`

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

Alwyn Young
London School of Economics
This draft: August 2017

Abstract

I follow R.A. Fisher's The Design of Experiments (1935), using randomization statistical inference to test the null hypothesis of no treatment effect in a comprehensive sample of 53 experimental papers drawn from the journals of the American Economic Association. Randomization tests of the significance of treatment coefficients find that 10 to 20 percent of conventionally significant coefficients are not significant at the same level. In joint tests for equations with multiple treatment measures, 30 to 40 percent of equations with an individually significant coefficient cannot reject the null of no treatment effect. An omnibus randomization test of overall experimental significance that incorporates all of the regressions in each paper finds that less than half of experimental papers are able to reject the null of no treatment effects anywhere at the .01 level. Bootstrap and jackknife methods support and confirm these results.

*I am grateful to Larry Katz, Alan Manning, Ben Olken, Steve Pischke, Jonathan de Quidt, Eric Verhoogen and anonymous referees 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 tests that are both exact, i.e. with a distribution that is known no matter what the sample size or the characteristics of errors, and intrinsically resilient to outliers. Randomization also facilitates the performance of joint tests and the use of higher powered multiple testing methods that depend upon the estimation of the joint distribution of coefficients across different estimating equations. Randomized experiments rarely make use of such methods, by and large only presenting conventional econometric tests of individual coefficients using asymptotically accurate clustered/robust covariance estimates. In this paper I apply randomization tests to 53 randomized experiments, using them to construct counterparts to conventional tests of significance within regressions and, more ambitiously, omnibus tests of overall significance that combine all of the regressions within tables or in a paper in a manner that is, practically speaking, largely infeasible in conventional econometrics. At the coefficient level, randomization tests reduce the number of significant coefficients by 10 to 20 percent. Joint tests of statistical significance in multi-treatment equations reduce the number of regression specifications with statistically significant treatment effects by 30 to 40 percent, while the omnibus test finds that, when all treatment outcome equations are combined, only about 45 to 50 percent as many tables and 50 to 60 percent as many 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 and jackknife methods.

Two factors lie behind the discrepancy between the results reported in journals and those reported in this paper. First, the regression design lying behind published results is typically very poor, with a few observations playing a key role in the determination of coefficients and standard errors. Remarkably, with the removal of just one cluster or observation, 26 percent of reported .01 significant results can be rendered insignificant, while the average gap between the highest and lowest p-value found by deleting individual observations in all regressions is an astonishing .24. Sensitivity to outliers comes from the interplay between extreme residuals and regression design which concentrates leverage in a few observations. The concentration of leverage also renders the clustered and robust variance estimates more volatile for a given set of regressors,

generating t-statistic distributions that are more dispersed than recognized by their putative degrees of freedom. Thus, regression design plays a role in two interrelated problems: sensitivity to the change in the sample and accurate inference for a given sample. Randomization inference addresses both. Randomization inference systematically raises the p-values of conventionally significant results which depend upon outliers, because it evaluates their extreme outcomes, which depend upon the treatment allocated to critical observations, as being more likely. Randomization inference also provides test statistics with finite sample exact size, without appeal to asymptotic theorems or worries about whether finite sample nominal degrees of freedom actually approximate the distribution of the t-statistic.

I find that where randomization inference and conventional inference disagree, the concentration of leverage is much higher. In ¾ of the cases where randomization inference disagrees with conventional inference the conventional results are sensitive to the deletion of just one or two clusters or observations. Thus, to disagree with randomization inference, in the context of the results presented in this paper, is to demand that readers accept as relevant results that depend upon one or two observations out of the hundreds or thousands in the sample. It is doubtful that this is something that authors, had they been aware of the problem, would have chosen to do.

The problems of conventional inference found in this paper are not unique to experimental papers. In Young (2017) I use the bootstrap to examine a comprehensive sample of 1400 instrumental variable regressions published in 32 papers and find significance rates for individual coefficients that are ⅓ to ½ lower than those reported by conventional methods. These reductions are likely to be conservative, as I find the bootstrap, itself an asymptotically accurate method, retains some size distortions in these samples. Fully 47 percent of reported .01 significant results are rendered insignificant by the removal of just one cluster or observation. While other fields must struggle to address the interrelated problems of finite sample size distortions and sensitivity to outliers, randomized experiments already have at their disposal a method that allows for exact inference and is inherently resistant to outliers.

The second factor behind the discrepancy between published results and the analysis presented here is that 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 5.7 treatment measures, of which only 1.6 are significant. However, only 31 of 998 regressions with multiple treatment measures report a conventional F- or Wald-test of the joint significance of all treatment variables within the regression.[1] In a similar vein, the typical paper reports 10 regressions with a treatment coefficient that is significant at the .01 level, and 29 regressions with no treatment coefficient that is significant at this level. When the multiple tests implicit in the many coefficients within regressions and the many regressions within papers are combined, significance levels fall.

Critics might claim that these tests are unfair, weakening power by combining relevant treatment measures and regressions with treatment measures or regressions designed to show that certain effects are not present, i.e. to implicitly confirm the validity of nulls.[2] Such explanations are inconsistent with the pattern of my results. I find that multiple testing procedures, which relative to joint tests increase power on the axes at the expense of within quadrants, generally produce lower rejection rates, so more evidence of significance is found when tests emphasize power for moderate values of all coefficients rather than extreme values of individual coefficients. Authors presumably present their most important results in their very first table of main results.[3] I find, however, that first tables often provide weak evidence in favour of treatment effects. More evidence against the null is found when joint and multiple tests are expanded to include the many alternative specifications presented in later tables. My emphasis in this paper is on finding *any* rejection of the null in a paper, not the number of such rejections. In that context, expanding the number of included measures beyond the key baseline results authors present first and foremost need not, and in practice does not, reduce power.

---

[1]These occur in two papers. In an additional 7 regressions in two other papers the authors make an attempt to test the joint significance of multiple treatment measures, but accidentally leave out some treatment measures.

[2]No paper in my sample reports a pre-analysis plan, so there is no way to verify this objectively. I should note, however, that in my analysis I naturally exclude regressions related to randomization balance (participant characteristics), pre-treatment differences in treatment and control, and attrition, which are designed to confirm the internal validity of the randomized experiment by demonstrating orthogonality with treatment.

[3]Main results are identified as the first section with this title (a frequent feature) or a title describing an outcome of the experiment (e.g. “Does treatment-name affect interesting-outcome?”).

In my study of instrumental variables regressions mentioned above, I find 313 estimating equations with more than one instrument, but authors only use these to increase the number of instrumented measures in 41 cases. In contrast, almost half of all regressions in my experimental sample contain more than one treatment variable, with an average of 4.8 treatment measures. Much of "data snooping" (White 2000) is probably unobservable, as instruments or ideas which fail to produce significant results are quietly abandoned and new approaches tried. In the case of experimental economics, where authors invest heavily in the creation of data sets and openly and forthrightly walk the reader through their investigation of their results, it is less so. In implementing joint and multiple tests in this paper I follow transparent rules rather than opaque discretion, combining coefficients mechanically as they appear in regressions and tables. There is little doubt this leads to the inclusion of some treatment measures which authors deemed unimportant or used to confirm nulls. The joint and multiple tests presented in this paper are intended to be suggestive, not definitive. They highlight the potential magnitude of the problem and provide procedures that authors might find useful.

Randomization tests have what some consider to be a major weakness: they provide exact tests, but only for sharp nulls, i.e. nulls which specify a precise treatment effect for each participant. Thus, in testing the null of no treatment effects, randomization inference is not testing whether the average treatment effect is zero, but rather whether the treatment effect is zero for all participants.[4] In the presence of unaccounted for heterogeneity in treatment effects, randomization tests are no longer exact and can have substantial size distortions (Chung & Romano 2013, Bugni, Canay & Shaikh 2017). It should be noted, however, that average treatment effects generate intrinsic treatment dependent heteroskedasticity, which renders conventional tests inaccurate in finite samples as well. While in such circumstances conventional estimates with robust covariance estimates will have asymptotically correct size, asymptotic accuracy in the face of average treatment effects is equally a feature of randomization inference, provided treatment is balanced or studentized statistics are used in the analysis (Janssen 1997, Chung & Romano 2013, Bugni, Canay & Shaikh 2017). Moreover, while a null of the same non-zero treatment effect for all is unpalatable, as it is hard to conceive of a treatment having exactly

[4]The sharp null merely has to be precise, and hence does not have to be the same treatment effect for all, but in practice it is difficult to specify heterogeneity that is not accounted for in the regression design.

the same effect on everyone, the sharp null of zero treatment effects is not, as this amounts to simply stating that the experimental treatment is, from the perspective of participants, irrelevant. In results below I find that conventionally significant regressions that randomization inference finds to be insignificant are much less likely to have detectable treatment related heteroskedasticity, suggesting that treatment heterogeneity is not present and supporting the notion of a sharp null. Finally, a focus on the perils of sharp nulls overemphasizes what motivates randomization tests at the expense of considering what they actually do. Randomization inference's consideration of "potential outcomes" mechanically raises the p-values of conventional results with low p-values and lowers the p-values of conventional results with high p-values when the conventional results, in either case, depend upon only a few observations. This robustness to outliers is extremely valuable.

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. Randomization statistical inference provides a solution to the problems identified in this paper, avoiding a dependence on asymptotic theorems that produce inaccurate finite sample statistical inference that is sensitive to outliers and allowing the simple calculation of omnibus tests and multiple testing procedures that incorporate all of the tests run in an analysis. While to date it rarely appears in experimental papers, which generally rely upon traditional econometric methods,[5] it can easily be incorporated into their analysis. Randomized experiments solve the problem of identification; they can also easily solve the problem of accurate statistical inference that is robust to outliers and facilitates joint and multiple testing procedures, making them even more powerful as an investigative tool.

This paper takes well-known issues and explores them in a broad practical sample. Consideration of whether randomization inference yields different results than conventional inference is not new. Lehmann (1959) showed that in a simple test of binary treatment a randomization t-test has an asymptotic distribution equal to the conventional t-test, and Imbens

[5]Of the 54 experimental papers that otherwise meet the criteria for inclusion in my sample (discussed below), only one uses randomization statistical inference throughout (and hence is not included in the final sample), while one other uses randomization inference to analyse results in some regressions and one more indicates that they confirmed the significance of results with randomization tests. Wilcoxon rank sum tests for some treatment results are reported in four other papers.

and Wooldridge (2009) found little difference between randomization and conventional tests for binary treatment in a sample of 8 program evaluations. Limiting the influence of outliers was a prominent econometric concern some decades ago (e.g. Huber 1981, Cook and Weisberg 1982). The size distortions produced by robust and clustered covariance estimates are well known and have been explored in, for example, Bertrand, Duflo and Mullainathan (2004) and Donald and Lang (2007). I find that these issues are all interrelated. Randomization and conventional inference differ substantially in situations where regression design produces unbalanced leverage which makes coefficients sensitive to outliers, while rendering the robust and clustered covariance estimates dependent upon a small set of residuals, producing size distortions. Several recent papers (Anderson 2008, Heckman et al 2010, Lee & Shaikh 2014, List, Shaikh & Xu 2016) have emphasized the problem of multiple testing in experimental papers and explored the robustness of results to step-down multiple-testing procedures in a few experiments. Step-down procedures require subset pivotality (Westfall & Young 1993), i.e. that the distribution of randomization p-values is identical for different subsets of the family of nulls, which generally becomes unavailable once multiple treatment measures are placed in a given regression. I address the issue of multiple testing by using joint testing procedures to test the overall null of whether any (and not which) treatment matters and combine them with single-step multiple testing methods as a means of shifting power to different regions of the parameter space.

The paper proceeds as follows: Section II explains the criteria used to select the 53 paper sample, which is as comprehensive and non-discriminatory as possible, using virtually every paper revealed by a search on the American Economic Association (AEA) website that provides data and code and allows for randomization inference. About 75 percent of the 2027 regressions are ordinary least squares (OLS)[6] and close to 70 percent use clustered or robust covariance estimates. Section III provides background information, including an evaluation of sensitivity to outliers and its relation to regression design in my sample, a thumbnail review of randomization inference, controlled simulations that show how outliers determine differences between randomization and conventional inference, and a simple exposition of the different emphasis of joint and single-step multiple testing procedures. Section IV lays out the results already

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

described above, while Section V concludes.

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 (in preparation) that shows how each paper was analysed, but the reader eager to know how a particular paper fared will have to execute this code themselves. A Stata ado file, available on my website, provides a canned routine for randomization inference suitable for most Stata estimation commands, allowing users to call for coefficient p-values and omnibus tests of the type presented in this paper in their own research.

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 (AER), American Economic Journal (AEJ): Applied Economics and AEJ: Microeconomics which yielded papers up through the March 2014 issue of the AER. I then dropped papers that:

(a) did not provide public use data files and Stata do-file code;
(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 is by far the most popular regression programme in this literature.

My definition of a randomized experiment excluded natural experiments (e.g. based upon an administrative legal change), but included laboratory experiments (i.e. experiments taking place in universities or research centres or recruiting their subjects from such populations). 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 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 sessions, and hence decided to include all laboratory experiments that met the other criteria.[7] Laboratory experiments account for 15 of the 53 papers but only 193 of the 2027 regressions.

The requirement that the experiment contain data on participant characteristics was designed to filter out a sample that would use mainstream multivariate regression techniques with estimated coefficients and standard errors. This removed a number of laboratory experiments, which tend to not have data on participant characteristics, use atypical econometric methods and whose passive randomization, if they dominated the sample, might raise concerns. Conditional on a paper having public use data on participant characteristics, however, I included all estimating equations as long as they produce a coefficient 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 four papers present some exact Wilcoxon rank sum tests. These papers and their non-randomization regressions were retained in the sample.

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

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

can be analysed by re-randomizing treatment T across participants, repeatedly estimating the coefficients β_T and β_{T*age}, and comparing their distribution to the experimentally estimated

[7] 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.

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.[8] I make it a point to always, without exception, adhere to the precise regression presented in tables.

Within papers, regressions were selected if they allow for randomization inference and:

(f) appear in a table and involve a coefficient estimate and standard error or a p-value;
(g) pertain to treatment effects and not to an analysis of randomization balance, sample attrition, or non-experimental cohorts;

while tests were done on the null that:

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

In many tables means are presented, without standard errors or p-values, i.e. without any attempt at statistical inference. I do not consider these regressions. Alternative specifications for regressions presented in tables are often discussed in surrounding text, but catching all such references, and ensuring that I interpret the specification correctly is extremely difficult (see the discussion of do-file inaccuracies below). Consequently, I limited myself to specifications presented in tables. 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. Similarly, I drop regressions projecting the behaviour of non-treatment cohorts on treatment measures, which are typically used by authors to, again, reinforce the internal validity of the experiment. In difference in difference equations, I only test the treatment coefficients associated with differences during the treatment period.

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

I test, universally, the null of no randomized treatment effect, while allowing non-randomized elements to influence behaviour. Consider, for example, the regression

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

where T is a 0/1 measure of treatment and T_0 and T_1 are dummies for the different treatment regimes. The null of no treatment effect is given by re-expressing the regression as (1) earlier above and testing $\beta_T = \beta_{T*age}=0$, while allowing α and β_{age} to take on any value.[9] In more complicated situations the paper might contain randomized overall treatments (e.g. the environmental information provided to participants) combined with other experimental conditions which were not randomized (e.g. whether the participant is offered a convex or linear payoff in each round). As long as the action space is the same under the different randomized treatments, I am able to test the null of no randomized treatment effect by re-randomizing this aspect across participants, while keeping the non-randomized elements constant.[10] 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 and bootstrap for some papers could be performed in less than an hour; for others, because of sample sizes and procedures, it took more than a year of dedicated workstation computing power. The do-files for many papers are remarkably clear and produce, exactly, the regressions reported in the papers. Other do-files produce regressions that are utterly different from those reported in the published paper, while yet others involve extraordinarily convoluted code (aimlessly loading, formatting, dropping, reloading and reformatting data again and again) that could never be implemented 10000 times (in randomization). In between, there are gradations of error and complexity. Rather than allowing myself to choose which papers were “too hard” to work

[9] 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 complex and 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.

[10] 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.

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, 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.[11]

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, 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 these issues 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

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

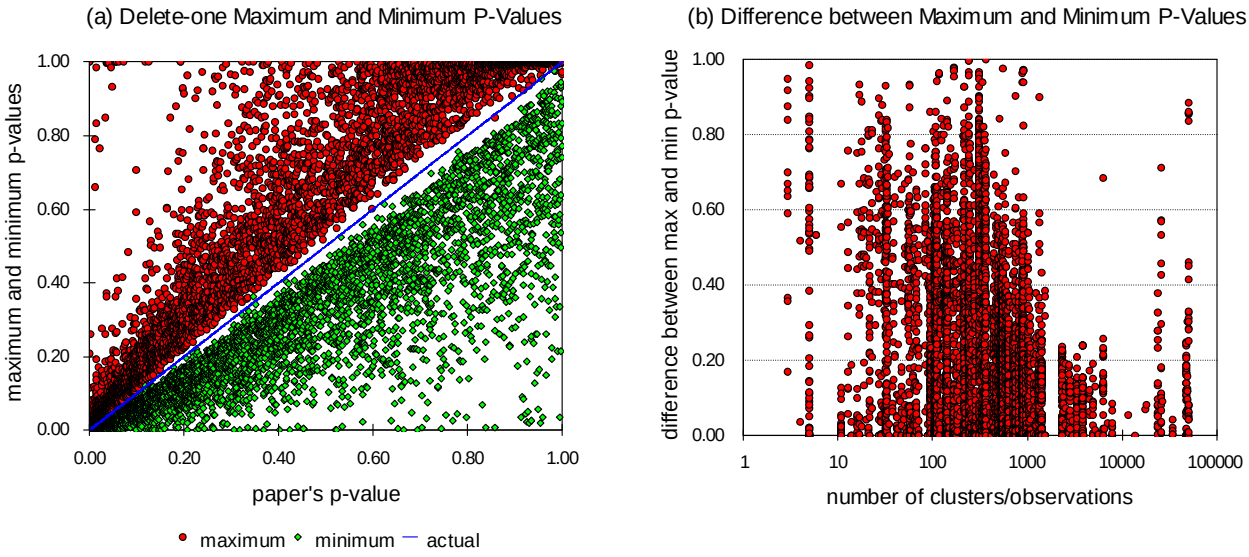
Table I: Characteristics of the Sample

53 papers			2027 regressions		
location		journal	Type		covariance
38	field	29 AER	1513	ordinary least squares	447 default
15	lab	20 AEJ: Applied Economics	327	maximum likelihood	1363 clustered/robust
		4 AEJ: Microeconomics	187	other	126 bootstrap
					91 other

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

Table I above 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 AER, 20 in the AEJ: Applied Economics, and only 4 in the AEJ: Microeconomics. Turning to the 2027 regressions, 75 percent of these are ordinary least squares regressions and 16 percent are maximum likelihood estimates (mostly discrete choice models), while the remaining 9 percent includes handfuls of generalized least squares, quantile regressions, t-tests with unequal variances, two-step Heckman models, and other methods. A little over one-fifth of the regressions in my sample make use of Stata's default covariance matrix calculation. Close to 70

Figure I: Sensitivity to Outliers



percent, however, avail themselves of the cluster estimate of covariance or its single observation robust counterpart. Clustering below treatment level is fairly common. In 170 regressions in 12 papers (9 lab, 3 field) treatment is applied to groups, but the authors either do not cluster at all or cluster at a lower level of aggregation. In another 5 papers (3 lab, 2 field), the authors generally cluster at treatment level, but fail to cluster 120 regressions.[12] This is not considered best practice, as correlations between the residuals for individuals playing games together in a lab or living in the same geographical region are quite likely. Bootstrap and "other" covariance estimation methods (such as the jackknife and the hc3 and brl corrections of the robust and cluster options) make up the remainder of the sample.[13]

III: Issues and Methods

(a) Problems of Conventional Inference in Practical Application

One of the central characteristics of conventional inference, in practical application, is its remarkable sensitivity to outliers. Panel (a) of Figure I above plots the maximum and minimum coefficient p-values found when one deletes one cluster or observation from each regression in my sample against the p-value found with the full sample.[14] With the removal of just one

[12]In my study of instrumental variables papers noted earlier above, I find that when instrument "treatment" is applied in blocks of observations, authors always cluster at that level.

[13]These are, unfortunately, often applied with gross programming errors, and hence do not improve matters.

[14]As elsewhere in this paper, I follow authors' methods. Where they do not cluster, I delete individual observations; where they cluster, I delete clusters. In the case of 120 regressions in papers which otherwise cluster (as described above), I delete cluster groupings.

observation, .26 of reported results that are .01 significant can be rendered insignificant at that level, while .08 of reported results that are .01 insignificant can be found to be significant. Panel (b) of the figure graphs the difference between these maximum and minimum values against the number of clusters/observations in the regression. The average difference between the maximum and minimum delete-one p-values is .24. To be sure, the problem is more acute in smaller samples, but surprising sensitivity can be found in samples with 1000 clusters or observations and even in the very largest regressions in my sample, with more than 50000 observations.

A few simple formulas identify the sources of delete-one sensitivity. In OLS regressions, which make up most of my sample, the coefficient estimate with observation i removed ($\hat{\beta}_{\sim i}$) is related to the coefficient estimate from the full sample ($\hat{\beta}$) through the formula:

$$(3)\ \hat{\beta}_{\sim i} = \hat{\beta} - \frac{\tilde{x}_i}{\sum_i \tilde{x}_i^2} \frac{\hat{\varepsilon}_i}{(1-h_{ii})}$$

where $\tilde{x}_i$ denotes the i[th] residual from the projection of independent variable **x** on the other regressors in the n x k matrix of regressors **X**, $\hat{\varepsilon}_i$ the i[th] residual of the full regression, and h_{ii}, commonly known as leverage, is the i[th] diagonal element of the hat matrix $\mathbf{H} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$.[15] The robust variance estimate can be expressed as

$$(4)\ \frac{1}{\sum \tilde{x}_i^2} \sum \tilde{h}_{ii} \left(\hat{\varepsilon}_i^2 \frac{n}{n-k} \right), \text{ where } \tilde{h}_{ii} = \frac{\tilde{x}_i^2}{\sum_i \tilde{x}_i^2}.$$

$\tilde{h}_{ii}$ might be termed coefficient leverage, because it is the i[th] diagonal element of the hat matrix $\tilde{\mathbf{H}} = \tilde{\mathbf{x}}(\tilde{\mathbf{x}}'\tilde{\mathbf{x}})^{-1}\tilde{\mathbf{x}}'$ for the partitioned regression. As seen in (3) and (4), when coefficient leverage is concentrated in a few observations, coefficient and standard error estimates, depending upon the realization of residuals, are potentially sensitive to the deletion of those observations. Moreover, when standard error estimates depend heavily on a reduced subset of stochastic disturbances they become intrinsically more volatile, producing t-statistic distributions that are more dispersed than recognized by nominal degrees of freedom. Thus, sensitivity to the modification of the sample and accuracy of inference for a given sample are related problems.

[15] So-called because $\hat{\mathbf{y}} = \mathbf{H}\mathbf{y}$. When h_{ii} equals 1, one of the regressors is in effect determined by observation i. As long as this is not the regressor of interest, its coefficient estimate is unaffected by the deletion of the observation. The formula for the deletion of vector **i** of clustered observations is $\hat{\beta}_{\sim \mathbf{i}} = \hat{\beta} - \tilde{\mathbf{x}}_\mathbf{i}'(\mathbf{I}_\mathbf{i} - \mathbf{H}_{\mathbf{ii}})^{-1}\hat{\boldsymbol{\varepsilon}}_\mathbf{i} / (\tilde{\mathbf{x}}'\tilde{\mathbf{x}})$. In this case, if there are cluster dummies so that $\mathbf{I_i} - \mathbf{H_{ii}}$ is singular, one modifies the formula by calculating $\mathbf{H_{ii}}$ using the residuals of the other regressors projected on the cluster dummies.

Table II: Shares of Squared Residuals & Coefficient Leverage of Treatment Variables

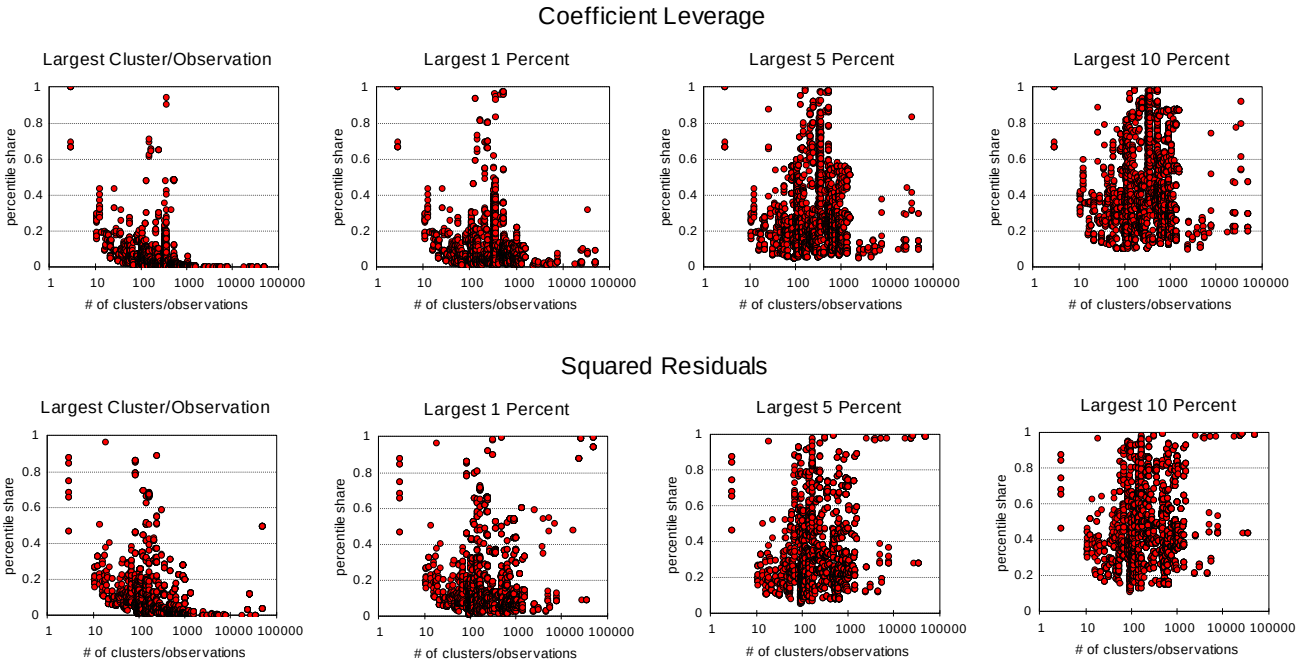
	squared residuals				coefficient leverage				coefficient leverage without covariates			
	max	1%	5%	10%	max	1%	5%	10%	max	1%	5%	10%
all OLS (N=3914)	.087	.163	.336	.464	.050	.100	.268	.409	.048	.098	.269	.413
binary dependent (1541)	.044	.087	.236	.368	.049	.100	.267	.417	.048	.098	.269	.419
other dependent (2373)	.114	.213	.400	.526	.051	.101	.269	.405	.048	.098	.270	.410
with interactions (1940)	.101	.185	.359	.486	.063	.128	.329	.480	.064	.133	.351	.511
without inter. (1974)	.073	.142	.313	.442	.038	.074	.209	.340	.033	.064	.189	.317
first table (766)	.065	.129	.318	.451	.028	.069	.214	.358	.025	.065	.208	.356
other tables (3148)	.092	.171	.340	.467	.056	.108	.282	.422	.054	.106	.284	.427
default covariance (819)	.056	.148	.314	.442	.035	.088	.238	.368	.032	.083	.225	.352
other covariance (3095)	.095	.167	.341	.469	.055	.104	.277	.420	.052	.102	.281	.430

Notes: Numbers in parentheses denote number of coefficient observations. Otherwise, reported figures represent averages across treatment coefficients. The third column calculates treatment coefficient leverage without considering non-treatment covariates other than the constant term.

Table II reports the average shares of squared residuals and coefficient leverage accounted for by the largest and the 1st, 5th and 10th largest percentiles of the clusters or observations of the OLS regressions in my sample. On average, the largest residual accounts for .087 of squared residuals, with the largest 1st percentile accounting for .163 and the largest 10th .464. Coefficient leverage is similarly unbalanced, with the largest observation having a .050 average share and the top 1st and 10th percentiles accounting for .100 and .409 of total leverage, respectively.[16] Figure II below graphs these shares against the number of clusters/observations. Once again, the problem is more acute in small samples, but it is surprising how severe it can be in large samples, with 5th and 10th percentile shares in excess of .5 and even approaching 1.0 appearing routinely in regressions with 1000 to 50000 observations. For delete-one sensitivity, only the shares of the maximum observation matter, but these high percentile shares point to a more general delete-n sensitivity and to the unrecognized volatility of robust and clustered covariance estimates, which depend heavily upon a *very* reduced subset of residuals.

[16]High percentile shares can be generated by a small number of observations, e.g. the max,1st, 5th and 10th percentile shares for a sample with four observations are all .25. This is not the cause of the results presented above. With equal shares for each observation in each regression, the average max, 1st, 5th and 10th percentile shares for all regressions reported in the table would be .009, .016, .055, and .105.

Figure II: Shares of Coefficient Leverage & Squared Residuals



The different rows and columns of Table II explore the causes of the concentrated leverage and dominant residuals found in my sample. In the second and third rows we see that the share of the largest observation and 1st percentile of squared residuals in regressions with a binary dependent variable is close to ⅓ of that found where the dependent variable takes on more values, suggesting that winsorizing or otherwise censoring data could substantially alleviate this problem. The largest and 1st percentile shares of coefficient leverage of treatment variables is almost twice as high in regressions which interact treatment with non-treatment covariates than it is in those which do not (rows 4 & 5), and roughly ½ as large in the first table where authors present their main results than it is in other tables (rows 6 & 7). Nevertheless, even in first tables and regressions without interactions, just .1 of observations on average account for more than ⅓ of total leverage. Regressions which use the default covariance estimate have both less extreme residuals and less extreme concentration of coefficient leverage. The third column of the table calculates coefficient leverage without covariates, i.e. using the residuals of the regression of each treatment variable on other treatment measures (including treatment interactions with other covariates) and a constant, but leaving out other non-treatment regressors. These measures are at

best only slightly lower than those calculated using all covariates, so one can conclude that high coefficient leverage is principally due to treatment design in the regression.

A few examples illustrate how regression design can lead to uneven coefficient leverage. Binary treatment, applied 50/50 to the entire sample, produces uniform leverage, with shares equal to percentile values. Apply three binary treatments and control each to ¼ of the population, and in a joint regression each treatment arm concentrates the entirety of leverage in ½ of the observations. The clustered/robust covariance estimate is now based on only half of the residuals and consequently has a volatility (degrees of freedom) and sensitivity to outliers consistent with half the sample size. Run, as is often done, a regression using only one of the three treatment measures as a right hand side variable, so that binary treatment in the regression is applied in 25/75 proportions, and .25 of observations account for .75 of leverage. Apply 50/50 binary treatment, and create a second treatment measure by interacting it with a participant characteristic that rises uniformly in even discrete increments within treatment and control, and .20 of observations account for about .60 of coefficient leverage for the binary treatment measure (even without the non-treatment characteristic in the regression). Seemingly innocuous adjustments in regression design away from the binary 50/50 baseline generate substantially unbalanced leverage, producing coefficient estimates and clustered/robust covariance estimates that are sensitive to residuals and t-statistic distributions which are more dispersed than recognized.

In the OLS sample of Table II, the share of .01 significant results that are delete-one sensitive in regressions with binary dependent variables, in first tables, without interactions or with the default covariance estimate is .21, .21, .25 and .14, respectively, while the share of significant results that are delete-one sensitive in regressions with non-binary dependent variables, in other tables, with interactions or using other covariance estimates is .35, .32, .34 and .30, respectively. In regressions presented in the on-line appendix, however, I find that none of these variables is a robust determinant of delete-one sensitivity once coefficient leverage, the concentration of residuals and sample size, or paper fixed effects, are introduced into the regression. Coefficient leverage is the only variable that is a consistently .01 significant determinant of delete-one sensitivity across all specifications. Regression design is superior in certain types of regressions and this produces results that are less sensitive to outliers and, as shown further below, more robust to the use of alternative inferential procedures.

(b) Randomization Statistical Inference

Randomization statistical inference provides exact tests of sharp (i.e. precise) hypotheses no matter what the sample size, regression design or characteristics of the disturbance term. The typical experimental regression can be described as $y_i = \mathbf{t}_i'\boldsymbol{\beta}_{\mathbf{t}} + \mathbf{x}_i'\boldsymbol{\beta}_{\mathbf{x}} + \varepsilon_i$, where $\mathbf{t}_i$ is a vector of treatment variables (including interactions with non-treatment covariates) 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 hypotheses which specify the treatment effect for each participant, because sharp hypotheses of this sort allow the calculation of what the estimated $\boldsymbol{\beta}$s would be for any potential random allocation of treatment. The Fisherian null of no treatment effects 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.

An exact test of a Fisherian null can be constructed by calculating all possible realizations of a test statistic and rejecting if the observed realization in the experiment itself is extreme enough. Specifically, let the matrix $\mathbf{T}_E$ composed of the row vectors $\mathbf{t}_i'$ denote the treatment allocation in the experiment. In the typical experiment this matrix has a finite universe $\boldsymbol{\Omega}$ of potential realizations. Say there are S elements in $\boldsymbol{\Omega}$, with $\mathbf{T}_n$ denoting a particular element. Let $f(\mathbf{T}_n)$ be a statistic calculated by inserting matrix $\mathbf{T}_n$ into the estimating equation given earlier above, and let $f(\mathbf{T}_E)$ denote the same statistic calculated using the actual treatment applied in the experiment. Under the null of no treatment effect, $y_i = \mathbf{x}_i'\boldsymbol{\beta}_{\mathbf{x}} + \varepsilon_i$ is the same no matter which treatment is applied, i.e. experimental outcomes would have been exactly the same regardless of the specific randomized draw of $\mathbf{T}_E$ from $\boldsymbol{\Omega}$, so $f(\mathbf{T}_n)$ can be calculated by regressing the fixed observed values of y_i on the fixed regressors $\mathbf{x}_i$ and randomly varied treatment vector $\mathbf{t}_i$.[17] The

[17]More generally, for non-zero nulls, one subtracts the sharp null values times treatment measures from the dependent variable, re-randomizes, adds back in the null treatment effects, and re-estimates the equation.

p-value of the experiment's test statistic is given by:

$$(5)\ \text{randomization p - value} \;=\; \frac{1}{S}\sum_{n=1}^{S} I_n(>T_E) \;+\; U * \frac{1}{S}\sum_{n=1}^{S} I_n(=T_E)$$

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

Calculating (5), evaluating $f(\mathbf{T}_n)$ for all possible treatment realizations in $\mathbf{\Omega}$, is generally impractical. However, under the null random sampling with replacement from $\mathbf{\Omega}$ allows the calculation of an equally exact p-value provided the original treatment result is automatically counted as a tie with itself. Specifically, with N additional draws (beyond the original treatment) from $\mathbf{\Omega}$, the p-value of the experimental result is given by:

$$(6)\ \text{sampling randomization p - value} \;=\; \frac{1}{N+1}\sum_{n=1}^{N} I_n(>T_E) \;+\; U * \frac{1}{N+1}\left[1+\sum_{n=1}^{N} I_n(=T_E)\right]$$

In the on-line I appendix I show that this p-value is uniformly distributed regardless of the number of draws N used to evaluate the test statistic.[18] This establishes that size always equals its nominal value. However, power, provided it is a concave function of the nominal size of the test, is increasing in N (Jockel 1986). Intuitively, as the number of draws increases the procedure is better able to identify what constitutes an outlier outcome in the distribution of the test statistic f(). In my analysis, I use 10000 draws to evaluate (6). When compared with results calculated with fewer draws, I find no appreciable change in rejection rates beyond 2000 draws.

I make use of two randomization based test statistics, which find counterparts in commonly used bootstrap tests. The first is based upon the comparison of the Wald statistics of the conventional test of no treatment effects, as given by $\hat{\boldsymbol{\beta}}_t'(\mathbf{T}_n)\mathbf{V}(\hat{\boldsymbol{\beta}}_t(\mathbf{T}_n))^{-1}\hat{\boldsymbol{\beta}}_t(\mathbf{T}_n)$, where $\hat{\boldsymbol{\beta}}_t$ and $\mathbf{V}(\hat{\boldsymbol{\beta}}_t)$ are the regression's treatment coefficients and the estimated variance of those coefficients. This method in effect calculates the probability

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

$$(7)\quad \hat{\boldsymbol{\beta}}_t'(\mathbf{T_n})\mathbf{V}(\hat{\boldsymbol{\beta}}_t(\mathbf{T_n}))^{-1}\hat{\boldsymbol{\beta}}_t(\mathbf{T_n}) \geq \hat{\boldsymbol{\beta}}_t'(\mathbf{T_E})\mathbf{V}(\hat{\boldsymbol{\beta}}_t(\mathbf{T_E}))^{-1}\hat{\boldsymbol{\beta}}_t(\mathbf{T_E}).$$

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

An alternative test of no treatment effects 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 $\hat{\boldsymbol{\beta}}_t$ across the universe of potential treatment draws in $\boldsymbol{\Omega}$. 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

$$(8)\quad \hat{\boldsymbol{\beta}}_t'(\mathbf{T_n})\mathbf{V}(\hat{\boldsymbol{\beta}}_t(\boldsymbol{\Omega}))^{-1}\hat{\boldsymbol{\beta}}_t(\mathbf{T_n}) \geq \hat{\boldsymbol{\beta}}_t'(\mathbf{T_E})\mathbf{V}(\hat{\boldsymbol{\beta}}_t(\boldsymbol{\Omega}))^{-1}\hat{\boldsymbol{\beta}}_t(\mathbf{T_E}).$$

In the univariate case, this reduces to the square of the coefficients divided by a common variance and, after eliminating the common denominator, is simply a comparison of squared coefficients. Hence, I refer to this comparison as the randomization-c. It corresponds to bootstrap tests which use the distribution of bootstrapped coefficients to calculate the covariance matrix. I use the coefficient covariance across the 10000 randomization draws to approximate $\mathbf{V}(\hat{\boldsymbol{\beta}}_t(\boldsymbol{\Omega}))$.[19]

The randomization-c allows for an easy omnibus test of overall statistical significance across multiple or all regressions in an experimental paper. One simply stacks all treatment coefficients into $\hat{\boldsymbol{\beta}}_t$, draws repeated randomization treatments $\mathbf{T}_n$ from $\boldsymbol{\Omega}$, and calculates (8) above. The estimated covariance of these coefficients in the universe $\boldsymbol{\Omega}$ is calculated from their joint realizations. An omnibus version of the randomization-t is much more difficult, as it requires iteration specific estimates of the covariance of coefficients across multiple equations.[20]

While randomization inference is exact ex ante the allocation of treatment, it is important to note that it is not exact ex post the allocation of treatment in a population sampling world with

[19]Strictly speaking, in multi-coefficient tests the test statistics $f(T_n)$ are now a function of the joint realization of the randomization draw, so the proof of exactness for a finite number of draws in the on-line appendix is no longer valid. My intent, however, is to provide a counterpart to the common bootstrap technique; with 10000 draws I should have a fairly close approximation of the true coefficient covariance matrix; and in tests of an individual coefficient the variance cancels from both sides of (8), so the proof of exactness remains valid.

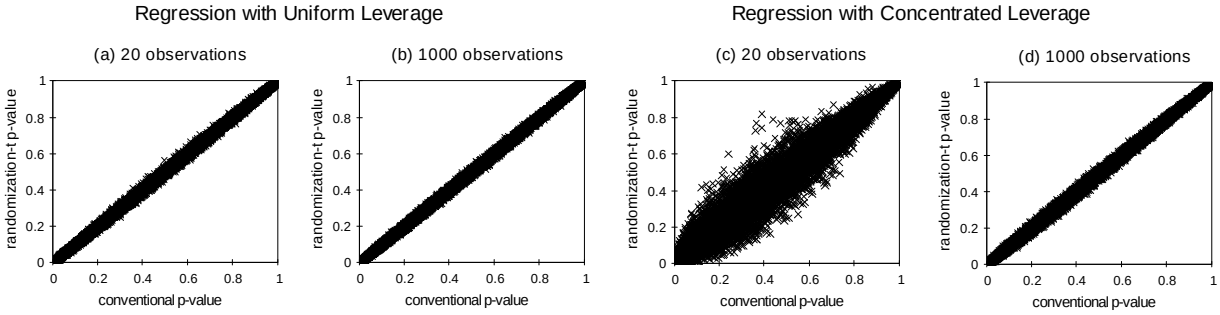
[20]White (1982) showed that an asymptotically valid estimate of the covariance matrix for coefficients estimated in multiple equations is given by a multi-equation version of the robust covariance estimate that uses equation level scores (see also Weesie 1999). However, most papers present the relevant data in multiple, differently organized, data files, so the cross-product of scores is extraordinarily difficult to form. Moreover, when scores can be calculated within a single data file, the resulting covariance matrices are often hopelessly singular.

stochastic disturbances. Consider the case of the randomization-c. The variance of a treatment coefficient under population sampling depends upon the sum of the squared residuals of the treatment measure projected on the other covariates. Consequently, with any other covariates in the regression, each potential realization of randomized treatment will lead to a different variance of treatment results under population sampling from stochastic disturbances. If the allocation of treatment is such that the residual variance of the treatment regressor is high, the coefficients will have a lower variance than is the case for other treatment outcomes. As the randomization-c randomizes across those other potential outcomes, it will find that on average, across the realizations of the disturbances, they generate more dispersed coefficients, leading to a rejection rate of the null lower than nominal value. Conversely, for an allocation of treatment where the residual variance of treatment is unusually low, rejection rates will be higher than nominal value. This problem is clearly greater in the randomization-c, which does not divide by the regression specific covariance estimate, but with inaccurate population-sampling covariance estimation it can arise in the randomization-t as well. For a sharp null, randomization inference is exact ex ante the stochastic allocation of treatment no matter what the residuals. Consequently, it is exact ex ante the allocation of treatment across all stochastic realizations of the residuals. It is not, however, necessarily exact ex post the allocation of treatment, where treatment is now fixed and the distribution is integrated only across the potential realizations of the disturbances.

The on-line appendix provides practical details on how I executed randomization inference for my sample. One issue, however, deserves mention here. As noted above, 12 papers systematically clustered at a level of aggregation below that in which treatment was applied. These papers in effect make the statement that the grouping of observations in laboratory sessions or geographical areas is meaningless. For these papers, I randomize treatment at the level at which authors clustered (or didn't), treating the actual treatment grouping as irrelevant. Results with randomization at the treatment level, reported in the on-line appendix, find far fewer significant coefficients. For papers which generally clustered at treatment level but did not cluster subsets of equations, I consistently randomize at the treatment level as this is necessary to calculate the joint distribution of coefficients across equations.[21]

[21]Seven papers clustered at least some of their regressions at a level of aggregation greater than treatment, but the units so described contain heterogeneous treatment in uneven numbers, so it is not possible to re-randomize in a

Figure III: Randomization Inference vs. Conventional Inference with IID Normal Errors



Notes: Conventional p-value calculated with default covariance estimate. Uniform leverage: regression with random 50/50 binary treatment and a constant. Concentrated leverage: regression with random 50/50 binary treatment, uniformly distributed subject characteristic, binary treatment interacted with characteristic, and constant. All p-values refer to coefficient on binary treatment. Dependent variable is standard normal.

(c) Randomization and Conventional Inference in the Context of Outliers

In this section I use two examples to show how differences between randomization and conventional inference centre on the issue of outliers. In the first, binary treatment is applied 50/50 to the sample, so that coefficient leverage is uniform. In the second, binary treatment is once again applied 50/50 to the sample, but is also interacted with a participant characteristic that increases uniformly within the sample in discrete increments. Binary treatment is randomly applied but, as noted earlier above, the presence of the characteristic interacted with treatment in the regression can produce substantially concentrated coefficient leverage. In both examples the dependent variable is standard normal, i.e. independent of all regressors, and I focus on the statistical significance of the coefficient on the binary treatment variable. I perform 10,000 simulations with samples of 20, 100 and 1000 observations.

Figure III shows that in the case of uniform leverage randomization and conventional inference with the default covariance matrix produce virtually identical results, whether in small or large samples. With concentrated leverage, however, the two differ markedly with only 20 observations, but these differences disappear in large samples. Table III identifies the reason behind these similarities and differences. The columns of the table divide the results into groups depending upon whether the conventional-t and randomization-t reject the null of no treatment effects at the .01 level. Within each category I report the fraction of results which are conventionally delete-one sensitive, i.e. where the deletion of one observation overturns the conventional significance or insignificance of the coefficient, and the number of observations that fit in the category. The sum of observations in columns (1) and (2) is the number of times

---

fashion consistent with the clustering. Two papers clustered across portions of different treatment units, mainly due to coding errors (e.g. the cluster designations were not unique).

Table III: Shares That Are Delete-One Sensitive By Rejection of Null on Binary Treatment

	statistically significant using conventional-t, randomization-t			
	(1) yes, yes	(2) yes, no	(3) no, no	(4) no, yes
Binary treatment is the only regressor				
default-t, randomization-t				
20 observations	.43 (N=93)	1.0 (N=8)	.03 (N=9888)	1.0 (N=11)
100 observations	.45 (N=94)	1.0 (N=9)	.01 (N=9883)	1.0 (N=14)
1000 observations	.20 (N=84)	.80 (N=10)	.00 (N=9895)	.82 (N=11)
Regression includes uniformly distributed participant characteristic interacted with binary treatment				
default-t, randomization-t				
20 observations	.74 (N=54)	.83 (N=42)	.03 (N=9863)	.71 (N=41)
100 observations	.47 (N=77)	1.0 (N=24)	.02 (N=9874)	.88 (N=25)
1000 observations	.28 (N=83)	.90 (N=10)	.00 (N=9890)	.88 (N=17)
robust-t, randomization-t				
20 observations	.52 (N=87)	.90 (N=144)	.09 (N=9762)	.86 (N=7)
100 observations	.27 (N=91)	1.0 (N=30)	.03 (N=9874)	1.0 (N=5)
1000 observations	.28 (N=90)	1.0 (N=10)	.01 (N=9892)	1.0 (N=8)

Notes: Dependent variable is standard normal and independent of all regressors. N = number of observations in the cell, i.e. underlying the calculation of the share that are delete-one sensitive.

conventional inference rejects the null, while the sum of columns (1) and (4) represents the frequency with which randomization inference rejects the null.

As shown in the table, in all cases when randomization inference disagrees with conventional inference the conventional result is usually delete-one sensitive. Thus, the similarities and differences graphed in Figure III are largely a function of whether or not the conventional results depend upon outliers. With concentrated leverage, conventional results are more frequently sensitive to outliers, and so randomization inference disagrees more frequently with conventional results. Comparing the sum of columns (1) and (2) to the sum of columns (1) and (4) in Table III, we see that when the default covariance estimate is used, so that both randomization inference and conventional inference are exact, both methods reject roughly .01 of the time (allowing for simulation variation). Randomization inference simply substitutes some cases where the conventional failure to reject is dependent upon one observation (column 4) for cases where the conventional rejection of the null depends upon only one observation (column 2). As the number of observations grows, the sensitivity of conventional results to outliers falls, and randomization and conventional inference are in closer agreement. In the case of conventional

inference with the robust covariance estimate, with disproportionate weight placed on a small subset of residuals the distribution of the t-statistic is more dispersed than recognized, so in small samples the conventional test rejects more than .01 of the time.[22] Randomization inference once again disagrees mostly in cases where conventional rejections are delete-one sensitive.

The preceding results can be understood by considering the mechanics underlying randomization inference. Randomization inference operates by considering the universe of potential experimental outcomes. In both examples above, this consists of the N choose N/2 possible allocations of the 50/50 binary treatment across the sample of N observations. When a significant conventional result depends upon relatively few observations, however, most of these allocations do not matter; all that matters is the allocation given to the critical observations, and there are relatively few such possible outcomes. The actual allocation of treatment is now one of only a few possible outcomes, so randomization inference concludes that a large t-statistic or coefficient is a fairly common occurrence and, relative to the conventional result, raises the p-value. In contrast, when an insignificant conventional result depends upon relatively few observations, randomization inference, based upon the limited universe of outcomes generated by allocations to those observations, finds that the potential distribution of t-statistics is shifted toward smaller t-statistics, producing a lower p-value than the conventional result.

(d) Joint vs Multiple Hypothesis Testing

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

[22]As the sample size increases, even if the proportionate reduction in the number of residuals used to calculate the covariance matrix remains constant, size distortions fall because the gap between the t-distribution with N and cN ($c < 1$) degrees of freedom falls as N increases.

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

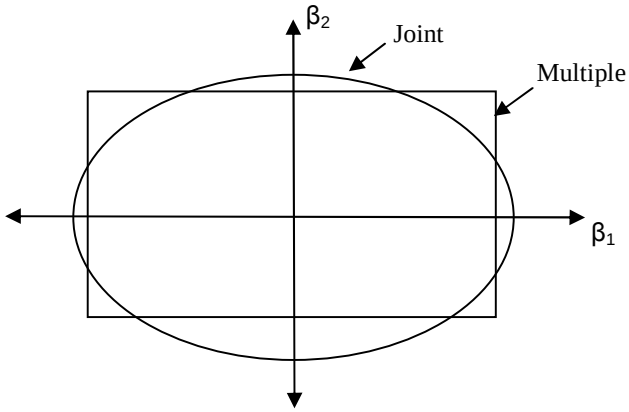


Figure IV illustrates the case where one is interested in testing the significance of two coefficients whose distribution is known to be normal and independent of each other.[23] The oval drawn in the figure is the Wald acceptance region for the joint significance of the two coefficients, while the rectangle is the acceptance region for the two coefficients tested individually. In the multiple testing framework, to keep the probability of one or more Type I errors across the two tests at level α, one could select a size η for each test such that $1-(1-\eta)^2 = \alpha$. The probability of no rejections, under the null, given by the integral of the probability density inside the rectangle, then equals $1-\alpha$. The integral of the probability density inside the Wald ellipse is also $1-\alpha$. The Wald ellipse, however, has the property that it is the minimum translation-invariant[24] area such that the probability of falling in the acceptance region is $1-\alpha$. It achieves this, relative to the multiple testing rectangle, by dropping corners, where the probability of two extreme outcomes is low, and increasing the acceptance region along the axes. If one thinks of alternatives as randomly falling anywhere in the two-dimensional space, the ability (power) of the joint testing framework to achieve a rejection when the joint null is false is higher, because of its smaller overall acceptance region. If one thinks of alternatives as falling along the axes, i.e. some nulls are true while others are false, the ability of the multiple testing framework to achieve a rejection when the joint null is false is higher, because of the smaller length of the acceptance region along the axes. As shown shortly, multiple testing frameworks may possess

[23] A version of this diagram can be found in Savin (1984).

[24] That is, such that if one adds a constant to both the point estimate and the null one remains in the confidence region. Stein (1962) provides examples of smaller areas that do not satisfy this requirement.

more or less power than suggested by this stylized example, but the basic intuition provided by the diagram, that Wald joint tests try to maximize power for general alternatives and multiple testing frameworks try to maximize power for alternatives that lie along the axes, both while controlling the probability of a Type I error under the null, carries through.

Multiple testing is an evolving literature. The classical Bonferroni method is a single step procedure that keeps the probability of a Type I error in N tests at or below α by appealing to Boole's probability inequality and evaluating each test at the α/N level. Holm (1979) introduced a step-down procedure that sorts p-values in ascending order $p_1 \leq p_2 \ldots \leq p_N$ and moves through the list, evaluating each p_j against $\alpha/(N+j-1)$, and stopping on the first failure to reject. Step-downs of this sort, however, are not used in this paper. With randomization inference the distribution of each treatment test statistic generally depends upon the null for others in the same equation, as these values are subtracted from the dependent variable as treatment is re-randomized. Moving through the step-down procedure requires either inconsistently maintaining the belief that the parameters just rejected equal their null values or the costly or infeasible recalculation of p-values based upon the reduced subset of hypotheses.[25] Step-down procedures are more easily applied when p-values are subset pivotal (Westfall & Young 1993), i.e. do not depend upon the null for other tests, as in when each regression contains only one treatment measure and samples do not overlap across regressions.[26] Unfortunately, this is generally not the case in the papers I study. My interest in multiple testing is as a comparison, with different power properties, to joint tests. For this purpose, it is enough to note whether any rejection occurs, not how many or which rejections occur. For its very first rejection decision, Holm's method uses the same α/N cutoff as Bonferroni.[27] The α/N cutoff, however, provides a

[25]In the case of multiple treatments in one equation, one can follow Romano & Wolf's (2005a) suggestion and permute only the treatments within the set of null hypotheses currently tested (but this is only consistent if the dropped treatments do not appear in other equations). Otherwise, or when a given treatment appears multiple times (through interactions with participant characteristics) in a given regression, one needs to recalculate the p-values of remaining hypotheses for least favourable values of the parameters whose nulls have been rejected, including allowing for alternatives such as sharp heterogeneous effects, i.e. effects that vary by observation.

[26]Romano & Wolf (2005a) show that a more general requirement, trivially met by subset pivotality, is that critical values are weakly monotonic in subsets of hypotheses, but in most cases this is hard to confirm.

[27]Similarly, control of the false discovery rate at rate α using Benjamini & Hochberg's (1995) step-down procedure imposes a rejection criterion of α/N for the first step. While this procedure has greater power than Holm's method after the first rejection, with regards to the question of whether at least one rejection occurs, it provides exactly the same answer as Bonferroni-Holm.

substantial relaxation on the axes relative to joint tests. For example, with N equal to 2 or 10 the Bonferonni-Holm cutoffs at the .01 level are z-stats of 2.8 and 3.3, respectively, while the equivalent cutoffs along the axes for the joint testing ellipse are z-stats of 3.0 and 4.8.

Bonferroni's method does not make use of information on the covariance of p-values and hence is conservative, i.e. the probability of a Type I error is generally below the nominal size of the test. Westfall & Young (1993) suggested using bootstrap or randomization inference to calculate the joint-distribution of p-values and then using the α^{th} percentile of the minimum as the single-step cutoff value. It is instructive to apply this method to the problem examined in Figure IV. Since the coefficient estimates are independent, the uniform distributions of the calculated p-values will be independent. The probability their minimum is less than or equal to η is thus given by $1-(1-\eta)^2$. To attain size α, one must select an η such that $1-(1-\eta)^2 = \alpha$. This is precisely the cutoff described earlier. It is easily confirmed that, for all $N > 1$, $\eta = 1-(1-\alpha)^{1/N} > \alpha/N$, so the rejection region for the coefficient with the lowest p-value is wider than in Bonferroni's method. Practically speaking, however, these differences are miniscule. The value of the Westfall-Young procedure is greater when the p-values are correlated, because in this case a larger rejection region can be used. For example, if the p-values are perfectly correlated, then, under the null, α is the α^{th} percentile of their minimum and hence provides an α probability of a Type I error. As N increases, adjustments of this sort can provide a very substantial improvement over the Bonferroni α/N cutoff. Randomization or bootstrap inference to calculate the joint distribution of p-values is essential in this improved multiple testing procedure. Westfall & Young also allowed for step-down methods (see also Romano & Wolf 2005b), but, as noted above, they are not relevant for this paper.

While multiple testing has a reputation for reducing power, this is not necessarily the case for the manner in which I use it, where the focus is on the existence of any rejection of the null of treatment irrelevance. Bonferroni and Westfall & Young's single-step cutoffs, or the ordered p-value cutoff values in Holm's step-down procedure, mechanically ensure that the probability of rejecting the null for a given coefficient cannot increase as additional tests are added. The probability of a rejection of the null by *any* coefficient, however, can rise or fall, depending upon whether the power of the additional tests offsets the loss of power in existing tests. The same is true in joint tests, where adding additional dimensions to the test increases the acceptance region

for pre-existing variables, but may nevertheless, depending upon power in the added dimensions, increase the probability of rejecting the null. In results below I show that the frequency with which the null of no treatment effects is rejected rises when the poor quality, but voluminous, information in later tables is added to the high quality, but limited, information in first tables.

Joint and multiple tests should be organized around a common theme or null hypothesis where the finding of significant results will change beliefs. There is no point in combining tests of unrelated nulls into a joint test, as, in the event of a positive result, one does not know which has been rejected. Similarly, in multiple testing there is no point of saddling the investigation of one null hypothesis with tighter rejection regions to maintain overall Type I error rate control with an unrelated null hypothesis. Control of Type I error is most meaningful, and clearly interpretable, in the investigation of a unified theme. This is all the more true when the emphasis is on the existence of any rejection of the coefficient nulls, and not the number per se. Later in this paper I organize joint and multiple tests around the themes presented by the authors themselves in their regressions and in their organization of results into tables.

IV: Results

(a) Significance Rates at the Coefficient Level

Table IV reports the statistical significance of treatment effects using different methods. In the upper left-hand panel we see that of the 5792 treatment coefficients appearing in the 53 papers, using authors' methods 719 and 1411 are found to be significant at the .01 and .05 levels, respectively. When evaluated using the randomization-t and randomization-c, the number of significant coefficients falls to .81 and .86, respectively, of the original rate at the .01 level and .90 and .89 at the .05 level. Significance rates fall less for coefficients presented in the first table with main results, and somewhat more in later tables. Treatment coefficients in regressions with covariate interactions do particularly badly, with significance rates falling to .7 of the original level at the .01 level. In OLS regressions with the default covariance matrix there is no systematic difference between conventional and randomization results. Table IV also shows that the bootstrap-t and jackknife[28] produce significance rates that are generally similar to those of randomization inference, while the bootstrap-c universally rejects more often, albeit usually less

[28] I conduct two sided tests using the bootstrapped percentiles of the squared values of t-statistics and coefficients, and the jackknifed standard error estimate using the t-distribution with N -1 degrees of freedom.

Table IV: Statistical Significance of Treatment Effects at the Coefficient Level

	.01	.05	.01	.05	.01	.05
	all tables (5792 coefficients)		first table (1127 coefficients)		other tables (4665 coefficients)	
authors' method	719	1411	271	411	448	1000
randomization-t	.81	.90	.86	.96	.78	.88
randomization-c	.86	.89	.86	.96	.86	.86
bootstrap-t	.86	.88	.87	.92	.85	.86
bootstrap-c	.92	.94	.96	.98	.89	.93
jackknife	.83	.86	.91	.92	.78	.84
	covariate interactions (2333 coefficients)		without interactions (3459 coefficients)		OLS default covariance (819 coefficients)	
authors' method	220	529	499	882	73	143
randomization-t	.70	.86	.86	.93	.97	.99
randomization-c	.73	.82	.92	.93	1.05	1.05
bootstrap-t	.82	.81	.88	.92	1.16	1.09
bootstrap-c	.88	.92	.93	.96	1.26	1.13
jackknife	.75	.84	.87	.88	1.03	1.01

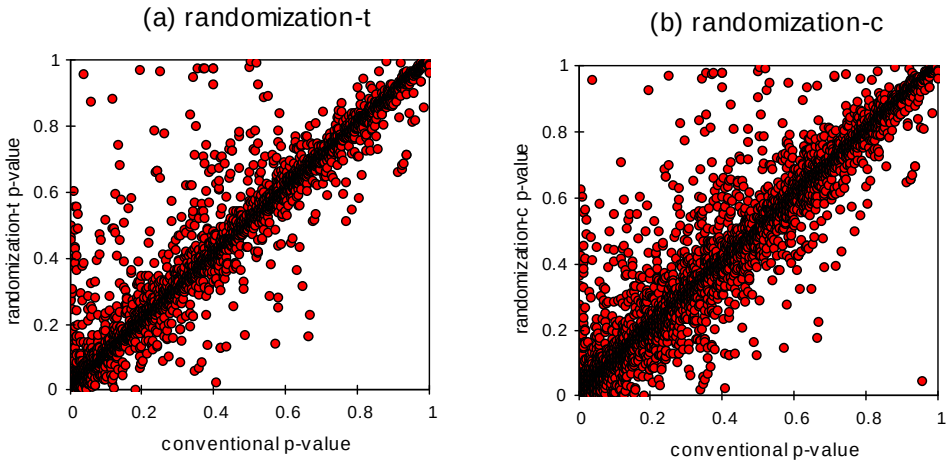
Notes: .01/.05 = level of the test. Top row reports number of significant results evaluated using authors' methods; values in lower rows are number of significant results evaluated using indicated method divided by the top row. Authors' method = p-value evaluated using authors' chosen covariance estimate and distribution (t or normal). Interactions = treatment measures in regressions which interact randomized treatment with participant characteristics or other non-treatment covariates.

than conventional tests. Bootstrapping using the pivotal t-statistic is generally considered to be superior to bootstrapping using the non-pivotal coefficient distribution, as the coverage error of the former converges to zero more rapidly (Hall 1992).

Figure V graphs the randomization-t and -c p-values against those calculated using authors' methods. As shown, randomization inference often changes p-values dramatically. In the case of results which are conventionally significant, in particular, the changes are substantial. Among coefficients whose .01 (.05) conventionally significant p-value is found to be insignificant using randomization methods, the average p-value rises from .004 (.024) with conventional inference to .058 (.138) using the randomization-t and .069 (.129) using the randomization-c.[29] Using the randomization-c, where it is less costly to calculate multiple

[29]In the frequentist world of the *s, **s and ***s reported in economics journals, such comparisons have no meaning. A p-value of .011 is no more significant at the .01 level than a p-value of .11, so all that matters are the 0/1 changes in significance rates reported in Table IV. Seminar participants and referees, however, often ask whether p-value changes are substantial, reflecting, presumably, quasi-Bayesian calculations involving the likelihood of

Figure V: Randomization and Conventional P-values



versions of tests, I find that ⅓ of coefficients whose significance is changed at the .01 level actually cannot reject the null that the treatment effect is equal in magnitude and *opposite in sign* to that reported by authors. Given the sensitivity to outliers shown earlier in Figure I, the magnitude of these changes is perhaps not all that surprising.

Table V examines the characteristics of regressions where the randomization test agrees and disagrees with the conventional test, focusing on tests at the .01 level using the randomization-t (comparisons for the .05 level and the randomization-c are similar and are reported in the on-line appendix). As shown, samples sizes tend to be smaller when the randomization test disagrees with the conventional test. In ¾ of the cases where the randomization test reverses a conventionally significant result, the conventional result can be changed by deleting only one or two clusters/observations, while the same is true for virtually all cases where the randomization test reverses an insignificant conventional result.[30] Focusing on OLS regressions alone, Table V shows that the concentration of coefficient leverage is higher when the randomization test renders significant conventional results insignificant, and the concentration of residuals is much larger when the randomization test finds conventionally insignificant results to be significant. Tests for treatment related heteroskedasticity, regressing

---

outcomes under different hypotheses. In a similar vein, in a frequentist world a failure to reject doesn't actually confirm the null, but in a Bayesian world large p-values increase its posterior probability, which explains why authors emphasize the statistical insignificance of treatment coefficients in regressions related to randomization balance and sample attrition.

[30]As calculation of the full range of delete-two p-values is extremely costly, I only do so for cases where the randomization and conventional results disagree. When I examine delete-three sensitivity in computationally cheap equations with less than 500 observations (about .15 of the remaining cases), I find that all are delete-three sensitive. Of the remaining cases where results disagree and are not delete-one or two sensitive, about ½ are in a single paper which uses 10 to 18 treatment measures per equation.

Table V: Average Characteristics by Change of Significance at the .01 Level

	statistically significant using authors' method, randomization-t			
	yes, yes	yes, no	no, no	no, yes
All Regressions	N = 548	N = 171	N = 5041	N = 32
number of clusters/observations	1702	441	1440	418
delete-one sensitive	.159	.585	.075	.719
delete-one or -two sensitive		.731		.938
OLS Regressions	N = 399	N = 90	N = 3401	N = 24
number of clusters/observations	522	378	845	196
delete-one sensitive	.188	.667	.083	.833
delete-one or -two sensitive		.778		.917
share of coefficient leverage				
largest cluster/observation	.040	.101	.050	.027
1st percentile	.076	.161	.102	.046
5th percentile	.216	.351	.273	.164
10th percentile	.359	.520	.413	.266
share of squared residuals				
largest cluster/observation	.057	.078	.088	.489
1st percentile	.122	.144	.166	.548
5th percentile	.315	.307	.336	.723
10th percentile	.454	.432	.463	.789
treatment related heteroskedasticity	.546	.378	.273	.083

Notes: N = number of coefficients in the cell, otherwise each number is the average for observations in that cell. The following are 0/1 indicators: delete-one or two sensitive (conventional statistical significance changes with such deletions); treatment related heteroskedasticity (coefficient resides in a regression which rejects the homoscedastic null at the .01 level).

the squared residuals on treatment variables, find that the homoskedastic null is rejected less frequently in regressions where the randomization test disagrees with the conventional.[31] Since average but varying treatment effects produce heteroskedasticity that is associated with treatment variables, this makes it hard to dismiss differences between randomization and conventional results as reflecting problems of randomization inference in the presence of heterogeneous treatment effects.

Regression analysis, in the on-line appendix, finds that delete-one sensitivity and the

[31]I report results using Koenker's (1981) test. Results with Wooldridge's (2013) F-test are virtually identical. The Breusch-Pagan (1979) test produces rejection rates of (moving left to right in the columns of the table) .67, .56, .46 and .96, but its distribution is based upon the null of iid *and* normal disturbances.

percentile shares of coefficient leverage are robust determinants of disagreement between randomization and conventional results when the latter are significant, while delete-one sensitivity and the percentile shares of residuals are robust determinants of disagreement when conventional results are insignificant. When included alongside these measures, the number of clusters/observations and dummy variables for a first table, treatment interaction with participant characteristics or a default covariance matrix are easily rendered statistically insignificant. Thus, differences along such dimensions reported in Tables IV and V can be ascribed to the concentration of leverage and residuals and delete-one sensitivity.

It would be nice to argue that the agreement of randomization and conventional results in OLS regressions with default covariance estimates in Table IV indicates that Type I errors on the part of conventional inference explain the difference between randomization and conventional results. When authors believe that the nature of residuals allows them to apply methods that are accurate in small samples, randomization and conventional results agree, not only on average, but in the specifics as well.[32] The problem with this argument is that regression design is systematically superior in OLS regressions of this type and, as shown by the analysis mentioned in the previous paragraph, this can explain the similarity of randomization and conventional results in this case. Concentrated leverage produces a sensitivity to outliers, generating disagreement between conventional and randomization results, *and* systematic bias in favour of rejection when using robust and clustered covariance estimates. Given that the problems are highly correlated, regression analysis of the causes of conventional/randomization disagreement does not generate results that are robust enough to specification changes to convince a sceptic.

One can argue that the bootstrap shows that average differences between randomization and conventional inference can largely be attributed to the Type I errors of conventional tests under population sampling. The bootstrap samples from the experimental data, where the population moment is known, to calculate the distribution of test statistics around null values. Insofar as these distributions, when used to evaluate the original results, reduce the frequency of significant results close to those found using randomization tests, they suggest that false

[32]In OLS regressions that use the default covariance matrix, the randomization-t only renders a conventionally significant result insignificant in 4 of 72 cases at the .01 level and 3 of 143 cases at the .05 level, and these are offset, in both cases, by a randomization finding that 2 results that were conventionally insignificant are significant (with all cases being delete-one or two sensitive).

rejections of the null can explain the discrepancy between randomization and conventional results. The bootstrap-t, which is the asymptotically more accurate version of the two bootstraps (Hall 1992), produces average rejection rates that are very close to those of randomization inference. However, a sceptic might note that using an asymptotically accurate method, albeit one that is asymptotically more accurate, to establish the small sample distortions of other asymptotically accurate methods is, if not humorous, at the very least not utterly compelling.

Another approach is to re-evaluate significance in clustered/robust OLS regressions using "effective degrees of freedom" adjustments which correct for the excess variability in standard error estimates brought on by concentrated leverage and, as shown in Young (2016) in simulations with iid normal disturbances using the OLS equations of this sample, render statistical inference using these methods nearly exact. I find that these corrections lower significance rates in clustered/robust OLS regressions by the same amount as the randomization-t. Moreover, the two methods agree as to which of the authors' significant coefficients should be reclassified as insignificant in more than 70 percent of cases. Of course, the actual disturbances in the data are not iid-normal, so one can question these results as well.

The alternative is to argue that the average differences between conventional and randomization results in Table IV reflect Type II errors, i.e. a relative lack of randomization power. The size distortions of authors' methods, a flaw when the null is true, are a positive feature when it is false, as they increase power. If randomization tests are less sensitive to outliers, it is also likely that they will be less able to detect heterogeneous treatment effects. The problem with this argument is that most of the cases where randomization inference disagrees with conventionally significant results are delete-one or –two sensitive. If these are Type II randomization errors, they indicate extremely skewed outcomes, raising very serious questions about the positive and normative interpretation of results. It is doubtful that these are the sort of "average" treatment effects that most authors, had they been aware of this sensitivity, would want to hang their hats on.

(a) Significance Rates using Joint and Multiple Testing Procedures

In this section I report the frequency with which any rejection of any null is found using joint and multiple testing procedures. I organize this testing in the way results are presented by authors. First, at the regression level, I ask whether any treatment measure had a non-zero effect

Table VI: Statistical Significance of Joint Tests at the Regression Level
(regressions with multiple treatment regressors)

	.01	.05	.01	.05	.01	.05
	all tables (998 regressions)		first table (218 regressions)		other tables (780 regressions)	
significant coef.	348	576	109	144	239	432
conventional F	.89	.75	.83	.84	.91	.72
randomization-t	.69	.62	.74	.75	.66	.57
randomization-c	.73	.61	.77	.74	.71	.57
bootstrap-t	.66	.60	.74	.75	.62	.55
bootstrap-c	.84	.70	.83	.81	.85	.66
jackknife	.78	.65	.81	.77	.77	.62
	covariate interactions (464 regressions)		without interactions (534 regressions)		default covariance (201 regressions)	
significant coef.	151	278	197	298	72	122
conventional F	.93	.74	.86	.75	.85	.66
randomization-t	.59	.58	.76	.65	.76	.58
randomization-c	.66	.55	.78	.67	.86	.66
bootstrap-t	.57	.53	.73	.66	.81	.61
bootstrap-c	.80	.66	.87	.74	.97	.69
jackknife	.72	.59	.83	.71	.90	.66

Notes: top row in each panel indicates number of regressions with one or more conventionally significant coefficients at the level specified; lower rows reports number of regressions which reject the joint null as a fraction of top row.

on the outcome of the regression. This corrects for the fact that authors are, often quite explicitly, examining whether any of multiple treatment measures had an impact on the dependent variable. Second, at the table level, I ask whether one can reject the null that all treatment measures, in all regressions presented in a given table, have zero effects. Authors organize their tables around a theme, e.g. "Did treatment have any effect on any of the following list of outcomes?" or "What is the evidence that treatment, examined using a variety of regression specifications, had an effect on a key outcome of interest?" Readers need some way to evaluate the multiple results that are presented, taking into account their joint distribution within and across regressions. Finally, at the paper level, I stack all treatment coefficients in all regressions and ask whether it is possible to reject the null that treatment had no effect anywhere, i.e. is simply irrelevant.

Table VI presents the statistical significance of treatment effects in joint tests at the

regression level in regressions with more than one treatment regressor. In the top-line of each panel I report the number of regressions where at least one significant coefficient is found at the level specified, leading the reader, given the almost universal absence of a reported F-test, to conclude that treatment was having a significant effect on the dependent variable. When a conventional F-test is applied, the number of regressions that can reject the null of no treatment effects falls by .11 at the .01 level and .25 at the .05 level. Randomization versions of these tests lower significance rates further, by about .3 and .4 of the original level at the .01 and .05 levels, respectively.[33] Regressions in first tables do better, with only a .25 reduction in significance rates at both levels using randomization tests. Regressions outside of first tables, or in tables which interact treatment with non-treatment characteristics, do especially badly, with .3 to .4 and .4 to .45 reductions in significance rates at the .01 and .05 levels, respectively. In regressions which use the default covariance estimate, the conventional F-test reduces significance rates substantially with results that, once again, are quite close to those found using randomization inference. Bootstrap and jackknife results are in rough agreement with randomization inference, with the bootstrap-c once again providing the most generous results.

The greater resiliency to joint testing of regressions in first tables can be attributed to two factors. First, to begin with, these regressions have less extreme coefficient leverage and residuals, and consequently have conventional p-values which are less subject to size distortions and less sensitive to outliers. In Table VI the rejection rate of joint randomization tests is approximately .9 that of the conventional F test in regressions from first tables, whereas it is between .65 and .8 of the conventional F in regressions from other tables or with covariate interactions. Second, there is, quite simply, much less multiple testing in first tables. When they have at least one significant result and more than one treatment variable, regressions outside of first tables have an average of 6.1 treatment variables. In contrast, in such cases regressions in first tables have an average of 4.8 treatment variables. Results presented in first tables are more reliable, with less extreme leverage and residuals, sensitivity to outliers and multiple testing.

Table VII presents significance rates for joint tests at the table and paper level. Once

[33]As in the case of coefficients earlier, most changes are not trivial. Where a regression has a minimum conventional p-value below .01 (.05), but the randomization joint test does not reject at that level, the average randomization-t p-value rises to .12 (.21), while that of the randomization-c rises to .18 (.25).

Table VII: Statistical Significance of Joint Tests at Table & Paper Level by Level of Test

	.01	.05	.01	.05	.01	.05
	all tables (203 tables)		first table (53 tables)		other tables (150 tables)	
significant coef.	126	168	37	45	89	123
randomization-c	.45	.51	.46	.53	.44	.50
bootstrap-c	.54	.50	.54	.56	.54	.48
	no interactions (120 tables)		with interactions (83 tables)		paper level (53 papers)	
significant coef.	72	97	54	71	50	52
randomization-c	.46	.54	.44	.48	.48	.58
bootstrap-c	.63	.56	.43	.42	.54	.71

Notes: as in Table VI, except the unit of analysis is the table or paper.

again the top row of each panel reports the number of tables or papers where at least one conventionally significant result can be found, and lower rows report the relative number of significant results found using the randomization-c and bootstrap-c. All of these tests are "omnibus" tests, in that they stack the coefficients from multiple regressions, including, depending upon how the authors have organized their results into tables, multiple dependent variables (i.e. outcomes). As shown, no matter whether in all tables, first tables, other tables, or tables without or with interactions, the frequency with which randomization tests reject the null of zero treatment effects at the .01 and .05 level is only about .45 and .50, respectively, of the rate at which at least one conventionally significant treatment effect is found somewhere in the table. As elsewhere, the bootstrap-c is often more generous, but nevertheless still dramatically reduces the number of significant results.

Table VII brings to light three interesting facts. First, in only 37 of 53 first tables with main results do authors report a statistically significant effect at the .01 level. Overall, however, authors find .01 significant effects in 50 of the 53 papers. Thus, in many cases strong conventional evidence of statistical significance at the coefficient level is only found later in regressions with more leverage and more treatment measures. Second, randomization inference finds that in only 17 (.46 of 37) of the 53 papers is there sufficient evidence in the first table to

reject the null that treatment is having no effect at the .01 level.[34] Thus, in the tables where authors address, presumably, the issues of greatest economic significance using regressions which are econometrically most reliable, there is limited evidence that treatment matters. Third, while the information presented in later tables is less reliable on a case by case basis than that given in first tables, it is cumulatively important. First tables typically have 21.3 treatment regressors in 9.1 regressions covering 2.8 outcomes. Later tables have an average of 31.1 treatment regressors in 10.3 regressions covering 3.8 different outcomes. While each coefficient p-value is less reliable, there is a greater quantity of information and this leads to similar relative randomization and bootstrap rejection rates of the null that treatment is having no effects.

With this background, it is possible to understand why the omnibus test applied at the paper level, as is done in the last panel of Table VII, is not prejudiced against authors' results. The randomization omnibus test at the paper level finds that 24 of the 53 papers reject the null at the .01 level. 15 of the 17 rejections of treatment irrelevance using first table information alone are retained, and 9 additional rejections are found.[35] The many regressions presented in later tables do not saddle the joint test with low powered results or tests of patently true nulls. Rather, while individually unreliable, they cumulatively provide additional evidence that experimental treatment is not irrelevant. A joint hypothesis test merely asks if anything is going on, not what is going on, and in this context the addition of all of the information presented on the common theme of treatment effects in a paper is a boon, not a tax. Although authors find .01 significant treatment effects in 50 of the 53 papers, and randomization inference finds them in only 24, this is not because the omnibus test dilutes strong evidence with noise. Rather, it is because in their most important and highest quality regressions, in their first tables, authors present relatively little evidence of significant treatment effects, so that additional support must come from poorer quality, highly leveraged, regressions which explore treatment interactions with participant characteristics. Such was, no doubt, the original intent of authors in presenting these additional specifications, and the omnibus test of experimental significance, at the paper level, makes use of this information in precisely that manner.

[34]Once again, the p-value changes are very large. Where a first table has a conventional p-value that is below .01 (.05), but the omnibus test does not reject at that level, the average omnibus randomization p-value is .27 (.39).

[35]However, in papers that report at least one result with a p-value less than .01 (.05) but randomization inference does not reject at that level, the omnibus paper randomization p-value remains high, averaging .27 (.37).

Table VIII: Presence of at Least One Significant Coefficient at Regression Level
with Multiple Testing Type I Error Rate Control
(regressions with multiple treatment measures)

	.01	.05	.01	.05	.01	.05
	all tables (998 regressions)		first table (218 regressions)		other tables (780 regressions)	
significant coef.	348	576	109	144	239	432
Bonferroni:						
randomization-t p-value	.58	.56	.62	.77	.56	.49
randomization-c p-value	.62	.58	.68	.78	.59	.52
bootstrap-t p-value	.63	.59	.69	.76	.60	.54
bootstrap-c p-value	.67	.64	.74	.83	.63	.58
jackknife p-value	.62	.58	.72	.78	.57	.51
Westfall-Young:						
randomization-t p-value	.58	.59	.62	.78	.56	.52
randomization-c p-value	.63	.60	.69	.80	.61	.54
bootstrap-t p-value	.64	.61	.71	.79	.62	.55
bootstrap-c p-value	.70	.68	.77	.85	.66	.62

Notes: as in Table VI, except lower rows indicate number of regressions with at least one significant coefficient using multiple testing procedures as a fraction of top row.

One possible criticism of the joint tests presented above is that they dissipate power by combining treatment regressors that the authors believed might have positive effects with treatment regressors that were introduced, along with control, merely to demonstrate their insignificance. Thus, while accepting that the many regressions and specifications presented in tables after the first contribute to the rejection of the null, one might still argue that the joint tests listed above reduce power by combining treatment placebos with genuine treatment. Absent pre-analysis plans, which are not reported in any paper, there is no way to objectively evaluate the validity of this argument. One can give it maximum credence, however, by using multiple testing procedures, increasing, as noted earlier above, power to reject the null along the axes at the expense of power when multiple coefficients take on non-zero values. Such tests are presented in Tables VIII and IX. As usual, the top row reports the number of regressions, tables or papers where at least one significant coefficient is found, leading readers to conclude, absent multiple testing corrections, that treatment was having significant effects. The lower rows report the number of cases where at least one significant coefficient is found using multiple testing procedures as a share of this baseline value. I concern myself only with whether any significant

Table IX: Presence of at Least One Significant Coefficient at Table & Paper Level with Multiple Testing Type I Error Rate Control

	.01	.05	.01	.05	.01	.05	.01	.05
	all tables (203 tables)		first table (53 tables)		other tables (150 tables)		paper (53 papers)	
significant coef.	126	168	37	45	89	123	50	52
Bonferroni:								
randomization-t p	.37	.49	.41	.64	.35	.43	.42	.65
randomization-c p	.40	.50	.46	.64	.38	.45	.48	.65
bootstrap-t p	.40	.48	.46	.58	.37	.44	.42	.58
bootstrap-c p	.52	.55	.68	.64	.45	.51	.58	.63
jackknife p	.43	.50	.51	.62	.39	.46	.46	.63
Westfall-Young:								
randomization-t p	.44	.56	.51	.71	.40	.50	.44	.69
randomization-c p	.45	.58	.51	.73	.43	.52	.50	.69
bootstrap-t p	.43	.53	.51	.62	.39	.50	.44	.65
bootstrap-c p	.55	.60	.68	.73	.49	.54	.58	.77

Notes: as in Table VIII, except the unit of analysis is the table or paper.

effect is found, not how many, as the finding of even one significant coefficient is sufficient to reject the null that all treatment was irrelevant.

Tables VIII and IX show that, by and large, multiple testing procedures produce fewer rejections of the null of no treatment effects than joint tests. At the regression level (Table VIII), Bonferroni and Westfall-Young multiple testing procedures using randomization inference p-values produce around .6 of the number of rejections suggested by the minimum reported conventional p-value in all tables, .6 to .8 of authors' rejections in first tables, and .5 to .6 in other tables. This generally compares poorly with the ratios of .6 to .7 in all tables, .75 in first tables, and .6 to .7 in other tables reported earlier in Table VI's joint tests at the regression level. At the table and paper level (Table IX), multiple tests based upon randomization inference have .01 relative rejection rates that often fall below the minimum of .44 for the randomization-c found earlier in Table VII's joint-analysis. The only appreciable improvement is an increase in the .05 relative rejection rate for first tables and in the omnibus paper-level test from approximately .55 in Table VII to .65 or .7 with multiple testing procedures. As expected, Westfall-Young's procedure, which substitutes the estimated joint distribution of p-values for Bonferonni's conservative use of Boole's probability inequality, produces systematically higher rejection rates,

particularly when large numbers of comparisons are made in tables and papers. Outside of .05 size for first tables and complete papers, however, these remain at or below the level found with joint tests. In sum, with some exceptions, shifting power to the axes either lowers or does not improve the rate at which significant treatment effects are found. Dismissing ex-post joint and multiple testing adjustments as inappropriate on the grounds that ex-ante some treatment measures and regressions were intended to matter while all others were considered irrelevant placebos, mischaracterizes the actual pattern of results found in my sample, where more often than not rejections of the null are found when tests emphasize the possibility of multiple treatment measures having moderate, if individually insignificant, effects.

V. Conclusion

The results presented above do not reflect negatively, in any way, on the authors in my sample or randomized experiments in general. Conventionally significant results which are sensitive to outliers and overturned with resampling methods is not a problem unique to these papers, and may be much worse in other fields, as shown by my study of instrumental variables regressions mentioned earlier. The virtual universal absence of joint or multiple tests at any level in papers that have passed through numerous seminar presentations and the scrutiny of knowledgeable referees points, again, to problems with professional practice, rather than individual deficiencies. That said, there is no reason not to address these issues as their practical relevance becomes increasingly apparent. Randomized experiments aspire to achieve the highest standards of identification. They can equally aspire for the highest standards of inference which, in finite samples, absent the chimera of iid normal errors, can only be attained through randomization tests of individual, joint and multiple sharp nulls.

Beyond advocating randomization inference, the results of this paper, and the on-going research of others, suggest there is value added in paying attention to regression and experimental design. Balanced designs lead to uniform leverage with conventional results that are less sensitive to outliers, less subject to size distortions with robust/clustered covariance estimates, and more similar to randomization inference. Balanced designs also enhance the asymptotic properties of randomization inference in the presence of heterogeneous treatment effects (Janssen 1997, Chung & Romano 2013, Bugni, Canay & Shaikh 2017). As noted earlier, regressions with

multiple treatments and treatment interactions with participant characteristics generate concentrated leverage, producing coefficients and clustered/robust standard errors that depend heavily upon a limited set of observations. Rather than maintain the fiction that identification and inference comes from the full sample, more robust finite sample and asymptotic results can be achieved by breaking the experiment and regression into groups (e.g. based on treatment regime or participant characteristics), each with a balanced treatment design.

Consideration of experimental and regression design can also play a role in joint and multiple testing. While the omnibus and single-step procedures used in this paper can evaluate the general question of treatment relevance, more discerning results can be achieved if regressions are designed in a fashion that allows step-down procedures to control the Type I error or false discovery rate (e.g. Holm 1979, Westfall & Young 1993, and Benjamini & Hochberg 1995). Practically speaking, to allow for this tests have to be set up in a fashion that allows subset pivotality (Westfall & Young 1993), where the distribution of each randomization test statistic is independent of the nulls for other treatment results. Dividing regressions into mutually exclusive samples is a trivially easy way to ensure this. Where it is desirable to explore multiple related specifications in a single sample, an omnibus joint test can be used to generate a combined p-value and these results then incorporated into step-down procedures.

Pre-analysis plans, as in Neumark (2001) and Casey, Glennerster & Miguel (2012), are increasingly seen as a means of enhancing research credibility by figuratively tying researchers' hands. Randomization inference both liberates and burdens the design of such plans. On the one hand, it is not actually necessary to pre-specify the details of each regression, as explicit search algorithms that learn from and adapt to the data can be used, with the finite sample distribution of such results under sharp nulls reliably uncovered by rerandomizing and repeating the algorithm. On the other hand, in pre-specifying an analytical approach careful consideration needs to be given to whether the resulting regression designs will be balanced and whether it is advantageous to fashion them so as to allow step-down procedures.

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

Anderson, Michael L. 2008. "Multiple Inference and Gender Differences in the Effects of Early Intervention: A Reevaluation of the Abecedarian, Perry Preschool and Early Training Projects." *Journal of the American Statistical Association* 103 (484): 1481-1495.

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

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.

Breusch ,T. S. and A. R. Pagan. 1979. "A Simple Test for Heteroscedasticity and Random Coefficient Variation." *Econometrica* 47 (5): 1287–1294.

Bugni, Federico A., Ivan A. Canay, and Azeem M. Shaikh. 2017. "Inference under Covariate-Adaptive Randomization." Manuscript: Duke University, Northwester University & University of Chicago.

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

Chung, Eun Yi and Joseph P. Romano. 2013. "Exact and Asymptotically Robust Permutation Tests." *The Annals of Statistics* 41 (2): 484-507.

Cook, R. Dennis and Sanford Weisberg. 1982. "Residuals and Influence in Regression." NewYork: Chapman and Hall, 1982.

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.

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

Heckman, James, Seong Hyeok Moon, Rodrigo Pinto, Peter Savelyev, and Adam Yavitz. 2010. *Quantitative Economics* 1 (1): 1-46.

Holm, Sture. 1979. "A Simple Sequentially Rejective Multiple Test Procedure." *Scandinavian Journal of Statistics* 6 (2): 65-70.

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

Imbens, Guido W. and Jeffrey M. Wooldridge. 2009. "Recent Developments in the Econometrics of Program Evaluation." *Journal of Economic Literature* 47 (1): 5-86.

Janssen, Arnold. 1997. "Studentized Permutation Tests for Non-iid Hypotheses and the Generalized Behrens-Fisher Problem." *Statistics & Probability Letters* 36 (1): 9-21.

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

Koenker, Roger. 1981. "A Note on Studentizing a Test for Heteroskedasticity." *Journal of Econometrics* 17 (1): 107–112.

Lee, Soohyung and Azeem M. Shaikh. 2014. "Multiple Testing and Heterogeneous Treatment Effects: Re-Evaluating the Effect of Progresa on School Enrollment." *Journal of Applied Economics* 29 (4): 612-626.

List, John A., Azeem M. Shaikh and Yang Xu. 2016. "Multiple Hypothesis Testing in Experimental Economics." Manuscript, 2016.

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

Neumark, David. 2001. "The Employment Effects of Minimum Wages: Evidence from a Prespecified Research Design." Industrial Relations 40 (1): 121-144.

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

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

Savin, N.E. 1984. "Multiple Hypothesis Testing." In Zvi Griliches and Michael D. Intriligator, eds. Handbook of Econometrics, Vol. II. Amsterdam: North Holland, 1984.

Stein, C.M. 1962. "Confidence Sets for the Mean of a Multivariate Normal Distribution." *Journal of the Royal Statistical Society, Series B (Methodological)* 24 (2): 265-296.

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

Westfall, Peter H. and S. Stanley Young. 1993. Resampling-Based Multiple-Testing: Examples and Methods for p-Value Adjustment. New York: John Wiley & Sons, 1993.

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

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

Wooldridge, Jeffrey M. 2013. Introductory Econometrics: A Modern Approach. 5th ed. Mason, OH: South-Western, 2013.

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.

Young, Alwyn. 2017. "Consistency without Inference." Manuscript, London School of Economics.