, avail themselves of the cluster estimate of covariance and about another 16 percent use the robust option, a single observation version of the clustered matrix. I discuss and analyse robust covariance estimates separately from clustered because the grouping of observations in clusters makes the sampling distribution of the test statistic dependent upon a somewhat different measure of maximal leverage, as explained in the next section. Bootstrap and "other" methods (consisting of the jackknife and the hc3 and brl bias corrections of the robust and cluster options) make up the remainder of the sample.

[15] I include t-tests with equal variances, as well as any other Stata command that can be re-expressed as an ordinary least squares regression, under the OLS category.

## III: Theory

In this section I provide a thumbnail sketch of the econometric issues and techniques that underlie later empirical results, focusing on statistical inference. First, I lay out the argument that the design of the typical experimental regression invalidates appeals to asymptotic theorems. In particular, I argue that maximal leverage provides a metric of how "asymptotic" the sample is and that, on this measure, the typical experimental regression is indeed very far from asymptopia.[16] I link maximal leverage to variation in the bias and variance of the clustered and robust covariance estimates which I show later on explains all of the average empirical coverage bias of conventional tests. The discussion in this part is limited to OLS regressions, which account for 70 percent of all regressions in my sample. Extensions to some non-OLS frameworks are possible, but involve additional complexity.

Second, having established that there are problems with conventional statistical inference in my sample papers, I present a thumbnail sketch of the theory and methods underlying randomization statistical inference which, given randomization, allows test statistics with distributions that are exact (i.e. known) regardless of sample size, regression design or the characteristics of the error term. I establish terminology and describe alternative measures whose relative power has been theoretically explored. Third, as the bootstrap also features in experimental papers and provides an alternative sampling-based procedure for inference, I review this method as well. As in the case of randomization inference, the bootstrap can be calculated in a number of ways. I note that the method implemented by Stata and its users is theoretically known to be less accurate and, in application, is systematically biased in favour of rejecting null hypotheses. Finally, I review the difference between joint and multiple hypothesis testing procedures, explaining why the latter might have greater power to uncover alternatives of relevance to authors of experimental papers.

### (a) Leverage and the Road to Asymptopia

Consider the regression model $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$, where $\boldsymbol{\varepsilon}$ is an n x 1 vector of (possibly non-normal) disturbances with covariance matrix $\boldsymbol{\Sigma}$.[17] The hat matrix (Hoaglin and Welsch 1978) is

[16] With a respectful tip of the hat to Leamer (2010).

[17] I follow conventional notation, using bold capital letters to denote matrices, bold lower case letters to denote column vectors and lower case letters with subscripts to denote elements of vectors and matrices.

given by $\mathbf{H} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$ and derives its name from the fact that it puts a hat on **y** as $\hat{\mathbf{y}} = \mathbf{X}\hat{\boldsymbol{\beta}} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y} = \mathbf{H}\mathbf{y}$. The element $h_{ij}$ is the derivative of the predicted value of $y_i$ with respect to observation $y_j$. $h_{ii}$, the influence of observation $y_i$ on its own predicted value, is known as the leverage of observation i. **H** is symmetric and idempotent, so we have

$$(3)\ h_{ii} = \sum_j h_{ij}^2 = h_{ii}^2 + \sum_{j \neq i} h_{ij}^2 \geq h_{ii}^2,$$

from which it follows that $1 \geq h_{ii} \geq 0$. Average leverage is given by k/n as

$$(4)\ \bar{h}_{ii} = \frac{1}{n}\sum_i h_{ii} = \frac{1}{n}trace(\mathbf{H}) = \frac{1}{n}trace(\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}') = \frac{1}{n}trace(\mathbf{X}'\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}) = \frac{k}{n}.$$

As the number of observations increases, average leverage falls. However, the maximum leverage in the sample, $h_{ii}^{\max}$, need not. For example, if the regression contains a dummy variable for a particular cluster, the $h_{ii}$ in that cluster (and by extension the maximum $h_{ii}$) always remains above 1/n$_g$, where n$_g$ equals the number of observations in the cluster group.[18] Since $h_{ii}^{\max} \geq \bar{h}_{ii} = k/n$, maximum leverage cannot go to zero unless n → ∞, but n → ∞ does not guarantee $h_{ii}^{\max} \to 0$. As can be seen from (3), when maximum leverage goes to zero all off-diagonal elements in **H** go to zero as well.

Maximum leverage plays a critical, if largely unseen, role in standard econometric theorems. Textbook proofs of the asymptotic consistency or normality (in the presence of non-normal disturbances) of $\hat{\boldsymbol{\beta}}$, for example, typically start by assuming that the limit as n → ∞ of **X′X**/n = **Q**, a positive definite matrix. As shown in the on-line appendix, a necessary condition for this is that $h_{ii}^{\max}$ go to 0. When this condition does not hold, no alternative proof of consistency and normality exists, as Huber (1981) showed that if $\lim_{n\to\infty} h_{ii}^{\max} > 0$ then at least one element of $\hat{\boldsymbol{\beta}}$ is in fact not a consistent estimator of the corresponding element in $\boldsymbol{\beta}$ and, in the presence of non-normal disturbances, is not asymptotically normally distributed.[19] The intuition for these results is trivial. With non-negligible maximum leverage, the predicted value

[18] Removing the dummy variable for cluster g from the list of regressors, let **Z** denote the residuals of the remaining regressors projected on that dummy (in practice, this means that the values within cluster g have their cluster mean removed and all other non-g values are unchanged). Then, using results on partitioned matrices, we find that for any i in cluster g $h_{ii} = 1/n_g + \mathbf{z}_i'(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{z}_i \geq 1/n_g$, where $\mathbf{z}_i'$ is the i$^{th}$ observation row of **Z**.

[19] Huber actually showed that that some of the fitted (predicted) values of $y_i$ will neither be consistent nor, in the event of non-normal disturbances, normal. Since the fitted values are a fixed linear combination of the coefficients, it follows that at least one coefficient must not be consistent or normal.

for some observations is moving with the error terms for those observations. This can only happen if some of the estimated parameters in $\hat{\boldsymbol{\beta}}$ are moving as well. Consequently, it is not possible for the probability that all elements of $\hat{\boldsymbol{\beta}}$ deviate from $\boldsymbol{\beta}$ by more than epsilon to fall to zero, as some must always remain dependent upon the stochastic realization of a small number of disturbances. Moreover, the dependence upon a small number of disturbances eliminates the averaging implicit in central limit theorems, so some elements of $\hat{\boldsymbol{\beta}}$ retain the distributional characteristics of non-normal errors.

Maximum leverage also plays a role in determining the finite sample behaviour of the robust and clustered covariance estimates. With non-stochastic regressors, the estimated coefficients of the regression model described above have the well-known covariance matrix

$$(5)\ \mathbf{V} = E(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta})(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta})' = E[(\mathbf{X'X})^{-1}\mathbf{X'}\boldsymbol{\varepsilon\varepsilon}'\mathbf{X}(\mathbf{X'X})^{-1}] = (\mathbf{X'X})^{-1}\mathbf{X'\Sigma X}(\mathbf{X'X})^{-1}$$

The robust and clustered covariance matrices are calculated using the formulas:

$$(6)\ \ \mathbf{V_R} = c_R(\mathbf{X'X})^{-1}\mathbf{X'}\{\hat{\varepsilon}_i^2\}\mathbf{X}(\mathbf{X'X})^{-1} \quad \mathbf{V_{Cl}} = c_{Cl}(\mathbf{X'X})^{-1}\mathbf{X'}\{\hat{\boldsymbol{\varepsilon}}_\mathbf{g}\hat{\boldsymbol{\varepsilon}}_\mathbf{g}'\}\mathbf{X}(\mathbf{X'X})^{-1}$$

where $c$ denotes a finite sample adjustment, subscript i an observation and g a cluster, $\hat{\varepsilon}_i$ and $\hat{\boldsymbol{\varepsilon}}_\mathbf{g}$ the estimated residuals of observation i and cluster g, respectively, and where I use the notation {a} to denote a diagonal or block diagonal matrix with diagonal elements a. White (1980) argued that, under certain assumptions, $\mathbf{V_R}$ is a consistent estimator of $\mathbf{V}$ when $\mathbf{\Sigma}$ is diagonal, and $\mathbf{V_{Cl}}$ is a natural extension of his work to the case where $\mathbf{\Sigma}$ is block diagonal by cluster. White (1980) assumed that the limit as $n \rightarrow \infty$ of $\mathbf{X'X}/n = \mathbf{Q}$, a positive definite matrix, so it is perhaps not surprising to find that leverage plays a key role in determining the bias and variance of the robust and clustered covariance estimates.

In Young (2016) I show that when $\boldsymbol{\varepsilon}$ is distributed iid normal bounds on the bias of the robust and clustered estimates of the variance of any linear combination $\mathbf{w}$ of the estimated coefficients $\hat{\boldsymbol{\beta}}$ are given by:

$$(7)\ \ c_R(1-h_{ii}^{\max}) \leq \frac{E[\mathbf{w'V_Rw}]}{\mathbf{w'Vw}} \leq c_R(1-h_{ii}^{\min}),\ c_{Cl}(1-\lambda^{\max}(\{\mathbf{H_{gg}}\})) \leq \frac{E[\mathbf{w'V_{Cl}w}]}{\mathbf{w'Vw}} \leq c_{Cl}(1-\lambda^{\min}(\{\mathbf{H_{gg}}\}))$$

where $h_{ii}^{\max}$ and $h_{ii}^{\min}$ are the maximum and minimum diagonal elements of the hat matrix (leverage) and $\lambda^{\max}(\{\mathbf{H_{gg}}\})$ and $\lambda^{\min}(\{\mathbf{H_{gg}}\})$ the maximum and minimum eigenvalues of the block diagonal matrix made up of the sub-matrices of the hat matrix associated with the cluster observations. In referring to clustered covariance estimates, I shall use the term "maximal leverage", somewhat loosely, to denote $\lambda^{\max}(\{\mathbf{H_{gg}}\})$, as in theoretical results this is the cluster counterpart of $h_{ii}^{\max}$.

Intuition for this bias result can be found by considering the way in which least squares fitting results in an uneven downward bias in the size of residuals. To this end, let the symmetric and idempotent matrix **M = I – H** denote the "residual maker", as the estimated residuals are given by $\hat{\boldsymbol{\varepsilon}} = \mathbf{y} - \hat{\mathbf{y}} = \mathbf{My} = \mathbf{M(X\boldsymbol{\beta} + \boldsymbol{\varepsilon})} = \mathbf{M\boldsymbol{\varepsilon}}$. Consequently, $\hat{\varepsilon}_i = \mathbf{m}_\mathbf{i}'\boldsymbol{\varepsilon}$, where $\mathbf{m}_\mathbf{i}'$ is the i$^{\text{th}}$ row of $\mathbf{M}$[20] and the expected value of the i$^{\text{th}}$ squared residual has a downward bias determined by leverage as $E(\hat{\varepsilon}_i^2) = E(\mathbf{m}_\mathbf{i}'\boldsymbol{\varepsilon\varepsilon}'\mathbf{m}_\mathbf{i}) = \mathbf{m}_\mathbf{i}'\{\sigma^2\}\mathbf{m}_\mathbf{i} = \sigma^2 m_{ii} = \sigma^2(1-h_{ii})$, where I have made use of the fact that **M** is idempotent and the assumption that **ε** is distributed iid. The conventional OLS estimate of variance treats all residuals symmetrically, summing them and dividing by n-k. This yields an unbiased estimate of $\sigma^2$ as $(n\text{-}k)^{-1}\Sigma\sigma^2(1\text{-}h_{ii}) = (n\text{-}k)^{-1}\sigma^2(n\text{-}k) = \sigma^2$. The robust covariance estimate, however, is an unevenly weighted function of the residuals, which allows a bias that is determined by the range of the bias of the residuals. In the case of the clustered covariance estimate, which places uneven weight on clusters of residuals, the range of bias is greater as $\lambda^{\max}(\{\mathbf{H_{gg}}\}) \geq h_{ii}^{\max}$ and $\lambda^{\min}(\{\mathbf{H_{gg}}\}) \leq h_{ii}^{\min}$ (as proven in Young 2016). In practice, $\lambda^{\min}(\{\mathbf{H_{gg}}\})$ and $h_{ii}^{\min}$ vary little, as they must lie between 0 and k/n, while $\lambda^{\max}(\{\mathbf{H_{gg}}\})$ and $h_{ii}^{\max}$ vary a lot, as they lie between k/n and 1. Thus, variation in maximal leverage is the principal determinant of the range of potential bias.[21]

Leverage also plays a role in determining the variance of the robust or clustered covariance estimate and hence, by extension, the thickness of the tails of the distribution of the test statistic. Consider the linear combination of coefficients given by $\mathbf{w} = \mathbf{x}_\text{i}$, where $\mathbf{x}_\text{i}'$ is the i$^{\text{th}}$ observation row of **X**. In this case, the robust estimate of the variance of $\mathbf{w}'\hat{\boldsymbol{\beta}}$ is given by $\mathbf{w'V_Rw}$ $= c_R\mathbf{x}_i'\mathbf{(X'X)}^{-1}\mathbf{X}'\{\hat{\varepsilon}_i^2\}\mathbf{X(X'X)}^{-1}\mathbf{x}_i = c_R\mathbf{h}_i'\{\hat{\varepsilon}_i^2\}\mathbf{h}_i$, where $\mathbf{h}_\text{i}'$ is the i$^{\text{th}}$ row of **H**. As $h_{ii}$, the leverage of observation i, increases, all the other $h_{ij}$ (j ≠ i) elements of $\mathbf{h}_\text{i}$ go to zero, as can be seen in (3)

[20] $\mathbf{m}_\mathbf{i}$ is the i$^{\text{th}}$ column of **M**, but as **M** is symmetric, $\mathbf{m}_\mathbf{i}'$ is also the i$^{\text{th}}$ row.

[21] Across my sample of 1371 OLS regressions, the standard deviation of $h_{ii}^{\min}$ is .027, while that of $h_{ii}^{\max}$ is .383; similarly, across the 824 OLS regressions which cluster, the standard deviation of $\lambda^{\min}(\{\mathbf{H_{gg}}\})$ is .001, while that of $\lambda^{\max}(\{\mathbf{H_{gg}}\})$ is .615.

above. Consequently, the covariance estimate places weight on a smaller and smaller subset of residuals and, in the limit, depends upon only one residual. This reduced dimensionality increases the variance of the variance estimate, as an estimate made up of a smaller number of random variables is more variable. In Young (2016) I establish the following bounds on the "effective degrees of freedom" that characterize the distribution of the t-statistic for any hypothesis test based upon a linear combination of the estimated coefficients using the robust and clustered covariance estimates when the error disturbances are iid normal:

$$(8)\ \ n-k \geq \mathrm{edf}_{\mathrm{R}} \geq \max(1,(h_{ii}^{\max})^{-1}-1), \quad \min(n_c\text{-}1,n-k) \geq \mathrm{edf}_{\mathrm{Cl}} \geq \max(1,\lambda^{\max}(\{\mathbf{H}_{\mathbf{gg}}\})^{-1}-1)$$

where $n_c$ is the number of clusters.[22] The n-k and $n_c$-1 degrees of freedom typically used to evaluate test statistics based upon these covariance matrices are the upper bound on the realized distributions, i.e. actual tails can only be thicker than is customarily assumed.

Equations (7) and (8) describe bounds. If, however, one thinks of different hypothesis tests as producing results that randomly range within these bounds, it is easy to see why high maximal leverage leads to biased statistical inference. As maximal leverage rises, the range of bias increases, producing over and under estimates. With good finite sample corrections $c_R$ and $c_{Cl}$, the covariance estimates may remain, on average, unbiased.[23] However, since they appear in the denominator of test statistics, their variation, by Jensen's inequality, increases the average absolute value of test statistics which tends to raise the average rejection rate across hypothesis tests. High maximal leverage also allows effective degrees of freedom to fall, i.e. the higher variance of the covariance estimate produces thicker tails than indicated by the n-k or $n_c$-1 degrees of freedom used to evaluate the test statistics, raising rejection rates above the putative nominal size of the test. This effect tends to produce higher than nominal rejection rates in each and every hypothesis test. Because of a strongly positively biased covariance estimate, it is

[22] The bounds for robust covariance estimates in (7) and (8) can be found in Chesher and Jewitt (1987) and Chesher (1989). Those for the clustered case are my extension of their results.

[23] Since, with iid errors, the $i^{th}$ residual underestimates its own variance by 1-$h_{ii}$, the average residual underestimates its own variance by $n^{-1}\Sigma(1\text{-}h_{ii})$ = (n-k)/n. This suggests an n/(n-k) finite sample correction, which is what is typically used for $\mathbf{V_R}$. In the case of $\mathbf{V_{Cl}}$, Stata applies an (n-1)$n_c$/(n-k)($n_c$-1) correction in the case of the reg or areg clustered commands, which, for large n and $n_c$, is approximately equal to n/(n-k). In the case of the xtreg fe clustered command, however, Stata uses (n-1)$n_c$/(n-k+$k_{fe}$)($n_c$-1), where $k_{fe}$ is the number of fixed effects. This produces systematically lower p-values than the otherwise identical areg command. Three papers in my sample use the xtreg fe clustered command in 100 regressions and I find that the alternative degrees of freedom adjustment reduces the variance estimate by .85 on average and .5 in one instance.

Table II: Regression Design in the Sample Papers (1371 OLS regressions)

| | mean | min | .01 | .05 | .10 | .25 | .50 | .75 | .90 | .95 | .99 | max |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Cumulative distribution of y values and normality of residuals | | | | | | | | | | | | |
| # of y values | 326 | 2 | 2 | 2 | 2 | 2 | 12 | 135 | 500 | 2548 | 4225 | $13e^{3}$ |
| Modal share | .468 | $3e^{-4}$ | .001 | .005 | .021 | .125 | .530 | .736 | .917 | .963 | .993 | .9999 |
| Normality of $\hat{\varepsilon}$ | .012 | 0 | 0 | 0 | 0 | 0 | 0 | $1e^{-10}$ | $6e^{-4}$ | .011 | .422 | .848 |
| Cumulative distribution of leverage | | | | | | | | | | | | |
| $\bar{h}_{ii}$ | .051 | $2e^{-5}$ | $1e^{-4}$ | $1e^{-3}$ | .002 | .008 | .026 | .058 | .148 | .207 | .330 | .533 |
| $h_{ii}^{\max}$ | .384 | $4e^{-5}$ | $2e^{-4}$ | .002 | .008 | .025 | .198 | .729 | 1 | 1 | 1 | 1 |
| V$_{R}$: $h_{ii}^{\max}$ | .493 | .001 | .001 | .002 | .014 | .170 | .404 | 1 | 1 | 1 | 1 | 1 |
| V$_{Cl}$: $\lambda^{\max}(\{\mathbf{H}_{\mathbf{gg}}\})$ | .615 | $9e^{-4}$ | .016 | .038 | .053 | .207 | .701 | 1 | 1 | 1 | 1 | 1 |

Notes: $ae^{b}$ stands for $a*10^{b}$. Normality = p-value in Stata's sktest of normality of residuals based on skewness and kurtosis. V$_{R}$ & V$_{Cl}$ = leverage distribution measures calculated for the 160 and 824 OLS regressions, respectively, which use the robust or clustered estimate of covariance. $\bar{h}_{ii}$ and $h_{ii}^{\max}$, average and maximum leverage. $\lambda^{\max}(\{\mathbf{H}_{\mathbf{gg}}\})$= maximum eigenvalue of the block diagonal matrix made up of the elements of the hat matrix associated with the cluster groups.

possible that actual size remains less than nominal value in any particular test, but, on average, across all hypothesis tests, the variation in bias and excess variation in the covariance estimate produce higher than nominal rejection rates.

The practical relevance of the theoretical issues discussed above is shown in Table II, which summarizes key features of OLS regression design in my sample of experimental papers. As shown in the top row, the dependent variable typically takes on very few values and in 43 percent of regressions is, in fact, a 0/1 dichotomous variable.[24] The share of the modal y value is also typically quite high, exceeding .530 in ½ of regressions and, extraordinarily, .963 in 1/20th of the sample.[25] Not surprisingly, tests of the normality of residuals reject the null, at the 1 percent level, 95 percent of the time. Using bootstrap samples to simulate the sampling distribution of

[24] These are linear probability models, not probits or logits.

[25] Including two regressions with 33103 observations in which the y variable takes on an alternate value for only 4 observations and 7 other regressions, with 217 to 840 observations each, in which the y variable takes on an alternate value in 1 observation alone. The sensitivity of results to a few observations in cases with high modal shares is obviously an issue, but in this paper I focus on inference alone, taking samples and regression specifications as given.

the coefficients (further below) I find that these reject the null of the normal distribution 43 percent of the time at the .01 level, a consequence of the non-normality of the residuals and high leverage. Practically speaking, however, these non-normal distributions have tails which, variance adjusted, are not systematically thicker or thinner than those of the normal distribution, so the pervasive non-normality, while highlighting how far the typical regression is from the asymptotic ideal, does not lead to systematic bias in test statistics.

Moving to the independent variables, we see that the typical paper has an average leverage of .051, indicating about 20 observations per regressor, with about 5 percent of the sample showing an average leverage greater than .2, i.e. less than 5 observations per regressor. Despite having an average of 5300 seemingly asymptotic observations per regression, maximal leverage tends to be quite high, averaging .383 and exceeding .729 in one quarter of regressions. These results have implications for the normality of estimated coefficients. The third row examines the 160 OLS regressions which use the robust estimate of covariance ($\mathbf{V_R}$), where leverage affects both normality and the accuracy of the covariance estimate. Here, unfortunately, maximal leverage is higher, averaging .493 and equalling 1 in 33 percent of the robust covariance estimate sample. In the 824 OLS regressions which use the clustered estimate of covariance ($\mathbf{V_{Cl}}$), the situation is, effectively, much worse as the average maximal eigenvalue of the blocks of the hat matrix associated with the cluster groups, which is what matters for these matrices, is .615, with 39 percent of the sample showing a maximum eigenvalue of 1.

Readers familiar with leverage will know that it is possible to make too much of high leverage values. Just as the influence of leverage on estimated coefficients depends upon its interaction with residuals,[26] so too does its influence on consistency, normality and covariance estimation. Consider the case where regressor $\mathbf{x}_1$ takes on the value of 1 for observation #1, and 0 for all others. The estimated residual for observation #1 will always be zero and its leverage, and the maximum leverage in the regression, equals 1. The estimated coefficient $\hat{\beta}_1$ on $\mathbf{x}_1$ will be inconsistent and, if the disturbance is non-normal, non-normal as well. However, none of this matters at all for the remainder of the regression, where the estimated coefficients, residuals and standard errors (robust or otherwise) are completely independent of observation #1, $\mathbf{x}_1$ and $\hat{\beta}_1$.

[26] For an intuitive discussion see Fox (2008).

Consequently, assuming asymptotically vanishing leverage in the remaining observations, they are consistent and normal with a covariance matrix that is asymptotically consistently estimated by the unbiased robust or clustered covariance matrix. This extreme example, while instructive, is not of much relevance, as regressions generally do not contain regressors of this type. In my analysis further below I find that regressions with a maximal leverage of 1 suffer a reduction in effective degrees of freedom consistent with the influence of the lower bound in other equations and a variance estimate downward bias which, while not attaining the potential bias of 100 percent implied by (7), is nevertheless substantial.

Huber (1981, p. 162), in his study of robust statistics, advised

> ...large values of $h_{ii}$ should serve as warning signals that the i$^{th}$ observation may have a decisive, yet hardly checkable, influence. Values $h_{ii} \leq 0.2$ appear to be safe, values between 0.2 and 0.5 are risky, and if we can control the design at all, we had better avoid values above 0.5.

Huber's concern was the sensitivity of coefficient estimates to particular observations. In this paper I take coefficient estimates as inviolate, and focus on the accuracy of tests of significance. The bounds presented above show how badly leverage can bias statistical inference. With a maximal leverage of .5, the downward bias of the covariance estimate can be as high 50 percent and the effective degrees of freedom reduced to 1, i.e. distributional tails with a thickness equal to that reached when n-k = 1. In this context, Huber's cautionary advice is perhaps worth considering.

### (b) Randomization Statistical Inference

Randomization statistical inference provides exact tests of sharp (i.e. precise) hypotheses no matter what the sample size, regression design or characteristics of the disturbance term. The typical experimental regression can be described as $y_i = \mathbf{t}_i'\boldsymbol{\beta}_\mathbf{t} + \mathbf{x}_i'\boldsymbol{\beta}_\mathbf{x} + \varepsilon_i$, where $\mathbf{t}_i$ is a vector of treatment variables[27] and $\mathbf{x}_i$ a vector of other causal determinants of $y_i$, the dependent variable of interest. Conventional econometrics describes the statistical distribution of the estimated $\boldsymbol{\beta}$s as coming from the stochastic draw of the disturbance term $\varepsilon_i$, and possibly the regressors, from a population distribution. In contrast, in randomization inference the motivating thought experiment is that, given the sample of experimental participants, the only stochastic element

[27] Which may contain interactions with non-treatment characteristics, as in the case of $\beta_{T*age}T*age$ in (1) earlier above.

determining the realization of outcomes is the randomized allocation of treatment. For each participant, $y_i$ is conceived as a determinate function of treatment $y_i(\mathbf{t}_i)$ following the equation given above and the stochastic realization of $\mathbf{t}_i$ determines the statistical distribution of the estimated $\boldsymbol{\beta}$s. As such, it allows the testing of sharp hypotheses which specify the treatment effect for each participant, because sharp hypotheses of this sort allow the calculation of the realization of the estimated $\boldsymbol{\beta}$s for any potential random allocation of treatment. The Fisherian null hypothesis of no treatment effect is that $y_i(\mathbf{t}_i) = y_i(\mathbf{0})$ for all i and all treatment vectors $\mathbf{t}_i$, i.e. the experiment has absolutely no effect on any participant. This is not a null of zero average treatment effect, it is a null of no effect whatsoever on any participant.

An exact test of the Fisherian null can be constructed by calculating all of the possible realizations of a test statistic and rejecting if the observed realization in the experiment itself is extreme enough. Specifically, let the matrix $\mathbf{T}_E$ composed of the row vectors $\mathbf{t}_i'$ denote the treatment allocation in the experiment. In the typical experiment this matrix has a finite universe $\boldsymbol{\Omega}$ of potential realizations. Say there are S elements in $\boldsymbol{\Omega}$, with $\mathbf{T}_n$ denoting a particular element. Let $f(\mathbf{T}_n)$ be a statistic calculated by inserting matrix $\mathbf{T}_n$ into the estimating equation given earlier above, and let $f(\mathbf{T}_E)$ denote the same statistic calculated using the actual treatment applied in the experiment. Under the null of no treatment effect, $y_i = \mathbf{x}_i'\boldsymbol{\beta}_\mathbf{x} + \varepsilon_i$ is the same no matter which treatment is applied, i.e. experimental outcomes would have been exactly the same regardless of the specific randomized draw of $\mathbf{T}_E$ from $\boldsymbol{\Omega}$, so $f(\mathbf{T}_n)$ can be calculated by regressing the fixed observed values of $y_i$ on the fixed regressors $\mathbf{x}_i$ and randomly varied treatment vector $\mathbf{t}_i$. The p-value of the experiment's test statistic is given by:

$$\text{(9) randomization p-value} \;=\; \frac{1}{S}\sum_{n=1}^{S} I_n(>T_E) \;+\; U*\frac{1}{S}\sum_{n=1}^{S} I_n(=T_E)$$

where $I_n(>\mathbf{T_E})$ and $I_n(=\mathbf{T_E})$ are indicator functions for $f(T_n) > f(T_E)$ and $f(T_n) = f(T_E)$, respectively, and $U$ is a random variable drawn from the uniform distribution. In words, the p-value of the randomization test equals the fraction of potential outcomes that have a more extreme test statistic added to the fraction that have an equal test statistic times a uniformly distributed random number. In the on-line appendix I prove that this p-value is always uniformly distributed, i.e. the test is exact, regardless of the sample size or the characteristics of $y_i$, $\mathbf{x}_i$ and $\varepsilon_i$.

Calculating (9), evaluating f($\mathbf{T}_n$) for all possible treatment realizations in $\mathbf{\Omega}$, is generally impractical. However, under the null random sampling with replacement from $\mathbf{\Omega}$ allows the calculation of an equally exact p-value provided the original treatment result is automatically counted as a tie with itself. Specifically, with N additional draws (beyond the original treatment) from $\mathbf{\Omega}$, the p-value of the experimental result is given by:

$$\text{(10) sampling randomization p-value} \;=\; \frac{1}{\mathrm{N}+1}\sum_{n=1}^{N}\mathrm{I}_{\mathrm{n}}(>\mathrm{T}_{\mathrm{E}}) \;+\; U*\frac{1}{\mathrm{N}+1}\left[1+\sum_{n=1}^{N}\mathrm{I}_{\mathrm{n}}(=\mathrm{T}_{\mathrm{E}})\right]$$

In the on-line I appendix I show that this p-value is uniformly distributed regardless of the number of draws N used to evaluate the test statistic. [28] This establishes that size always equals its nominal value. Power, however, as shown by Jockel (1986), is increasing in N.[29] Intuitively, as the number of draws increases the procedure is better able to identify what constitutes an outlier outcome in the distribution of the test statistic f(). In my analysis of the experimental papers I use 10000 draws to evaluate (10). When compared with results calculated with fewer draws, I find no appreciable change in rejection probabilities beyond 2000 draws, suggesting that increasing N beyond 10000 would have no effect on the results.

In the analysis below, for theoretical reasons associated with both randomization inference and the bootstrap, I make use of two randomization based test statistics. The first is based upon the comparison of the Wald statistics of the conventional econometric two-sided test of the null hypothesis of no treatment effect. The Wald statistic for the conventional test is given by $\hat{\boldsymbol{\beta}}_t'(\mathbf{T}_n)\mathbf{V}(\hat{\boldsymbol{\beta}}_t(\mathbf{T}_n))^{-1}\hat{\boldsymbol{\beta}}_t(\mathbf{T}_n)$, where $\hat{\boldsymbol{\beta}}_t$ and $\mathbf{V}(\hat{\boldsymbol{\beta}}_t)$ are the regression's treatment coefficients and the estimated variance of those coefficients, so this method in effect calculates the probability

$$\text{(11)}\quad \hat{\boldsymbol{\beta}}_t'(\mathbf{T}_n)\mathbf{V}(\hat{\boldsymbol{\beta}}_t(\mathbf{T}_n))^{-1}\hat{\boldsymbol{\beta}}_t(\mathbf{T}_n) \geq \hat{\boldsymbol{\beta}}_t'(\mathbf{T}_E)\mathbf{V}(\hat{\boldsymbol{\beta}}_t(\mathbf{T}_E))^{-1}\hat{\boldsymbol{\beta}}_t(\mathbf{T}_E)\,.$$

I use the notation ($\mathbf{T}_n$) to emphasize that both the coefficients and covariance matrix are calculated for each realization of the randomized draw $\mathbf{T}_n$ from $\mathbf{\Omega}$. In the univariate case the statistic reduces to a comparison of the squared values of the t-statistics, and consequently I dub this test the randomization-t.

[28] The proof is a straightforward generalization of Jockel's (1986) result for nominal size equal to an integer multiple of 1/(N+1).

[29] Provided power itself is a concave in the nominal size of the test.

An alternative test of no treatment effects, similar to some bootstrap techniques, is to compare the relative values of $\hat{\boldsymbol{\beta}}_{\mathbf{t}}'(\mathbf{T}_{\mathbf{n}})\mathbf{V}(\hat{\boldsymbol{\beta}}_{\mathbf{t}}(\boldsymbol{\Omega}))^{-1}\hat{\boldsymbol{\beta}}_{\mathbf{t}}(\mathbf{T}_{\mathbf{n}})$, where $\mathbf{V}(\hat{\boldsymbol{\beta}}_{\mathbf{t}}(\boldsymbol{\Omega}))$ is the covariance of $\hat{\boldsymbol{\beta}}_{\mathbf{t}}$ across the universe of potential treatment draws in **Ω**. In this case, a fixed covariance matrix is used to evaluate the coefficients produced by each randomized draw $\mathbf{T}_n$ from **Ω**, calculating the probability

(12) $\hat{\boldsymbol{\beta}}_{\mathbf{t}}'(\mathbf{T}_{\mathbf{n}})\mathbf{V}(\hat{\boldsymbol{\beta}}_{\mathbf{t}}(\boldsymbol{\Omega}))^{-1}\hat{\boldsymbol{\beta}}_{\mathbf{t}}(\mathbf{T}_{\mathbf{n}}) \geq \hat{\boldsymbol{\beta}}_{\mathbf{t}}'(\mathbf{T}_{\mathbf{E}})\mathbf{V}(\hat{\boldsymbol{\beta}}_{\mathbf{t}}(\boldsymbol{\Omega}))^{-1}\hat{\boldsymbol{\beta}}_{\mathbf{t}}(\mathbf{T}_{\mathbf{E}})$.

In the univariate case, this reduces to the square of the coefficients divided by a common variance and hence, after eliminating the common denominator of both sides, is basically a comparison of squared coefficients. Hence, I refer to this comparison as the randomization-c. I use the coefficient covariance across the 10000 randomization draws to approximate $\mathbf{V}(\hat{\boldsymbol{\beta}}_{\mathbf{t}}(\boldsymbol{\Omega}))$.[30]

I use the two versions of the randomization test, the -t and -c, to provide counterparts to commonly used bootstrap tests. Lehmann (1959) showed that in the simple test of mean differences between treatment and control with iid errors the randomization-t is uniformly most powerful and asymptotically identical to the conventional t-test of no treatment effect. Despite this result, I find that in practical application, with the type of errors and treatment effects present in my experimental sample, the randomization-c is, if anything, actually more powerful than the randomization-t. Across most tables, however, the two methods produce very similar results. The same cannot be said, however, for analogous bootstrap tests, which produce systematically different results for reasons that are explained below.

The randomization-c allows for an easy omnibus test of the overall statistical significance of all of the regressions in an experimental paper. One simply stacks all the treatment coefficients from all of the regression equations, draws repeated randomization treatments $\mathbf{T}_n$ from **Ω**, and calculates (12) above, with $\hat{\boldsymbol{\beta}}_{\mathbf{t}}$ denoting all treatment coefficients in the paper. The estimated covariance of these coefficients in the universe **Ω** is simply calculated from their joint realizations. An omnibus version of the randomization-t is much more difficult, as it requires an iteration by iteration estimate of $\mathbf{V}(\hat{\boldsymbol{\beta}}_{\mathbf{t}}(\mathbf{T}_{\mathbf{n}}))$, including the covariance of coefficients across

[30] Strictly speaking, in multi-coefficient tests the test statistics $f(T_n)$ are now a function of the joint realization of the randomization draw, so the proof of exactness for a finite number of draws in the on-line appendix is no longer valid. My intent, however is to provide a counterpart to a common bootstrap technique; with 10000 draws I should have a fairly close approximation of the true coefficient covariance matrix; and in tests of an individual coefficient the variance cancels from both sides of (12), so the proof of exactness remains valid.

equations.[31] In a single equation setting, as already noted, I find very little difference in the simulated power and test outcomes of the randomization-t and -c.

I conclude this presentation by noting some of the details of my methods. First, in calculating the $\mathbf{T_n}$ specific coefficient covariance matrix, I defer to the decisions made by authors and use their covariance estimation methods no matter how complex, computationally intensive or, to my eye, flawed they may be.[32] This maintains my rule of following author methods as closely as possible in assessing their results. Second, in producing the randomization distribution I do not calculate one equation at a time, but rather apply the randomized experimental treatment draw $\mathbf{T}_n$ to the entire experimental data set, and then calculate all equations together. This allows the calculation of the cross-equation covariance of all regression coefficients that allows me to calculate the omnibus randomization test described above. As I apply the randomized treatment outcome to the sample, I recalculate all variables that are contingent upon that realization, e.g. participant characteristics interacted with treatment outcomes. I also reproduce any coding errors in the original do-files that affect treatment measures, e.g. a line of code that unintentionally drops half the sample or another piece that intends to recode individuals of a limited type to have a zero x-variable but unintentionally recodes all individuals in broader groups to have that zero x-variable. All of this follows the Fisherian null: all procedures and outcomes in the experiment are invariant with respect to who received what treatment.

[31] White (1982) showed that an asymptotically valid estimate of the covariance matrix for all of the coefficients estimated in multiple equations is given by yet another sandwich covariance matrix, with the block diagonal default covariance matrix of the individual equations as the bread and the outer product of the equation level scores as the filling (see also Weesie 1999). The practical barriers to its implementation, however, are staggering. Many estimation procedures do not produce scores. Many papers present the relevant data in multiple, differently organized, data files, so the cross-product of scores is extraordinarily difficult to form. When scores can be calculated within a single data file, the resulting covariance matrices, calculated across all of the equations and their coefficients, often exceed the 11k x 11k limitations of Stata and, when they do not, are often hopelessly singular, even within the sub-matrices defined only by treatment variables. Finally, I find that the use of this multi-equation covariance estimate to test joint hypotheses on treatment measures in just 10 to 14 equations (let alone dozens) in one of my sample papers, using Stata's suest command, produces some of the most extraordinary size distortions, with an average rejection probability of .25 at the .01 level. This suggests that White's multi-equation covariance estimation procedure, where it can be implemented, is not very accurate in finite samples.

[32] Thus, in the three papers where authors bootstrap 100s of iterations for their estimate of covariance, I do the same for each of the 10000 iterations of the randomization-t. In another case, the authors use an incorrect code for calculating the biased-reduced linearization (brl) estimate of covariance which unfortunately also executes extraordinarily slowly. Rather than substitute my own faster brl code (which I wrote to confirm the errors) I implement their code time and again. Producing the randomization estimates for each of these papers takes 6 months of workstation time.

Third, in executing randomization iterations[33] I accept an iteration, even if the covariance matrix is singular, as long as Stata produces a coefficient estimate and standard error for the treatment variable. I state this to avoid criticism that I use inappropriate coefficient estimates. In my sample no less than one-fifth of the original regressions have singular covariance matrices. This generally arises because of maximal leverage of 1 in robust and clustered covariance matrices, but it also regularly occurs because maximum likelihood procedures do not converge and/or authors estimate equations that are guaranteed to have singular covariance matrices. Stata usually warns the user that the procedure did not converge, or when the covariance matrix is highly singular and suspect. Coefficients and standard errors produced by these methods are accepted and reported in journal tables. In order to be able to analyse the sample, and in the spirit of the Fisherian null that all procedures and outcomes are invariant with respect to randomization, I follow authors' procedures and accept results if Stata is able to deliver them, no matter how badly conditioned the covariance matrix is.[34]

Fourth, in making randomization draws from the universe of potential treatments $\mathbf{\Omega}$ I restrict my draws to the subset $\mathbf{\Omega}$ that has the same treatment balance as $\mathbf{T}_E$, the experimental draw. This subtle distinction, irrelevant from the point of view of the exactness of the randomization test statistic, avoids my making unnecessary, and potentially inaccurate, inferences about the alternative balance of treatments that might have arisen. For example, a number of experiments applied treatment by taking random draws from a distribution (e.g. drawing a chit from a bag). Rather than trying to replicate the underlying distribution, I take the realized outcomes and randomly reallocate them across participants. I adopted this procedure after observing that in some papers the distribution of outcomes does not actually follow the description of the underlying process given in the paper. A few papers note problems in implementation, and some authors, in correspondence, noted that even after they selected a particular randomized allocation of treatment, field agents did not always implement it

[33] Or bootstrap iterations or size and power simulations of the bootstrap and randomization statistics.

[34] Throughout this paper, in noting the number of bootstrap or randomization "iterations" I refer to the number of attempted iterations. It is occasionally the case that Stata cannot manage its customary miracle of producing Wald tests from hopelessly singular matrices. These iterations are dropped. This is why (in the on-line appendix) I generalize Jockel's (1986) proof of the exactness of the randomization test statistic (10) for α equal to an integer multiple of 1/N+1 to any number α, because it is not possible to guarantee the same number of successful iterations N for all equations and papers.

accurately. I follow the papers in taking all of these errors in implementation as part of the random allocation of treatment. Under the randomization hypothesis, strongly maintained in every paper, treatment quantities, even if not in the proportions intended by the authors, could in principle have been applied to any participant. Thus, subject only to the stratification scheme, clarified by detailed examination of the data and correspondence with the authors, I shuffle *realized* treatment outcomes across participants. This shuffling amounts to drawing the treatment vectors $\mathbf{T}_n$ in $\mathbf{\Omega}$ that share the same treatment balance as $\mathbf{T}_E$.[35]

Finally, I should note that I test instrumental variables regressions using the implied intent to treat regressions. In these regressions treatment variables are used as instruments, most of the time representing an opportunity that is offered to a participant that is then taken up or not. The null here cannot be that the treatment instrument has no effect on the instrumented variable, as this is obviously false (e.g. one can only take up an opportunity if one is offered the chance to do so). Consequently, one cannot shuffle treatment and rerun the first stage regression. However, a reasonable null, and the relationship being tested in the second-stage regression, is that the instrumented variable has no effect on final outcomes of interest. Combined with the exogeneity assumption used to identify the regression, in an iv setting this implies that there exists no linear relationship between the outcome variable and the treatment variables themselves, i.e. no significant relation in the intent to treat regression. Consequently, I test the significance of instrumental variables regressions by running the implied intention to treat regression for the experiment and then comparing its coefficients and p-values to those produced through the randomization distribution under the null that final outcomes are invariant with respect to the actual realization of treatment.[36]

[35] All of this is done, of course, in units of treatment, e.g. field villages or lab sessions. To keep the presentation familiar, I have described randomization tests as sampling from a population of potential outcomes. A more general presentation (e.g. Romano 1989) argues that under the null outcomes are invariant with respect to all transformations G that map from $\mathbf{\Omega}$ to $\mathbf{\Omega}$. The shuffling or rearranging of outcomes across participants is precisely such a mapping.

[36] In using the bootstrap, and reporting original authors' results, I continue to use the second stage iv regression itself. To keep the number of iv and intent to treat coefficients equal across methods, I only examine exactly identified iv regressions (i.e. exclude a small number of overidentified two stage least squares). I should also note that I have found that many of the intent to treat regressions implied by iv regressions duplicate regressions found elsewhere in the paper. I drop these duplicates from the analysis. This, plus eliminating other duplicate regressions within papers, explains the reduction in the number of regressions analysed in this draft relative to earlier versions of this paper. Sometimes authors present first-stage regressions along with iv results. I skip these if they

### (c) Bootstrap Statistical Inference

While randomization statistical inference is based on thought experiments concerning the stochastic allocation of treatment to a fixed experimental population, conventional statistical inference revolves around the notion of stochastic variation brought about by random sampling from a larger population. To forestall the mistaken conclusion that the results of this paper stem from this philosophical difference, I complement the randomization analysis below with results based on the bootstrap. Conventional econometrics uses assumptions and asymptotic theorems to infer the distribution of a statistic f calculated from a sample with empirical distribution $F_1$ drawn from an infinite parent population with distribution $F_0$, which can be described as $f(F_1|F_0)$. In contrast, the bootstrap estimates the distribution of $f(F_1|F_0)$ by drawing random samples $F_2$ from the population distribution $F_1$ and observing the distribution of $f(F_2|F_1)$ (Hall 1992). If f is a smooth function of the sample, then asymptotically the bootstrapped distribution converges to the true distribution (Lehmann and Romano 2005), as, intuitively, the outcomes observed when sampling $F_2$ from an infinite sample $F_1$ approach those arrived at from sampling $F_1$ from the actual population $F_0$. The bootstrap is another asymptotically accurate method which in finite samples has problems of its own, but I make use of it because it allows me to provide supporting evidence, based on sampling rather than randomization methods, regarding statistical significance and the cross-equation correlation of results. Moreover, I use bootstrapped samples from the experimental population to examine issues of power and size, thereby contrasting conventional and randomization methods using simulated populations that contain the specific characteristics, such as correlated or heteroskedastic errors and heterogeneous treatment effects, that exist in the actual experimental populations themselves.

As in the case of randomization tests, there are many possible ways of calculating the bootstrap. I use two which, parallel to the randomization tests described above, can be called the bootstrap-t and -c. Let $B_n$ denote the bootstrap sample randomly drawn from the experimental population $F_1$, $\hat{\boldsymbol{\beta}}(B_n)$ and $\mathbf{V}(\hat{\boldsymbol{\beta}}(B_n))$ the coefficient and coefficient covariance estimates for that

involve a dependent variable that is never used as a treatment outcome elsewhere in the paper. In total, this leads me to drop 14 first stage regressions in three papers, which are all of form described above, where the dependent variable is trivially determined by treatment. On the other hand, I retain first stage regressions where the authors, having used the dependent variable as a treatment outcome elsewhere in the paper, now use it as an instrumented variable in determining some other treatment outcome.

sample, and $\Omega_1$ the universe of potential sample draws from $F_1$. We are interested in a two-sided test of the null hypothesis $\boldsymbol{\beta}(F_0) = \mathbf{0}$, i.e. in evaluating the distribution of the Wald test statistic $\hat{\boldsymbol{\beta}}(B_E)'\mathbf{V}(\hat{\boldsymbol{\beta}}(B_E))^{-1}\hat{\boldsymbol{\beta}}(B_E)$, where $B_E = F_1$ is the experimental sample. We know that $\boldsymbol{\beta}(F_1) = \hat{\boldsymbol{\beta}}(B_E)$, i.e. the average treatment effect in population $F_1$ is that given by the experimentally estimated coefficients, so we evaluate the distribution of the experimental Wald statistic $f(F_1|F_0)$ under the null $\boldsymbol{\beta}(F_0) = \mathbf{0}$ by examining the distribution of the bootstrapped Wald statistics $f(F_2|F_1)$ around the null that we know to be true for $F_1$, calculating the probability

$$(13)\quad [\hat{\boldsymbol{\beta}}(B_n) - \hat{\boldsymbol{\beta}}(B_E)]'\mathbf{V}(\hat{\boldsymbol{\beta}}(B_n))^{-1}[\hat{\boldsymbol{\beta}}(B_n) - \hat{\boldsymbol{\beta}}(B_E)] \geq \hat{\boldsymbol{\beta}}(B_E)'\mathbf{V}(\hat{\boldsymbol{\beta}}(B_E))^{-1}\hat{\boldsymbol{\beta}}(B_E)$$

In the univariate case this reduces to a comparison of squared t-statistics, so I refer to it as the bootstrap-t. An alternative measure involves using a common covariance estimate on both sides, calculating the probability

$$(14)\quad [\hat{\boldsymbol{\beta}}(B_n) - \hat{\boldsymbol{\beta}}(B_E)]'\mathbf{V}(\hat{\boldsymbol{\beta}}(\Omega_1))^{-1}[\hat{\boldsymbol{\beta}}(B_n) - \hat{\boldsymbol{\beta}}(B_E)] \geq \hat{\boldsymbol{\beta}}(B_E)'\mathbf{V}(\hat{\boldsymbol{\beta}}(\Omega_1))^{-1}\hat{\boldsymbol{\beta}}(B_E).$$

where $\mathbf{V}(\hat{\boldsymbol{\beta}}(\Omega_1))$ is the covariance of $\hat{\boldsymbol{\beta}}(B_n)$ across the entire universe of draws from $F_1$. Again, in the univariate case one can cancel the common denominator on both sides of the equation and see that this reduces to a comparison of squared coefficient deviations from the null, so I refer to this as the bootstrap-c. As in the case of randomization tests above, I use the coefficients of the 10000 bootstrap samples to approximate $\mathbf{V}(\hat{\boldsymbol{\beta}}(\Omega_1))$.

As explained by Hall (1992), while the coverage error in a one-sided hypothesis test of a single coefficient of the bootstrap-t converges to its nominal size at a rate $O(n^{-1})$, the coverage error of the bootstrap-c converges at a rate of only $O(n^{-1/2})$, i.e. no better than the standard root-n convergence of asymptotic normal approximations. The reason for this is that the distribution of the studentized coefficient (the t-statistic) is asymptotically pivotal, i.e. does not depend upon unknowns. In contrast, the distribution of the coefficient itself is not pivotal, as it depends upon the estimation of its variance, which imparts additional inaccuracy. In the finite sample, this translates into a rejection bias borne of the unaccounted for sampling variation of $\mathbf{V}(\hat{\boldsymbol{\beta}}(\Omega_1))$, as can be seen more clearly by considering Stata's default bootstrap.

In the three papers where authors use the bootstrap in my sample, they defer to Stata's default approach, which is a form of the bootstrap-c with the addition of a normality assumption. Stata draws random samples from the regression sample, calculates the covariance matrix of the bootstrapped coefficients and uses it to report the standard errors and, based on the normal distribution, p-values of individual and joint coefficient tests. If the bootstrapped coefficients are actually normally distributed, these p-values amount to calculating the probability

$$(15)\ [\hat{\boldsymbol{\beta}}(\mathbf{B_n}) - E(\hat{\boldsymbol{\beta}}(\mathrm{F}_1))]'\mathbf{V}(\hat{\boldsymbol{\beta}}(\Omega_1))^{-1}[\hat{\boldsymbol{\beta}}(\mathbf{B_n}) - E(\hat{\boldsymbol{\beta}}(\mathrm{F}_1))] > \hat{\boldsymbol{\beta}}(\mathbf{B_E})'\mathbf{V}(\hat{\boldsymbol{\beta}}(\Omega_1))^{-1}\hat{\boldsymbol{\beta}}(\mathbf{B_E})$$

which is essentially the bootstrap-c.[37] As I show in the size analysis of authors' methods further below, this leads to systematic over-rejection of the null. The problem is not the normality assumption, but the fact that just as $\hat{\boldsymbol{\beta}}$ based on F$_1$ drawn from F$_0$ has sampling variation, so too does $\mathbf{V}(\hat{\boldsymbol{\beta}}(\Omega_1))$ based on F$_1$ drawn from F$_0$.[38] In essence, if one wishes to use the bootstrap-c, a degrees of freedom adjustment is necessary, which can be arrived at by (the rather costly) bootstrapping of the bootstrap to determine the sampling variation of $\mathbf{V}(\hat{\boldsymbol{\beta}}(\Omega_1))$. This problem does not exist in the randomization-c, because the distribution of the test statistic is not motivated by a sampling framework: $\Omega$ (the universe of potential randomization outcomes) and $\mathbf{V}(\hat{\boldsymbol{\beta}}(\Omega))$ are fixed in the thought experiment that generates the distribution of the test statistic. Given its default popularity, I make use the bootstrap-c in this paper and show that it generates systematically higher rejection rates than the bootstrap-t. It still, however, registers fewer rejections of the null than authors' methods.

Regarding practical methods, I implement the bootstrap in a manner that follows the error structure indicated by authors' methods. Thus, if authors' cluster, I draw in clusters, but if they do not, I sample individual observations. In evaluating the results of individual equations, the bootstrap is limited to the observations and clusters present in that equation alone, so that each iteration has the same number of observations and clusters. However, to implement an omnibus version of the bootstrap-c, to complement that done with the randomization-c, I bootstrap the entire experimental sample to allow the estimation of the empirical covariance of all coefficients in the paper. In this case the number of observations or clusters used in each equation varies as sampling results in variation in the number of "not-available" entries for the covariates used in different equations. As in the case of the randomization-c, iteration specific covariance estimates

[37] The only difference being the centering of the distribution around its mean rather than the parameter of the parent population, but in practice I find the difference is usually negligible.

[38] It is important to emphasize that the problem here is not the number of bootstrap draws used to calculate $\mathbf{V}(\hat{\boldsymbol{\beta}}(\Omega_1))$. More draws provide a more accurate measure of $\mathbf{V}(\hat{\boldsymbol{\beta}}(\Omega_1))$, but do not change the fact that it varies with F$_1$, i.e. has a sampling distribution.

for the bootstrap-t are calculated using authors' methods.[39] P-values are evaluated using the experimental test statistic and the bootstrapped draws by applying equation (10) earlier. This provides p-values that are uniformly distributed, subject to the maintained (but only asymptotically accurate) assumption that the bootstrap distribution is the true distribution of the test statistic under the null.

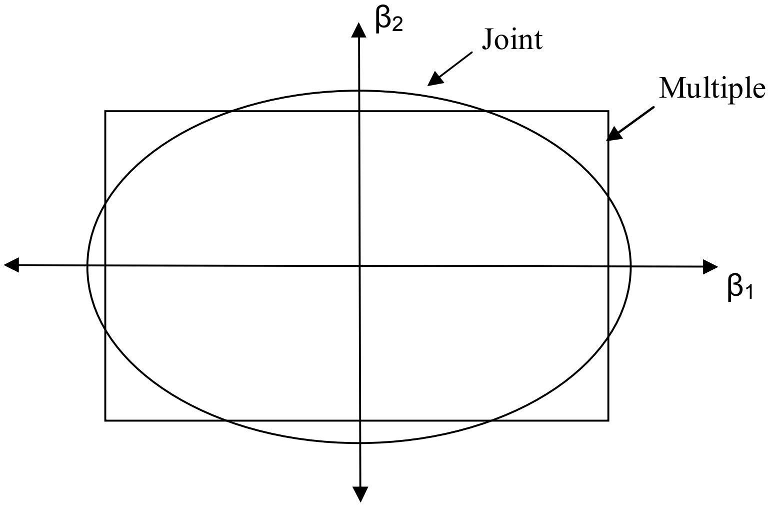
**(d) Joint vs Multiple Hypothesis Testing**

I use Wald statistics to test the joint hypothesis that all of the treatment coefficients in an equation or paper are equal to zero. This test either cannot reject the null, allowing the conclusion that all coefficients are zero, or rejects the null, allowing the conclusion that some unspecified subset of the coefficients is not equal to zero. An alternative approach is to simultaneously test whether each coefficient is equal to zero, allowing for the rejection or acceptance of the null for each coefficient individually, but taking into account the growing possibility of Type I errors created by the repeated drawing of test statistics. Since multiple testing of this sort increases power in the identification of alternatives that might be of greater interest to authors, I use it as a complement to joint tests in the analysis below.

Figure I illustrates the case where one is interested in testing the significance of two coefficients whose distribution is known to be normal and independent of each other. The oval drawn in the figure is the Wald acceptance region for the joint significance of the two coefficients, while the rectangle is the acceptance region for the two coefficients tested individually. In the multiple testing framework, to keep the probability of one or more Type I errors across the two tests at level $\alpha$, one could select a size $\eta$ for each test such that $1\text{-}(1\text{-}\eta)^2 = \alpha$. The probability of no rejections, under the null, given by the integral of the probability density inside the rectangle, then equals $1\text{-}\alpha$. The integral of the probability density inside the Wald ellipse is also $1\text{-}\alpha$. The Wald ellipse, however, has the property that it is the minimum area in the two dimensional space such that the probability of falling in the acceptance region is $1\text{-}\alpha$. It achieves this, relative to the multiple testing rectangle, by dropping corners, where the probability of two extreme outcomes is low, and increasing the acceptance region along the axes. If one

[39] Thus, in the particular case where the authors use the bootstrap, I bootstrap the bootstrap. This ends up revealing the shortcomings of the bootstrap-c, as shown further below.

Figure I: Acceptance Regions for Joint and Multiple Testing with Independent Estimates



thinks of alternatives as randomly falling anywhere in the two-dimensional space, the ability (power) of the joint testing framework to achieve a rejection when the joint null is false is higher, because of its smaller overall acceptance region. If one thinks of alternatives as falling along the axes, i.e. some nulls are true while others are false, the ability of the multiple testing framework to achieve a rejection when the joint null is false is higher, because of the smaller length of the acceptance region along the axes. As shown shortly, multiple testing frameworks may possess less or more power than suggested by this stylized example, but the basic intuition provided by the diagram, that Wald joint tests try to maximize power for general alternatives and multiple testing frameworks try to maximize power for alternatives that lie along the axes, both while controlling the probability of a Type I error under the null, carries through.

Multiple testing is an evolving literature. The classical Bonferroni method keeps the probability of any Type I error in N tests at or below $\alpha$ by evaluating each test at the $\alpha/N$ level. Holm's (1979) refinement increases power while maintaining the upper bound on size by sorting the p-values of the tests in ascending order $p_1 \leq p_2 \dots \leq p_N$ and rejecting all hypotheses s for which $p_j \leq \alpha/(N-j+1)$ for all $j \leq s$, i.e. moving down the list, evaluating each $p_j$ against $\alpha/(N+j-1)$, and stopping on the first failure to reject. Neither the Bonferroni nor the Holm method makes use of information on the covariance of the test statistics, and hence both are conservative, i.e. the probability of a Type I error is generally below the nominal size of the test. Romano and Wolf

(2005a, 2005b), extending earlier work of White (2000), have recently introduced a method that attempts to increase power by using covariance information to control the probability of a Type I error at exactly α. The method involves the following steps:

(1) Order the absolute value of the individual t-statistics in descending order.

(2) Use bootstrap or randomization inference to calculate the distribution of the maximum t-statistic of the individual tests.

(3) Reject the hypothesis with the maximum t-statistic if its t-statistic exceeds the $\alpha^{th}$ percentile of the upper tail of the maximum distribution.

(4) If (3) involves a rejection, repeat (1)-(3) using the remaining test statistics; if not, stop.

Since the method evaluates the realized experimental maximum or minimum of a test statistic against the distribution of that maximum or minimum under the null, it has size α at each step.[40] Cumulated across all steps the probability of a Type I error is still α because at each step either the null was true and the procedure generated a Type I error with probability α, or the null was false and a Type I error has not occurred.

While attempting to improve on other multiple testing methods by using covariance information, in practice the power of Romano-Wolf's approach may be no better or even inferior. The reason for this is that in practical application the distribution of t-statistics varies dramatically across estimating equations and *within* estimating equations.[41] If one coefficient in the test has a particularly extreme distribution, then it will determine the maximum distribution. If that coefficient, however, has a modest realized t-statistic within its own distribution, it will fail to reject and the testing procedure will stop there, even if the other t-statistics attain critical values

[40] In finite samples in the case of randomization inference, and asymptotically in the case of bootstrap inference. An additional distinction is that bootstrap inference has strong control of the error rate, i.e. the probability of a Type I error is α even if some of the nulls are false, while randomization inference generally has weak control of the error rate, i.e. the calculated distribution of the test statistics and hence probability of a Type I error depends upon all nulls being true. Romano and Wolf (2005a) discuss conditions under which randomization inference allows for strong control, but these generally do not apply in my sample papers. These distinctions aside, in my sample bootstrap and randomization inference produce very similar results.

[41] While in the ideal OLS case the degrees of freedom associated with any individual coefficient test within an equation is the same, this is not generally true. As discussed above, and shown more fully in Young (2016), the effective degrees of freedom of hypothesis tests using the robust or clustered covariance matrix varies with the test. Hence, the distribution of the t-statistic varies across coefficients within equations. Not using the robust or clustered covariance matrices in situations where the disturbances are not ideal does not improve matters either. For example, if there are unaccounted for cluster level random effects, the default OLS standard error estimate will understate the standard error more for regressors that are more correlated at the cluster level. Consequently, the actual (bootstrap or randomization inferred) distribution of t-statistics will vary at the coefficient level.

within their own individual (less extreme) distributions. In contrast, if the Bonferroni method had been used each t-statistic would have been evaluated on its own merits and more rejections might have occurred.

In order to avoid this problem, and give individual experimental results the best chance of rejecting the null of no effects, in this paper I develop and use a modification of the Romano-Wolf method which solves the problem of varying distributions. The procedure is as follows:

(1) Use bootstrap or randomization inference to evaluate the p-value of each coefficient t-statistic. Similarly, evaluate the p-value of the t-statistics for each randomization or bootstrap draw using their randomized or bootstrapped distribution producing, by construction, uniformly distributed p-values.

(2) Order the p-values of the estimated coefficients in ascending order.

(3) Calculate the distribution of the minimum p-value of the individual tests based upon the joint distribution of p-values indicated by the randomization and bootstrap draws.

(4) Reject the hypothesis with the minimum p-value if its p-value is less than the $\alpha^{th}$ percentile of the lower tail of the minimum distribution.

(5) If (4) involves a rejection, repeat (2)-(4) using the remaining test statistics; if not, stop.

In this framework, each test statistic (calculated p-value) has a uniform distribution, so the procedure is not dominated by the extreme distributions of individual coefficients. It is critical, in implementing this procedure, to calculate the p-values using the bootstrapped or randomized distribution and not simply by evaluating the estimated t-statistics using their assumed asymptotic distribution since, as shown throughout this paper, this produces decidedly non-uniform p-values with distorted tail probabilities, i.e. retains the problem of extreme distributions that can affect the Romano-Wolf procedure.

It is instructive to apply the uniform p-value version of the Romano-Wolf method to the problem examined in Figure I. Since the coefficient estimates are independent, the uniform distributions of the calculated p-values will be independent. The probability their minimum is less than or equal to $\eta$ is thus given by $1-(1-\eta)^2$. To attain size $\alpha$, one must select an $\eta$ such that $1-(1-\eta)^2 = \alpha$. This is precisely the cutoff used in the analysis described earlier in Figure I. If, however, the first step rejects, then in the second step the acceptance region (on the remaining axis) is tightened, because, following the distribution of the minimum of a single p-value, a cutoff level $\eta$ such that $1-(1-\eta) = \eta = \alpha$ is used. Thus, this multi-step testing procedure actually has greater power than the single step multiple testing procedure described by the rectangle in Figure

I. I use this method in tables below, where in some cases it produces higher rejection rates than other multiple testing methods, although these, in the context of all of the seemingly significant coefficients reported in papers, remain remarkably low.

## IV: Results

This section reports the empirical results of the paper. I begin by laying out the relative number of significant results found using conventional, randomization and bootstrap methods in tests of individual coefficients and joint tests of treatment significance at the regression and paper level. Although the baseline results test all treatment outcomes and coefficients present in the paper, I use a variety of alternative samples and procedures to sift out measures that are clearly of primary interest to authors to show that the substantially reduced significance rates found in joint tests reflect the large amount of multiple testing implicitly taking place in experimental papers, and not some mistaken focus on testing irrelevant outcomes and experimental details. I then use bootstrap samples drawn from the experimental samples to explore the size and power characteristics of conventional and randomization tests. The size bias of conventional tests is linked to the bias and variance of the coefficient variance estimates, which in turn is linked to bounds determined by maximum leverage, confirming the theory sketched earlier above. The relative power of conventional and randomization tests is shown to be determined by the size bias of the former and differences between the clustering decisions of authors and the groupings in which treatment is actually applied in experiments. Power is virtually equalized when adjustments are made to clustering or the level at which randomization is putatively carried out. Such adjustments, however, have little effect on the results for the papers themselves, suggesting that power is not the key issue.

### (a) Significance Rates

Table III summarizes the statistical significance of treatment effects using different criteria. In the upper left-hand panel we see that of the 5880 treatment coefficients appearing in the 53 papers, using authors' methods 751 and 1459 are found to be significant at the .01 and .05 levels, respectively. When randomization tests are applied, the number of significant coefficients

Table III: Statistical Significance of Treatment Effects at Different Levels

| | coefficients | | multi-treatment regressions | | papers | |
|---|---|---|---|---|---|---|
| | .01 | .05 | .01 | .05 | .01 | .05 |
| Full Sample: 5880 coefficients, 1009 multi-treatment regressions, 53 papers | | | | | | |
| significant coefficient | 751 | 1459 | 355 | 586 | 50 | 52 |
| standard Wald test | | | .88 | .75 | | |
| randomization-t | .78 | .88 | .64 | .60 | | |
| randomization-c | .82 | .87 | .67 | .60 | .38 | .46 |
| bootstrap-t | .85 | .89 | .65 | .59 | | |
| bootstrap-c | .90 | .95 | .83 | .69 | .52 | .69 |
| Primary Treatment: 1701 coefficients, 245 multi-treatment regressions, 36 papers | | | | | | |
| significant coefficient | 340 | 581 | 98 | 144 | 31 | 34 |
| standard Wald test | | | .90 | .76 | | |
| randomization-t | .80 | .88 | .73 | .68 | | |
| randomization-c | .84 | .89 | .73 | .73 | .39 | .50 |
| bootstrap-t | .79 | .89 | .79 | .69 | | |
| bootstrap-c | .85 | .93 | .83 | .70 | .61 | .76 |

Notes: .01/.05 = level of the test. Top rows report number of significant results evaluated using authors' methods; values in lower rows are number of significant results evaluated using indicated method divided by the top rows. For multi-treatment regressions and papers top rows indicates number with at least one significant treatment coefficient at the level specified. Standard Wald test = p-value evaluated using authors' chosen covariance estimate and distribution (F or Chi$^2$); randomization and bootstrap-t = significance evaluated using studentized measures based upon authors' covariance estimation methods; randomization and bootstrap -c = significance evaluated using distribution of coefficients. Bootstrap and randomization-t cannot be calculated at the paper level as it is generally not possible to calculate an iteration specific cross-equation covariance matrix. See Section III above for a full description of randomization and bootstrap methods.

falls to .78 to .82 and .87 to .88 of the numbers found using authors methods at the two levels.[42] The middle panel examines the 1009 regressions in the papers with multiple treatment effects. Of these, 355 and 586 have at least one significant treatment coefficient at the .01 and .05 levels, which, given the almost universal absence of F-tests, might lead readers to conclude that treatment was having significant effects. When conventional F/Wald tests using the authors' covariance calculation methods are applied, the number of regressions that can reject the null of

[42] The results in lower rows are not proper subsets of the significant results reported in the top row, but are close to being so. Thus, for example, of the 586 coefficients the randomization-t finds to be significant at the .01 level, all but 35 are found to be significant at that level using authors' methods.

zero treatment effects falls to .88 and .75 of original authors' reports at the two levels. Randomization tests reduce the relative number of significant results further, to .64 to .67 at the .01 level and .60 at .05. The upper right hand panel of the table reports tests of the joint significance of all treatment coefficients in each paper. As shown, 50 papers have at least one .01 significant coefficient and 52 papers have at least one .05 significant coefficient, leading readers, in the absence of joint tests, to conclude that treatment was having statistically significant effects. When randomization tests are used to evaluate the joint null of no effect whatsoever anywhere in the experiment, the number of rejections falls to .38 and .46 of reported results at the .01 and .05 levels, respectively. As can be seen by the ranges just described, there is generally little variation in results across the -t and -c versions of the randomization tests, which produce similar rejection rates.

Table III also presents bootstrap counterparts of the randomization tests. Among these, the bootstrap-c tends to show the highest rejection rates. As shown further below in the size analysis of authors' methods, this approach is biased in favour of rejection, as it does not account for the sampling variation of the variance estimate. Even the bootstrap-c, however, shows substantially fewer significant results than indicated by conventional tests using authors' methods. The studentized refinement, the bootstrap-t, goes further, producing rejection rates that are very similar to those of the randomization tests. The bootstrap-t represents the best practical attempt to evaluate the actual finite sampling distribution of the test statistics given the population characteristics of the experimental sample, i.e. the non-normality, heteroskedasticity and cross-correlation (if any) of errors and the heterogeneity (if any) of treatment effects. It indicates that virtually **all** of the gap between conventional and randomization results can be attributed to the misspecification of the sampling distribution of the t and Wald statistics of conventional tests.

The lower panel of Table III focuses on a subset of "primary treatment" coefficients and regressions in the sample papers. In the full sample I comprehensively analyse all treatment measures that were randomized and whose coefficients can be tested using randomization inference. Authors might object that this includes trivial experimental details of little interest to them, such as, in the extreme, the "puzzle type" received by participants. The primary treatment sample addresses this objection by only looking at regressions where treatment measures divide

Table IV: Regression Equations by Joint Statistical Significance of Treatment Effects
(all regressions with more than 1 treatment variable)

| | | at .01 level | | | | at .05 level | | |
|---|---|---|---|---|---|---|---|---|
| | | Wald test with diagonalized covariance matrix | | | | | | |
| | | No | Yes | Total | | No | Yes | Total |
| Standard Wald test | No | 629 | 66 | 695 | No | 496 | 76 | 572 |
| | Yes | 74 | 240 | 314 | Yes | 47 | 390 | 437 |
| | Total | 703 | 306 | 1009 | Total | 543 | 466 | 1009 |

Note: Yes/No = significant or not at the level specified; diagonalized covariance matrix = off diagonal terms set to zero. All tests are conventional Wald tests, using authors' covariance calculation methods and evaluated using the F or Chi$^2$ distribution following authors' methods.

the experimental population into mutually exclusive treatment groups,[43] and only cases where such regressions account for at least 1/3 of all treatment regressions in the paper. Given the prominence they are being given in presentation and analysis, such treatment measures cannot be trivial in the minds of the authors. As shown in the table, relative rejection rates for the primary treatment sample are very much the same as in the full sample, particularly at the coefficient and full paper level. The only apparent difference is at the equation level, where relative rejection rates are perhaps .1 higher in the randomization and studentized bootstrap tests. As explained in the Introduction, throughout this paper I follow transparent rules rather opaque discretion. It is certainly true that this leads in some cases to the testing of trivial experimental details. These trivial details, however, are as often as not just as significant as primary treatment effects, so the main results reported above cannot be attributed to a failure to apply discretionary judgment.

Table IV provides some insight into the reduction in significance levels found in joint tests at the equation level. For the conventional Wald test, I recalculate p-values using diagonalized versions of the estimated covariance matrices, i.e. matrices where the off-diagonal

[43] Thus, if there are two binary treatment variables ($T_1$,$T_2$), and one group receives (1,0), another (0,1) and a control (0,0), the regression is a primary treatment regression as each group receives one and only one treatment. If a fourth population group receives (1,1), the regression is no longer a primary treatment regression, as it is not clear if there is any intended hierarchy of effects in the minds of authors (e.g. $T_1$ is the primary treatment and $T_2$ something they consider comparatively trivial). However, if authors code receiving both $T_1$ and $T_2$ as $T_3 = 1$, the regression is a primary treatment regression as the four groups in the regression are coded (1,0,0), (0,1,0), (0,0,0) and (0,0,1). In this, as elsewhere in this paper, I follow rules and defer to the judgement of authors: when do they decide to present treatment effects in a manner that divides the sample into mutually exclusive groups, each with a separately identified treatment regime represented by a separate coefficient.

Table V: Joint Statistical Significance at the Paper Level
(significant papers as a fraction of those reporting a significant treatment effect)

| | using .05 significant dependent variables only | | | | using equation (block) diagonalized covariance matrix | | | |
|---|---|---|---|---|---|---|---|---|
| | full sample | | primary sample | | full sample | | primary sample | |
| | .01 | .05 | .01 | .05 | .01 | .05 | .01 | .05 |
| randomization-c | .42 | .54 | .42 | .62 | .62 | .73 | .58 | .71 |
| bootstrap-c | .54 | .71 | .65 | .76 | .76 | .85 | .74 | .85 |

Notes: Unless otherwise noted, as in Table III. Reported figures are significant results as a fraction of the papers with at least one significant coefficient at the specified level (top rows, right hand panel, Table III).

covariance terms are set to zero. As shown, there is a remarkable similarity between the acceptance and rejection of the null hypothesis using these artificial covariance matrices and those calculated using the estimated covariance matrix. In 1009 regressions with more than one treatment variable, statistical significance only differs in 140 cases at the .01 level and overall rejection rates are roughly the same. The mean p-value using the estimated covariance matrix is .228, using the diagonalized covariance matrix it is .238, and the correlation between the two is .917. This tells us that the typical treatment covariance matrix is close to being diagonal, that is, the typical regression treatment design basically involves a series of (conditional on covariates) mutually orthogonal regressors producing a series of uncorrelated test statistics. In regressions with more than one treatment variable there are on average 4.9 treatment measures, with a median of 3 and with 25 percent of these regressions having 6 or more and 5 percent having 16 or more treatment measures. Table IV shows that, under the null the .01 and .05 coefficient significance levels reported in these papers represent multiple independent rolls of 100-sided and 20-sided dice, and should be discounted accordingly. The F/Wald test of joint significance does so with a minimum volume ellipse that maximizes power against a general null. I show below that multiple testing methods designed to maximize power along the axes produce similar results.

Table V provides insight into the large reduction in significance levels found in joint tests at the paper level using two variations. I begin, in the left hand panel, by only including regressions with a dependent variable that generates a .05 significant treatment coefficient in some regression in the paper. This addresses the critique that the omnibus paper level test produces low significance rates by including irrelevant outcomes that authors never believed

Table VI: Average Cross-Equation Wald P-Value Correlation

| | all equations | with .01 significant treatment coefficients | without .01 significant treatment coefficients |
|---|---|---|---|
| randomization | .208 | .390 | .159 |
| bootstrap | .223 | .360 | .163 |

Notes: Off-diagonal elements of the correlation matrix for equation level conventional Wald p-values calculated across 10000 randomization or bootstrap draws, averaged at the paper level and then averaged across papers and reported in the table. All 53 papers used in the first column, but only 46 and 44 papers in the second and third columns as some papers do not have more than one equation with the listed characteristic.

would be affected by experimental treatment. Despite the pre-selection on significance, relative significance rates generally only rise a few percentage points relative to those reported earlier in the right-hand panel of Table III. The low rates of rejection in the omnibus tests are not the result of mixing dependent variables that never generate a significant coefficient with variables that consistently generate significant results. Rather, it is the case that dependent variables that produce significant treatment coefficients in some regressions do not generate significant treatment effects in others. Once again, this shows that it is not my application of indiscriminate rules, testing all reported treatment outcomes, that generates the results of this paper.

The right hand panel of Table V recalculates the paper-level Wald statistics using equation (block) diagonalized covariance matrices, i.e. acting as if each equation contains independent information. As shown, this has a much larger effect on significance rates, raising them .19 to .27 percentage points above those reported in Table III in the case of the randomization-c. Table VI explains why this happens by calculating the cross-equation correlation of the equation level bootstrap or randomization draw conventional Wald p-value.[44] As shown, the average randomization correlation of p-values is .208, but between equations which have a .01 significant treatment coefficient it is .390, while between equations that have no .01 significant treatment coefficients it is .159. The bootstrap draws show a similar pattern. As noted earlier, the average paper has 10 equations with a .01 significant treatment effect and 27 equations without any .01 significant treatment effects. Unrecognized by readers, and probably authors as well, the small number of equations with significant results are highly correlated,

[44] The randomization p-value is the test of the null in each randomized sample that all treatment coefficients in the equation are equal to zero, the bootstrap p-value is the test of the null in each bootstrapped sample that they are equal to the originally estimated values (which is the true null for the population represented by the original data).

Table VII: Statistical Significance at the Coefficient Level
with Multiple Testing Type I Error Control

| | (1) Full sample:<br>5880 coefficients<br>in 53 papers | | (2) Primary Sample:<br>1701 coefficients<br>in 36 papers | | (3) Papers'<br>significant<br>coefficients only | |
|---|---|---|---|---|---|---|
| | .01 | .05 | .01 | .05 | .01 | .05 |
| significant coefficients | 751 | 1459 | 340 | 581 | 751 | 1459 |
| Holm: | | | | | | |
| paper's p-value | .34 | .23 | .43 | .35 | .52 | .33 |
| randomization-t p-value | .15 | .16 | .17 | .24 | .34 | .24 |
| randomization-c p-value | .18 | .19 | .23 | .30 | .38 | .27 |
| bootstrap-t p-value | .18 | .18 | .18 | .25 | .38 | .26 |
| bootstrap-c p-value | .24 | .21 | .25 | .29 | .46 | .29 |
| Romano-Wolf t-stat: | | | | | | |
| randomization t-stat | .22 | .18 | .31 | .30 | .38 | .28 |
| bootstrap t-stat | .19 | .16 | .28 | .27 | .35 | .26 |
| Romano-Wolf uniform p | | | | | | |
| randomization p-value | .17 | .19 | .22 | .30 | .43 | .30 |
| bootstrap p-value | .18 | .18 | .19 | .25 | .41 | .28 |

Notes: .01/.05 = probability of a Type I error. Top row reports number of significant coefficients evaluated individually using paper's methods at the .01 and .05 levels. Values in lower rows are number of significant coefficients found using the indicated multiple testing procedure with the probability of a Type I error controlled at the .01 and .05 levels. Holm, Romano-Wolf t-stat and Romano-Wolf uniform p as described earlier in Section III. Samples as described in text above.

indicating the repetition of a limited amount of information, while the much larger number of insignificant equations are relatively uncorrelated, presenting a broad mass of independent information. The low rejection rates found in the omnibus paper-level test of overall experimental significance reflect this evidence.

Table VII tests significance at the coefficient level using multiple testing procedures that control the overall probability of one or more Type I errors. I report results using the full sample and primary sample of Table III, and also a third sample composed only of treatment coefficients that were found to be significant at the .01 or .05 levels in the papers themselves. In this last I do my very best to narrow the focus to results that authors, almost surely, felt were relevant. In the top row I report the number of significant coefficients found in the paper at the level specified, while the lower rows report the relative number of significant coefficients found using a multiple testing procedure that controls the overall probability of a Type I error at the same level. I use

Holm's (1979) procedure, Romano-Wolf's t-statistic procedure and my modification of Romano-Wolf using uniform p-values, as described earlier above.

As shown in the table, taking into account the number of times the p-value dice are rolled dramatically reduces the statistical significance of reported results. Although the results vary slightly by method, with the bootstrap-c again showing the highest rejection rates, the overall pattern is fairly clear. Only about .20 of treatment effects are found to be significant in the full sample, between .20 and .30 in the primary sample and between .25 and .40 in the significant sample. The last result is particularly remarkable because the tests are pre-selected on the basis of having an individually significant result in the first place. Table VII highlights the importance of using randomization methods in evaluating p-values. If papers' p-values are used in the Holm procedure, rejection rates are about .2 to .25 higher at the .01 level than those found using randomization methods.[45]

Table VIII reports the relative number of multi-treatment equations or papers in which at least one significant coefficient is found, in effect rejecting the null of no effect anywhere, after Table VII's adjustment for multiple testing. Multiple testing finds a significant coefficient in only about 50 to 60 percent as many equations as report a conventionally significant coefficient in the full sample and 60 to 70 percent as many equations in the primary sample. This is comparable to the relative rejection rate found in joint tests in Table III earlier. At the paper level, multiple testing finds some significant effect somewhere somewhat more often than joint testing, but the relative rejections rates are still very low (40 to 50 percent as often at the .01 level and 60 to 70 percent as often at the .05 level). These results, as well as others in this section, show that the low relative significance rates I find using randomization and bootstrap tests do not come from testing trivial hypotheses of no interest to authors. I have used multiple testing procedures to increase power on the axes, examined only primary treatment measures that are emphasized in presentation, limited myself only to dependent variables that generate significant results somewhere in the paper, and even focused (in multiple testing) only on individually significant coefficients. There simply is a vast amount of multiple testing in the typical

[45] The Romano-Wolf procedure requires estimates of the covariance of t-statistics and p-values for individual coefficients drawn from multiple equations and hence is not directly implementable using conventional methods.

Table VIII: Significant Multi-Treatment Regressions and Papers

| | multi-treatment equations (1009 full, 245 primary) | | | | papers (53 full, 36 primary) | | | |
|---|---|---|---|---|---|---|---|---|
| | full sample | | primary sample | | full sample | | primary sample | |
| | .01 | .05 | .01 | .05 | .01 | .05 | .01 | .05 |
| significant coefficient | 355 | 586 | 98 | 144 | 50 | 52 | 31 | 34 |
| Holm: | | | | | | | | |
| paper's p-value | .68 | .65 | .82 | .74 | .64 | .81 | .61 | .74 |
| randomization-t p-value | .54 | .54 | .61 | .67 | .36 | .58 | .42 | .56 |
| randomization-c p-value | .58 | .55 | .70 | .69 | .42 | .56 | .39 | .53 |
| bootstrap-t p-value | .61 | .59 | .71 | .71 | .46 | .60 | .48 | .59 |
| bootstrap-c p-value | .66 | .63 | .74 | .72 | .52 | .67 | .52 | .62 |
| Romano-Wolf t-stat: | | | | | | | | |
| randomization t-stat | .52 | .55 | .62 | .66 | .42 | .63 | .45 | .62 |
| bootstrap t-stat | .56 | .58 | .70 | .72 | .36 | .54 | .42 | .56 |
| Romano-Wolf uniform p: | | | | | | | | |
| randomization p-value | .53 | .57 | .61 | .69 | .36 | .67 | .48 | .65 |
| bootstrap p-value | .59 | .62 | .70 | .73 | .44 | .69 | .52 | .62 |

Notes: .01/.05 = probability of a Type I error. Top row reports number of equations or papers with at least one significant coefficient (evaluated individually) using paper's methods at the .01 and .05 levels. Values in lower rows are number of equations or papers in which a significant coefficient is found using the indicated multiple testing procedure with the probability of a Type I error controlled at the .01 and .05 levels.

experimental paper, without any information, in the form of joint or multiple tests, given to readers to help them interpret and evaluate the multiple p-value die rolls presented to them in tables.

**(b) Size and Power**

Table IX reports the average rejection rates at the .01 and .05 levels of conventional tests, using authors' methods, applied to bootstrapped samples drawn from the original experimental sample. In each case, the calculated p-value is for the test that the bootstrapped coefficient estimates equal that of the parent sample, i.e. a test of a null that is known to be true. The t and F/Wald statistics produced by these draws were used to provide the studentized bootstrap-t evaluation of the distribution of the test statistics of the original papers in the preceding section.[46]

[46] As noted earlier, the bootstrap seeks to evaluate the distribution of test statistics for $F_1$ drawn from $F_0$, $t(F_1|F_0)$, by drawing samples $F_2$ from $F_1$ and calculating the distribution of $t(F_2|F_1)$. The experimental samples consist of $F_1$ drawn from parent populations $F_0$. In $t(F_1|F_0)$ we are interested in testing whether $\beta$ equals zero for $F_0$. In $F_1$ we know the parameter $\beta$ equals the originally estimated coefficient, so on each bootstrapped iteration our

Table IX: Size: Rejection Rates of the Null When True Using Authors' Methods
(distribution estimated using 10000 bootstraps)

| covariance estimate used | # | coefficients (t-tests) | | | | # | regressions (F/Wald tests) | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | .01 level | | .05 level | | | .01 level | | .05 level | |
| | | mean | sd | mean | sd | | mean | sd | mean | sd |
| all | 5088 | .020 | .035 | .066 | .044 | 1009 | .054 | .098 | .112 | .115 |
| default | 1228 | .015 | .019 | .058 | .030 | 201 | .022 | .030 | .069 | .047 |
| clustered | 2675 | .025 | .042 | .073 | .051 | 499 | .073 | .122 | .137 | .139 |
| robust | 1477 | .017 | .028 | .060 | .034 | 181 | .056 | .091 | .114 | .112 |
| bootstrap | 407 | .017 | .019 | .059 | .036 | 126 | .028 | .021 | .080 | .038 |
| other | 93 | .047 | .080 | .101 | .094 | 2 | .009 | .003 | .044 | .009 |

Notes: # = number of coefficients or multi-treatment regressions. Other data reported are the mean and standard deviation (sd) of the rejection probability at the stated level.

I now use authors' methods to calculate the reported p-value of these test statistics on each draw, arriving at an estimate of their size distortion.

As shown, given the data generating process behind their data, as incarnated in their data itself, the methods used by authors generate biased and highly inaccurate coverage. At the .01 level the average treatment coefficient t-test has a rejection probability of .020. Moreover, with rejection rates for individual coefficients varying from 0 to .886, the standard deviation of size is .035, which is 35 times greater than the predicted .001 of an exact statistic given the 10000 draws used in each regression simulation in the table.[47] Bias is substantial using the default covariance estimate, which is not surprising as errors are often far from the iid ideal, but also in papers using clustered and robust methods. The authors' bootstrap, the non-studentized bootstrap-c with the addition of a normality assumption, does no better on average, and "other" methods, principally attempts at correcting the bias but not the variance of the clustered and robust methods, do spectacularly poorly, although these are concentrated in one paper which is unlikely to be representative.[48]

$t(F_2|F_1)$ is the test of whether the bootstrapped coefficient equals that value. From this we infer the distribution of $t(F_1|F_0)$ when it is re-centered around a true null of $\beta$=0 for $F_0$.

[47] As the standard deviation of the average realization of n 0/1 draws, each with a probability p, is sqrt(p(1-p)/n) = .001 for p = .01 and n = 10000.

[48] In Young (2016) I find that bias adjustment methods generally improve the performance of the clustered and robust covariance estimation methods, but not by enough to produce accurate or unbiased coverage. Accurate coverage requires making an adjustment for degrees of freedom, i.e. for the variance of the variance estimate.

Bias and inaccuracy compound with the dimensionality of the test, as in joint F/Wald tests in multi-treatment equations average coverage is .054 at the .01 level with a standard deviation 98 times that of the exact ideal. This is shown further by noting that in equations with 5 or more treatment coefficients the average rejection probability at the .01 level is .077 with a standard deviation of .123, while in equations with 10 or more coefficients the average rejection probability is .122 with a standard deviation of .161. After the transformation afforded by the inverse of the coefficient covariance matrix, the Wald test statistic is interpreted as being the sum of independently distributed squared random variables. As the number of such variables increases, the critical value for rejection is naturally increased. This requires, however, an accurate assessment of the probability each squared random variable can, by itself, attain increasingly extreme values. Proportionate bias in the estimation of probabilities for individual coefficients is, however, greater at the more extreme tails. At the .01 level, as already noted, the average rejection probability at the coefficient level is 2 times nominal size. At the .001 and .0001 levels, however, I find it is 6.3 and 32 times nominal size, respectively. Greater proportionate bias in the estimation of increasingly extreme tail probabilities translates into greater bias for given nominal size as more coefficients are jointly evaluated.

Non-normality plays virtually no role in producing the results recorded above. To show this, I take the bootstrapped coefficients, remove their mean and divide by their standard error, square the resulting "test" statistic, and evaluate it using the chi-squared distribution. If the coefficients are distributed normally, this test statistic will have a tail probability equal to nominal size. Across the 5880 treatment coefficients I find average "rejection" rates at the .25, .1, .05 and .01 levels of .245, .098, .050 and .011, respectively. The standard deviation of tail size at the .01 level is .003, only three times the level predicted in 10000 simulations per coefficient for a normally distributed variable. Stata's test of normality based upon skewness and kurtosis rejects the null at the .01 level for 43 percent of the 5880 treatment coefficients, so most coefficients are definitely not normally distributed. Practically speaking, however, the deviation from the normal distribution is unimportant.

According to the theory underlying the t-statistic, the variance estimate of the coefficient has a mean $m$ and variance $v$ equal to $2m^2/dof$, where *dof* equals degrees of freedom. Multiply by $2m/v$ and it is a chi-squared variable with *dof* degrees of freedom, mean *dof* and variance 2**dof*.

Table X: Bias in Conventional Variance and Degrees of Freedom Estimates by Distribution used in Stata to Evaluate the Test Statistic

| | evaluated using t-distribution | | | | | evaluated using $\chi^2$ distribution | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | $\ln\left(\frac{m[\hat{\sigma}(\hat{\beta})^2]}{\sigma(\hat{\beta})^2}\right)$ | | $\ln\left(\frac{dof_a}{dof_n}\right)$ | | | $\ln\left(\frac{m[\hat{\sigma}(\hat{\beta})^2]}{\sigma(\hat{\beta})^2}\right)$ | | $\ln\left(\frac{dof_a}{obs}\right)$ | |
| | # | mean | sd | mean | sd | # | mean | sd | mean | sd |
| all | 4191 | -.021 | .421 | -1.70 | 1.55 | 1689 | .423 | 3.49 | -2.13 | 2.18 |
| default | 875 | .027 | .432 | -1.02 | 1.68 | 353 | 1.84 | 7.32 | -3.09 | 2.82 |
| clustered | 1905 | -.071 | .159 | -1.84 | 1.31 | 770 | -.010 | .972 | -1.08 | 1.12 |
| robust | 1318 | -.012 | .095 | -1.70 | 1.29 | 159 | -.047 | .083 | -1.13 | 1.72 |
| bootstrap | | | | | | 407 | .194 | .248 | -3.70 | 1.90 |
| other | 93 | .435 | 2.32 | -5.06 | 2.69 | | | | | |

Notes: # = number of coefficients; $\sigma(\hat{\beta})^2$ = bootstrapped (actual) variance of coefficient estimates; $m[\hat{\sigma}(\hat{\beta})^2]$ = bootstrapped mean of coefficient variance estimate; $dof_n$, $dof_a$ = degrees of freedom, nominal (used by Stata) and actual (equal to $2(m^2/v)$, where $m$ and $v$ denote the mean and variance of the bootstrapped variance estimates); *obs* = observations or number of clusters where clustered.

So, one can take the estimates of variance, calculate their moments, multiply by $2m/v$ and evaluate the tail probabilities of the transformed variables using the chi-squared distribution with *dof* degrees of freedom. At the .25, .1, .05 and .01 level I find average "rejection" rates of .248, .102, .054 and .015, respectively. Although there is much more variation, with for example a standard deviation of "size" at the .01 level of .053, with the exception of the .01 tail probability these are not substantial deviations from the hypothesized distribution. Again, these calculations suggest that the assumption of normally based distributions is not the principal source of coverage bias in conventional test statistics. The problem, instead, lies in the parameters used to characterize these distributions.

Table X compares the estimated coefficient variance and degrees of freedom implied by the moments of the bootstrapped variance estimates to the actual bootstrapped coefficient variance and the conventionally assumed degrees of freedom used in the evaluation of the test statistics. I present results separately for coefficients which Stata evaluates using the t-distribution and those evaluated using the chi-squared distribution, and by the form of the covariance estimate used by authors. As shown, variance estimates are extraordinarily but not systematically biased, while actual degrees of freedom are systematically and substantially lower than nominally assumed.[49] The average bias of the variance estimate is substantially positive in particular subsamples, but the very large standard deviations suggest that it is often negative as well. For example, despite a mean bias of .423 in 1689 coefficients evaluated using the chi-squared distribution, the bias is actually negative in 1051 of these cases. Degrees of freedom, however, are almost always lower than nominally assumed or, in the case of coefficients evaluated using the asymptotic chi-squared distribution, lower than might be suggested by sample size. Actual degrees of freedom are only greater than nominally assumed in 274 of 4191 coefficients evaluated using the t-distribution[50] and greater than the number of observations in 88 of 1689 coefficients evaluated using the chi-squared distribution. Since degrees of freedom, estimated from the moments of each variance estimate, equal $2m^2/v$, or 2 times the squared inverse of the coefficient of variation, low degrees of freedom shows that variance estimates are much more volatile than nominally assumed.

Of particular interest in Table X are the averages for Stata's bootstrap where Stata, as discussed earlier, uses the bootstrap to calculate an estimate of coefficient variance and then evaluates the resulting test statistic using the asymptotic chi-squared distribution. The sampling variance of these variance estimates, revealed by bootstrapping these bootstraps, is so high that average degrees of freedom are only 26, despite an average of 690 observations per coefficient estimate. This reveals the danger of using non-pivotal bootstrap statistics: the test statistic depends fundamentally upon a calculated measure whose sampling distribution is unknown, making it impossible to accurately evaluate the test statistic without further refinement of the procedure (such as a bootstrap of the bootstrap). In contrast, in bootstrapping the distribution of a pivotal statistic, such as the t-statistic, there is no parameter whose distribution remains unknown at the end of the initial bootstrap. This is the finite sample manifestation of Hall's (1992) asymptotic convergence result noted earlier.

[49] Nominal degrees of freedom are those selected by Stata at the regression level and applied to each coefficient. Actual degrees of freedom can, however, vary at the coefficient level through the interaction between hypothesis tests and the structure of the covariance matrix as discussed, with examples, in Young (2016). Consequently, I calculate them at that level using the moments of the variance estimate for each coefficient.

[50] About 70 percent of these occur for the default covariance estimate or in non-OLS settings. As noted earlier in (8), actual degrees of freedom with ideal OLS iid errors are always lower than nominally assumed for the clustered and robust covariance estimates, but the measures in Table X are point estimates based upon the moments of distributions in environments with less than ideal errors. Even so, actual degrees of freedom are found to be greater than nominal for only 87 of 2652 OLS coefficients with clustered or robust covariance estimates.

Table XI: Ln Coverage Divided by Nominal Size for Coefficients as Determined by Variance and Degrees of Freedom Biases

| | test statistics evaluated using t-distribution | | test statistics evaluated using Chi$^2$ distribution | |
|---|---|---|---|---|
| | .01 level | .05 level | .01 level | .05 level |
| $\ln\left(\frac{m[\hat{\sigma}(\hat{\beta})^2]}{\sigma(\hat{\beta})^2}\right)$ | -.440<br>(.024) | -.270<br>(.015) | -.051<br>(.007) | -.050<br>(.004) |
| $\ln\left(\frac{dof_a}{dof_n}\right)$ | -.210<br>(.006) | -.107<br>(.004) | | |
| $\ln(dof_a)$ | | | -.098<br>(.008) | -.046<br>(.005) |
| constant | .017<br>(.014) | .004<br>(.009) | .758<br>(.039) | .353<br>(.027) |
| N | 4178 | 4181 | 1680 | 1686 |
| R$^2$ | .247 | .194 | .088 | .081 |
| $\mu_y$ | .384 | .192 | .314 | .135 |

Notes: Unless otherwise noted, as in Table X. $\mu_y$ = mean of dependent variable.

Table XI shows that coverage bias is determined by the variance and degrees of freedom biases recorded above. In the left panel I regress the ln coverage (rejection rate), as estimated and reported earlier in Table IX, divided by nominal size on the ln variance estimate bias and ln degrees of freedom bias for test statistics evaluated using the t-distribution. When the variance estimate is unbiased and actual degrees of freedom equal nominal, these two regressors equal zero. Thus, the constant term in the regression indicates how much ln proportionate coverage bias remains when there is no bias in these two elements. As shown in the table, the answer is close to zero at both the .01 and .05 levels. The R$^2$s of .247 and .194 are also substantial. Thus, these two biases explain all of the mean and much of the variation of coverage bias in test statistics evaluated using the t-distribution. For test-statistics evaluated using the chi-squared distribution, in the right hand panel, there are no nominal degrees of freedom, so one cannot speak of what would constitute “unbiased” degrees of freedom. The constant term divided by the coefficient on ln actual degrees of freedom entered as the regressor indicates at what degrees of freedom coverage bias goes to zero. In all four specifications shown in the table this is attained when ln actual degrees of freedom is approximately 8, or at about 3000 effective observations,

Table XII: Degrees of Freedom and Covariance Biases in Robust and Clustered OLS Regressions as Determined by Leverage and Sample Size

| | dependent variables | | | | | | |
|---|---|---|---|---|---|---|---|
| regressors | $\ln\left(\frac{dof_a}{dof_n}\right)$ | | | | $\ln\left(\frac{m[\hat{\sigma}(\hat{\beta})^2]}{\sigma(\hat{\beta})^2}\right)$ | | |
| $\ln\left(\frac{dof_{\min}}{dof_n}\right)$ | .366<br>(.018) | | .349<br>(.020) | .357<br>(.019) | | | |
| ln(observations) | | -.248<br>(.026) | -.046<br>(.026) | | | .028<br>(.003) | .035<br>(.003) |
| $\ln(\sigma^2_{\text{biasmin}})$ | | | | | .117<br>(.008) | | .105<br>(.008) |
| $\ln(\sigma^2_{\text{biasmax}})$ | | | | | .396<br>(.068) | | .672<br>(.070) |
| $I(\lambda^{\max}=0)$ | | | | -.080<br>(.057) | -.035<br>(.009) | | -.025<br>(.009) |
| constant | .080<br>(.080) | -.196<br>(.135) | .243<br>(.121) | .073<br>(.080) | -.024<br>(.006) | -.204<br>(.018) | -.218<br>(.020) |
| N | 984 | 984 | 984 | 984 | 984 | 984 | 984 |
| $R^2$ | .294 | .084 | .296 | .295 | .212 | .062 | .288 |
| $\mu_y$ | -1.54 | -1.54 | -1.54 | -.154 | -.059 | -.059 | -.059 |

Notes: Unless otherwise noted, as in Tables X and XI. $dof_{\min}$ = lower bound on actual degrees of freedom, $\sigma^2_{\text{biasmin}}$ and $\sigma^2_{\text{biasmax}}$ = minimum and maximum bounds on variance estimate bias; $I(\lambda^{\max}=0)$ = indicator for maximum leverage of 1. Following (7) and (8) earlier, allowing $\lambda^{\min}$ and $\lambda^{\max}$ to denote the minimum and maximum eigenvalues of the block diagonal matrix made up of the cluster sub-matrices of the hat matrix (equivalently, the minimum and maximum leverage values ($h_{ii}$) in the case of the robust covariance estimate), and c Stata's finite sample adjustment, we have $dof_{\min} = \max(1,1/\lambda^{\max}-1)$, $\sigma^2_{\text{biasmin}} = c(1-\lambda^{\max})$, and $\sigma^2_{\text{biasmax}} = c(1-\lambda^{\min})$. Where $\lambda^{\max}=1$, $\ln \sigma^2_{\text{biasmin}}$ is entered as 0 and the effect is captured by $I(\lambda^{\max}=0)$.

where the squared t (i.e. F) distribution is very close to the chi-squared.

Table XII shows that the theory outlined earlier above explains the degrees of freedom and variance estimate biases of the clustered and robust covariance estimates. Based upon (8), we see that, depending upon the particular hypothesis test, actual degrees of freedom relative to the nominal n-k or $n_c$-1 chosen by Stata will vary between 1 and a lower bound determined by maximal leverage in the regression. Based upon (7), we see that the covariance estimate bias will vary between upper and lower bounds determined by minimum and maximum leverage. Since these bounds are characteristics of regressions not coefficients, all of the right-hand side variables are constant within regressions, so I compute regression averages of the ln coefficient degrees of freedom and covariance biases and run them on ln regression characteristics, weighting by the inverse of the square root of the number of treatment coefficients (unweighted regressions are very similar).

As shown in the first column of the table, the theoretical lower bound on degrees of freedom is an important determinant of degrees of freedom bias, with an $R^2$ of .294. As maximal leverage goes to zero and the lower bound equals nominal degrees of freedom, the degrees of freedom bias goes to zero, as can be seen by comparing the minimal and insignificant constant term (.080, se .080) to the mean of the dependent variable (-1.54). In the second and third column we see that the actual number of observations or clusters is a poor predictor of degrees of freedom bias, with a lower $R^2$ when entered by itself (where it has a perverse sign!), and statistically insignificant when entered along with the lower bound implied by leverage. In the fourth column we see, using an indicator for a maximal leverage of 1, that these regressions have, if anything, slightly worse effects on degrees of freedom than predicted by the lower bound. This shows that regressions with a maximum leverage of 1 are not of the innocuous sort described by the extreme example given in the preceding section. To summarize, high leverage creates the possibility, depending upon the hypothesis test, of a reduction in effective degrees of freedom. As leverage rises, on average, across the treatment coefficients tested in the sample papers, effective degrees of freedom follow the lower bound down.

By the theory described in (7) earlier above, the bias of the robust and clustered covariance estimate varies between upper and lower bounds determined by maximal and minimal leverage. Column (5) of Table XII shows that, empirically, the ln bias follows the lower bound down and the upper bound up. An indicator for regressions with a maximal leverage of 1, where the ln lower bound is undefined and entered as zero in the regression, when compared with the coefficient on ln lower bounds which are not undefined, shows that these operate as if they had a lower bound of bias of -.300. As leverage goes to zero, the ln lower and upper bias bounds converge to 0, and the bias in the variance estimate, as indicated by the constant term (-.024) is close to zero. The ln number of observations, as before, is a poorer predictor of variance bias than minimum and maximum leverage, with an $R^2$ of .062 versus the .212 achieved using leverage based theory. In sum, leverage determines the degrees of freedom and variance biases of robust and clustered covariance estimates, which in turn determine the size distortions of conventional techniques, as shown earlier above. The actual degrees of freedom and variance bias can be calculated and predicted for any specific hypothesis test (Young 2016), but knowing the maximal leverage of the regression and the bounds it implies provides a quick and more accurate guide than sample size as to how bad things might, and in fact do, become.

Turning to power, Table XIII uses bootstrapped samples from each experimental paper to compare the power of conventional and randomization techniques to reject the null of no treatment effects when it is false. Since the samples are drawn from the actual experimental population, it measures the power of different techniques when the alternative of average treatment effects in the amounts estimated by the authors is true. Because of the computational cost of randomization tests, I am only able to perform 50-100 bootstrap samples per paper,[51] and use only 2000 randomization iterations (1000 for the more computationally costly equations) to calculate the randomization p-values. Bootstrap samples are drawn at the observation or cluster level, depending upon the covariance calculation method used in the paper.[52]

As shown in panels (a) of Table XIII, randomization tests have less power than the conventional methods used by authors with average rejection rates that are .03 to .04 lower at the coefficient level and .08 to .11 lower at the multi-treatment equation level. This difference in power is, however, concentrated in a particular subset of papers. In 12 of 15 laboratory experiments and 5 of 38 field experiments authors cluster all or some of the covariance matrices at a level of aggregation below that at which treatment is applied (e.g. clustering at the subject level, when treatment is applied to subjects in laboratory sessions or field districts).[53] In these regressions, as shown in the table, the difference in power is staggering. In contrast, in equations in which authors cluster at a level of aggregation above that at which treatment is applied, the

[51] As time goes by, and calculations complete, I hope to raise the number of bootstrap samples to 250 per paper. The results should not change substantially.

[52] When drawing bootstrap samples to evaluate conventional techniques alone (i.e. all tables up to this point), I draw bootstrap samples at the equation level, for the sample of the equation alone, sampling based upon the covariance methods used in each equation. When using bootstrap samples to evaluate the power and efficacy of randomization tests, I draw combined bootstrap samples at the paper level, so that I can implement my randomization code jointly and consistently on the new experimental data. Thus, if the paper clusters some equations but not others, I draw an experimental sample at the cluster level for all equations, although the covariance matrices of unclustered equations are still evaluated in the unclustered fashion used by authors.

[53] For the purposes of brevity, in the discussion which follows a paper that uses the robust covariance estimate or does not cluster at all is said to be "clustering" at the observation level. If treatment is applied at the observation level, they are clustering at the treatment level, otherwise they are clustering below treatment level.

Table XIII: Power: Average Rejection Rates of the Null on No Effect
When False by Nominal Size of Test using Authors' and Randomization Methods
(estimated using 50-100 bootstrapped samples of each experiment)

| | .01 | .05 | .01 | .05 | .01 | .05 | .01 | .05 |
|---|---|---|---|---|---|---|---|---|
| coefficients: | | | (a) authors' clustering | | | | | |
| | all (N=5880) | | below (N=497) | | above (N=803) | | neither (N=4580) | |
| authors' methods | .193 | .314 | .226 | .358 | .241 | .361 | .181 | .301 |
| randomization-t | .156 | .274 | .083 | .182 | .232 | .353 | .150 | .270 |
| randomization-c | .168 | .287 | .073 | .168 | .234 | .354 | .166 | .288 |
| | (b) adjusting authors' clustering | | | | | | (c) adjusting randomization | |
| | below (N=497) | | above (N=803) | | neither (N=4580) | | below (N=497) | |
| authors' methods | .257 | .392 | .237 | .352 | .181 | .301 | .207 | .336 |
| randomization-t | .130 | .253 | .238 | .350 | .152 | .271 | .171 | .293 |
| randomization-c | .126 | .243 | .246 | .361 | .168 | .289 | .181 | .303 |
| multi-treatment regressions: | | | (a) authors' clustering | | | | | |
| | all (N=1009) | | below (N=88) | | above (N=148) | | neither (N=773) | |
| authors' methods | .467 | .602 | .497 | .594 | .543 | .637 | .449 | .596 |
| randomization-t | .358 | .512 | .216 | .337 | .515 | .621 | .344 | .511 |
| randomization-c | .375 | .523 | .231 | .337 | .524 | .621 | .363 | .525 |
| | (b) adjusting authors' clustering | | | | | | (c) adjusting randomization | |
| | below (N = 88) | | above (N=148) | | neither (N=773) | | below (N = 88) | |
| authors' methods | .657 | .756 | .534 | .625 | .449 | .596 | .497 | .596 |
| randomization-t | .387 | .520 | .522 | .624 | .350 | .515 | .436 | .539 |
| randomization-c | .388 | .496 | .553 | .658 | .366 | .528 | .464 | .545 |

Notes: above/below/neither = authors cluster at a level of aggregation below, above or across/equal to that at which treatment is applied. Panels: (a) bootstrap and covariance calculation is done at authors' selected level of clustering; (b) bootstrap and covariance calculation is done at treatment level; (c) where authors consistently cluster below treatment level, treatment randomization done at author's level, otherwise as in panel (b); N = number of coefficients or equations across which averages are calculated.

difference in power is negligible, while in equations where they neither cluster above nor below the treatment level it is relatively small, particularly for the randomization-c.[54] The level at which clustering is applied has two implications for power: first, with more aggregation the statistical procedure acts as if there is less independent information in the sample, which should

[54] This includes two papers where the authors, apparently as a result of coding errors, cluster across treatment groups (in one case because they don't realize the cluster identifier is not unique and in the other because they mistakenly switch the cluster variable in half of their results).

reduce power, but with greater aggregation maximal leverage rises[55] increasing the small sample size distortions of asymptotic methods, which raises power; second, since the samples are bootstrapped at the cluster level, if clustering is done at a level other than treatment, then there is an inconsistency between the data generating process as described by the clustering and treatment procedures.

In panels (b) and (c) of Table XIII I adjust the clustering, sampling and treatment levels. First, in panels (b) I cluster covariance matrices and bootstrap sample at the treatment level wherever this is possible.[56] Comparing with panels (a), we see that the power of randomization tests rises substantially when the data generating process in the bootstrap is consistent with the treatment process. In particular, in regressions where authors' originally clustered, and bootstrap data was generated, at a level of aggregation below that of treatment, the power of randomization techniques almost doubles at the .01 level. The power of conventional techniques systematically moves with the level of aggregation, falling where authors formerly clustered at a higher level of aggregation and rising where authors formerly clustered at a lower level of aggregation. These counterintuitive changes come from the changing size distortions of these methods. For example, for coefficients where authors cluster at a level of aggregation below that of treatment, the average rejection probability at the .01 level is .022 using authors' methods but rises to .082 when clustering is done at the treatment level.[57] In many laboratory experiments, in particular, there are less than 10 (and sometimes as few as 4) sessions and clustering at the session level, appealing to asymptotic theorems, is problematic.

Clustering at the treatment level seems natural and prudent. Treatment groups, whether in sessions in a lab or geographic groups in the field, may share common characteristics or, at the very least, interact with each other in ways that will generate correlations between errors.

[55] As noted earlier, Young (2016) shows that $\lambda^{\max}(\{\mathbf{H_{gg}}\}) \geq h_{ii}^{\max}$. The proof can be generalized, as off-diagonal elements are added to the cluster sub-matrix of the hat matrix, the maximum eigenvalue rises. In the extreme, when there is only one cluster, its eigenvalues are those of the hat matrix itself, i.e. all zeros and ones.

[56] For 92 of the 497 coefficients and 37 of the 88 multi-treatment regressions Stata does not have a cluster option or there are so many treatment variables and so few treatment clusters that it is impossible to calculate a Wald statistic (e.g. 20 treatment measures in a regression which has 8 treatment clusters). In these cases, the bootstrap samples are drawn at the treatment level, but covariance matrices continue to be calculated assuming independent observations (i.e. below treatment level).

[57] Calculated by running 10000 bootstrap simulations at the equation level for each conventional technique.

However, if authors maintain that observations are independent at a lower level of aggregation, then one can argue that treatment was effectively randomized at that level of aggregation. For example, if all observations within sessions are independent, then randomization can be seen as being executed at the observation level, as the presence of an observation in one session or another has, contingent on treatment, no significance. To this end, in panels (c) of Table XIII I adjust the randomization scheme to mimic authors' methods in the 10 experiments (284 coefficients and 65 multi-treatment equations) where authors consistently clustered at a level below treatment.[58] As can be seen in panel (c), this reduces the gap between the power of conventional and randomization techniques at the coefficient level in these equations, originally ranging from .15 to .19 in panel (a), to .03 to .04 percentage points.

Table XIV uses regressions to show how all of the difference between randomization and conventional power comes from the size distortions of conventional tests and differing assumptions regarding the data generating process. I run the ln relative rejection rate of randomization to conventional methods on the ln excess coverage of authors' methods. As this regressor has a value of zero when conventional size is unbiased, the constant term indicates the predicted difference when conventional and randomization methods are both exact. Panels (a), (b) and (c) use the power data shown in the corresponding panels of Table XIII.[59] As shown in panel (a), when clustering and bootstrap sampling (i.e. the data generating process) is done at a level that is consistent with authors' methods but often inconsistent with the treatment process, the constant term indicates 7 to 16 percent lower relative randomization power. When clustering and data generation is done at the treatment level (panel b) and randomization analysis is done

[58] In the remaining 213 coefficients and 23 multi-treatment equations where authors clustered below treatment level, they cluster at the treatment level elsewhere in the paper. The non-clustered equations arise because of coding errors or because Stata does not support that option for their chosen command. For these coefficients and equations, I follow panel (b), clustering at treatment level (if possible) and bootstrap sampling at the treatment level. For the case where authors clustered at a higher level of aggregation, it is not possible to adjust the randomization scheme up because the groups are of unequal size and treatment is mixed within them, so there is no way to reallocate the treatment received by one larger aggregation to another. Hence there is no version of panel (c) for "above" clustering.

[59] While the dependent variable, relative power, is calculated using 50 to 100 bootstrap samples, the regressor, the size distortion of conventional techniques, is calculated at the equation level using 10000 bootstrap samples to minimize measurement error. While randomization inference requires 1000 to 2000 iterations for each bootstrap sample, conventional inference requires only one regression per regression sample, and hence can be simulated extensively at relatively low cost.

Table XIV: Ln Randomization Power Relative to Authors' Methods
Adjusted for Coverage Bias and Clustering/Randomization Methods

| | 5880 coefficients | | | | 1009 multi-treatment regressions | | | |
|---|---|---|---|---|---|---|---|---|
| | randomization-t | | randomization-c | | randomization-t | | randomization-c | |
| | .01 | .05 | .01 | .05 | .01 | .05 | .01 | .05 |
| | (a) authors' clustering & treatment randomization | | | | | | | |
| $\ln\left(\frac{\text{coverage}}{\text{nominal size}}\right)$ | -.237 | -.290 | -.115 | -.211 | -.179 | -.166 | -.131 | -.140 |
| | (.014) | (.018) | (.017) | (.022) | (.024) | (.032) | (.029) | (.039) |
| constant | -.143 | -.130 | -.067 | -.069 | -.159 | -.133 | -.106 | -.101 |
| | (.011) | (.008) | (.014) | (.010) | (.034) | (.028) | (.041) | (.034) |
| N | 4862 | 5727 | 4792 | 5645 | 927 | 982 | 923 | 978 |
| $R^2$ | .053 | .044 | .009 | .016 | .058 | .027 | .022 | .012 |
| $\mu_y$ | -.236 | -.183 | -.112 | -.107 | -.329 | -.221 | -.230 | -.175 |
| | (b) adjusting clustering to treatment level | | | | | | | |
| $\ln\left(\frac{\text{coverage}}{\text{nominal size}}\right)$ | -.366 | -.408 | -.311 | -.366 | -.287 | -.312 | -.283 | -.345 |
| | (.011) | (.013) | (.014) | (.017) | (.018) | (.024) | (.025) | (.034) |
| constant | -.030 | -.049 | .073 | .023 | -.002 | .013 | .097 | .081 |
| | (.010) | (.007) | (.013) | (.009) | (.029) | (.023) | (.040) | (.032) |
| N | 4957 | 5738 | 4898 | 5671 | 955 | 998 | 946 | 993 |
| $R^2$ | .192 | .144 | .093 | .078 | .205 | .140 | .116 | .096 |
| $\mu_y$ | -.205 | -.143 | -.074 | -.061 | -.303 | -.175 | -.196 | -.125 |
| | (c) adjusting randomization to authors' level | | | | | | | |
| $\ln\left(\frac{\text{coverage}}{\text{nominal size}}\right)$ | -.291 | -.320 | -.192 | -.250 | -.249 | -.257 | -.213 | -.253 |
| | (.011) | (.014) | (.015) | (.017) | (.020) | (.026) | (.026) | (.035) |
| constant | -.048 | -.059 | .052 | .015 | -.031 | -.013 | .061 | .040 |
| | (.009) | (.007) | (.012) | (.008) | (.028) | (.022) | (.037) | (.030) |
| N | 5001 | 5765 | 4939 | 5705 | 951 | 999 | 942 | 994 |
| $R^2$ | .115 | .087 | .034 | .036 | .144 | .090 | .064 | .049 |
| $\mu_y$ | -.166 | -.118 | -.025 | -.031 | -.251 | -.142 | -.127 | -.086 |

Notes: Panels: (a) covariance estimates are clustered and samples bootstrapped at authors level of aggregation, treatment is randomized using experimental procedure; (b) covariance estimates are clustered and samples bootstrapped at treatment randomization level; (c) where authors consistently cluster below treatment level, treatment randomization adjusted to authors' level, otherwise clustering adjusted to treatment level as in panel (b). The dependent variable in each column is the ln ratio of the randomization rejection rate to the conventional rejection rate, for individual coefficients or multi-treatment joint tests, at the significance level (.01 or .05) specified. Because only 50 to 100 bootstrap samples are used per equation, both conventional and randomization rejection rates at the .01 level are often zero, and these observations are dropped from the sample (hence N falls below the total number of coefficients and equations). As can be seen from the regressions at the .05 level, where this happens rarely, this does not drive the results. Later drafts will include more bootstrap samples, so this will be less of an issue.

assuming independence at the lower levels of aggregation specified by authors (panel c), the gap between randomization and conventional techniques disappears. The randomization-t has similar power to conventional techniques at the equation level and perhaps 3 to 6 percent less power at the coefficient level. The randomization-c actually consistently shows higher power than conventional techniques, with the constant term as high as .097. These results are based upon estimates that suggest that a 1 percent ln size distortion of conventional methods translates into .2 to .4 percent greater ln power. Even without this adjustment, randomization methods can approach the power of biased conventional techniques. Examining the means of the dependent variable reported in the table ($\mu_y$), we see that when authors' clustering is done at the treatment level and/or randomization analysis is implemented at a level consistent with authors' clustering decisions, the randomization-c has only ln .025 to .074 less power than biased conventional techniques at the coefficient level, despite the size distortions of the latter.

Having discovered how clustering and randomization levels determine the relative power of conventional and randomization techniques, it is natural to explore their implications for the base results in the papers themselves. This is done in Table XV. As in other significance tables, the top row indicates the number of significant coefficients or multi-treatment equations with at least one significant coefficient, the lower rows indicate the relative number of significant results in single coefficient and multi-treatment joint coefficient tests, with the letters (a)-(c) denoting the application of the methods used in the corresponding panels of Tables XIII and XIV. As shown, when conventional tests are clustered at treatment level, there is a small drop in the number of significant coefficients and small rise in the number of jointly significant results. When randomization is implemented following the treatment regime, but with covariance estimates clustered at treatment, significance rates in the randomization-t (which depends upon the covariance estimate) rise slightly in multi-treatment joint tests. When, further, randomization is implemented at the authors' clustering level in the 10 papers which consistently cluster at a lower level of aggregation, the relative number of significant results rises by two to five percentage points. With methods (b) and (c), we are looking at randomization tests whose power is close to that of conventional tests, adjusted for their size distortions, but the number of significant results remains well below that reported in the experimental papers. With method (c)

Table XV: Statistical Significance of Treatment Effects using Different Clustering and Randomization Levels

| | 5880 coefficients | | 1009 multi-treatment regressions | |
|---|---|---|---|---|
| | .01 | .05 | .01 | .05 |
| significant coefficient | 751 | 1459 | 355 | 586 |
| conventional test: | | | | |
| (a) authors' methods | 1.00 | 1.00 | .88 | .75 |
| (b) clustering at treatment | .97 | 1.00 | .90 | .76 |
| randomization-t | | | | |
| (a) following treatment, authors covariance | .78 | .88 | .64 | .60 |
| (b) following treatment, cluster at treatment | .77 | .88 | .67 | .62 |
| (c) following authors' clustering | .81 | .90 | .68 | .61 |
| randomization-c | | | | |
| (a) following treatment, authors covariance | .82 | .87 | .67 | .60 |
| (b) following treatment, cluster at treatment | .82 | .87 | .67 | .60 |
| (c) following authors' clustering | .86 | .89 | .72 | .61 |

Notes: .01/.05 = level of the test. Top row reports number of significant results evaluated using authors' methods; values in lower rows are number of significant results evaluated using indicated method divided by the top rows. For multi-treatment regressions top row indicates number with at least one significant treatment coefficient at the level specified (suggesting non-zero effects), lower rows indicate relative number of rejections in the joint test of all coefficients in the regression.

in the randomization-c, we are looking at a technique whose relative power at the coefficient level is within 3 percent of conventional methods *without* adjustment for their excess size (see $\mu_y$ in panel (c) of Table XIV). And yet here the relative number of significant coefficient results at the .01 and .05 levels is still only .86 and .89, respectively.

The preceding results can seem contradictory, as experimental data is used to establish that the power of different techniques (adjusted if necessary) is not all that different, while in the very same experimental data they produce (even when adjusted) different results. A simple example, reviewing the methodology, can explain the apparent inconsistency. Consider a population $F_0$, with a mean $\mu_0 = 0$. As samples $F_1$ are drawn from this population, each has a mean $\mu_1 \neq 0$. If we repeatedly test the null $\mu_1 = 0$ using an exact test on each of these samples, we will reject with a frequency equal to the size of the test because in the parent population $F_0$ $\mu_0 = 0$. If, however, we draw bootstrap samples $F_2$ from each $F_1$, each time testing $\mu_2 = 0$, we will reject the null more frequently than nominal size, because in the parent population $F_1$ $\mu_1 \neq 0$. Thus, we can learn about power without necessarily implying that $\mu_0 = \mu_1 \neq 0$, i.e. that what is true for $F_1$ is true for $F_0$. Similarly, by drawing bootstrap samples $F_2$ and testing $\mu_2 = \mu_1$, i.e. recentering around what we know to be true for $F_1$, we learn about the size distortions of tests based on $F_1$ of $\mu_1 = \mu_0 = 0$, without it necessarily having to be true that $\mu_0 = 0$.

In Table III's analysis of relative significance rates, earlier above, the bootstrap-t results are very close to those produced by randomization inference. This shows that an adjustment for the actual distribution of the conventional t and Wald statistics under the null, i.e. an adjustment for size, can explain the discrepancy between conventional and randomization results in the experimental sample. In Table XV we see that adjustments for the degree of aggregation in clustering and randomization inference, which levels the relative power of conventional and randomization techniques, cannot explain the difference between conventional and randomization results in the experimental sample. This suggests that the size distortions of conventional techniques are the key problem. A shift to exact randomization inference, implemented, if authors' must insist, using assumptions about sub-treatment level independence, can provide power equal to unbiased conventional techniques in samples that are very far from the asymptotic ideal used to justify conventional methods. There seems to be little reason for randomized experiments not to do this.

## V. Conclusion

The discrepancy between randomization and conventional results in my sample of experimental papers is a natural consequence of how economists, as a profession, perform research. Armed with an idea and a data set, we search for statistically significant relations, examining the relationship between dependent and independent variables that are of interest to us. Having found a significant relation, we then work energetically to convince seminar participants, referees and editors that it is robust, adding more and more right-hand side variables and employing universal "corrections" to deal with unknown problems with the error disturbance. This paper suggests that this dialogue between our roles as authors and our roles as sceptical readers may be misdirected. Correlations between dependent and independent variables may reflect the role of omitted variables, but they may also be the result of completely random correlation. This is unlikely to be revealed by adding additional non-random right-hand side variables. Moreover, the high maximal leverage produced by these conditioning relations, combined with the use of leverage dependent asymptotic standard error corrections, produces a systematic bias in favour of finding significant results in finite samples. A much better indication of random correlation is the number of attempted insignificant specifications that accompanied the finding of a significant result. A large number of statistically independent insignificant results contain much more information than a sequence of correlated variations on a limited number of significant specifications. This fact is lost in our professional dialogue, with its focus on testing the robustness of significant relations.

The lack of omnibus tests that link equations in my sample of published papers is not surprising, as these tests are near nigh impossible to implement using conventional methods. The almost complete lack of F-tests within equations, however, is much more revealing of professional practice. Regressions with an individually .01 level significant coefficient have an average of 5.9 treatment measures, representing multiple treatments and the interaction of treatment with participant characteristics, of which on average 4.2 are insignificant. The fact that the multiple tests implicit in these regressions are almost never jointly evaluated cannot be blamed on authors, because these papers have all gone through the scrutiny of seminar participants, referees and editors. Instead, it must be seen as reflecting a professional focus on disproving significant results and inability to see all the information embodied in the insignificant results that are laid out in front of us.

Only one paper in my sample emphasizes the lack of statistically significant treatment effects. The present paper suggests that this is much more widespread than the results of individual regressions might lead one to believe, i.e. many experimental treatments appear to be having no effect on participants. I arrive at this conclusion not by modifying equations and testing the robustness of coefficients, but by combining the evidence presented honestly and forthrightly by the authors of these papers. A lack of statistically significant results is typically seen as a barrier to publication, but, as the aforementioned paper indicates, this need not be the case. To an economist reading these papers it seems prima facie obvious that the manipulations and treatments presented therein should have a substantial effect on participants. The fact that in so many cases there do not appear to be any (at least) statistically significant effects is, in many respects, much more stimulating than the confirmation of pre-existing beliefs. A greater emphasis on statistically insignificant results, both in the evaluation of evidence and in the consideration of the value of papers, might be beneficial. To quote R.A. Fisher (1935):

> The liberation of the human intellect must, however, remain incomplete so long as it is free only to work out the consequences of a prescribed body of dogmatic data, and is denied the access to unsuspected truths, which only direct observation can give.

Randomized experiments, with their potential for accurate and unbiased finite sample statistical inference, may reveal such truths.

## BIBLIOGRAPHY

Experimental Sample

Abeler, Johannes, Armin Falk, Lorenz Goette, and David Huffman. 2011. "Reference Points and Effort Provision." *American Economic Review* 101(2): 470–49.

Aker, Jenny C., Christopher Ksoll, and Travis J. Lybbert. 2012. "Can Mobile Phones Improve Learning? Evidence from a Field Experiment in Niger." *American Economic Journal: Applied Economics* 4(4): 94–120.

Alatas, Vivi, Abhijit Banerjee, Rema Hanna, Benjamin A. Olken, and Julia Tobias. 2012. "Targeting the Poor: Evidence from a Field Experiment in Indonesia." *American Economic Review* 102(4): 1206–1240.

Angrist, Joshua, Daniel Lang, and Philip Oreopoulos. 2009. "Incentives and Services for College Achievement: Evidence from a Randomized Trial." *American Economic Journal: Applied Economics* 1(1): 136–163.

Angrist, Joshua, and Victor Lavy. 2009. "The Effects of High Stakes High School Achievement Awards: Evidence from a Randomized Trial." *American Economic Review* 99(4): 1384–1414.

Ashraf, Nava. 2009. "Spousal Control and Intra-Household Decision Making: An Experimental Study in the Philippines." *American Economic Review* 99(4): 1245–1277.

Ashraf, Nava, James Berry, and Jesse M. Shapiro. 2010. "Can Higher Prices Stimulate Product Use? Evidence from a Field Experiment in Zambia." *American Economic Review* 100(5): 2383–2413.

Barrera-Osorio, Felipe, Marianne Bertrand, Leigh L. Linden, and Francisco Perez-Calle. 2011. "Improving the Design of Conditional Transfer Programs: Evidence from a Randomized Education Experiment in Colombia." *American Economic Journal: Applied Economics* 3(2): 167–195

Beaman, Lori and Jeremy Magruder. 2012. "Who Gets the Job Referral? Evidence from a Social Networks Experiment." *American Economic Review* 102(7): 3574–3593.

Burde, Dana and Leigh L. Linden. 2013. "Bringing Education to Afghan Girls: A Randomized Controlled Trial of Village-Based Schools." *American Economic Journal: Applied Economics* 5(3): 27–40.

Cai, Hongbin, Yuyu Chen, and Hanming Fang. 2009. "Observational Learning: Evidence from a Randomized Natural Field Experiment." *American Economic Review* 99(3): 864–882.

Callen, Michael, Mohammad Isaqzadeh, James D. Long, and Charles Sprenger. 2014. "Violence and Risk Preference: Experimental Evidence from Afghanistan." *American Economic Review* 104(1): 123–148.

Camera, Gabriele and Marco Casari. 2014. "The Coordination Value of Monetary Exchange: Experimental Evidence." *American Economic Journal: Microeconomics* 6(1): 290–314.

Carpenter, Jeffrey, Peter Hans Matthews, and John Schirm. 2010. "Tournaments and Office Politics: Evidence from a Real Effort Experiment." *American Economic Review* 100(1): 504–517.

Chen, Roy and Yan Chen. 2011. "The Potential of Social Identity for Equilibrium Selection." *American Economic Review* 101(6): 2562–2589.

Chen, Yan and Sherry Xin Li. 2009. "Group Identity and Social Preferences." *American Economic Review* 99(1): 431–457.

Chen, Yan, F. Maxwell Harper, Joseph Konstan, and Sherry Xin Li. 2010. "Social Comparisons and Contributions to Online Communities: A Field Experiment on MovieLens." *American Economic Review* 100(4): 1358–1398.

Cole, Shawn, Xavier Giné, Jeremy Tobacman, Petia Topalova, Robert Townsend, and James Vickery. 2013. "Barriers to Household Risk Management: Evidence from India." *American Economic Journal: Applied Economics* 5(1): 104–135.

Duflo, Esther, Pascaline Dupas, and Michael Kremer. 2011. "Peer Effects, Teacher Incentives, and the Impact of Tracking: Evidence from a Randomized Evaluation in Kenya." *American Economic Review* 101(5): 1739–1774.

Duflo, Esther, Michael Kremer, and Jonathan Robinson. 2011. "Nudging Farmers to Use Fertilizer: Theory and Experimental Evidence from Kenya." *American Economic Review* 101(6): 2350–2390.

Duflo, Esther, Rema Hanna, and Stephen P. Ryan. 2012. "Incentives Work: Getting Teachers to Come to School." *American Economic Review* 102(4): 1241–1278.

Dupas, Pascaline. 2011. "Do Teenagers Respond to HIV Risk Information? Evidence from a Field Experiment in Kenya." *American Economic Journal: Applied Economics* 3(1): 1–34.

Dupas, Pascaline and Jonathan Robinson. 2013. "Savings Constraints and Microenterprise Development: Evidence from a Field Experiment in Kenya." *American Economic Journal: Applied Economics* 5(1): 163–192.

Dupas, Pascaline and Jonathan Robinson. 2013. "Why Don't the Poor Save More? Evidence from Health Savings Experiments." *American Economic Review* 103(4): 1138–1171.

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): 1014–1039.

Erkal, Nisvan, Lata Gangadharan, and Nikos Nikiforakis. 2011. "Relative Earnings and Giving in a Real-Effort Experiment." *American Economic Review* 101(7): 3330–3348.

Fehr, Ernst and Lorenze Goette. 2007. "Do Workers Work More if Wages Are High? Evidence from a Randomized Field Experiment." *American Economic Review* 97(1): 298-317.

Fehr, Ernst, Holger Herz, and Tom Wilkening. 2013. "The Lure of Authority: Motivation and Incentive Effects of Power." *American Economic Review* 103(4): 1325–1359.

Field, Erica, Seema Jayachandran, and Rohini Pande. 2010. "Do Traditional Institutions Constrain Female Entrepreneurship? A Field Experiment on Business Training in India." *American Economic Review: Papers & Proceedings* 100(2): 125–129.

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): 2196–2226.

Fong, Christina M. and Erzo F. P. Luttmer. 2009. "What Determines Giving to Hurricane Katrina Victims? Experimental Evidence on Racial Group Loyalty." *American Economic Journal: Applied Economics* 1(2): 64–87.

Galiani, Sebastian, Martín A. Rossi, and Ernesto Schargrodsky. 2011. "Conscription and Crime: Evidence from the Argentine Draft Lottery." *American Economic Journal: Applied Economics* 3(2): 119–136.

Gerber, Alan S., Dean Karlan, and Daniel Bergan. 2009. "Does the Media Matter? A Field Experiment Measuring the Effect of Newspapers on Voting Behavior and Political Opinions." *American Economic Journal: Applied Economics* 1(2): 35–52.

Gertler, Paul J., Sebastian W. Martinez, and Marta Rubio-Codina. 2012. "Investing Cash Transfers to Raise Long-Term Living Standards." *American Economic Journal: Applied Economics* 4(1): 164–192.

Giné, Xavier, Jessica Goldberg, and Dean Yang. 2012. "Credit Market Consequences of Improved Personal Identification: Field Experimental Evidence from Malawi." *American Economic Review* 102(6): 2923–2954.

Guryan, Jonathan, Kory Kroft, and Matthew J. Notowidigdo. 2009. "Peer Effects in the Workplace: Evidence from Random Groupings in Professional Golf Tournaments." *American Economic Journal: Applied Economics* 1(4): 34–68.

Heffetz, Ori and Moses Shayo. 2009. "How Large Are Non-Budget-Constraint Effects of Prices on Demand?" *American Economic Journal: Applied Economics* 1(4): 170–199.

Ifcher, John and Homa Zarghamee. 2011. "Happiness and Time Preference: The Effect of Positive Affect in a Random-Assignment Experiment." *American Economic Review* 101(7): 3109–3129.

Karlan, Dean and John A. List. 2007. "Does Price Matter in Charitable Giving? Evidence from a Large-Scale Natural Field Experiment." *American Economic Review* 97(5): 1774-1793.

Kosfeld, Michael and Susanne Neckermann. 2011. "Getting More Work for Nothing? Symbolic Awards and Worker Performance." *American Economic Journal: Microeconomics* 3(3): 86–99.

Kube, Sebastian, Michel André Maréchal, and Clemens Puppe. 2012. "The Currency of Reciprocity: Gift Exchange in the Workplace." *American Economic Review* 102(4): 1644–1662.

Landry, Craig E., Andreas Lange, John A. List, Michael K. Price, and Nicholas G. Rupp. "Is a Donor in Hand Better than Two in the Bush? Evidence from a Natural Field Experiment." *American Economic Review* 100(3): 958–983.

Larkin, Ian and Stephen Leider. 2012. "Incentive Schemes, Sorting, and Behavioral Biases of Employees: Experimental Evidence." *American Economic Journal: Microeconomics* 4(2): 184–214.

Lazear, Edward P., Ulrike Malmendier, and Roberto A. Weber. 2012. "Sorting in Experiments with Application to Social Preferences." *American Economic Journal: Applied Economics* 4(1): 136–163.

Macours, Karen, Norbert Schady, and Renos Vakis. 2012. "Cash Transfers, Behavioral Changes, and Cognitive Development in Early Childhood: Evidence from a Randomized Experiment." *American Economic Journal: Applied Economics* 4(2): 247–273.

de Mel, Suresh, David McKenzie, and Christopher Woodruff. 2009. "Are Women More Credit Constrained? Experimental Evidence on Gender and Microenterprise Returns." *American Economic Journal: Applied Economics* 1(3): 1–32.

de Mel, Suresh, David McKenzie, and Christopher Woodruff. 2013. "The Demand for, and Consequences of, Formalization among Informal Firms in Sri Lanka." *American Economic Journal: Applied Economics* 5(2): 122–150.

Oster, Emily and Rebecca Thornton. 2011. "Menstruation, Sanitary Products, and School Attendance: Evidence from a Randomized Evaluation." *American Economic Journal: Applied Economics* 3(1): 91–100.

Robinson, Jonathan. 2012. "Limited Insurance within the Household: Evidence from a Field Experiment in Kenya." *American Economic Journal: Applied Economics* 4(4): 140–164.

Sautmann, Anja. 2013. "Contracts for Agents with Biased Beliefs: Some Theory and an Experiment." *American Economic Journal: Microeconomics* 5(3): 124–156.

Thornton, Rebecca L. 2008. "The Demand for, and Impact of, Learning HIV Status." *American Economic Review* 98(5): 1829–1863.

Vossler, Christian A., Maurice Doyon, and Daniel Rondeau. 2012. "Truth in Consequentiality: Theory and Field Evidence on Discrete Choice Experiments." *American Economic Journal: Microeconomics* 4(4): 145–171.

Wisdom, Jessica, Julie S. Downs, and George Loewenstein. 2010. "Promoting Healthy Choices: Information versus Convenience." *American Economic Journal: Applied Economics* 2(2): 164–178.

Sources Cited in the Paper

Bertrand, Mariaane, Esther Duflo and Sendhil Mullainathan. 2004. "How Much Should we Trust Differences-in-Differences Estimates?" *Quarterly Journal of Economics* 119 (1): 249-275.

Burtless, Gary. 1995. "The Case for Randomized Field Trials in Economics and Policy Research." *Journal of Economic Perspectives* 9(2): 63-84.

Chesher, Andrew and Ian Jewitt. 1987. "The Bias of a Heteroskedasticity Consistent Covariance Matrix Estimator." *Econometrica* 55(5): 1217-1222.

Chesher, Andrew. 1989. "Hajek Inequalities, Measures of Leverage and the Size of Heteroskedasticity Robust Wald Tests." *Econometrica* 57 (4): 971-977.

Deaton, Angus. 2010. "Instruments, Randomization and Learning about Development." *Journal of Economic Literature* 48(2): 424-455.

Donald, Stephen G. and Kevin Lang. 2007. "Inference with Difference-in-Differences and other Panel Data." *The Review of Economics and Statistics* 89 (2): 221-233.

Duflo, Esther, Rachel Glennerster and Michael Kremer (2008). "Using Randomization in Development Economics Research: A Toolkit." In T. Schultz and John Strauss, eds. Handbook of Development Economics, Vol.4. Amsterdam: North Holland, 2008.

Fisher, Ronald A. 1935, 6th edition 1951. The Design of Experiments. Sixth edition. Edinburgh: Oliver and Boyd, Ltd, 1951.

Fox, John. 2008. Applied Regression Analysis and Generalized Linear Models. Second edition. Los Angeles: Sage Publications, 2008.

Hall, Peter. 1992. The Bootstrap and Edgeworth Expansion. New York: Springer-Verlag, 1992.

Heckman, James J. and Jeffrey A. Smith. 1995. "Assessing the Case for Social Experiments." *Journal of Economic Perspectives* 9(2): 85-110.

Hoaglin, David C. and Roy E. Welsch. 1978. "The Hat Matrix in Regression and ANOVA." *The American Statistician* 32(1): 17-22.

Huber, Peter J. 1981. Robust Statistics. New York: John Wiley & Sons, 1981.

Imbens, Guido W. 2010. "Better LATE Than Nothing: Some Comments on Deaton (2009) and Heckman and Urzua (2009). *Journal of Economic Literature* 48(2): 399-423.

Jockel, Karl-Heinz. 1986. "Finite Sample Properties and Asymptotic Efficiency of Monte Carlo Tests." *The Annals of Statistics* 14 (1): 336-347.

Kremer, Michael and Edward Miguel. 2004. "Worms: Identifying Impacts on Education and Health in the Presence of Treatment Externalities." *Econometrica* 72(1): 159–217.

Leamer, Edward E. 1978. Specification Searches: Ad Hoc Inference with Nonexperimental Data. New York: John Wiley & Sons, 1978.

Leamer, Edward E. 2010. "Tantalus on the Road to Asymptopia." *Journal of Economic Perspectives* 24 (2): 31-46.

Lehmann, E.L. 1959. Testing Statistical Hypotheses. New York: John Wiley & Sons, 1959.

Lehmann, E.L and Joseph P. Romano. 2005. Testing Statistical Hypotheses. Third edition. New York: Springer Science + Business Media, 2005.

MacKinnon, James G. and Halbert White. 1985. "Some Heteroskedasticity-Consistent Covariance Matrix Estimators with Improved Finite Sample Properties." *Journal of Econometrics* 29 (3): 305-325.

Romano, Joseph P. 1989. "Bootstrap and Randomization Tests of Some Nonparametric Hypotheses." *The Annals of Statistics* 17 (1): 141-159.

Romano, Joseph P. and Michael Wolf. 2005a. "Exact and Approximate Stepdown Methods for Multiple Hypothesis Testing." *Journal of the American Statistical Association* 100 (469): 94-108.

Romano, Joseph P. and Michael Wolf. 2005b. "Stepwise Multiple Testing as Formalized Data Snooping." *Econometrica* 73 (4): 1237-1282.

Weesie, Jeroen. 1999. "Seemingly unrelated estimation and the cluster-adjusted sandwich estimator." Stata Technical Bulletin 52 (November 1999): 34-47.

White, Halbert. 1980. "A Heteroskedasticity-Consistent Covariance Matrix Estimator and a Direct Test for Heteroskedasticity." *Econometrica* 48 (4): 817-838.

White, Halbert. 1982. "Maximum Likelihood Estimation of Misspecificed Models." *Econometrica* 50 (1): 1-25.

White, Halbert. 2000. "A Reality Check for Data Snooping." *Econometrica* 68 (5): 1097-1126.

Young, Alwyn. 2016. "Improved, Nearly Exact, Statistical Inference with Robust and Clustered Covariance Matrices using Effective Degrees of Freedom Corrections." Manuscript, London School of Economics.