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

Alwyn Young
London School of Economics
This draft: February 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 2003 regressions in 53 experimental papers drawn from the journals of the American Economic Association. Randomization F/Wald tests of the significance of treatment coefficients find that 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 25 to 50 percent of experimental papers, depending upon the significance level and test, are able to reject the null of no treatment effect anywhere. Bootstrap and simulation methods support and confirm these results.

*I am grateful to Alan Manning, 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 F and Wald 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. I find that randomization F/Wald tests at the equation level 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 25 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, present empirical simulations of the bias of conventional methods, and show that the equation level power of randomization tests is virtually identical to that of conventional methods in idealized situations where conventional methods are also exact.

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 5.8 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 1036 regressions with multiple treatment measures report a conventional F- or Wald-test of the joint significance of all treatment

variables within the regression.[1] When tests of joint significance are applied, far fewer regressions show significant effects. I find that additional significant results appear, as additional treatment regressors are added to equations within papers, at a rate comparable to that implied by random chance under the null of no treatment effect. Specification search, as measured by the numbers of treatment regressors, produces additional significant results at a rate that is consistent with spurious correlation.

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 28 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

---

[1]These occur in two papers. In an additional 8 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 random experiment.

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 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. 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.

The reader looking for a definitive breakdown of the results between the contribution of the multiplicity of tests and the contribution of the finite sample bias of asymptotic methods

should be forewarned that a unique deconstruction of this sort simply does not exist. The reason for this is that the coverage bias, i.e. rejection probability greater than nominal size, of conventional tests increases with the dimensionality of the test.[3] I find, both in actual results and in size simulations, that the gap between conventional and randomization/bootstrap tests is small at the coefficient level, larger at the equation level (combining coefficients) and enormous at the paper level (combining all equations and coefficients, in the few instances where this is possible using conventional techniques). If one first uses conventional methods to move from coefficients to equations to paper level tests (where it is possible to implement them conventionally) and then compares the paper level results with randomization tests, one concludes that the issue of multiplicity is of modest relevance and the gap between conventional and randomization inference (evaluated at the paper level) explains most of the results. If, however, one first compares conventional and randomization results at the coefficient level and then uses randomization inference to move from coefficients to equations to paper level tests, one concludes that the gap between randomization and conventional inference is small, and multiplicity (as captured in the rapidly declining significance of randomization tests at higher levels of aggregation) is all important. The evaluation of these differing paths is further complicated by the fact that power also compounds with the dimensionality of the test, and that tests with excess size typically have greater power, which, depending upon whether one wishes to give the benefit of the doubt to the null or the alternative, alters ones view of conventional and randomization tests.

Although I report results at all levels of aggregation, I handle these issues by focusing on presenting the path of results with maximum credibility. F/Wald tests of the overall significance of multiple coefficients within an equation are eminently familiar and easily verifiable, so I take as the first step the conventional comparison of individual coefficient versus equation level significance. The application of conventional F/Wald tests to equations with multiple treatment

[3]A possible reason for this lies in the fact that coverage bias relative to nominal size for each individual coefficient is greater at smaller nominal probabilities, i.e. the ratio of tail probabilities is greater at more extreme outcomes. In the Wald tests below, after the transformation afforded by the inverse of the coefficient covariance matrix, the 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 increased. This requires, however, an accurate assessment of the probability each squared random variable can, by itself, attain increasingly extreme values. As the dimensionality of the test increases this assessment is proportionately increasingly wrong and the overall rejection probability rises.

measures finds that 12 and 26 percent of equations (at the .01 and .05 level, respectively) that have at least one significant treatment coefficient are found to have, overall, no significant treatment effect. Allowing for single treatment coefficient equations whose significance is unchanged, these conventional tests reduce the number of equations with significant treatment effects by 8 to 17 percent at the .01 and .05 levels, respectively. Moving further, from the equation to the paper level, using conventional covariance estimates for systems of seemingly unrelated equations is largely infeasible, as the covariance matrices produced by this method are usually utterly singular. I am able to calculate such a conventional test for only 9 papers, and simulations show that the test statistics have extraordinarily biased coverage (i.e. a .30 rejection probability at the .01 level). Hence, it is not credible to advance to the paper level analysis using conventional methods.

I find, in simulations, that the power of randomization tests at the equation level is almost identical to that of conventional methods. Consequently, the next step in the presentation of results is the movement from conventional F/Wald tests at the equation level to comparable randomization tests. I find that these reduce the number of significant treatment effects further, for a cumulative reduction (from the coefficient level) of 35 to 40 percent in equations with more than one treatment measure and 30 percent in the entire sample. Thus, up to this point, the difference between conventional and randomization results accounts for ½ to ¾ of the reduction in significance levels, with the multiplicity of tests, as embodied in the conventional F/Wald test, accounting for ½ to ¼. The final step in the path is the movement from equation to paper level, where I employ an omnibus Wald randomization test of overall significance which stacks all of the treatment coefficients of all outcome regressions. Here, between ½ and ¾ of papers cannot reject the null of no treatment effect. I compare the power of these tests to conventional counterparts for the easily simulated and calculated special case where the cross-equation correlation of coefficients is zero, and find them to be weaker. If, however, I restrict attention to the 36 to 40 papers where the power of the randomization test is very close to that of a conventional counterpart, I still find that ½ to ⅔ of papers cannot reject the null of no treatment effects at the .01 level, while .4 to .6 cannot reject the null at the .05 level. Based upon the path just described, about half of this final result can be attributed to accounting for the multiplicity of

tests implicitly present in reported results and about half to the difference between randomization and conventional results.

The bootstrap, which confirms the randomization results, 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 distributions not only confirm the randomization results on statistical significance, but also the within and cross equation correlation of coefficients that provides intuition for results, as noted above. In the paper I present simulations using ideal disturbances applied to the estimating equations of my sample papers which show that conventional methods produce rejection rates that are greater than nominal size. I then show that these simulated distributions, when used to evaluate the papers' conventional test statistics, produce significant results at a rate almost identical to that of randomization inference, i.e. that the error and bias in the evaluation of the distribution of conventional test statistics explains almost all of the discrepancy in results. Readers might doubt these, seeking more realistic simulations that reproduce the complex correlation of errors across observations and equations and the mixture of no treatment effects and actual treatment effects (nulls and alternatives) present in the data. For these, they need look no further than the bootstrap results, which use the experimental samples to simulate the distribution of the test statistics under the conditions actually present in the population, and, as noted already, reproduce the outcomes of the randomization tests.

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,[4] it can easily

[4]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),

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.[5] 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/2003 regression 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 randomization techniques). About 70 percent of the regressions are ordinary least squares (OLS)[6] 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, and later show, that maximal leverage determines the coverage bias of tests using the robust and clustered covariance matrices. The theory underlying randomization and bootstrap statistical inference is reviewed

---

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 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 (under the assumption of independence within groups), 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.

[5]This integrating result, proving the importance of Fisherian inference in a Neyman population sampling setting, is of relevance to a potential criticism of randomization inference. Randomization inference allows exact tests of sharp hypotheses, i.e. hypotheses that specify a precise treatment effect for each participant (in the case of this paper, a zero effect for all participants). It does not provide exact tests of non-sharp hypotheses, e.g. a mean average treatment effect of zero with unknown distribution across participants. That said, other methods do not provide distribution free exact tests either (as they depend upon knowledge of the distribution of the error term which, if the average treatment effect has an unknown distribution, becomes heteroskedastic in an unknown fashion).

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

and several alternative measures, with different theoretical properties, are presented. In particular, I note that Stata's (and authors') default method of calculating the bootstrap is based upon a method whose rejection probability converges to nominal size at a rate $O(n^{-1/2})$, i.e. no better than that of standard asymptotic normal approximations. In contrast, a well-known refinement of the bootstrap, based upon pivotal statistics, converges at a rate $O(n^{-1})$ and, in application, produces systematically higher p-values (i.e. lower significance levels).

Section IV presents the main empirical results. I begin by reviewing the results on statistical significance and within and across equation coefficient correlation discussed above. I then present size simulations, using data based upon a version of each regression in which treatment has no effect and the disturbances are ideal iid, and show that the robust and clustered covariance matrices produce test statistics that at the .01 level have average rejection probabilities of .02 to .03 in OLS samples and .045 in non-OLS regressions when the null is true. Rejection probabilities of .5 or .6, at the .01 nominal level, appear in particular regressions. These results carryover to environments where the disturbances are far from ideal, i.e. cluster correlated or heteroskedastic. I show that maximal leverage accounts for virtually all of the average size bias of tests based upon the robust and clustered covariance matrices in situations with ideal or non-ideal disturbances. When the distributions produced by size simulations are used to evaluate the conventional test statistics in my experimental sample, their significance levels move close to those of randomization tests. In sum, coverage bias, due to the unaccounted for excess thickness of the tails of distributions based upon robust and clustered covariance estimates, can explain virtually all of the discrepancy between randomization and conventional results.

While randomization inference is exact when the null is true, experimenters might be concerned about its performance when the null is false. To address these fears, Section IV includes simulations of power based upon data which take each regression's estimated treatment effects as part of the data generating process. I find that with ideal iid errors the equation level power of randomization tests is, for all intents and purposes, identical to that of OLS regressions using the default covariance matrix. Thus, in a situation where both conventional and randomization inference are exact and have accurate size, they have identical power as well. However, in non-OLS settings I find the power of randomization tests is slightly lower than that of conventional methods. When the authors' covariance estimation methods, with their

systematic underestimation of test statistic variation, are substituted, the gap in power becomes greater, but is still very small compared to the differences that appear when conventional and randomization tests are applied to the papers themselves. As noted earlier, power differences are much larger at the paper level. Nevertheless, if one restricts the implementation of the omnibus test of overall experimental significance to cases where the power of the randomization test is within .01 of that of conventional tests in simulation, one still finds that ½ to ⅔ of my sample cannot reject the null of no treatment effects at the .01 level. In sum, power does not explain the results reported in this paper. 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,[7] 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

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

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). This paper presents simulations, with iid and non-iid errors, for 2000 published regressions with the robust covariance matrix and 1000 regressions with the clustered covariance structure, affirming 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 showing how maximal leverage determines bounds on the variance of estimates 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 empirical coverage bias of test statistics based upon robust and clustered covariance estimates. The idea of evaluating these test statistics using the distributions generated under ideal iid disturbances is present, in various forms, in Kott (1996), Bell and McCaffrey (2002) and Young (2016), and has been endorsed by Imbens and Kolesar (2015). In Young (2016), in particular, I show that this generates nearly exact inference on individual coefficients in my experimental sample in simulated situations where the disturbances are far from ideal. In this paper I show that the actual conventional F/Wald statistics of my experimental sample, when evaluated using these ideal distributions, produce significance rates that are very close to those generated by randomization methods, as already noted above.

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 F/Wald tests of multiple treatment coefficients to combine the information implicit in multiple tests within equations. I examine the rate at which the progressive addition of more treatment measures and interactions, within papers, produces additional significant results and compare it to that generated by random chance under the null. I use an omnibus test to combine all the treatment information in all of the regressions run in a paper. In this last, the integrity of authors in the presentation of the many specifications they ran allows, through the

explicit consideration of the covariance of all the 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 and anonymized paper, allowing researchers to reproduce the tables in this paper and use the randomization, bootstrap and simulation 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 or Stata do-file code[8];
(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

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

place in universities or research centres or recruiting their subjects from such populations).[9] 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.[10] Laboratory experiments account for 15 of the 53 papers but only 197 of the 2003 regressions. The results for laboratory experiments are not substantially different than those for field experiments.

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 saved me the task of working through a large number of laboratory experiments, which tend to not have data on participant characteristics, use atypical econometric methods and whose passive randomization might raise concerns, while leaving enough lab papers and regressions to see if the results generalized to this sample as well. 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 test 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

[9]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.

[10]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.

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 β_T and β_{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.[11] 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, non-experimental cohorts, sample attrition or first-stage regressions that do not involve treatment outcomes analysed elsewhere in the paper;

while tests were done on the null that:

[11]Stacking the regressions very often also 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 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 agrees that the stacked regression was the authors' actual intent.

(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 tests 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. 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 later, 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) $y = \alpha + \beta_T T + \beta_{T0age} T_0 * age + \beta_{T1age} T_1 * age + \varepsilon$

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 β_{age} to take on any value.[12] In more complicated situations the paper might contain randomized overall treatments (e.g. the

[12]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.

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.[13] 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 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 (loading, 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,[14] 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.[15]

---

[13]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.

[14]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.

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

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).

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 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 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 2003 regressions, almost 70 percent of these are ordinary least squares regressions and an additional 15 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 3 percent more of the sample, and instrumental variables account for another 3 percent. I develop a method for randomization inference with instrumental variables, as explained below, but not for over-identified two stage least squares, so I exclude the latter from my sample. The final residual category, "other", accounts for 9 percent of regressions and includes quantile regressions, weighted average treatment effects, population weighted regressions, seemingly unrelated estimates, two step ordinary least squares regression estimates, non-maximum likelihood two step Heckman models, tests of difference of proportions and t-tests with unequal variances.[16]

A little under a quarter of the regressions in my sample make use of Stata's standard or default covariance matrix calculation. Almost half of all regressions, however, avail themselves of the cluster estimate of covariance and about another 20 percent use the robust option, which is

[16]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.

Table I: Characteristics of the Sample

53 papers			2003 regressions		
location		journal	Type		covariance
38	field	29 AER	1378	ordinary least squares	447 standard
15	lab	20 AEJ: Applied Economics	322	maximum likelihood	995 clustered
		4 AEJ: Microeconomics	67	generalized least squares	344 robust
			55	instrumental variables	126 bootstrap
			181	other	91 other

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

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.

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.[17] I link maximal leverage to variation in the bias and variance of the clustered and robust covariance estimates and later show that maximal leverage, and not sample size, explains nearly all of the average empirical coverage bias of tests based upon these covariance matrices. 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

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

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 and is borne out in later empirical simulations. Finally, 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 systematically biased in favour of rejecting null hypotheses.

(a) Leverage and the Road to Asymptopia

Consider the regression model $\mathbf{y} = \mathbf{X\beta} + \boldsymbol{\varepsilon}$, where $\boldsymbol{\varepsilon}$ is an n x 1 vector of (possibly non-normal) disturbances with covariance matrix $\boldsymbol{\Sigma}$.[18] The hat matrix (Hoaglin and Welsch 1978) is given by $\mathbf{H} = \mathbf{X(X'X)^{-1}X'}$ and derives its name from the fact that it puts a hat on **y** as $\hat{\mathbf{y}} = \mathbf{X}\hat{\boldsymbol{\beta}} = \mathbf{X(X'X)^{-1}X'y} = \mathbf{Hy}$. 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

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

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.[19] Since

[18] 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.

[19] 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 $h_{ii} = 1/n_g + \mathbf{z}_i'(\mathbf{Z'Z})^{-1}\mathbf{z}_i \geq 1/n_g$, where $\mathbf{z}_i'$ is the i^{th} observation row of **Z**.

$h_{ii}^{\max} \geq \bar{h}_{ii} = k/n$, 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 **β** and, in the presence of non-normal disturbances, is not asymptotically normally distributed.[20] The intuition for these results is trivial. With non-negligible maximum leverage, the predicted value 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 **β** 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

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

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

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

where *c* denotes a finite sample adjustment, subscript i an observation and g a cluster,

[20] 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.

$\hat{\varepsilon}_i$ 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 $\boldsymbol{\Sigma}$ is diagonal, and $\mathbf{V_{Cl}}$ is a natural extension of his work to the case where $\boldsymbol{\Sigma}$ is block diagonal by cluster. White (1980) assumed that the limit as $n \to \infty$ 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:

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

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 $\mathbf{M = I - H}$ denote the "residual maker", as the estimated residuals are given by $\hat{\boldsymbol{\varepsilon}} = \mathbf{y} - \hat{\mathbf{y}} = \mathbf{My} = \mathbf{M(X\beta + \varepsilon)} = \mathbf{M\varepsilon}$. Consequently, $\hat{\varepsilon}_i = \mathbf{m_i'\varepsilon}$, where $\mathbf{m_i'}$ is the i[th] row of $\mathbf{M}$[21] and the expected value of the i[th] squared residual has a downward bias determined by leverage as $E(\hat{\varepsilon}_i^2) = E(\mathbf{m_i'\varepsilon\varepsilon'm_i}) = \mathbf{m_i'}\{\sigma^2\}\mathbf{m_i} = \sigma^2 m_{ii} = \sigma^2(1-h_{ii})$, where I have made use of the fact that $\mathbf{M}$ is idempotent and the assumption that $\boldsymbol{\varepsilon}$ 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 σ^2 as $(n-k)^{-1}\Sigma\sigma^2(1-h_{ii}) = (n-k)^{-1}\sigma^2(n-k) = \sigma^2$. The robust covariance

[21] $\mathbf{m_i}$ is the i[th] column of $\mathbf{M}$, but as $\mathbf{M}$ is symmetric, $\mathbf{m_i'}$ is also the i[th] row.

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.[22]

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 $\mathbf{w} = \mathbf{x}_i$, where $\mathbf{x}_i'$ is the ith observation row of $\mathbf{X}$. In this case, the robust estimate of the variance of $\mathbf{w}'\hat{\boldsymbol{\beta}}$ is given by $\mathbf{w}'\mathbf{V_R}\mathbf{w} = c_R\mathbf{x}_i'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\{\hat{\varepsilon}_i^2\}\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}_i = c_R\mathbf{h}_i'\{\hat{\varepsilon}_i^2\}\mathbf{h}_i$, where $\mathbf{h}_i'$ is the ith row of $\mathbf{H}$. As h_{ii}, the leverage of observation i, increases, all the other h_{ij} ($j \neq i$) elements of $\mathbf{h}_i$ go to zero, as can be seen in (3) 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:

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

where n_c is the number of clusters.[23] 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

[22]Across my sample of 1378 OLS regressions, the standard deviation of $h_{ii}^{\min}$ is .027, while that of $h_{ii}^{\max}$ is .383; similarly, across the 827 OLS regressions which cluster, the standard deviation of $\lambda^{\min}(\{\mathbf{H_{gg}}\})$ is .001, while that of $\lambda^{\max}(\{\mathbf{H_{gg}}\})$ is .616.

[23]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.

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.[24] 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 all 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 all hypothesis tests. Because of a strongly positively biased covariance estimate, it is 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.[25] The share of the modal y value is also typically quite high, exceeding .5 in ½ of regressions and, extraordinarily, .96 in 1/20th of the sample.[26] Not surprisingly, tests of the normality of residuals reject the null, at the 1 percent

[24]Since, with iid errors, the ith residual underestimates its own variance by 1-h_{ii}, the average residual underestimates its own variance by $n^{-1}\Sigma(1-h_{ii}) = (n-k)/n$. This suggests an n/(n-k) finite sample correction, which is what is typically used for $\mathbf{V_R}$. I find, in Young (2016), that when applied to all OLS regressions in my sample this produces an average bias in the estimate of the variance of individual treatment coefficients of .99. 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). I find that this produces an average bias in the estimate of the variance of individual treatment coefficients in the clustered OLS regressions of my sample of .98. 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.

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

[26]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.

Table II: Regression Design in the Sample Papers (1378 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	325	2	2	2	2	2	12	135	500	2548	4225	$13e^{3}$
Modal share	.467	$3e^{-4}$	.001	.005	.021	.125	.528	.735	.917	.963	.993	.9999
Normality of $\hat{\varepsilon}$	.011	0	0	0	0	0	0	$1e^{-10}$	$6e^{-4}$	.011	.422	.848
Cumulative distribution of leverage												
$\overline{h}_{ii}$	.051	$2e^{-5}$	$1e^{-4}$	$1e^{-3}$	.002	.008	.026	.058	.148	.207	.330	.533
$h_{ii}^{\max}$	.383	$4e^{-5}$	$2e^{-4}$	.002	.008	.025	.196	.729	1	1	1	1
V_R: $h_{ii}^{\max}$	.491	.001	.001	.002	.011	.170	.404	1	1	1	1	1
V_{Cl}: $h_{ii}^{\max}$	.409	$4e^{-5}$	$2e^{-4}$	.002	.006	.031	.251	1	1	1	1	1
V_{Cl}: $\lambda^{\max}(\{\mathbf{H_{gg}}\})$	.616	$9e^{-4}$	.016	.038	.053	.221	.725	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 164 and 827 OLS regressions, respectively, which use the robust or clustered estimate of covariance. $\overline{h}_{ii}$ and $h_{ii}^{\max}$, average and maximum leverage. $\lambda^{\max}(\{\mathbf{H_{gg}}\})$= maximum eigenvalue of the block diagonal matrix made up of the elements of the hat matrix associated with the cluster groups.

level, 94.8 percent of the time. This discreteness and consequent non-normality of coefficient estimates impairs the accuracy of conventional significance tests as the F and Chi^2 distributions do not properly describe the distribution of test statistics. Using the simulated and bootstrapped distribution of coefficients, I find, however, that non-normality does not generate any systematic bias, as the variance adjusted tails of the non-normal coefficients are not systematically thicker than those implied by the normal distribution. Consequently, I relegate formal analysis of this issue, which concerns accuracy but not bias, to the on-line appendix. Discreteness also generates heteroskedasticity and within this paper I make use of this feature to highlight the failings of robust covariance estimates.

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 164 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 .491 and equalling 1 in 32 percent of the robust covariance estimate sample. In the 827 OLS regressions which use the clustered estimate of covariance ($\mathbf{V_{Cl}}$), the situation is, effectively, much worse. While the average maximum observation leverage is .409, the average maximal eigenvalue of the blocks of the hat matrix associated with the cluster groups, which is what matters for these matrices, is .616, 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,[27] so 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 $\hat{\beta}_1$. 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 may have some relevance, as I find that regressions with a maximal leverage of 1 in my sample do not have as much coverage bias as a linear extrapolation of effects might suggest. They remain, however, substantially biased in favour of rejecting the null.

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

> ...large values of h_{ii} should serve as warning signals that the ith 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.

[27] For a discussion see Fox (2008).

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.

(c) 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[28] 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 ε_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 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

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

extreme enough. Specifically, let the matrix $\mathbf{T}_E$ composed of the column vectors $\mathbf{t}_i$ denote the treatment allocation in the experiment. In the typical experiment this matrix has a countable universe $\mathbf{\Omega}$ of potential realizations. Say there are N elements in $\mathbf{\Omega}$, 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 $\mathbf{\Omega}$, 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:

$$(9)\ \text{randomization p-value} \;=\; \frac{1}{\mathrm{N}}\sum_{n=1}^{N} I_n(>T_{\mathbf{E}}) \;+\; U * \frac{1}{\mathrm{N}}\sum_{n=1}^{N} I_n(=T_{\mathbf{E}})$$

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.

The randomization p-value in (9) above is based solely on the null hypothesis of no treatment effect, and the fact that treatment was drawn randomly from $\mathbf{\Omega}$. Consequently, under the null its distribution does not depend upon sample size or any assumptions regarding the characteristics of y_i, x_i and ε_i. This makes it vastly more robust than conventional econometric technique. Moreover, its power against alternatives is virtually identical to that of conventional tests in situations where they are exact, as shown later in the paper. The use of random draws to resolve ties, however, can be awkward from a presentational perspective. In my sample, a presentationally relevant[29] number of ties appear in the case of only one and, in one test, two papers. Consequently, I remove those papers from the discussion and analysis in later sections

[29]By "presentationally relevant" I mean ties that cross the .01 or .05 levels used in the discussion. Ties that are above these cutoff points, e.g. the p-value lies somewhere between .7 and .8, are not presentationally relevant because, regardless of the draw of U, the test will not reject the null at the selected significance level.

and set U = 0 in all other cases, i.e. resolve the remaining (minimal) ties fully in favour of rejecting the null of no treatment effect.

In the analysis below, for theoretical reasons associated with both randomization inference and the bootstrap, I make use of three randomization based test statistics. The first is based upon the comparison of the Wald statistics of the conventional econometric 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 coefficient estimate of the treatment effect and the estimated variance of that coefficient estimate, so this method in effect calculates the probability

$$(10)\quad \hat{\boldsymbol{\beta}}_t'(\mathbf{T}_n)\mathbf{V}(\hat{\boldsymbol{\beta}}_t(\mathbf{T}_n))^{-1}\hat{\boldsymbol{\beta}}_t(\mathbf{T}_n) > \hat{\boldsymbol{\beta}}_t'(\mathbf{T}_E)\mathbf{V}(\hat{\boldsymbol{\beta}}_t(\mathbf{T}_E))^{-1}\hat{\boldsymbol{\beta}}_t(\mathbf{T}_E).$$

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 $\boldsymbol{\Omega}$. Since the p-values of Wald tests are monotonic in the test statistic, this is basically a comparison of the p-value of the Wald test of no effect performed on the original sample against the distribution of p-values produced by the universe of randomized iterations, and hence the I dub this test the randomization-p.

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

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

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. I refer to this comparison of coefficients as the randomization-c.

Between the randomization-p and randomization-c lies an intermediate case in which the standard errors of individual coefficients are calculated for each randomized draw $\mathbf{T}_n$, but their full covariance matrix is not. The test statistic in this case calculates the probability

$$(12)\quad \mathbf{t}_{\beta_t}'(\mathbf{T}_n)\boldsymbol{\rho}(\hat{\boldsymbol{\beta}}_t(\boldsymbol{\Omega}))^{-1}\mathbf{t}_{\beta_t}(\mathbf{T}_n) > \mathbf{t}_{\beta_t}'(\mathbf{T}_E)\boldsymbol{\rho}(\hat{\boldsymbol{\beta}}_t(\boldsymbol{\Omega}))^{-1}\mathbf{t}_{\beta_t}(\mathbf{T}_E)$$

where $\mathbf{t}_{\beta_t}(\mathbf{T}_n)$ is the vector of t-statistics for the treatment coefficients $\boldsymbol{\beta}_t$ associated with realization $\mathbf{T}_n$ and $\boldsymbol{\rho}(\hat{\boldsymbol{\beta}}_t(\boldsymbol{\Omega}))$ is the correlation matrix of $\hat{\boldsymbol{\beta}}_t$ produced by the universe of potential draws $\boldsymbol{\Omega}$. This intermediate case allows for studentized analysis in multi-equation cases where individual coefficient standard errors are easily computed for each $\mathbf{T}_n$ but the calculation of the full multi-equation coefficient covariance matrix is problematic.[30] In the univariate case, ρ is simply the scalar 1, and this reduces to a comparison of the squared t-statistics. Hence, I call this comparison of studentized coefficient estimates the randomization-t.

Any randomization test is exact, so the choice of test statistic should be based upon something other than size, such as power. Lehmann (1959) showed that in the simple test of mean differences between treatment and control[31] the randomization-t and randomization-p (which are identical in this univariate case) are uniformly most powerful and asymptotically identical to the conventional t-test of no treatment effect. This result may be more general than Lehmann's specific case suggests as I find, in simulations with iid errors further below, when the null of no treatment effect is actually false the randomization-p is more powerful than the randomization-c, while the power of the randomization-t (in multivariate cases) lies between the two. The differences, however, are quite small. Across most of the analysis, I find the different randomization tests produce very similar results. The same cannot be said, however, for analogous bootstrap tests, which have different biases and produce systematically different results.

The randomization-c and -t allow 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 $\boldsymbol{\Omega}$, and calculates (11) and (12) above, with $\hat{\boldsymbol{\beta}}_t$ and $\mathbf{t}_{\beta t}$ now denoting all treatment coefficients and t-statistics from all regressions. The estimated variance and correlation

[30]To clarify, calculation of the covariance matrix for coefficients within an equation or all coefficients in the paper using the universe of realized randomized (or, later, bootstrapped) iterations is always trivially possible. The issue here is calculation of the iteration specific estimate of the covariance of all coefficients. This is straightforward for coefficients within individual equations but near nigh impossible for coefficients across multiple equations estimated using different data files. The randomization-t, thus, is an intermediate computation, using the iteration specific equation standard errors to produce studentized coefficients and the universe of realized iterations to calculate the full multi-equation coefficient correlation matrix.

[31]Equivalently, the significance of treatment in a regression with a treatment dummy and a constant.

of these coefficients in the universe $\mathbf{\Omega}$ is simply calculated from their joint realizations. An omnibus version of the randomization-p is much more difficult, as it requires an iteration by iteration estimate of $\mathbf{V}(\hat{\boldsymbol{\beta}}_t(\mathbf{T}_n))$, including the covariance of coefficients across equations. In a single equation setting, as already noted, I find very little difference in the simulated power and test outcomes of the randomization-p, -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. However, in simulations I also present results using alternative covariance calculation methods. 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.

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 each treatment variable and can deliver a p-value of the conventional Wald test of overall treatment

[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 and -p. 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 alone for each of these papers takes 6 months of workstation time.

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

significance. 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 occurs because maximum likelihood procedures do not converge and/or authors estimate equations that are guaranteed to have singular covariance matrices (e.g. probit equations where they do not let Stata drop observations that are completely determined by regressors). 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.

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 one or two authors, in correspondence, noted that even after they selected a particular randomized allocation of treatment, field agents did not always implement it 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.[34] 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]

Calculation of $f(\mathbf{T}_n)$ for all possible realizations of $\mathbf{T}_n$ in $\mathbf{\Omega}$ is usually computationally infeasible. Random sampling with replacement from $\mathbf{\Omega}$ creates a simulated universe of potential outcomes. The p-values produced by using a finite number of draws to simulate $\mathbf{\Omega}$ are not necessarily uniformly distributed; that is, are no longer exact. However, the precision with which the actual p-value is uncovered by simulation can be controlled by the number of draws from $\mathbf{\Omega}$. I use 10000 iterations, and find almost no change in p-values beyond the 200th draw. Simple examples indicate that tail probabilities may be excessively large with a small number of draws[36] producing excessively large rejection probabilities, but this bias rapidly becomes negligible.

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 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 outcomes 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.[37] Consequently, I test the

---

[34]This, of course, is done in the units of treatment, e.g. field villages or lab sessions.

[35]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]Consider the simplest case, in which $\mathbf{T_E}$ is evaluated with one draw $\mathbf{T_1}$ from $\mathbf{\Omega}$. If the number of potential draws is large (so the probability of ties is negligible), there is a 50/50 chance that $f(\mathbf{T_E})$ will be greater or smaller than $f(\mathbf{T_1})$. Consequently, 50 percent of the time the p-value will be 0 and 50 percent of the time it will be 1. At the .01 and .05 level, this procedure rejects the null ½ of the time.

[37]In other words, dropping (for brevity) error terms, the iv regression $y_i = \boldsymbol{\beta}_z'\mathbf{z}_i + \boldsymbol{\beta}_x'\mathbf{x}_i$, where $\mathbf{z}_i$ is an mx1 vector of endogenous variables that are linearly related to an mx1 treatment vector $\mathbf{t}_i$ through $\mathbf{z}_i = \mathbf{\Gamma}\mathbf{t}_i$ (with $\mathbf{\Gamma}$ an mxm non-zero matrix of coefficients), implies the linear relation $y_i = \boldsymbol{\beta}_z'\mathbf{\Gamma}\mathbf{t}_i + \boldsymbol{\beta}_x'\mathbf{x}_i = \boldsymbol{\beta}_t'\mathbf{t}_i + \boldsymbol{\beta}_x'\mathbf{x}_i$. The null $\boldsymbol{\beta}_z = \mathbf{0}$ implies $\boldsymbol{\beta}_t = \mathbf{0}$ in the regression of y_i on $\mathbf{t}_i$ and $\mathbf{x}_i$, and this is what I test. I should note that this approach is not applicable to

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.[38]

(d) 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-coefficient and cross-equation correlation of coefficients.

The finite sample difficulties faced by the bootstrap become clearer if one explicitly

---

overidentified two stage least squares, where no simple linear projection of y_i on $\mathbf{t}_i$ is implied, so I do not include these in my analysis.

[38]Sometimes authors present first-stage regressions along with iv results. I skip these if they 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.

characterizes the samples at each level. Consider that, in the ideal OLS regression model, the regression sample is drawn from a population characterized by

(13) $F_0 : \; y_i = \boldsymbol{\beta}'\mathbf{x_i} + \varepsilon_i ,$

$$\text{with } E(\varepsilon_i) = 0, \; E(\mathbf{x_i}\varepsilon_i) = \mathbf{0}, \; E(\varepsilon_i^2) = \sigma^2, \; E(\varepsilon_i \varepsilon_j) = 0,$$

whereas the bootstrapped sample is drawn from a population characterized by

(14) $F_1 : \; y_i = \boldsymbol{\gamma}'\mathbf{x_i} + \eta_i , \quad \boldsymbol{\gamma} = \hat{\boldsymbol{\beta}}, \quad \eta_i = \hat{\varepsilon}_i$

$$\text{with } E(\eta_i) = 0, \; E(\mathbf{x_i}\eta_i) = \mathbf{0} \, , E(\eta_i^2) = m_{ii}\sigma^2, \; E(\eta_i \eta_j) = m_{ij}\sigma^2$$

where m_{ii} and m_{ij} are the elements of **M** (the residual maker). The original sample is drawn from a population F_0 in which the linear relation between y_i and $\mathbf{x}_i$ is an unknown vector $\boldsymbol{\beta}$; the bootstrapped sample is drawn from a population F_1 in which the linear relation is known to equal $\hat{\boldsymbol{\beta}}$, the estimated coefficients of the original regression. Both the original and bootstrapped samples are drawn from populations in which the errors are mean zero and uncorrelated with the regressors. However, while the errors in F_0 are homoskedastic and uncorrelated, in F_1 they are heteroskedastic and correlated. Most importantly, as m_{ii} = 1- h_{ii} is less than 1, the error process in F_1 has less variation than that in F_0, producing bootstrapped coefficients which vary less than the coefficients estimated in the original regression. These problems are once again related to leverage, as when maximal leverage goes to zero, **M** = **I** (the identity matrix), and the error process in F_1 is homoskedastic, uncorrelated and has the same variance as that of F_0.

As in the case of randomization tests, there are multiple possible ways of calculating the bootstrap. I shall focus on three which, parallel to the randomization tests described above, can be called the bootstrap-c, -p and -t. Let $\mathbf{B_n}$ denote the bootstrap sample randomly drawn from population F_1, $\hat{\boldsymbol{\beta}}(\mathbf{B_n})$ and $\mathbf{V}(\hat{\boldsymbol{\beta}}(\mathbf{B_n}))$ the coefficient and coefficient covariance estimates for that sample, and $\mathbf{V}(\hat{\boldsymbol{\beta}}(F_1))$ the covariance of $\hat{\boldsymbol{\beta}}(\mathbf{B_n})$ across the entire universe of bootstrapped samples from F_1. The original experimental sample, the parent population, can be denoted by $\mathbf{B_E}$. In testing the null hypothesis $\boldsymbol{\beta} = \mathbf{0}$, the bootstrap-c calculates the probability

(15) $[\hat{\boldsymbol{\beta}}(\mathbf{B_n}) - \hat{\boldsymbol{\beta}}(\mathbf{B_E})]'\mathbf{V}(\hat{\boldsymbol{\beta}}(F_1))^{-1}[\hat{\boldsymbol{\beta}}(\mathbf{B_n}) - \hat{\boldsymbol{\beta}}(\mathbf{B_E})] > \hat{\boldsymbol{\beta}}(\mathbf{B_E})'\mathbf{V}(\hat{\boldsymbol{\beta}}(F_1))^{-1}\hat{\boldsymbol{\beta}}(\mathbf{B_E}) .$

The test is centred around $\hat{\boldsymbol{\beta}}(\mathbf{B_E})$ because this is the null that is true for the parent population F_1. Again, for the special case of just one variable, one can ignore the fixed variance term that appears in both denominators and see that this amounts to comparing the squared variation

around the population value of the bootstrapped coefficient to the square of the originally estimated coefficient. The bootstrap-p, in contrast, compares the distributions of the estimated Wald statistics (or Wald p-values)

$$(16)\ [\hat{\boldsymbol{\beta}}(\mathbf{B_n}) - \hat{\boldsymbol{\beta}}(\mathbf{B_E})]' \mathbf{V}(\hat{\boldsymbol{\beta}}(\mathbf{B_n}))^{-1} [\hat{\boldsymbol{\beta}}(\mathbf{B_n}) - \hat{\boldsymbol{\beta}}(\mathbf{B_E})] > \hat{\boldsymbol{\beta}}(\mathbf{B_E})' \mathbf{V}(\hat{\boldsymbol{\beta}}(\mathbf{B_E}))^{-1} \hat{\boldsymbol{\beta}}(\mathbf{B_E}),$$

while the bootstrap-t compares the distribution of the estimated t-statistics around the bootstrapped coefficient correlation matrix

$$(17)\ [\mathbf{t}(\mathbf{B_n}) - \mathbf{t}(\mathbf{B_E})]' \boldsymbol{\rho}(\hat{\boldsymbol{\beta}}(F_1))^{-1} [\mathbf{t}(\mathbf{B_n}) - \mathbf{t}(\mathbf{B_E})] > \mathbf{t}(\mathbf{B_E})' \boldsymbol{\rho}(\hat{\boldsymbol{\beta}}(F_1))^{-1} \mathbf{t}(\mathbf{B_E}).$$

As in the case of the randomization-t, this comparison of studentized coefficient estimates allows for computation in situations where standard errors for individual coefficients are easily estimated on a sample-by-sample basis, but calculating sample specific estimates of the full covariance of all coefficients is much more difficult. I will use the bootstrap-t, in particular, to implement, in parallel with the omnibus randomization test, an omnibus bootstrap test of the significance of all coefficients estimated in an experimental paper.

As explained by Hall (1992), while the coverage error in a one-sided hypothesis test of a single coefficient of the bootstrap-t (and equivalently, in this univariate case, the bootstrap-p) converges to its nominal size at a rate $O(n^{-1})$, the coverage error of the bootstrap-c converges at a rate of only $O(n^{-½})$, 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 is asymptotically pivotal, i.e. has a distribution which does not depend upon unknowns. In contrast, the distribution of the coefficient is not pivotal, as it depends upon the estimation of its variance, which imparts additional inaccuracy. Asymptotics aside, because the error process in F_1 has less variation than that in F_0, bootstrapped coefficients tend to have less sampling variation and hence inferring the variance of the original coefficients from the bootstrapped coefficient distribution, as is done by the bootstrap-c, biases tests in favour of rejecting the null. This bias is absent in the bootstrap-t and -p, which compare the distribution of the original t statistic or Wald p-value, i.e. pivotal values which do not depend upon the level of an estimated variance. In my bootstrap analysis of the sample regressions further below, I find that the bootstrap-c produces a lower p-value than the bootstrap-p in 1230 of 2003 instances, and a higher p-value in only 616 cases.[39]

[39]In contrast, the bootstrap-t produces a lower p-value than the -p in 971 cases and a higher p-value in 876 cases.

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 coefficients. The covariance matrix is also used, along with the Chi^2 distribution, to perform Wald tests of linear hypotheses. If the bootstrapped coefficients are actually normally distributed, these p-values amount to calculating the probability

$$(18)\quad [\hat{\boldsymbol{\beta}}(\mathbf{B_n}) - E(\hat{\boldsymbol{\beta}}(\mathrm{F}_1))]'\mathbf{V}(\hat{\boldsymbol{\beta}}(\mathrm{F}_1))^{-1}[\hat{\boldsymbol{\beta}}(\mathbf{B_n}) - E(\hat{\boldsymbol{\beta}}(\mathrm{F}_1))] > \hat{\boldsymbol{\beta}}(\mathbf{B_E})'\mathbf{V}(\hat{\boldsymbol{\beta}}(\mathrm{F}_1))^{-1}\hat{\boldsymbol{\beta}}(\mathbf{B_E})$$

which is essentially the bootstrap-c.[40] Thus, Stata's default approach employs an asymptotically less accurate, systematically biased method and then adds to its inaccuracy by assuming normality[41] and dropping all of the information collected on other moments of the bootstrap distribution.[42] This unfortunate mixture stems from Stata's need to provide flexibility to users. Not knowing what the user wants to test, Stata calculates a bootstrap covariance matrix which is retained in memory for whatever post-estimation tests the user wants to perform. The "saving" option in the bootstrap line command allows users to save elements calculated from bootstrap iterations, and the "bsample" line command, as part of a larger user written programme, allows the flexibility to calculate more complicated test statistics in general. Unfortunately, in the one case in my sample where authors avail themselves of these commands, they use them to calculate Stata's bootstrap, i.e. the bootstrap-c evaluated using a normal distribution. In sum, the default bootstrap method in Stata, adopted for reasons of programme flexibility, appears to have been absorbed by its users. As noted, I find it systematically produces lower p-values than other well-known alternatives.[43]

---

[40]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.

[41]Even if the disturbances in the original regression are normal, the regressors need not be, and this can produce non-normally distributed bootstrapped coefficients.

[42]Additional inaccuracy is imparted by another feature of Stata's bootstrap. When a regressor is dropped (for example, because of colinearity in the bootstrapped sample) Stata usually drops that iteration. In the case of some commands, however, it stores a coefficient value of 0 for that iteration and retains the iteration in the calculation of the covariance matrix. When the original coefficient is near zero, this tends to reduce its estimated variance; when it is far from zero, it tends to increase its variance estimate.

[43]Cameron and Trivedi (2010), for example, in their text on Microeconometrics Using Stata, specifically suggest using the percentiles of studentized coefficients (i.e. the bootstrap-t).

Regarding practical methods, I implement the bootstrap in a manner that allows me to calculate the covariance of coefficients across equations for the omnibus test of significance. Thus, rather than bootstrap each equation separately, on each iteration I bootstrap an entire experimental sample and run all of the estimating equations for that sample. I sample at the cluster level, if the paper clusters, or the treatment level, if that is a higher level of aggregation. Thus, for example, in the case of laboratory experiment papers, where treatment varies by session, I sample sessions whether or not the paper clusters at that level. If a paper administers treatment at the level of the individual observation, and does not cluster, I sample individual observations. Iteration specific covariance estimates for the bootstrap-t and -p are calculated using authors' methods.[44]

IV: Empirical Results

(a) Statistical Significance

Table III summarizes the statistical significance of the regressions in experimental papers calculated using different criteria. It identifies two factors that lead to a systematic overstatement of statistical significance in these papers: (a) the absence of a summary evaluation of the combined significance of the multiplicity of tests that produce individually significant results; (b) the systematic overstatement of significance brought about by the use of test statistics whose distribution is only asymptotically known. In combination, these two effects produce a dramatic overstatement of the statistical significance of experimental treatments, with 30 to 40 percent of seemingly significant regressions, i.e. regressions in which at least some aspect of experimental treatment is reported as having a statistically significant effect, failing to reject the overall null of no effect whatsoever. Although there are 53 papers in my sample, the table drops one paper in which there are a large numbers of presentationally relevant "ties" in the randomization test.[45] The remaining 10 or so relevant ties are resolved in favour of rejecting the null (setting U in (9) equal to 0).

---

[44]Thus, in the particular case where the authors use the bootstrap, I bootstrap the bootstrap.

[45]As noted earlier, by presentationally relevant I mean cases where the lower and upper bounds of the probability defined by (9) cross over the .01 and .05 cutoffs. Ties which occur above these cutoffs (e.g. the p-value lies between .333 and .666) do occur, but I do not consider these "relevant", as regardless of the draw of U these regressions are insignificant. I should note that in the case of the one paper I drop the lower and upper bounds are 0 to .333, so if I were to apply (9) and draw a random number U for each regression most would be found to be insignificant at the .01 and .05 levels.

Table III: Statistical Significance of Treatment Effects Based Upon Different Criteria
(number of significant regressions by level of test)

	all 1965 regressions		1036 regressions with > 1 treatment variable	
	.01	.05	.01	.05
(1) significant coefficient	530	918	366	604
(2) standard Wald test	485	760	321	446
(3) randomization-p	374	643	239	364
(4) randomization-t	370	641	235	362
(5) randomization-c	389	648	252	366
(6) bootstrap-p	342	629	221	351
(7) bootstrap-t	347	639	224	362
(8) bootstrap-c	418	699	287	410

Notes: "significant coefficient": at least one treatment coefficient significant at the level specified. Standard Wald test: test using the F or Chi^2 distribution (depending on which Stata uses) of a Wald statistic calculated using the regression covariance matrix computed in the paper. Randomization-t and -p and bootstrap-t and -p calculated using the covariance matrix method used in the paper.

Beginning with the top row of the table, we see that of the 1965 regressions in the 52 papers, 530 contain at least one treatment coefficient which is significant at the .01 level and 918 contain at least one treatment coefficient which is significant at the .05 level. As noted in the Introduction, virtually none of these regressions include tests of the overall significance of multiple treatment coefficients, i.e. of the overall significance of experimental treatment. When conventional F/Wald tests of this sort are applied, in the second row of the table, one finds that fewer regressions are significant. As the Wald test yields the same significance level as that of the individual coefficient when there is only one treatment variable, this is most clearly illustrated in the right-hand panel of the table, where I restrict attention to regression specifications with more than one treatment variable. Of the 1036 regressions of this sort, 604 contain at least one coefficient which is significant at the .05 level, leading the reader to believe that experimental treatment was having some significant effect, at least on some dimension. However, when the overall significance of treatment is tested, only 446 specifications reject the null of no overall treatment effect at the .05 level. Thus, fully ¼ of seemingly .05 significant regression specifications cannot reject the null of no overall treatment effect at the same level.[46] At the .01

[46]Regressions in which no individual coefficient is significant can be found to be overall significant. However, this is only true for 16 of the 446 significant Wald tests in the right-hand panel. Thus, by and large here, and in most other cases in the table, the number of significant results found to be insignificant is vastly greater than

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	644	71	715	No	509	81	590
	Yes	76	245	321	Yes	45	401	446
	Total	720	316	1036	Total	554	482	1036

level, the gap is smaller, with 12 percent fewer regression specifications found to be significant.

Tables IV and V provide some insight into the relation between individually significant results, the F/Wald test results and the multiplicity of treatment measures. In Table IV I recalculate the Wald p-values using diagonalized versions of the estimated covariance matrices, i.e. matrices where the off-diagonal covariance terms are set to zero. As shown, there is a remarkable similarity between the p-values calculated using these artificial covariance matrices and those calculated using the estimated covariance matrix. In 1036 regressions with more than one treatment variable, statistical significance only differs in 147 cases at the .01 level and 126 cases at the .05 level. The mean p-value using the estimated covariance matrix is .227, using the diagonalized covariance matrix it is .237, and the correlation between the two is .911. 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 mutually orthogonal[47] 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 converse. So, it does not do too much injustice to the results to speak of the lower rows of the table as representing subsets of the top row, although this is not strictly correct.

[47]More precisely, the residuals of the treatment measures regressed on the non-treatment right-hand side variables are mutually orthogonal.

Table V explores the relationship between the existence of a .01 or .05 level significant coefficient in a regression and the number of treatment regressors using a simple model of specification search. Let P_{paper} represent the fundamental probability a paper's experiment will yield detectably significant results, said probability being the product of the probability the treatment has real effects and the underlying power of the sample to detect such effects. Further, let P_{paper} be independent of the number of treatment regressors that are added by the investigator; true effects are detectable or not, but manipulation of the regression does not improve the chances of doing so. As treatment measures are added to the regression, the probability of rejecting the null of no effects rises, but this is driven purely by the level of the test, and has no relation to the underlying significance of the experiment. Thus, the probability of finding at least one significant result at the .0i level (i = 1 or 5) as mutually orthogonal treatment regressors are added is given by

(19) $P_{paper} + (1-P_{paper})*(1-(1-.0i)^{\text{number of treatment regressors}})$,

with the second term representing the growing probability of rejecting the null brought on by specification search. Linearizing around P_{paper} = number = 0, we find that the expected value of a 0/1 indicator I_i of at least one significant result is given by $I_i = P_{paper} - \ln(1-.0i)*$number of treatment regressors. This motivates the linear regression of such indicators on paper fixed effects and the number of treatment regressors in Table V. A natural extension of the model is to consider the possibility that treatment has significant effects on some dependent variables and not others, so Table V also presents regressions with paper x dependent variable fixed effects.

Table V shows that as additional treatment variables are added to regressions significant effects appear at a rate approximately equal to that indicated by the level of the test. In column (1) we see that the coefficient on treatment regressors for .01 significant results is .012, which is not significantly larger than the .010 = -ln(1-.01) predicted by random chance. With paper x dependent variable fixed effects, in column (3), the point estimate falls to .006. For .05 significant effects, the point estimates with paper or paper x variable fixed effects, .028 and .025 in columns (2) and (4), are both substantially less than the predicted .051 = -ln(1-.05). In all four columns, however, the "constant term" or average y-intercept (equal to the mean of I_i minus β_{number} times the mean of the number of treatment regressors) is well above zero, indicating that P_{paper} is typically substantially positive. If one linearizes (19) around a positive value of P_{paper}, the

Table V: Significant Results Regressed on Number of Treatment Measures (N=1965)

	actual results				random probability			
	(1) I_1	(2) I_5	(3) I_1	(4) I_5	(5) P_1	(6) P_5	(7) P_1	(8) P_5
number	.012	.028	.006	.025	.007	.019	.006	.019
	(.008)	(.007)	(.006)	(.007)				
y-intercept	.233	.382	.252	.392	.234	.400	.252	.409
	(.023)	(.021)	(.019)	(.021)				
paper f.e.	yes	yes			yes	yes		
variable f.e.			yes	yes			yes	yes
R^2	.206	.239	.605	.613	1.00	.995	1.00	.997

Notes: I_i = indicator for a significant coefficient at the .0i level; P_i = calculated probability of a significant coefficient at the .0i level using (19) above with P_{paper} given by the estimated fixed effects of columns (1)-(4); N = number of observations; number = number of treatment regressors; y-intercept = mean of the dependent variable minus β_{number} times mean of number; paper f.e. = paper fixed effects; variable f.e. = paper x dependent variable fixed effects. Standard errors clustered at the paper level.

coefficient on the number of treatment regressors falls to $-(1-P_{paper})\ln(1-.0i)$. To see how far this goes in explaining the results, in columns (5)-(8) I run regressions using artificial probability data that exactly matches equation (19), with P_{paper} set at the values estimated by the fixed effects of columns (1)-(4). The "R^2"s of columns (5)-(8) are all nearly 1, showing that the linearized equation fits the non-linear data generating process extremely well.[48] The estimated y-intercepts are also quite close to those in columns (1)-(4). Finally, and most significantly, the coefficients on the number of treatment regressors are all not far below those found in the actual data, particularly in the specifications with paper x variable fixed effects. In sum, the rate at which significant effects appear as additional treatment measures are added to regressions within papers is approximately consistent with a data generating process in which results are either significant or not and additional regressors add "significance" at a rate equal to, or only slightly better than, that predicted by the level of the test.

Returning to Table III, I use randomization and bootstrap tests to show that accepted econometric practice leads to an overstatement of statistical significance. As shown in the third row of the table, when randomization Wald-tests are applied, experimental treatment appears to

[48]This is not a consequence of fitting the variation in the large number of fixed effects. If I set P_{paper} for all papers and dependent variables in the data generating process equal to the y-intercepts in columns (1)-(4) and run the regressions without fixed effects, the R^2s for columns (5)-(8) are .999, .982, .999 and .982, respectively. The coefficients on number in these regressions (.007, .022, .007, and .021, respectively) are larger than those reported in columns (5)-(8) and hence even closer to the estimates in columns (1)-(4).

be much less significant than indicated by Wald tests using the estimated covariance matrices. While 485 regressions reject the null at the .01 level using the conventionally calculated Wald test, only 374 reject the null of no effect at the same level using the randomization-p criterion. Similarly, while 760 regressions reject the null of no overall treatment effect at the .05 level using conventional econometrics, only 643 reject the same null when the percentiles of realized Wald p-values are calculated using the randomization-p. Comparing the 1st and 3rd rows in the left-hand panel of the table, one sees that in total about 30 percent of all regressions in which the reader might be led, by the presence of a conventionally significant treatment coefficient, to believe that the experiment was having some effect, one cannot reject the Fisherian null that the experiment had absolutely no effect, whatsoever, on any participant. Comparing the 1st and 3rd rows in the right-hand panel of the table, one sees that this discrepancy rises to 35 to 40 percent if one focuses on regressions with more than one treatment coefficient.[49]

Table III also reports results based upon the non-studentized randomization-c, which relies upon the covariance of the randomized coefficients, and the randomization-t, whose use of studentized coefficient estimates and the coefficient correlation matrix places it between the randomization-p and randomization-c (see Section III). As noted earlier, Lehmann (1959) found that the randomization-p was uniformly most powerful in the simple comparison of treatment means. In simulations below, I find that when the null is false the randomization-p on average produces lower p-values and higher rejection probabilities than the randomization-c, with the randomization-t lying between the two, although the differences are very small. In application to the actual experiments themselves, however, there is no such ranking of rejection probabilities, as shown in Table III, or the mean p-values of the -p, -t, and -c (.278, .283, and .278, respectively). Of course, if the null is often true the three tests have the same distribution and any minor power ranking of p-values and rejection probabilities will tend to disappear.

[49]Although Table III uses 10000 draws from the universe of potential randomized outcomes to simulate the randomization distribution as accurately as possible, it appears that far fewer are necessary. The mean p-values across the 1965 regressions calculated using 100, 200, 500, 1000, and 10000 draws in the randomization-p are .278, .279, .279, .278 and .278, respectively. The correlations of the p-values with 100, 200, 500 and 1000 draws with those computed with 10000 draws are .994, .997, .999, and .999, respectively. The number of rejections with 100, 200, 500, 1000 and 10000 draws at the .01 (.05) level are 411 (625), 384 (635), 381 (640), 372 (632), and 375 (641), respectively. Based upon these results, I use 200 draws in power simulations further below.

Randomization tests of regressions in randomized experiments should be compelling, as they are exact (i.e. have precisely known size) under any possible distribution of the error term and, as shown further below, have power which is similar to that of conventional econometric tests. Nevertheless, to provide additional evidence, in the bottom rows of Table III I report supporting results based upon bootstrap statistical inference. Tests of the overall significance of treatment based upon Wald tests using the bootstrapped percentiles of the Wald test, the bootstrap-p, or the studentized bootstrap-t, reject the null of no treatment effect even less often than randomization tests, while tests based upon the coefficient bootstrap, the bootstrap-c, reject the null somewhat more often. Theoretically, as noted earlier, the bootstrap-p and -t are more accurate than the bootstrap-c. All of the bootstrap tests, however, reject the null of no treatment effect less often than Wald tests calculated using each regression's estimated covariance matrix. Hence, whatever their relative merits, together they strongly support randomization tests in indicating that statistical inference based upon the covariance matrices calculated in the sample papers systematically overstates the significance of results.

Table VI provides a breakdown of the discrepancy between the significance of conventionally calculated Wald tests and their randomized and bootstrapped counterparts by the form of the covariance matrix used in the regression. While results with the default covariance matrix, i.e. the standard covariance matrix for that regression type computed by Stata, are fairly close to randomization and bootstrap outcomes, all of the optional covariance forms selected by authors do poorly, be they clustered, robust, bootstrapped or "other".[50] The last category encompasses bias adjustments of the clustered or robust covariance matrices designed to correct their known failings which unfortunately do not appear to do particularly well.[51] Authors' choice of covariance matrix is, however, endogenous to the problems they face in their regressions, so this table, while reporting results readers would probably like to see, is not necessarily

[50]The reader might note that the bootstrap-c produces lower rejection rates than Stata's bootstrap, which is a variant of the bootrstrap-c with the normality assumption. In using the bootstrap I make full use of the percentiles of the bootstrapped coefficient distribution. In the regressions in the bootstrap column, variance adjusted tails are thicker than implied by the normal distribution, producing lower rejection rates.

[51]In Young (2016) I apply the hc3, brl and other bias adjustments in simulations with the entire sample, and show that by themselves they do poorly. Bias and "effective degrees of freedom" corrections, adjusting for the excess thickness of distribution tails, however, render test statistics with the robust and clustered covariance estimates nearly exact. I use such distributional adjustments further below to re-evaluate the conventional test statistics calculated using non-default methods.

Table VI: Significance by Form of the Wald Test & Regression Covariance Estimate

	default N = 417		clustered N = 987		robust N = 344		bootstrap N = 126		other N = 91	
	.01	.05	.01	.05	.01	.05	.01	.05	.01	.05
(1) standard Wald	84	133	273	436	99	132	16	31	13	28
(2) randomization-p	72	119	218	388	73	96	8	24	3	16
(3) randomization-t	74	120	226	401	58	77	10	28	2	15
(4) randomization-c	77	129	245	403	58	84	8	20	1	12
(5) bootstrap-p	81	131	173	339	73	113	10	23	5	23
(6) bootstrap-t	84	138	168	338	86	122	5	22	4	19
(7) bootstrap-c	89	141	231	399	82	111	11	25	5	23

Notes: N = number of regressions; .01/.05 = significant at this level. "Bootstrap" in the column title is Stata's bootstrap, i.e. the bootstrap-c with the addition of a normality assumption.

informative. To establish more clearly the problems of conventionally calculated covariance matrices, in the next section I use size simulations with known error processes. These show that statistical inference based upon the covariance matrices used in these papers is systematically biased in favour of rejecting the null and that this bias can explain most of the difference between randomization and conventional results.

Table VII presents the results of the omnibus (stacked) randomization and bootstrap tests of overall experimental significance. When this test is applied to the 53 papers in my experimental sample, presentationally relevant randomization ties occur in the case of the paper noted above and one further paper (both laboratory experiments).[52] The results for the remaining 51 papers are presented in the table. In the first column of panel A I apply the omnibus test to all treatment outcome regression specifications. With the randomization-t only 14 papers are found to be significant at the .01 level and 20 at the .05 level. The randomization-c provides better results, with 19 and 24 papers significant at those levels, but still leaves the conclusion that only 40 to 50 percent of experiments, when all of the treatment regressions are evaluated together, reject the null of no experimental effect whatsoever. The bootstrap confirms these results, with the different bootstraps once again bracketing the randomization results. A critic might argue that these tests are somewhat unfair, as they include outcomes that the authors did not

[52]Based on the ties, the randomization p-values for these papers lie between 0 and .333/.335. Thus, as in the case of Table III, if I resolved these ties with the draw of a random number, the two papers would rarely reject at the .01 or .05 level.

Table VII: Omnibus Test of Experimental Significance
(51 papers, excluding 2 papers with substantial ties)

(A) # of papers meeting significance level

	(1) all treatment outcomes		(2) .05 significant treatment outcomes		(3) .01 significant treatment outcomes		(4) block diagonalized covariance	
	.01	.05	.01	.05	.01	.05	.01	.05
randomization-t	14	20	15	21	17	24	28	36
randomization-c	19	24	21	27	23	28	31	38
bootstrap-t	7	15	9	20	9	18	26	32
bootstrap-c	30	34	31	35	29	35	38	44
	aer (N = 28)		~ aer (N = 23)		lab (N = 13)		~ lab (N = 38)	
	.01	.05	.01	.05	.01	.05	.01	.05
randomization-t	8	12	6	8	2	5	12	15
randomization-c	9	12	10	12	3	3	16	21
bootstrap-t	3	7	4	8	3	3	4	12
bootstrap-c	15	18	15	16	10	11	20	23

(B) average cross-coefficient correlation

	all			within equations			between equations		
	all	≤ .01	>.01	all	≤ .01	>.01	all	≤ .01	>.01
randomization	.119	.226	.100	.016	.039	-.022	.124	.236	.108
bootstrap	.111	.196	.104	-.009	.004	-.029	.118	.204	.113

(C) average cross-equation Wald p-value correlation

	all	≤ .01	> .01
randomization	.218	.406	.153
bootstrap	.218	.366	.169

Notes: .05 and .01 treatment outcomes = only including regressions with a dependent variable that generates a significant treatment coefficient at that level in some regression specification; block diagonalized covariance = cross equation coefficient covariances set to 0; 01/.05 = # of papers significant at this level; N = total number of papers in this category; all = all estimating equations; ≤.01 (>.01) = equations with at least one (no) coefficient reported significant at the.01 level.

actually expect, ex ante, to be influenced by experimental treatment.[53] To allow for this, in the second and third columns of panel (A) I restrict the analysis to equations that include dependent variables that show at least one statistically significant treatment effect at the .05 or .01 level in

[53]As noted earlier above, I limit the analysis in column (1) to regressions with treatment effects as the dependent variable, and hence exclude regressions associated with randomization balance or attrition, as well as 14 first stage regressions where the instrumented variable is not used as a treatment outcome elsewhere in the paper and treatment affects the instrumented independent variable virtually by construction (see the discussion of iv methods earlier above). I also exclude regressions associated with non-experimental cohorts.

some specification. Thus, these columns exclude treatment outcomes that are never associated with a significant treatment coefficient anywhere in the paper. As shown, this improves the results only minimally, adding 3 or 4 significant papers at best. The omnibus finding that most papers do not have statistically significant treatment effects is not due to the presence of outcome variables that are never found to be significant. Rather, it is a consequence of the fact that outcomes that generate statistically significant treatment coefficients in some specifications generate insignificant coefficients in many others. The lower part of panel A shows that significance rates, using all treatment outcomes, do not differ much between AER and non-AER papers, and that laboratory papers show somewhat lower significance rates than non-laboratory papers, but the sample sizes are so small the differences are meaningless. I present this information only to assure the reader that the results are not concentrated in one particular sub-group or the other.

The fourth column of panel A uses an equation-level block diagonalized covariance matrix to calculate the omnibus test of statistical significance with all treatment outcomes. In stark contrast with the similar within-equation exercise in Table IV, this artificial calculation substantially increases the number of significant papers. Thus, cross-equation correlation plays a very big role in producing lower statistical significance, while cross-coefficient correlation within equations (earlier in Table IV) played no role in determining significance. Panels B and C of the table explain why. In panel B I examine the average cross (treatment) coefficient correlation calculated across randomization or bootstrap iterations. For the randomization iterations, this is .119 across all coefficients, .016 when calculated within equations, and .124 when calculated across equations (i.e. excluding the within equation coefficient correlations). Thus, on average treatment coefficients within equations are uncorrelated, while treatment coefficients between equations are correlated. Performing the same calculations for equations that have a coefficient that is significant at the .01 level and, separately, equations that do not have a coefficient that is significant at the .01 level, we see that treatment coefficients are much more correlated between significant equations (.236) than between insignificant equations (.108). The bootstrap shows very similar patterns. Panel C of the table examines the correlation across randomization and bootstrap iterations of the p-values of the equation level Wald tests of statistical significance (calculated, in each case, using the covariance method chosen by the authors). Again, in the case

of randomization iterations, the average correlation of p-values across equations is .218, but this rises to .406 if one restricts attention to equations with a statistically significant treatment coefficient, and falls to .153 if one only considers equations without a statistically significant treatment coefficient. Bootstrap correlations confirm this in a population sampling framework.

The average paper in the 51 paper sample of Table VII has 10 treatment outcome equations with at least one .01 level statistically significant treatment coefficient and 28 equations without a statistically significant coefficient. As panels B and C of Table VII show, the treatment results produced in equations with statistically significant coefficients are highly correlated, while the treatment results in equations without significant coefficients are relatively uncorrelated. Thus, the large number of insignificant equations provides much more information about the lack of experimental significance than the small number of significant equations. The results of the omnibus randomization and bootstrap tests reflect this.

An equivalent, conventional, test of overall experimental significance is theoretically possible, but practically difficult to implement. White (1982) showed that an asymptotically accurate estimate of the covariance matrix for all of the coefficients estimated in multiple estimation 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). Stata's suest (seemingly unrelated estimation) command provides such estimates. The practical barriers to its implementation, however, are staggering. Many estimation procedures do not produce scores and hence are not supported by Stata's command. 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,[54] 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. In all, I am able to calculate a suest test which does not drop treatment

[54]In contrast, in implementing the randomization and bootstrap omnibus tests I need only calculate the covariance of the realized (in each randomization or bootstrap iteration) *treatment* coefficients and t-statistics, as opposed to trying to calculate a measure that, in standard conventional fashion, depends upon *all* regressors (recall, for example, $\mathbf{X}'\mathbf{X}^{-1}$).

Table VIII: Aggregation and Statistical Significance (at the.01 level)

	52 papers of Table III		9 papers with successful suest	
	5041 coefficients	1036 equations	154 equations	9 papers
conventional	605	321	96	7
randomization-t	474	235	87	3
bootstrap-t	509	224	88	1
rand-t/conventional	.78	.73	.91	.43
boot-t/conventional	.84	.70	.92	.14

variables (because of singularity in the subset of the covariance matrix associated with treatment coefficients) for only 9 of the 51 papers examined in Table VII. The conventional omnibus test finds that 7 of these 9 papers are significant at the .01 and .05 levels. In contrast, in the omnibus test of the upper-left hand corner of Table VII, at the .01 level only 3 of these papers reject the null in the randomization-t and 1 in the bootstrap-t. A conventional omnibus test is generally impossible to implement and, where possible, disagrees dramatically with randomization and bootstrap inference. In the size simulation of the next section I show that the suest test is extraordinarily biased, rejecting the null of no treatment effect, when it is true, an average of .30 of the time at the .01 level in this sample of 9 papers.

The results of the preceding paragraph highlight a point made in the Introduction, namely that the gap between conventional and randomization and bootstrap tests rises with the level of aggregation. This is illustrated in Table VIII. Examining the 5041 coefficients that appear in 1036 equations with more than one treatment variable in the 52 paper sample of Table III earlier, significance rates at the .01 level in the randomization-t and bootstrap-t are .78 and .84, respectively, of those achieved by the conventional t-test. When the significance of these coefficients is tested at the equation level, however, the randomization-t and bootstrap-t only show .73 and .70 the significance rate of the conventional F or Wald test. In the 154 equations of the 9 papers in which it is possible to successfully complete a conventional seemingly unrelated omnibus test of overall significance, the relative significance rate of the randomization-t and bootstrap-t Wald tests is just above .9. At the paper level, however, their relative significance rate is just .43 and .14, respectively. Excess size or, from a different perspective, greater power compounds as multiple tests are implicitly combined. I present systematic evidence of this in

both the size and power simulations below. For this reason, after evaluating the relative power of the randomization omnibus test further below, I ultimately restrict attention to those papers where simulation evidence shows that it has power equal to that of conventional tests. I still find that ½ to ⅔ of papers cannot reject the null of no treatment effect at the .01 level.

(b) Size

In this section I use simulations to show that the asymptotically-valid covariance estimation methods typically used in published papers generate empirical coverage, i.e. rejection rates when the null is true, that is systematically greater than nominal size. My focus is on the clustered and robust covariance matrices, which are the most popular methods. I show that coverage using these techniques is biased, both when errors are ideal and when they contain the pathologies these methods are supposed to address. This stems from the way, as noted earlier above and developed further in Young (2016), hypothesis tests interact with regression design in these methods to place uneven weight on a limited number of residuals (reducing effective degrees of freedom) which have, due to leverage, higher or lower than average variance (generating variation in bias that, through Jensen's inequality, raises average t-statistics). I show that maximum leverage is a better predictor of coverage bias than sample size, establishing that leverage is a clearer measure of how "asymptotic" the sample is. When simulations designed to capture the impact of regression design on the distribution of the test statistic are used to evaluate the conventional test statistics of my sample papers, significance rates move close to those of randomization inference.

I begin by simulating size with ideal errors, using baseline no-treatment regressions to parameterize a data generating process in which the null hypothesis of no treatment effect is literally true. To illustrate with OLS, as noted earlier the typical experimental regression is of the form $y_i = \mathbf{t}_i'\boldsymbol{\beta}_\mathbf{t} + \mathbf{x}_i'\boldsymbol{\beta}_\mathbf{x} + \varepsilon_i$, where $\mathbf{t}_i$ and $\mathbf{x}_i$ are vectors of treatment and non-treatment variables and $\boldsymbol{\beta}_\mathbf{t}$ and $\boldsymbol{\beta}_\mathbf{x}$ their associated coefficients. Under the assumption of no treatment effect, I run the regression $y_i = \mathbf{x}_i'\boldsymbol{\beta}_\mathbf{x} + \varepsilon_i$, and then use the predicted values and normally distributed iid disturbances with a variance equal to the estimated variance of ε_i to create 10000 simulated sets of data. On these I then run the full regression with the treatment measures added and perform Wald tests of their significance using the covariance matrix selected by the authors. If, instead, the authors run a random effects regression of the form $y_{ij} = \mathbf{t}_{ij}'\boldsymbol{\beta}_\mathbf{t} + \mathbf{x}_{ij}'\boldsymbol{\beta}_\mathbf{x} + v_i + \varepsilon_{ij}$, where v_i

denotes a group random effect and ε_{ij} an individual observation error, I run $y_{ij} = \mathbf{x}_{ij}'\boldsymbol{\beta}_{\mathbf{x}} + v_i + \varepsilon_{ij}$ and use the predicted values plus random shocks based upon the estimated variances of v_i and ε_{ij} to produce simulated data. If the authors use maximum likelihood techniques, I simulate the data using the likelihood equation. If they use weighted regression to correct for known heteroskedasticity, I use the estimated variance of the regression and the weights to create heteroskedastic error terms which are then made homoskedastic by the GLS transform. In each case I create simulated data based upon a regression run without treatment effects and error terms that are ideal within the context of the regression model, i.e. are iid and normally distributed except insofar as the regression equation itself explicitly posits an alternative distribution. A full description of methods is provided in the on-line appendix. I drop 103 regressions where it is unclear, from the method used, what would constitute "ideal" errors.[55]

The results of these simulations are reported in Table IX. For each regression I calculate the frequency with which the null of no treatment effect is rejected at the .01 and .05 levels and then report the mean, standard deviation, minimum and maximum, across all of the regressions, of this statistic. The first line of panel A provides a useful benchmark, as it relates to OLS regressions calculated using the default covariance matrix which in this case, with fixed regressors and normally distributed iid residuals, is exact. As shown, average rejection probabilities at the .01 and .05 level are precisely what they should be and the standard deviation of size, at .001 and .002 at the two levels, is minimal and exactly what should be expected from simulations with 10000 iterations per equation.

Moving down the table to the other OLS regressions, one immediately sees the rejection bias of the clustered and robust covariance matrices, which on average reject .024 and .028 of the time, respectively, at the .01 level, with the most spectacular outcomes showing rejection probabilities of the null (when it is actually true) of .662 and .502. At the .01 level, the standard deviation of size using these methods is 48 or 76 times higher than the simulation induced error

[55]These include 26 quantile regressions, 16 seemingly unrelated estimations of equation systems, and 58 two step regressions, as well as 3 more equations that I could not simulate because of technical problems with the original estimating equation. If I simulate the quantile regressions using an OLS iid normal baseline equation and the seemingly unrelated estimations by simulating independent OLS errors by equation, they have average rejection rates at the .01 level of .013 and .154, respectively. The two step regressions and many of the quantile regressions were bootstrapped, so I combine the remaining OLS bootstrapped results (all from one paper) with the smorgasbord of "other" techniques in the tables below.

Table IX: Empirical Coverage/Size by Type of Regression and Covariance Estimate
(model specific ideal iid error terms, 10000 simulations per regression)

			at .01 level				at .05 level			
		#	mean	sd	min	max	mean	sd	min	max
(A) based on covariance matrix used in paper										
OLS	default	328	.010	.001	.007	.013	.050	.002	.044	.056
	clustered	827	.024	.048	.002	.662	.077	.066	.017	.744
	robust	164	.028	.076	.008	.502	.077	.092	.041	.636
	other	59	.017	.004	.008	.026	.063	.009	.039	.080
~OLS	default	110	.018	.027	.000	.226	.065	.041	.006	.308
	clustered	152	.045	.068	.000	.389	.102	.088	.009	.508
	robust	180	.011	.004	.001	.031	.054	.011	.018	.113
	other	80	.018	.003	.008	.028	.064	.005	.045	.076
(B) substituting default covariance matrix										
OLS	clustered	827	.010	.001	.007	.014	.050	.002	.045	.061
	robust	164	.010	.001	.008	.014	.050	.002	.045	.056
	other	59	.010	.001	.007	.013	.050	.002	.044	.057
~OLS	clustered	152	.010	.002	.001	.015	.051	.005	.023	.067
	robust	180	.010	.002	.002	.030	.051	.007	.030	.111
	other	80	.010	.001	.006	.013	.050	.003	.041	.055

Notes: # = number of regression specifications. Other data reported are the mean, standard deviation, minimum and maximum rejection probability at the specified level. ~OLS = not ordinary least squares. "other" includes bootstrap, jackknife, robust hc3 and clustered bias reduced linearization.

for the default OLS covariance matrix. "Other" methods, made up of the bootstrap, jackknife hc3 and brl corrections of the robust and clustered covariance estimates, do better, but are still biased in favour of rejection. The results of the non-OLS simulations are even more discouraging. Here, the default covariance matrix, despite the fact that the errors exactly match the regression model, has a mean rejection probability of .018 at the .01 level, reaching a maximum of .226 in one case. The clustered covariance matrix is strongly biased, averaging a .045 rejection probability at the .01 level with a standard deviation 68 times that of the default method OLS simulations at the top of the table. "Other" methods continue to do poorly, while the robust method, in this sample, is variable but on average only slightly biased.

Panel B of Table IX highlights the role the choice of covariance matrix plays in these results by recalculating, for the regressions that did not use default methods, rejection probabilities when Wald tests are calculated using the default covariance matrix. As expected,

average coverage is exact, with the predicted simulation variation, in each group of OLS papers. More interestingly, in non-OLS papers average coverage is generally very close to nominal size and the standard deviation of coverage is quite close to that for OLS simulations. Thus, the coverage bias in non-OLS regressions that used the default covariance matrix (shown in panel A) is unusual in that generally default methods, in the face of ideal errors, perform quite well.[56]

Table X below extends the analysis by considering the type of departure from ideal errors that the clustered covariance matrix is designed to correct, namely correlation within clusters, which I model with cluster level random effects. The sample excludes regressions where the baseline specification includes cluster fixed effects or explicitly models cluster level random effects and the results are divided by the share of the total error variance accounted for by the random effect. Once again I use a data generating process parameterized by baseline no-treatment equations. Thus, if the original regression is a regular OLS specification $y_i = \mathbf{t}_i'\boldsymbol{\beta}_\mathbf{t} + \mathbf{x}_i'\boldsymbol{\beta}_\mathbf{x} + \varepsilon_i$, I run the preliminary regression $y_i = \mathbf{x}_i'\boldsymbol{\beta}_\mathbf{x} + \varepsilon_i$ to gather predicted values and an estimate of the error variance σ^2. I then divide σ^2 into an iid component ε_i and a cluster level component u_g (g denoting the cluster group) with $\rho=\sigma_u^2/(\sigma_u^2+\sigma_\varepsilon^2)$ indicating the share of the total residual variance accounted for by the cluster level random effect. If the baseline regression included non-cluster random effects, I include these, but again take the estimate of the residual variance σ^2 and divide it into iid and cluster disturbances. If the original specification was maximum likelihood with iid errors, I take the distributionally imposed observation level variance and divide it into an iid shock and a cluster level disturbance. In each case, the cluster level disturbance represents an unknown shock from the point of view of the regression specification, the type of unknown shock that clustering is supposed to address.

As shown in the table, the excess size of clustered regressions does not improve, in any way, in the presence of unknown cluster level disturbances. Moving down the table, we see that as ρ goes from 0 to .8 the mean rejection probability, and its standard deviation, actually rises ever so slightly as more extreme outcomes, with rejection probabilities of .753 and .908 at the .01

[56]The non-OLS regressions using default methods that produced unusually high rejection probabilities in panel A are interval regressions, heckman selection bias models, maximum likelihood estimated random effects regressions and multinomial logit models. The non-OLS regressions not using the default methods contain a much greater representation of probit & logit models (.49 of the regressions versus .28 in the earlier group) and these have default mean coverage that is almost identical to nominal size.

Table X: Coverage of Clustered Regressions with Random Effects at the Cluster Level
(653 OLS & 127 non-OLS Regressions, 1000 simulations per equation)

	at .01 level				at .05 level			
$\rho = \sigma_u^2/(\sigma_u^2+\sigma_\varepsilon^2)$	mean	sd	min	max	mean	sd	Min	max
				with clustered covariance matrix				
.0	.027	.051	.002	.410	.081	.070	.020	.597
.2	.030	.053	.001	.497	.085	.073	.032	.670
.4	.030	.056	.001	.753	.086	.075	.030	.857
.6	.030	.057	.001	.862	.087	.076	.031	.927
.8	.030	.057	.001	.908	.087	.075	.030	.961
				with default covariance matrix				
.0	.010	.003	.002	.021	.050	.007	.033	.072
.2	.215	.200	.003	.816	.313	.216	.029	.880
.4	.313	.246	.001	.905	.411	.252	.022	.944
.6	.377	.268	.001	.943	.472	.267	.021	.969
.8	.425	.281	.000	.963	.515	.276	.015	.983

Notes: ρ = share of total error variance due to random effect. Otherwise, see Table IX.

level, appear. While noting the shortcomings of the clustered covariance estimate, it is nevertheless important to recognize the even greater perils of not clustering or explicitly modelling random effects in situations where random effects might arise. In the bottom panel of Table X I run the same regressions using the default covariance matrix. As shown, average rejection probabilities when the null is true are extraordinarily high. This point was made by Moulton (1986), who showed that the proportionate underestimation of coefficient variation using the default OLS covariance matrix was increasing in the within group correlation of the independent variable. Within group correlation of treatment is greatest when treatment is, literally, applied to groups of observations. This happens in 36 papers in my sample.[57] Of these, 25 cluster or explicitly model random effects at that grouped level or at a greater degree of data aggregation in at least some of their regressions. 11 of these papers, however, either do not cluster at all or cluster at a lower level of aggregation. While the clustered covariance matrix

[57]By which I mean treatment is applied in groups of data as appear in regressions. Thus, if treatment is applied to a male and female couple, but each regression is done one sex at a time, treatment is not applied to grouped data. Conversely, if treatment is applied to individuals, but multiple observations for each individual appear in a regression, then treatment is applied to grouped data.

generally overstates the statistical significance of results, using the default covariance estimate, in situations where random effects at the group treatment level might arise, is not necessarily a better choice, as shown in Table X.

Table XI below illustrates the ineffectiveness of the robust covariance matrix in situations it is designed to handle using two examples. The first is population weighted regressions. Population weighted regressions are similar to generalized least squares weighted regressions, in that both the dependent and independent variables are multiplied by the inverse of the weights. However, in GLS the weights represent an estimate of the proportional heteroskedasticity of the different observations, so the GLS transform produces homoskedastic errors. In contrast, in population weighted regressions weights are used with the intention of arriving at a population-weighted average estimate of the coefficients. If the original error was homoskedastic, the error in the weighted regression becomes heteroskedastic. Consequently, the standard approach in population weighted regressions is to use the robust covariance estimate. One can easily, however, calculate the same regression with the default OLS covariance estimate, i.e. ignoring the heteroskedasticity in the error term.[58]

One paper in my sample makes extensive use of population weighted regressions. In Table IX earlier I followed my baseline procedure, simulating size for this paper by adding homoskedastic disturbances to the point estimates produced by a non-treatment estimating equation. When subsequently estimated using population weights, these regressions contain intrinsically heteroskedastic errors. There are basically two population groups in this paper, and the weights used create an error variance in the minority group (about 20 percent of the sample) that is 1.6 to 1.7 times that of the majority group. Panel A of Table XI extracts, from the grouped data of Table IX, the default and robust simulations for this paper. As shown, the default OLS estimate of covariance produces coverage which is slightly biased in favour of rejection, but still close to nominal value. When the robust covariance estimate is applied to solve the inference problem, rejection probabilities actually rise and become more variable.

A second example, again motivated by my sample, is the linear regression model. As noted earlier, in approximately 40 percent of the OLS regressions in my sample the dependent

[58]Specifying aweights rather than pweights in Stata produces the same weighted coefficients but with the default covariance estimate.

Table XI: Coverage of Baseline and Robust Methods with Heteroskedasticity
by Type of Covariance Estimate (OLS regressions, 10000 simulations per regression)

		at .01 level				at .05 level			
	#	mean	sd	min	max	mean	sd	min	max
(A) population weighted regressions									
default	44	.011	.002	.008	.016	.052	.005	.044	.065
robust	44	.015	.003	.011	.022	.064	.007	.054	.084
(B) regressions with binary y values (linear probability model)									
default	591	.014	.044	.000	.976	.057	.050	.000	.986
robust	591	.026	.078	.000	.694	.070	.080	.000	.706

Notes: See Table IX.

variable is dichotomous, taking on 0/1 values. Linear probability models of this type are inherently heteroskedastic as, given the underlying model $y_i = \mathbf{x}_i'\boldsymbol{\beta}_\mathbf{x} + \varepsilon_i$, the error variance of observation i is given by $(1- \mathbf{x}_i'\boldsymbol{\beta}_\mathbf{x})\mathbf{x}_i'\boldsymbol{\beta}_\mathbf{x}$. In the size simulations in Table IX above, however, I added homoskedastic normal disturbances based upon the standard error of the non-treatment equation to point estimates, producing normally distributed dependent variables y_i. I now correct this, explicitly simulating a 0/1 dichotomous variable based upon the non-treatment regression,[59] producing heteroskedastic errors.

Panel B of Table XI shows how the default and robust covariance matrices fare in these circumstances.[60] The standard OLS covariance matrix is biased in favour of rejection, with coverage of .014 at the .01 level, and a standard deviation of coverage (.044) which is 44 times larger than that found when using standard methods on homoskedastic errors (Table IX earlier). However, when the robust covariance matrix is applied in these regressions, things get much

[59]It is not possible to generate simulated data using the authors' linear probability models, as the point estimates these produce grossly violate the assumptions of the linear model. To illustrate, I note that the maximum absolute residual error exceeds 1 in 205 of the 591 original regressions the authors performed. In order to be able to simulate a random 0/1 dependent variable with the linear model it is necessary that the point estimates of the baseline equation lie between the dependent variable values 0 and 1, and this restriction is clearly not met. My approach is to estimate a probit equation using the non-treatment regressors and use the predicted probabilities of the probit to produce simulated data. This simulation is somewhat different than the others, as the data no longer strictly satisfy the non-treatment equation $y_i = \mathbf{x}_i'\boldsymbol{\beta}_\mathbf{x} + \varepsilon_i$ (instead the $\mathbf{x}_i'\boldsymbol{\beta}_\mathbf{x}$ must be taken as the linearization of the underlying probit equation). However, it is still the case that in the data generating process the treatment variables have no effect on the dependent variable, so it is still a simulation of size.

[60]The robust covariance matrix was not necessarily originally used in these particular regressions. I am simply applying it to illustrate its effects in a broader sample.

worse. The rejection probability at the .01 level almost doubles, to .026, as does the standard deviation of coverage (to .078). In sum, as shown by the two examples in Table XI, when confronted with the type of problem it is putatively designed to solve, i.e. heteroskedastic errors which are not accounted for in the regression specification, the performance of the robust covariance matrix is, both in an absolute sense and relative to the default estimate of covariance, highly variable and systematically biased in favour of rejecting the null.

Tables XII and XIII below show that maximum leverage explains most of the average coverage bias of test statistics based upon the robust and clustered covariance matrices. In each panel the dependent variable is ln empirical (simulated) size at the .01 level in the Wald test of treatment coefficients divided by .01, so a value of 0 represents unbiased coverage. Comparison of the constant term or intercept of these regressions with the mean value of the dependent variable shows how much of mean bias is explained by the regressors, excluding any fixed effects.[61] For regressions with maximum leverage, in particular, the intercept indicates the bias when maximum leverage attains its minimum value of zero. 827 OLS regressions in my sample use the clustered covariance matrix, but only 164 use the robust covariance matrix, so I expand that sample by running simulations with the robust matrix for all 1378 OLS regressions.

Table XII begins with an analysis of the bias of tests based on the robust covariance matrix in simulations with normally distributed ideal iid errors (the entire OLS sample in Table IX) or heteroskedastic errors produced by dichotomous dependent variables (as featured in Table XI above). As shown in columns 1-3 and 6-8, sample size has no significant relation to coverage bias. Point estimates are often of the wrong sign (positive), and when negative are easily rendered utterly insignificant with the addition of paper fixed effects or maximal leverage as regressors. Columns 3-5 show that maximal leverage is significantly correlated with coverage bias, especially when paper fixed effects are added to the regression in column 5, where it accounts for pretty much all of mean coverage bias (i.e. the intercept is near zero). As noted earlier, a maximal leverage of 1 might be the result of extreme regressors with little relevance for areas of interest in the regression. There is evidence of this in the sample. When the regressions are run in the right-hand panel without observations with $h_{ii}^{\max} = 1$, the estimated effect of

[61] In columns with fixed effects the reported intercept is the mean value of y minus the mean value of the independent variables (excluding fixed effects) times their coefficients.

Table XII: Determinants of Coverage Bias in OLS Regressions with Robust Covariance Matrix
(dependent variable = ln(rejection probability at .01 level/.01)

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
					(A) ideal iid errors					
			1378 observations					1032 observations with $h_{ii}^{max} < 1$		
# of obs	$-1.1e^{-6}$	$-3.0e^{-7}$	$-8.3e^{-7}$			$-1.1e^{-6}$	$-2.7e^{-7}$	$-5.6e^{-7}$		
	($5.7e^{-7}$)	($1.2e^{-6}$)	($5.7e^{-7}$)			($5.1e^{-7}$)	($1.0e^{-6}$)	($4.8e^{-7}$)		
h_{ii}^{max}			.135	.142	.462			.683	.691	.986
			(.036)	(.036)	(.042)			(.062)	(.062)	(.062)
intercept	.163	.159	.110	.103	-.020	.172	.166	.037	.031	-.026
	(.015)	(.013)	(.020)	(.020)	(.019)	(.015)	(.013)	(.019)	(.018)	(.015)
paper f.e.	no	yes	no	no	yes	no	yes	no	no	yes
μ_y	.157	.157	.157	.157	.157	.165	.165	.165	.165	.165
R^2	.003	.414	.013	.011	.464	.005	.464	.108	.107	.571
					(B) dichotomous y, heteroskedastic errors					
			589 observations					452 observations with $h_{ii}^{max} < 1$		
# of obs	$7.5e^{-7}$	$1.9e^{-6}$	$1.1e^{-6}$			$6.2e^{-7}$	$1.8e^{-6}$	$1.5e^{-6}$		
	($8.9e^{-7}$)	($2.1e^{-6}$)	($9.0e^{-7}$)			($8.8e^{-7}$)	($2.1e^{-6}$)	($8.2e^{-7}$)		
h_{ii}^{max}			.212	.197	.554			1.34	1.30	1.57
			(.084)	(.083)	(.100)			(.153)	(.152)	(.155)
intercept	.247	.238	.167	.181	.049	.281	.269	.034	.055	.008
	(.034)	(.032)	(.047)	(.045)	(.046)	(.039)	(.037)	(.046)	(.044)	(.039)
paper f.e.	no	yes	no	no	yes	no	yes	no	no	yes
μ_y	.253	.253	.253	.253	.253	.287	.287	.287	.287	.287
R^2	.001	.368	.012	.010	.400	.001	.375	.147	.141	.495

Notes: intercept = mean of dependent variable minus coefficient times mean of independent variables (excluding fixed effects). paper f.e. = paper fixed effects; μ_y = mean of dependent variable. F test statistics calculated using Stata's n/(n-k) finite sample adjustment and n-k denominator degrees of freedom.

maximal leverage rises dramatically and accounts for almost all of coverage bias, even without paper fixed effects (columns 8-10). Once regressions with $h_{ii}^{max} = 1$ are dropped, the R^2 with just maximal leverage in the regression is .107 in the iid sample and .141 in the heteroskedastic sample. I should note that coverage bias in regressions with $h_{ii}^{max} = 1$ is greater than in regressions in which this condition does not hold (.023 vs. .015 at the .01 level with ideal errors), it is simply less than a ln-linear extrapolation of effects might otherwise suggest.

Table XIII examines the bias of tests based on the clustered covariance estimate in simulations with iid errors (the clustered sample of Table IX) and unknown random effects

Table XIII: Determinants of Coverage Bias in Regressions with Clustered Covariance Matrix
(dependent variable = ln(rejection probability at .01 level/.01)

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
(A) ideal iid errors										
	827 observations					506 observations with $\lambda^{max}(\{\mathbf{H_{gg}}\}) < 1$				
# of clusters	$-1.6e^{-4}$	$-3.6e^{-5}$	$4.2e^{-5}$			$-1.7e^{-4}$	$-8.8e^{-4}$	$-2.4e^{-5}$		
	($5.9e^{-5}$)	($1.4e^{-4}$)	($1.4e^{-4}$)			($4.2e^{-5}$)	($9.9e^{-5}$)	($8.1e^{-5}$)		
h_{ii}^{max}			-.333					-.153		
			(.100)					(.123)		
$\lambda^{max}(\{\mathbf{H_{gg}}\})$			.752	.261	.495			1.30	.893	1.23
			(.108)	(.065)	(.077)			(.100)	(.069)	(.078)
intercept	.441	.414	.070	.245	.101	.466	.444	-.030	.089	-.035
	(.028)	(.035)	(.061)	(.047)	(.050)	(.025)	(.030)	(.039)	(.032)	(.032)
paper f.e.	no	yes	yes	no	yes	no	yes	yes	no	yes
μ_y	.406	.406	.406	.406	.406	.422	.422	.422	.422	.422
R^2	.009	.458	.492	.019	.485	.032	.499	.671	.248	.669
(B) cluster correlated errors (random effects)										
	655 observations					506 observations with $\lambda^{max}(\{\mathbf{H_{gg}}\}) < 1$				
	$\rho = 0$	$\rho = .2$	$\rho = .4$	$\rho = .6$	$\rho = .8$	$\rho = 0$	$\rho = .2$	$\rho = .4$	$\rho = .6$	$\rho = .8$
$\lambda^{max}(\{\mathbf{H_{gg}}\})$	.686	.692	.702	.742	.760	.967	1.11	1.11	1.14	1.16
	(.068)	(.067)	(.068)	(.068)	(.068)	(.085)	(.081)	(.083)	(.084)	(.082)
intercept	.070	.182	.176	.152	.134	.000	.079	.076	.054	.036
	(.043)	(.042)	(.043)	(.043)	(.042)	(.040)	(.038)	(.039)	(.039)	(.038)
paper f.e.	no	no	no	no	no	no	no	no	no	no
μ_y	.423	.539	.538	.534	.525	.360	.492	.488	.478	.467
R^2	.134	.140	.140	.155	.162	.204	.272	.262	.268	.281

Notes: intercept = mean of dependent variable minus coefficient times mean of independent variables (excluding fixed effects). paper f.e. = with paper fixed effects; μ_Y = mean of dependent variable. F test statistics calculated using Stata's $n_c(n-1)/(n_c-1)(n-k)$ finite sample adjustment and n_c-1 denominator degrees of freedom, where n_c is the number of clusters.

(Table X).[62] When entered on its own, the number of clusters is consistently negatively correlated with bias (panel A, columns 1, 2, 6 and 7), although in the broadest sample its statistical significance does not survive the addition of paper fixed effects. In a horse race with

[62]Stata's xtreg fe cluster command, as noted earlier, applies a different finite sample adjustment than the reg/areg commands. In Table IX I accept Stata's choices, as my objective is to illustrate size for the methods used by authors (xtreg fe cluster regressions do not appear in Table X as they all have cluster fixed effects). In Table XII, to analyse the role of leverage in creating bias, I put everything on a equal footing and use the reg/areg cluster finite sample adjustment for all regressions and Stata's default n_c-1 F denominator degrees of freedom, which is the theoretical maximum effective degrees of freedom for inference with clustered covariance matrices (see (8) above).

paper fixed effects between the number of clusters, maximal leverage $h_{ii}^{\max}$, and the maximum eigenvalue of the block diagonal cluster elements of the hat matrix $\lambda^{\max}(\{\mathbf{H_{gg}}\})$ (panel A, columns 3 and 8), only the last remains significant and of the correct sign. As argued earlier, these eigenvalues, and not maximal leverage per se, are what determines bias in the clustered covariance estimate. Columns 4, 5, 9 and 10 show that this relation exists with and without fixed effects and, as in the previous table, is stronger once extreme cases with $\lambda^{\max}(\{\mathbf{H_{gg}}\}) = 1$ are removed where, without paper fixed effects, $\lambda^{\max}(\{\mathbf{H_{gg}}\})$ alone produces an R^2 of .248. Coverage bias in regressions with $\lambda^{\max}(\{\mathbf{H_{gg}}\}) = 1$ is much higher than in those without (.032 vs. .013 at the .01 level with ideal errors), but less than might be indicated by a ln-linear extrapolation of effects. Panel B of the table shows that the maximum eigenvalue of the block diagonal elements of the hat matrix is also a significant determinant of the coverage bias of the clustered covariance matrix in the presence of unknown (or unmodelled) random effects, with R^2s ranging from .140 to .281. As in the earlier table, excluding extreme cases with $\lambda^{\max}(\{\mathbf{H_{gg}}\}) = 1$, coverage bias pretty much disappears when maximal leverage for clustered regressions, $\lambda^{\max}(\{\mathbf{H_{gg}}\})$, is zero (intercepts in panel A, columns 9-10, panel B, columns 6-10, compared with the mean of y). In sum, maximum leverage, and not sample size, provides the best indicator of the accuracy of statistical inference based upon the robust and clustered covariance matrices. Despite an average of 5300 observations and 216 clusters, in the typical experimental regression this is rather poor.

The preceding tables and regressions show that regression design affects the distribution of test statistics using the robust and clustered covariance matrices in a systematic fashion. This suggests the possibility, promoted by Kott (1992), Bell and McCaffrey (2002), Imbens and Kolesar (2015) and Young (2016), of using the known distribution of these test statistics with ideal errors to evaluate values calculated in practical applied situations.[63] In Table XIV I apply this idea to all test statistics calculated with non-default methods (most of which are robust, clustered or corrections thereof), using the ideal size simulations of Table IX, where the null of no treatment effect in the regression is true by construction, to evaluate the Wald test statistics of no treatment effect in each regression. As shown, in OLS regressions switching from Stata's

[63]These papers actually concern themselves with using an analytical "effective degrees of freedom" approximation of the distribution of a single linear combination of coefficients when disturbances are ideal iid normal. In the case of the F/Wald grouped tests of multiple linear constraints examined in this paper, no such simple approximation exists, but simulation provides an alternative approximation of the distribution.

Table XIV: Statistical Significance of Treatment Effects
(number of significant regressions by level of test)

	1042 OLS regressions		412 non-OLS regressions	
	.01	.05	.01	.05
standard Wald test				
evaluated by Stata	264	437	121	168
evaluated by simulation	212	386	99	147
randomization Wald test				
randomization-p	210	393	85	115
randomization-t	215	397	75	106
randomization-c	235	401	71	102

evaluation of the conventional test statistic to the p-value implied by the statistics' position in the ideal simulated distribution immediately brings conventional results completely in line with those produced by randomization statistical inference. In the case of the smaller number of non-OLS regressions, the adjustment is less successful, eliminating between ⅓ and ⅔ of the gap between conventional and randomization results,[64] reflecting, perhaps, the fact that leverage is not as clearly an immutable characteristic of regression design in non-OLS settings.[65] Altogether, however, the results of Table XIV show that the coverage bias reported in Table IX can explain most of the difference between conventional and randomization results in regressions using non-default covariance estimation methods which, as determined earlier in Table VI, account for most of the observed differences between conventional and randomization tests.

I conclude this section by presenting, as promised, summary statistics that show how coverage bias compounds with the number of implicit tests. In Table XV I compare average rejection rates at the coefficient, equation and paper level using the simulations with ideal errors of Table IX earlier. As seemingly unrelated estimation techniques are generally impossible to

[64]These results relate to regressions reported in Table IX which, as noted earlier, exclude regression specifications where the notion of an "ideal" error is less clear. For the quantile regressions and seemingly unrelated systems which I also simulated but did not include in Table IX (as noted in an earlier footnote), in the 36 regressions which do not use default covariance methods, 10 are found to be significant at the .01 level using the conventionally evaluated Wald statistic. If evaluated using the simulated distribution, however, the number of such significant results falls to 5, which agrees closely with the 5 found by the randomization-p and –t and 6 found by the randomization-c.

[65]Thus, for example, in a random effects regression one can define a leverage measure using the GLS transformed independent variables that functions much the same way, allowing the robust or clustered covariance estimates, depending upon the hypothesis test, to place uneven weight on particular residuals. This measure, however, now depends upon the estimated random effects and is not an intrinsic feature of regression design.

Table XV: Average Coverage Bias at Different Levels of Aggregation

		using default covariance estimation method		using paper's covariance estimation method	
	# of obs	.01	.05	.01	.05
coefficient	5552	.0103	.0507	.0142	.0578
equation	1900	.0105	.0511	.0217	.0706
paper	53	.0200	.0675	.1845	.2740
		seemingly unrelated estimation			
paper	9	.2972	.3902		

Note: Coverage calculated using 10000 ideal simulations and then averaged across the reported number of coefficient, equation or paper observations.

implement, I generate a conventional omnibus test for all 53 papers by using a block diagonal covariance matrix with the equation covariance matrices on the diagonal, evaluating the test statistic with a chi-squared distribution. This covariance matrix incorporates the fact that the equation level disturbances in the simulations are drawn independently, so the true covariance matrix is known to be block diagonal. As shown, default covariance methods, in this sample of OLS and non-OLS regressions, produce slightly biased coverage at the coefficient level, which increases as one moves up to the equation and paper level. However, as the bias at the coefficient level is miniscule (e.g. .0103 at.01 putative size), the bias at the paper level remains small (.0200 at .01). In contrast, the more substantial bias at the coefficient level found using each paper's covariance estimation method (.0142 at.01), compounds to an average .1845 rejection rate for .01 nominal size at the paper level. Biases which appear small at the coefficient level become shockingly large when on average more than 100 coefficients are combined to evaluate the complete paper. The bottom row of the table evaluates the bias of Stata's seemingly unrelated estimation technique in the 9 papers where I was able, in the earlier section, to apply the method to the papers themselves. With independent equation level errors, the suest command, in attempting to calculate the cross-equation covariance using the equation level scores, produces average rejection rates of .2972 and .3902 at .01 and .05 nominal size, respectively.[66]

[66]Confirmation of this substantial bias is given by the 16 seemingly unrelated estimation systems in one of my sample papers which, with an average of 11 treatment variables per system, produce average rejection rates of .154 at the .01 level, as noted in an earlier footnote.

(c) Power

Table XVI below shows that the equation level power of randomization tests is very close to that of unbiased conventional econometric methods. As in the case of the explorations of size in Table IX earlier, I produce simulated data based on equations estimated using the actual data in the papers, but this time I include all treatment effects in the predicted values. Thus, these simulations represent cases were, by construction, the null of no treatment effects is false and the alternative, of treatment effects in the amounts indicated by the authors' equations, is true. As before, all error terms are ideal within the context of the regression model, i.e. are iid and normally distributed except insofar as the regression equation itself explicitly posits an alternative distribution. The variance of each equation's disturbance is determined by the original estimating equation and the disturbances are independent across equations and iterations. Because randomization simulations are computationally intensive, I conduct only 100 simulations per regression with 200 randomizations used in each simulation to determine randomization p-values. My work with the original data found little change in p-values or rejection probabilities between 200 and 10000 randomization iterations, as discussed in an earlier footnote.

The left-hand side of Table XVI uses Stata's default covariance estimation method to calculate test statistics for both the conventional Wald test and the randomization-p and -t. These covariance matrices, without robust, clustered or other corrections, are the correct covariance matrices, insofar as the error terms are ideal and satisfy the baseline assumptions of the estimating equation. We see that when the null is false the average p-value of the Wald test (using the F-distribution) for OLS regressions is .163, while for the randomization-p it is .165. The correlation of the average p-value of the Wald test and the randomization-p by estimating equation is .998; the correlation at the iteration level of the p-value (i.e. in each individual simulation) is, extraordinarily, .990. The Wald test rejects .462 of the time at the .01 level and .586 at the .05 level, while the randomization-p probabilities of rejection are .471 and .585, respectively. For all intents and purposes, the power of the randomization-p is identical to that of conventional econometrics in OLS regressions. Lehmann’s result regarding the asymptotic equality of the power of t- and randomization tests in a simple single coefficient regression, noted earlier, is, for all intents and purposes, a finite sample result that extends to a vast array of

Table XVI: Power, 1340 OLS and 522 Non-OLS Regressions in 52 papers
(100 simulations per regression, 200 randomizations per simulation)

	using default covariance estimation method						using paper's covariance estimation method					
	(A) OLS			(B) ~ OLS			(C) OLS			(D) ~ OLS		
	p-values when H_1 is true											
	μ	ρ_E	ρ_I	μ	ρ_E	ρ_I	μ	ρ_E	ρ_I	μ	ρ_E	ρ_I
Wald test	.163			.182			.159			.178		
randomization-p	.165	.998	.990	.191	.995	.971	.168	.989	.985	.193	.991	.970
randomization-t	.167	.978	.976	.216	.844	.870	.170	.970	.971	.216	.847	.872
randomization-c	.172	.954	.960	.237	.758	.801	.172	.954	.958	.236	.761	.804
	rejection probabilities when H_1 is true (by nominal size of the test)											
	.01		.05		.01		.05		.01		.05	
Wald test	.462		.586		.445		.556		.475		.596	
randomization-p	.471		.585		.415		.525		.455		.574	
randomization-t	.468		.581		.374		.479		.457		.573	
randomization-c	.462		.575		.351		.450		.462		.574	

Notes: ~ OLS: not ordinary least squares; μ = mean (calculated across means by equation); ρ_E = correlation of mean p-values at the equation level with those of the conventional Wald test; ρ_I = correlation of p-values at the equation x iteration level with those of the conventional Wald test. This table excludes the one paper which was excluded earlier from Table III's analysis of significance at the regression level because of the large number of significant randomization ties at the equation level. Rejection probabilities with randomization techniques differ between columns devoted to default vs paper's covariance methods because (a) the -p and -t use the different methods in their calculation and (b) averages are calculated across iterations where Stata, using the covariance method described, is able to deliver a test of all treatment measures in the regression.

regression specifications and sample sizes.[67] However, column (B) of Table XVI shows that Lehmann's result does not quite generalize to non-OLS settings, as the power of randomization tests appears to be slightly lower than that achieved using conventional tests based upon the default covariance estimate, with the randomization-p showing an average p-value that is .009 greater and a rejection probability that is .030 less at the .01 and .05 levels than that achieved by conventional tests.

Columns (A) and (B) of Table XVI compare randomization tests to Stata's default conventional estimate of covariance which, as shown in earlier size simulations, even in non-OLS cases generally has very little bias in favour of rejection. To give the papers in my sample the benefit of the doubt, columns (C) and (D) of the table compare the power of randomization tests

[67]Sample sizes in these regression range from 40 to 450000 observations. In 19200 iterations across 192 regressions with less than 100 observations, the mean p-value of the conventional Wald test and the randomization-p are .360 and .361, respectively, with a correlation (at the iteration level) of .972.

to that achieved using the actual covariance matrix method (default, robust, clustered, etc) used by the papers. As shown earlier, these methods produce rejection probabilities higher than nominal size. This "weakness", when the null is true, becomes a "feature" when the null is false. As might be expected, in Table XVI these methods produce systematically lower p-values and higher rejection probabilities than their default covariance counterparts.

Columns (C) and (D) show, however, that even if one wishes to give experimental papers the benefit of the doubt, power alone cannot come anywhere near to explaining the discrepancy between conventional and randomization results. At the .01 level, using the covariance methods selected by authors, conventional OLS methods reject .475 of the time, while the randomization-p only rejects .455 of the time. Thus, if treatment has the effects estimated in the papers and ideal errors with variance based on that present in the estimating equations, randomization methods reject the null when it is false .96 as often as the methods selected by the authors. In my analysis of actual OLS regressions in the papers (Table III earlier) I find that 310 regression specifications reject the null at the .01 level using the conventional Wald test of treatment significance, but only 251 reject the null at that level using the randomization-p. This is a relative rejection frequency of .81. Table XV shows that in non-OLS settings the covariance methods selected by authors reject the null when it is false .458 of the time, while the randomization-p only rejects .404 of the time, for a relative rejection rate of .88. Analysing the actual non-OLS regressions in the papers, I find that conventional Wald tests of treatment significance reject the null in 175 cases at the .01 level, while the randomization-p only rejects 123 times at that level of significance. This is a relative rejection probability of .70. Even if one wishes to give the benefit of the doubt to each paper by selecting their covariance calculation methods, the large discrepancy between the rejection rates achieved in conventional tests and those recorded by the randomization-p cannot be attributed to a lack of power in randomization statistical inference.

Table XVI also allows a comparison of the relative power of different randomization methods. In Table XVI the power of the randomization-p is systematically greater than that of the -t, which in turn is more powerful than the randomization-c, as average p-values are always lower and rejection probabilities, in almost every case, higher, although the differences are relatively small. This, again, suggests that Lehmann's result that the randomization-t/-p was

Table XVII: Number of Omnibus Rejections When H_1 is True
(100 iterations per paper, tests based upon block diagonal covariance matrix)

	default covariance estimate		paper's covariance estimate	
	.01 level	.05 level	.01 level	.05 level
(1) conventional Wald test	.959	.973	.963	.975
(2) randomization-t	.835	.884	.825	.876
(3) randomization-c	.860	.895	.861	.895

Notes: Tests are calculated using a covariance matrix (for all coefficients in the paper) that is block diagonal by estimating equation, as the disturbances for each equation are by construction independent. This table excludes the two papers which were earlier excluded from Table VIII's analysis of overall experimental significance because of presentationally relevant randomization ties. Nevertheless, there are a large number of small ties (with an average width of .0028, but frequently extending across the .01 and .05 boundaries). I resolve all of these by drawing uniformly distributed random numbers, as in (9) earlier.

uniformly most powerful in the case of a single coefficient OLS regression may be more general than his limited example suggests.

Table XVII evaluates the relative power of randomization tests in the omnibus test of experimental significance. Since it is generally impossible to implement the multi-equation seemingly unrelated estimation test of overall significance as a benchmark, I use a conventional chi-squared test based upon a block diagonal covariance matrix made up of the individually estimated equation level covariance matrices, as was done in the simulations of size earlier in Table XV. Since the simulation disturbances for each equation are independent of each other, the cross equation covariance matrix is, actually, block diagonal. I make use of the same information/restriction in the randomization and bootstrap Wald calculations, to place them on an equal footing with the conventional benchmark.

As shown in the table, whether evaluated against Wald tests calculated using the default estimate of covariance or using the authors' own methods, the relative power of omnibus randomization tests is substantially lower. While conventional tests reject the null of no treatment effect between .96 and .98 of the time at the .01 and .05 levels, the power of randomization tests ranges from .83 to .90. This result is merely the power version of the aggregation problem discussed earlier above. As already seen in Table XVI, the power of randomization tests is slightly less than that of conventional tests at the equation level, particularly in non-OLS settings. As multiple equations are aggregated to the paper level, this difference becomes more pronounced.

The average differences in power reported in Table XVII reflect large differences in a limited number of papers. Consequently, I consider the impact of restricting the sample to those papers where the rejection frequencies at the .01 level of the randomization tests are no more than .01 less than those attained in Table XVII using the authors' own covariance estimation methods. For the randomization-t, this level of power is reached in 36 papers. Among these, only 12 and 16 reject the null of no treatment effect at the .01 and .05 levels, respectively, in the omnibus randomization-t analysis of the actual experiments. These are rejection rates of .33 and .44. For the randomization-c, relative power within .01 of an omnibus conventional Wald test is reached in 40 papers. Only 19 and 24 of these reject the null of no treatment effect at the .01 and .05 levels, respectively, in the randomization-c analysis of the actual experiments. These are rejection rates of .48 and .60. While power might explain some of the omnibus results, it does not explain them all. Between ½ and ⅔ of papers cannot reject the null of no treatment effect at the .01 level, even when the sample is restricted to papers where randomization tests have power equal to that of conventional tests.

The power simulations in Table XVII, based upon disturbances that are independent across equations, do not reproduce the high cross-equation correlation of coefficients and p-values that is present in my sample. This artificiality, however, highlights the principal reason for low rejection rates in the actual sample. The power of the randomization-t and -c at the .01 level, when restricted to those papers where they are within .01 of the power of conventional tests is, on average .96. These are rejection rates simulated using the coefficient estimates of the papers and the equation specific estimates of disturbance variation. The fact that these tests in application to the actual experiments themselves have remarkably lower rejection rates reflects the large discounting of significant results created by the strong cross-equation correlation of significant coefficients discussed earlier above. The reader seeking more realistic simulations producing cross-equation correlations that match those in the data need look no further than the omnibus bootstrap test results presented earlier. The bootstrap distributions, in effect, simulate the sampling variation and cross equation correlation actually present in the data. These simulations confirm the results of the omnibus randomization tests.

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 5.8 treatment measures, representing multiple treatments and the interaction of treatment with participant characteristics, of which on average 4.1 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

Bell, Robert M. and Daniel F. McCaffrey. 2002. "Bias Reduction in Standard Errors for Linear Regression with Multi-Stage Samples." *Survey Methodology* 28 (2): 169-181.

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.

Cameron, A. Colin and Pravin K. Trivedi. 2010. Microeconometrics Using Stata. Revised Edition. College Station, Texas: Stata Press, 2010.

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.

Imbens, Guido W. and Michal Kolesar. 2015. "Robust Standard Standard Errors in Small Samples: Some Practical Advice." *Review of Economics and Statistics*, forthcoming.

Kott, Phillip S. 1996. "Linear Regression in the Face of Specification Error: A Model-Based Exploration of Randomization-Based Techniques." *Proceedings of the Survey Methods Section, Statistical Society of Canada*: 39-47.

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.

Moulton, Brent R. 1986. "Random Group Effects and the Precision of Regression Estimates." *Journal of Econometrics* 32 (3) : 385-397.

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

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.

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.