On-Line Appendix for:

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

Alwyn Young
November 2018

Contents:

I: Proofs that the Randomization P-values of (4) and (5) are Exact.

In this section I present proofs that the randomization p-values of (4) and (5) in the text are exact for any arbitrary size α in [0,1]. Equation (4) is the case where the test statistic values for the entire universe of potential experimental realizations $\mathbf{T}_i$ in $\mathbf{\Omega}$ are known; (5) is the case where the p-value is evaluated by drawing N additional random treatments from $\mathbf{\Omega}$. The proof of (4) is trivial. The proof of (5) is an extension of Jockel's (1986) result that the p-value is exact for size α which is an integer multiple of 1/(N+1) to general α, accomplished by treating the experimental test statistic as part of the distributional sample and, consequently, as an observed tie with itself. To ease presentation and avoid confusion in discussing cumulative distribution functions, I define rejection probabilities based upon the number of potential outcomes with test statistics **less than or equal** to that of the experimental draw, rather than greater than or equal, as expressed in (4) and (5). The two approaches are interchangeable, as with a simple sign change of the test statistic one can be substituted for the other. In working through the following proof, the reader will find it helpful to regularly refer to equations (4) and (5) in the paper to confirm what the p-value will be under various circumstances.

Beginning with (4), index the N equally probable[1] elements $\mathbf{T}_i$ of $\mathbf{\Omega}$ so that $f(\mathbf{T}_1) \leq ... \leq f(\mathbf{T}_N)$ represent the ordered potential values of the test statistic $f(\mathbf{T}_i)$ and let E_i represent the number of outcomes whose test statistic ties with $f(\mathbf{T}_i)$ and L_i the number whose test statistic is less than $f(\mathbf{T}_i)$. Then, the p-value associated with a random draw $\mathbf{T}_E$ from the universe $\mathbf{\Omega}$ is given by $L_E/N + U*E_E/N$. Let α be a number between 0 and 1, inclusive, and let $[\alpha N]$ denote the largest integer less than or equal to αN. Whenever $f(\mathbf{T}_E) < f(\mathbf{T}_{[\alpha N]})$, the p-value is strictly less than α. There are $L_{[\alpha N]}$ such events. Whenever $f(\mathbf{T}_E) = f(\mathbf{T}_{[\alpha N]})$ the p-value is less than or equal to α if the random draw U is less than or equal to $(\alpha N - L_{[\alpha N]})/E_{[\alpha N]}$. There are $E_{[\alpha N]}$ such events. Whenever $f(\mathbf{T}_E) > f(\mathbf{T}_{[\alpha N]})$ the p-value is strictly greater than α. Consequently, the probability the p-value is less than or equal to α is given by:

[1] As noted in the paper, if outcomes are not equally likely simply duplicate them in $\mathbf{\Omega}$ according to their relative frequency. More generally, one can define a probability density across outcomes and use it in the proof, but this introduces additional notation.

$$\text{(a1)}\quad \text{Prob(p - value} \le \alpha) \;=\; \frac{L_{\lfloor \alpha N \rfloor}}{N} + \frac{E_{\lfloor \alpha N \rfloor}}{N} \int_0^{(\alpha N - L_{\lfloor \alpha N \rfloor})/E_{\lfloor \alpha N \rfloor}} dU$$

$$= \frac{1}{N}\left(L_{\lfloor \alpha N \rfloor} + E_{\lfloor \alpha N \rfloor}\frac{\alpha N - L_{\lfloor \alpha N \rfloor}}{E_{\lfloor \alpha N \rfloor}}\right) = \frac{1}{N}\alpha N = \alpha$$

which establishes that the p-value is uniformly distributed and exact.

Turning to (5), we begin by deriving a useful result by assuming, for now, that $\mathbf{\Omega}$ is such that $f(\mathbf{T}_i)$ is continuously distributed. Let $F(f(\mathbf{T}_x))$, or F_x for short, denote the cumulative distribution function of $f(\mathbf{T}_x)$, i.e. the probability $f(T_i)$ is less than or equal to $f(\mathbf{T}_x)$ for given $\mathbf{T}_x$. Obviously, F_x is uniformly distributed. Let T_E denote the draw associated with experimental treatment, integer N the additional N draws used to evaluate its p-value using (5) in the paper, and again the notation $[\alpha(N+1)]$ indicate the largest integer less than or equal to $\alpha(N+1)$. Since $f(\mathbf{T}_i)$ is continuously distributed, in making N additional draws from the randomization distribution there will be no ties, so the only tie in the calculation of (5) is the tie of the experimental test statistic $f(\mathbf{T}_E)$ with itself. Consequently, for a given treatment $f(\mathbf{T}_E)$ and associated F_E, the probability the p-value in (5) will be less than or equal to α is given by the probability $[\alpha(N+1)]-1$ draws or less have a test statistic less than $f(T_E)$ plus, given that $f(\mathbf{T}_E)$ provides a tie with itself, the probability exactly $[\alpha(N+1)]$ draws have a test statistic less than $f(\mathbf{T}_E)$ times the probability the uniformly distributed variable U in (5) is less than or equal to $\alpha(N+1) - [\alpha(N+1)]$, or:

$$\text{(a2)}\quad \sum_{x=0}^{[\alpha(N+1)]-1} \frac{N!}{x!N-x!} F_E^x (1-F_E)^{N-x} + \frac{N!}{[\alpha(N+1)]!N-[\alpha(N+1)]!} F_E^{[\alpha(N+1)]} (1-F_E)^{N-[\alpha(N+1)]} (\alpha(N+1)-[\alpha(N+1)])$$

where I have assumed (for the moment) that $1 \le [\alpha(N+1)] \le N$. Since F_E is distributed uniformly, the unconditional probability the experimental test statistic will be less than or equal to α is in these circumstances given by:

$$\text{(a3)}\quad \int_0^1 \left\{ \sum_{x=0}^{[\alpha(N+1)]-1} \frac{N!}{x!N-x!} F_i^x (1-F_i)^{N-x} + \frac{N!}{[\alpha(N+1)]!N-[\alpha(N+1)]!} F_i^{[\alpha(N+1)]} (1-F_i)^{N-[\alpha(N+1)]} (\alpha(N+1)-[\alpha(N+1)]) \right\} \mathrm{dF_i}$$

$$= \sum_{x=0}^{[\alpha(N+1)]-1} \int_0^1 \left\{ \frac{N!}{x!N-x!} F_i^x (1-F_i)^{N-x} \right\} \mathrm{dF_i} + \int_0^1 \left\{ \frac{N!}{[\alpha(N+1)]!N-[\alpha(N+1)]!} F_i^{[\alpha(N+1)]} (1-F_i)^{N-[\alpha(N+1)]} (\alpha(N+1)-[\alpha(N+1)]) \right\} \mathrm{dF_i}$$

$$= \sum_{x=0}^{[\alpha(N+1)]-1} \frac{N!}{x!N-x!} \frac{x!N-x!}{(N+1)!} + \frac{N!}{[\alpha(N+1)]!N-[\alpha(N+1)]!} \frac{[\alpha(N+1)]!N-[\alpha(N+1)]!}{N+1!} (\alpha(N+1)-[\alpha(N+1)])$$

$$= \frac{[\alpha(N+1)]}{N+1} + \frac{\alpha(N+1)-[\alpha(N+1)]}{N+1} = \alpha.$$

Provided $\alpha < 1$, the condition $[\alpha(N+1)] \leq N$ holds. Since the p-value in (5) is always less than or equal to 1, if $\alpha = 1$ the rejection probability is 1, as specified. If $[\alpha(N+1)] < 1$, the p-value (5) rejects at level α if none of the N draws has a test statistic less than $f(\mathbf{T}_E)$ and U in (5) is less than $\alpha(N+1)$, that is with probability:

$$(a4)\ (1-F_E)^N \alpha(N+1)$$

Integrating across the distribution of F_E

$$(a5)\ \int_0^1 (1-F_i)^N \alpha(N+1)\, dF_i = \alpha$$

Together, these results establish that when $f(\mathbf{T}_i)$ is continuously distributed the probability the p-value is less than or equal to α actually equals α for all α in [0,1], i.e. the test is exact.

I now turn to the case where $f(\mathbf{T}_i)$ is not continuously distributed and in particular takes on discrete values. As before, let F_x denote the probability $f(\mathbf{T}_i)$ is less than or equal to $f(\mathbf{T}_x)$ and define p_x as the probability $f(\mathbf{T}_i)$ exactly equals $f(\mathbf{T}_x)$, where $\mathbf{T}_x$ is an element in $\mathbf{\Omega}$. Define the new "test statistic" $g(\mathbf{T}_i) = F_i - u_i * p_i$, where u_i is a draw, for each outcome T_i, from the uniform distribution on [0,1]. By construction, $g(\mathbf{T}_i)$ is continuously and uniformly distributed on [0,1] and consequently, by the results above, is exact if evaluated using (5). I will now show that, conditional on $\mathbf{T}_E$ and the realized draws $\mathbf{T}_1$, $\mathbf{T}_2$, ... $\mathbf{T}_N$ from $\mathbf{\Omega}$, the p-value calculated using $g(\mathbf{T}_E)$ has the same probability of rejecting at level α as the p-value calculated using $f(\mathbf{T}_E)$. From this it follows that even though $g(\mathbf{T}_E)$ is never observed, the p-value calculated using $f(\mathbf{T}_E)$ is uniformly distributed, so the test using $f(\mathbf{T}_E)$ is exact.

I begin by noting that $f(\mathbf{T}_i) < f(\mathbf{T}_j)$ implies $g(\mathbf{T}_i) < g(\mathbf{T}_j)$, as then $F_j \geq F_i + p_j$, which lets us see that with probability one $F_j - u_j * p_j > F_j - p_j \geq F_i > F_i - u_i * p_i$, as the probability $u_j = 1$ and $u_i = 0$ is zero. Let the draw $\mathbf{T}_1$, $\mathbf{T}_2$, ... $\mathbf{T}_N$ from $\mathbf{\Omega}$ contain L draws with values of f() strictly less than $\mathbf{T}_E$ and E draws with values of f() equal to $\mathbf{T}_E$. Select an α. If $L \geq [\alpha(N+1)]+1$ the p-value calculated using (5) for both f() and g() is greater than α, i.e. neither test rejects. If $L + E + 1 < \alpha(N+1)$ the p-value calculated using (5) for both f() and g() is less than α, i.e. both tests reject. Consequently, we need only concern ourselves with the case where $L \leq [\alpha(N+1)]$ and $L + E + 1 \geq \alpha(N+1)$. In these circumstances, the p-value calculated using (5) for f() is given by:

$$(a6)\ \frac{L}{N+1} + U * \frac{E+1}{N+1}$$

where, as usual, I use the fact that $\mathbf{T}_E$ ties with itself. As U is uniformly distributed, this rejects (i.e. is less than α) with probability (α(N+1)-L)/(E+1), which is between 0 and 1 by the conditions stated above. In contrast, the p-value calculated using (5) for g() is given by

$$(a7)\ \frac{L}{N+1}+\frac{E_L}{N+1}+U*\frac{1}{N+1}$$

where $E_L \leq E$ denotes the number of draws $\mathbf{T}_i$ with $f(\mathbf{T}_i) = f(\mathbf{T}_E)$ which end up with $g(\mathbf{T}_i) < g(\mathbf{T}_E)$ after the realization of the draws u_i which determine $g(\mathbf{T}_i)$. Given the u_E which determined $g(T_E) = f(T_E) - u_E*p_E$, the probability a given element of the set of draws that have $f(\mathbf{T}_i) = f(T_{\mathbf{E}})$ ends up with $g(\mathbf{T}_i) < g(\mathbf{T}_E)$ is $1-u_E$. Say $L = [\alpha(N+1)]$. For given u_E, the probability (a7) is less than or equal to α equals the probability $E_L = 0$ times the probability U in (a7) is less than α(N+1)-L, or:[2]

$$(a8)\ u_E^E(\alpha(N+1)-L)$$

Integrating across the uniform distribution of u_E yields:

$$(a9)\ \int_0^1 u_E^E(\alpha(N+1)-L)\,du_E = (\alpha(N+1)-L)/(E+1)$$

which is the same probability as the test statistic using f(). If $L \leq [\alpha(N+1)] - 1$, for given u_E the probability (a7) is less than or equal to α is given by the probability $E_L \leq [\alpha(N+1)] - L-1$ plus the probability $E_L =[\alpha(N+1)] - L$ times the probability U in (a7) is less than or equal to α(N+1)-[α(N+1)], or:

$$(a10)\ \sum_{x=0}^{[\alpha(N+1)]-L-1}\frac{E!}{x!E-x!}(1-u_E)^x u_E^{E-x} + \frac{E!}{[\alpha(N+1)]-L!E-[\alpha(N+1)]+L!}(1-u_E)^{[\alpha(N+1)]-L}u_E^{E-[\alpha(N+1)]+L}(\alpha(N+1)-[\alpha(N+1)])$$

Again, integrating across the uniform distribution of u_E

[2] As a reminder, subscript E in u_E in the formula (and elsewhere) refers to the u_i associated with the calculation of $g(T_E)$. Superscript E in u_E^E and E in the factorials refers to the number of T_i draws with $f(T_i) = f(T_E)$.

$$(a11)\ \int_0^1 \sum_{x=0}^{[\alpha(N+1)]-L-1} \frac{E!}{x!E-x!}(1-u_E)^x u_E^{E-x} du_E$$

$$+ \int_0^1 \frac{E!}{[\alpha(N+1)]-L!E-[\alpha(N+1)]+L!}(1-u_E)^{[\alpha(N+1)]-L} u_E^{E-[\alpha(N+1)]+L}(\alpha(N+1)-[\alpha(N+1)])\,du_E$$

$$= \sum_{x=0}^{[\alpha(N+1)]-L-1} \frac{E!}{x!E-x!}\frac{x!E-x!}{E+1!} + \frac{E![\alpha(N+1)]-L!E-[\alpha(N+1)]+L!}{[\alpha(N+1)]-L!E-[\alpha(N+1)]+L!E+1!}(\alpha(N+1)-[\alpha(N+1)])$$

$$= \frac{[\alpha(N+1)]-L}{E+1} + \frac{(\alpha(N+1)-[\alpha(N+1)])}{E+1} = \frac{(\alpha(N+1)-L)}{E+1}$$

which again is the same as in the case of the f() statistic. Since these examples cover all possible cases, we see that for any possible set of realized draws $\mathbf{T}_1$, $\mathbf{T}_2$, ... $\mathbf{T}_N$ from $\mathbf{\Omega}$, the p-value calculated using $f(\mathbf{T}_E)$ has the exact same probability of rejecting at level α as the p-value calculated using $g(\mathbf{T}_E)$. Consequently, the test statistic based on $f(\mathbf{T}_E)$ in (5) is also exact.

II: Further Details on Randomization, Bootstrap and Jackknife Methods

This appendix provides further details on how I executed randomization, bootstrap and jackknife inference for my sample. First, in each case the regression specification I analyse is the regression specification that reproduces the coefficients and standard errors reported in published tables. This is often different from what is described in the paper or given in do-file code. Published results, however, can almost always be closely approximated through an investigation of the public use data file. Second, in calculating the coefficient covariance matrix for each randomization or bootstrap draw, I defer to the decisions made by authors and use their covariance estimation methods. Third, in producing the randomization distribution I apply the randomized experimental treatment draw $\mathbf{T}_S$ to the entire experimental dataset, recalculate all variables that are contingent upon that realization, e.g. participant characteristics interacted with treatment outcomes, and also reproduce all coding errors in the original do-files that affect treatment measures, e.g. a line of code that unintentionally drops half the sample. All of this follows the Fisherian null: all procedures and outcomes in the experiment are invariant with respect to who received what treatment. In executing the bootstrap, I also draw entire experimental samples (drawing clusters if the authors cluster their regressions), so as to parallel the randomization methods and be able to calculate the joint distribution of coefficients for multi-equation joint testing procedures.

Fourth, in executing randomization or bootstrap iterations I accept an iteration as long as Stata produces a coefficient estimate and standard error for the treatment variable. Some of the procedures authors use do not converge and in some cases Stata warns users that the covariance matrix is highly singular. Coefficients and standard errors produced by these methods are accepted and reported in journal tables. In order to be able to execute the analysis, and following the spirit of the Fisherian null, I duplicate authors' methods and accept results if Stata is able to deliver them, no matter how badly conditioned the covariance matrix is. In all cases I reproduce, as closely as possible, the data manipulation, equation specification and practical estimation that produced the results reported in published tables. I state all of this to forestall criticism that I have analysed inappropriate specifications and results. I reproduce and follow the coding and estimation methods that generate the results published in journals.

Fifth, in making randomization draws from the universe of potential treatments $\mathbf{\Omega}$ I restrict my draws to the subset $\mathbf{\Omega}$ that has the same treatment balance as $\mathbf{T}_E$, the experimental draw. This subtle distinction, irrelevant from the point of view of the exactness of the randomization test statistic, avoids my making unnecessary and potentially inaccurate inferences about the alternative balance of treatments that might have arisen. For example, a number of experiments applied treatment by taking random draws from a distribution (e.g. drawing a chit from a bag). Rather than trying to replicate the underlying distribution, I take the realized outcomes and randomly reallocate them across participants. I adopted this procedure after observing that the distribution of outcomes often does not follow the description of the underlying process given in the paper. A few papers note problems in implementation, and some authors, in correspondence, noted that even after they selected a particular randomized allocation of treatment, field agents did not always implement it accurately. I follow the papers in taking all of these errors in implementation as part of the random allocation of treatment. Under the randomization hypothesis, strongly maintained in every paper, treatment quantities, even if not in the proportions intended by the authors, could in principle have been applied to any participant. Thus, subject only to the stratification scheme, clarified by detailed examination of the data and helpful correspondence with the authors, I shuffle *realized* treatment outcomes across participants. This shuffling amounts to drawing the treatment vectors $\mathbf{T}_S$ in $\mathbf{\Omega}$ that share the same treatment balance as $\mathbf{T}_E$.[3]

Finally, I should note that I test instrumental variables regressions using the implied intent to treat regressions. In these regressions treatment variables are used as instruments, most of the time representing an opportunity that is offered to a participant that is then taken up or not. The null here cannot be that the treatment instrument has no effect on the instrumented variable, as this is obviously false (e.g. one can only take up an opportunity if one is offered the chance to do so). However, a reasonable null, and the relationship being tested in the second-stage regression,

---

[3] All of this is done, of course, in units of treatment, e.g. field villages or lab sessions. To keep the presentation familiar, here and in the paper I have described randomization tests as sampling from a population of potential outcomes. A more general presentation (e.g. Joseph P. Romano, 1989, "Bootstrap and Randomization Tests of Some Nonparametric Hypotheses," *The Annals of Statistics* 17 (1): 141-159) argues that under the null outcomes are invariant with respect to all transformations G that map from $\mathbf{\Omega}$ to $\mathbf{\Omega}$. The shuffling or rearranging of outcomes across participants is precisely such a mapping.

is that the instrumented variable has no effect on final outcomes of interest. Combined with the exogeneity assumption used to identify the regression, in an iv setting this implies that there exists no linear relationship between the outcome variable and the treatment variables themselves, i.e. no significant relation in the intent to treat regression. Consequently, I test the significance of instrumental variables regressions by running the implied intention to treat regression for the experiment and then comparing its coefficients and p-values to those produced through the randomization distribution under the null that final outcomes are invariant with respect to the actual realization of treatment.[4]

In the case of the jackknife, I calculate the covariance matrix using the formula:

$$\text{(a12)}\ \frac{N-1}{N}\sum_{\mathbf{i}}(\hat{\boldsymbol{\beta}}_{\sim\mathbf{i}}-\hat{\boldsymbol{\beta}})(\hat{\boldsymbol{\beta}}_{\sim\mathbf{i}}-\hat{\boldsymbol{\beta}})'$$

where $\hat{\boldsymbol{\beta}}_{\sim\mathbf{i}}$ denotes the vector of coefficients with the cluster group of observations **i** (or individual observation i if the paper does not cluster) deleted, and N represents the number of distinct coefficient vectors so estimated. I use this particular formula for the jackknife because, putting aside the standard (N-1)/N jackknife correction, it equals the hc3 adjustment of clustered/robust covariance matrices in the case of OLS regressions. As is customary, jackknife results are evaluated using the t and F distributions with N-1 degrees of freedom.

[4]In using the bootstrap, the jackknife, and reporting original authors' results, I continue to use the iv regression itself. To keep the number of iv and intent to treat coefficients equal across methods, I only examine exactly identified iv regressions (i.e. exclude a small number of overidentified two stage least squares). I should also note that many of the intent to treat regressions implied by iv regressions duplicate regressions found elsewhere in the paper. I drop these duplicates from the analysis. Sometimes authors present first-stage regressions along with iv results. I skip these if they involve a dependent variable that is never used as a treatment outcome elsewhere in the paper. In total, this leads me to drop 14 first stage regressions in three papers, which are all of form described above, where the dependent variable is trivially determined by treatment. On the other hand, I retain first stage regressions where the authors, having used the dependent variable as a treatment outcome elsewhere in the paper, now use it as an instrumented variable in determining some other treatment outcome.

III: Results for Bootstrap/Randomization -c and all Treatment Effects

The bootstrap and randomization results reported in the paper are based upon -t methods, as these are asymptotically superior in the case of the bootstrap and less sensitive to deviations away from the sharp null in the case of randomization inference, and are restricted to treatment effects reported by authors. This appendix presents expanded tables which include -c results and tests of all treatment effects, reported and unreported.

In Appendix Tables V, VI and VII below, results reported in Tables V, VI or VII of the paper are highlighted in bold, while additional results are reported in regular typeface. Bootstrap-c results consistently show higher rejection rates than the bootstrap-t, while randomization-c results vary around those of the randomization-t, with higher or lower rejection rates depending upon the test and level. Results for all treatment effects in regressions with reported treatment coefficients are very similar to those found in testing reported treatment effects alone in terms of rejection rates of alternative methods relative to conventional tests and the concentration of differences in high leverage papers, other tables than the first, and regressions with covariate interactions. Absolute rejection rates, for both conventional inference and alternative methods, are generally slightly higher when tests are expanded to include all treatment effects, as emphasized in the paper.

Appendix Table V: Statistical Significance of Individual Treatment Effects

	.01	.05	.01	.05	.01	.05	.01	.05
	all papers (53 papers)		low leverage (18 papers)		medium leverage (17 papers)		high leverage (18 papers)	
	based on 4044 reported treatment coefficients							
authors' p-value	**.216**	**.354**	**.199**	**.310**	**.164**	**.313**	**.283**	**.437**
randomization-t	**.78**	**.87**	**.96**	**.98**	**.79**	**.96**	**.65**	**.74**
bootstrap-t	**.79**	**.84**	**.99**	**.98**	**.87**	**.89**	**.60**	**.70**
jackknife	**.78**	**.83**	**.95**	**.89**	**.87**	**.91**	**.61**	**.73**
randomization-c	.75	.85	.93	.93	.79	.96	.60	.73
bootstrap-c	.88	.93	.96	.96	.92	.95	.79	.91
based on all 5740 treatment coefficients in regressions with reported treatment coefficients								
authors' p-value	.213	.345	.195	.306	.159	.301	.283	.425
randomization-t	.78	.87	.97	.98	.79	.95	.64	.73
bootstrap-t	.78	.84	.99	.98	.87	.89	.60	.70
jackknife	.78	.83	.95	.89	.86	.91	.61	.73
randomization-c	.75	.85	.94	.93	.81	.96	.59	.73
bootstrap-c	.87	.93	.97	.95	.91	.95	.79	.91
	first table (45 papers)		other tables (45 papers)		interactions (29 papers)		no interactions (29 papers)	
	based on 4044 reported treatment coefficients							
authors' p-value	**.303**	**.446**	**.188**	**.338**	**.148**	**.292**	**.310**	**.450**
randomization-t	**.82**	**.97**	**.81**	**.84**	**.76**	**.82**	**.87**	**.97**
bootstrap-t	**.85**	**.91**	**.90**	**.80**	**.86**	**.80**	**.87**	**.88**
jackknife	**.91**	**.94**	**.81**	**.79**	**.80**	**.83**	**.93**	**.89**
randomization-c	.79	.94	.84	.83	.76	.83	.87	.96
bootstrap-c	.96	.97	.90	.89	.90	.89	.92	.93
based on all 5740 treatment coefficients in regressions with reported treatment coefficients								
authors' p-value	.306	.448	.188	.329	.147	.293	.314	.438
randomization-t	.82	.98	.82	.84	.76	.82	.88	.97
bootstrap-t	.85	.91	.89	.80	.87	.81	.86	.90
jackknife	.91	.94	.80	.79	.81	.83	.90	.90
randomization-c	.79	.94	.85	.83	.76	.82	.88	.97
bootstrap-c	.96	.98	.90	.89	.90	.88	.90	.94

Notes: As in Table V in the paper.

Appendix Table VIa: Joint Statistical Significance of Treatment Effects (Regression Level)
(joint tests based on F and Wald statistics)

	all papers (47 papers)		low leverage (16 papers)		medium leverage (16 papers)		high leverage (15 papers)		first table (29 papers)		other tables (29 papers)	
	.01	.05	.01	.05	.01	.05	.01	.05	.01	.05	.01	.05
significant coef. (reported)	**.431**	**.643**	**.353**	**.596**	**.450**	**.607**	**.495**	**.731**	**.469**	**.620**	**.413**	**.584**
significant coef. (all)	.461	.682	.407	.655	.473	.665	.505	.730	.510	.643	.443	.610
reported treatment effects (922 regressions with > 1 reported treatment effect)												
authors' method	**.438**	**.546**	**.435**	**.508**	**.392**	**.539**	**.490**	**.595**	**.383**	**.528**	**.400**	**.473**
randomization-t	**.76**	**.83**	**1.01**	**1.00**	**.90**	**.94**	**.42**	**.58**	**.84**	**.92**	**.84**	**.86**
bootstrap-t	**.72**	**.81**	**.96**	**.96**	**.84**	**.91**	**.39**	**.57**	**.93**	**.84**	**.74**	**.81**
jackknife	**.90**	**.88**	**.98**	**.96**	**.97**	**.91**	**.76**	**.79**	**.98**	**.96**	**.90**	**.92**
randomization-c	.84	.88	.99	1.00	.87	.95	.66	.70	.96	.91	.90	.93
bootstrap-c	.94	.95	.95	.98	1.00	.93	.87	.95	1.01	1.01	.89	.96
all treatment effects (990 regressions with > 1 treatment effect)												
authors' method	.457	.574	.473	.561	.408	.555	.494	.607	.410	.567	.431	.506
randomization-t	.76	.82	.99	.99	.90	.93	.41	.55	.85	.87	.83	.86
bootstrap-t	.74	.82	1.00	.99	.82	.90	.39	.55	.96	.87	.77	.83
jackknife	.91	.88	.98	.96	.97	.93	.77	.76	.98	.90	.88	.90
randomization-c	.84	.88	.99	1.01	.89	.98	.65	.67	.95	.90	.90	.93
bootstrap-c	.95	.95	.97	.98	1.01	.95	.86	.91	1.02	.95	.85	.93

Notes: As in Table VI in the paper.

Appendix Table VIb: Joint Statistical Significance of Treatment Effects (Regression Level)
(presence of at least one significant effect in multiple testing using Bonferroni (B) and Westfall-Young (WY) methods)

	all papers (47 papers)		low leverage (16 papers)		medium leverage (16 papers)		high leverage (15 papers)		first table (29 papers)		other tables (29 papers)	
	.01	.05	.01	.05	.01	.05	.01	.05	.01	.05	.01	.05
significant coef. (reported)	**.431**	**.643**	**.353**	**.596**	**.450**	**.607**	**.495**	**.731**	**.469**	**.620**	**.413**	**.584**
significant coef. (all)	.461	.682	.407	.655	.473	.665	.505	.730	.510	.643	.443	.610
reported treatment effects (922 regressions with > 1 reported treatment effect)												
authors' p-value (B)	**.335**	**.494**	**.274**	**.426**	**.322**	**.526**	**.415**	**.533**	**.340**	**.501**	**.306**	**.442**
randomization-t (B)	**.73**	**.85**	**.99**	**1.02**	**.59**	**.88**	**.66**	**.67**	**.81**	**.96**	**.78**	**.82**
bootstrap-t (B)	**.76**	**.88**	**1.06**	**1.03**	**.80**	**.87**	**.51**	**.76**	**1.01**	**.90**	**.86**	**.85**
jackknife (B)	**.80**	**.85**	**1.03**	**.93**	**.78**	**.94**	**.64**	**.68**	**.98**	**.97**	**.85**	**.84**
randomization-c (B)	.74	.79	1.02	1.02	.76	.79	.52	.59	.87	1.00	.85	.89
bootstrap-c (B)	.82	.92	.91	.94	.84	.88	.75	.93	1.11	1.01	.81	.90
randomization-t (WY)	**.76**	**.89**	**1.00**	**1.08**	**.68**	**.92**	**.66**	**.70**	**.78**	**.96**	**.85**	**.89**
bootstrap-t (WY)	**.77**	**.92**	**1.07**	**1.03**	**.80**	**.93**	**.52**	**.81**	**1.01**	**.94**	**.87**	**.87**
randomization-c (WY)	.75	.81	1.06	1.03	.76	.81	.53	.61	.87	1.03	.88	.92
bootstrap-c (WY)	.86	.96	.93	.99	.86	.91	.82	.97	1.11	1.01	.85	.93
all treatment effects (990 regressions with > 1 treatment effect)												
authors' p-value (B)	.352	.512	.320	.465	.330	.534	.411	.537	.386	.534	.322	.456
randomization-t (B)	.74	.86	.97	1.03	.61	.90	.67	.66	.82	.98	.80	.84
bootstrap-t (B)	.78	.89	1.05	1.03	.83	.91	.52	.75	1.01	.92	.86	.84
jackknife (B)	.81	.86	1.01	.94	.80	.96	.65	.67	.97	.98	.84	.83
randomization-c (B)	.75	.80	.99	1.02	.80	.83	.52	.57	.86	1.01	.88	.90
bootstrap-c (B)	.84	.93	.92	.95	.87	.92	.75	.92	1.08	1.02	.80	.89
randomization-t (WY)	.77	.90	.98	1.08	.68	.93	.66	.69	.80	.98	.87	.90
bootstrap-t (WY)	.79	.93	1.06	1.03	.83	.97	.53	.79	1.01	.96	.87	.86
randomization-c (WY)	.77	.82	1.04	1.03	.81	.85	.52	.59	.86	1.03	.92	.92
bootstrap-c (WY)	.88	.97	.94	1.00	.88	.95	.83	.96	1.09	1.02	.84	.91

Notes: As in Table VI in the paper.

Appendix Table VIIa: Joint Statistical Significance of Treatment Effects (Table Level)
(joint tests based on Wald statistics)

	all papers (53 papers)		low leverage (18 papers)		medium leverage (17 papers)		high leverage (18 papers)		first table (45 papers)		other tables (45 papers)	
	.01	.05	.01	.05	.01	.05	.01	.05	.01	.05	.01	.05
significant coef. (reported)	**.662**	**.818**	**.617**	**.788**	**.602**	**.753**	**.764**	**.908**	**.711**	**.889**	**.630**	**.786**
significant coef. (all)	.680	.836	.658	.816	.613	.780	.764	.908	.733	.889	.650	.813
reported treatment effects (198 tables)												
conventional	**.493**	**.622**	**.337**	**.487**	**.431**	**.522**	**.706**	**.850**	**.422**	**.556**	**.483**	**.606**
randomization-t	**.51**	**.67**	**.92**	**1.00**	**.40**	**.74**	**.38**	**.45**	**.79**	**.80**	**.62**	**.78**
bootstrap-t	**.21**	**.33**	**.33**	**.51**	**.18**	**.41**	**.18**	**.19**	**.38**	**.45**	**.23**	**.40**
jackknife	**.77**	**.84**	**.92**	**.86**	**.76**	**.91**	**.71**	**.78**	**.89**	**.84**	**.81**	**.84**
randomization-c	.62	.70	.99	.90	.64	.77	.46	.55	.78	.79	.66	.71
bootstrap-c	.62	.68	.68	.58	.57	.72	.62	.73	.79	.64	.65	.67
all treatment effects (198 tables)												
conventional	.545	.654	.448	.571	.469	.534	.714	.850	.444	.533	.564	.676
randomization-t	.52	.67	.85	.95	.43	.75	.36	.44	.80	.88	.60	.74
bootstrap-t	.21	.33	.32	.49	.17	.40	.17	.19	.41	.52	.21	.36
jackknife	.78	.86	.91	.90	.72	.92	.75	.80	.90	.92	.80	.84
randomization-c	.62	.70	.86	.87	.66	.78	.45	.54	.79	.83	.61	.68
bootstrap-c	.58	.67	.63	.59	.48	.68	.61	.73	.80	.71	.57	.62

Notes: As in Table VII in the paper.

Appendix Table VIIb: Joint Statistical Significance of Treatment Effects (Table Level)
(presence of at least one significant effect in multiple testing using Bonferroni (B) and Westfall-Young (WY) methods)

	all papers (53 papers)		low leverage (18 papers)		medium leverage (17 papers)		high leverage (18 papers)		first table (45 papers)		other tables (45 papers)	
	.01	.05	.01	.05	.01	.05	.01	.05	.01	.05	.01	.05
significant coef. (reported)	**.662**	**.818**	**.617**	**.788**	**.602**	**.753**	**.764**	**.908**	**.711**	**.889**	**.630**	**.786**
significant coef. (all)	.680	.836	.658	.816	.613	.780	.764	.908	.733	.889	.650	.813
reported treatment effects (198 tables)												
authors' p-value (B)	**.377**	**.542**	**.329**	**.489**	**.275**	**.475**	**.521**	**.659**	**.400**	**.556**	**.349**	**.491**
randomization-t (B)	**.61**	**.81**	**.88**	**.98**	**.63**	**.75**	**.43**	**.72**	**.78**	**1.00**	**.69**	**.84**
bootstrap-t (B)	**.69**	**.78**	**1.00**	**1.06**	**.62**	**.73**	**.54**	**.62**	**.78**	**.92**	**.79**	**.82**
jackknife (B)	**.71**	**.85**	**1.00**	**.97**	**.66**	**.87**	**.56**	**.73**	**.89**	**1.00**	**.74**	**.84**
randomization-c (B)	.67	.79	1.02	1.01	.73	.89	.41	.55	.94	1.00	.64	.79
bootstrap-c (B)	.96	.90	1.00	1.01	.94	.77	.94	.91	1.06	1.08	.82	.91
randomization-t (WY)	**.77**	**.91**	**1.18**	**1.06**	**.67**	**.96**	**.55**	**.77**	**1.00**	**1.12**	**.79**	**.92**
bootstrap-t (WY)	**.79**	**.87**	**1.20**	**1.09**	**.73**	**.91**	**.55**	**.68**	**.89**	**1.00**	**.89**	**.95**
randomization-c (WY)	.75	.89	1.06	1.09	.88	.97	.48	.70	1.00	1.16	.74	.92
bootstrap-c (WY)	1.04	1.01	1.13	1.03	1.11	1.02	.94	.99	1.17	1.12	.90	1.02
all treatment effects (198 tables)												
authors' p-value (B)	.406	.553	.398	.545	.290	.463	.521	.645	.422	.578	.386	.500
randomization-t (B)	.64	.83	.94	.98	.60	.77	.43	.74	.74	1.00	.74	.87
bootstrap-t (B)	.72	.81	1.03	1.03	.63	.81	.54	.63	.79	.92	.82	.86
jackknife (B)	.73	.87	.97	.97	.71	.92	.56	.75	.89	1.00	.76	.88
randomization-c (B)	.71	.83	1.05	1.03	.79	.98	.41	.55	.95	.96	.72	.87
bootstrap-c (B)	.98	.92	1.03	1.01	.97	.82	.94	.91	1.05	1.04	.86	.95
randomization-t (WY)	.79	.92	1.19	1.06	.69	.96	.54	.78	1.00	1.12	.82	.94
bootstrap-t (WY)	.80	.89	1.17	1.05	.74	.99	.55	.69	.89	1.00	.90	.99
randomization-c (WY)	.78	.92	1.09	1.08	.94	1.03	.47	.71	1.00	1.12	.79	.97
bootstrap-c (WY)	1.04	1.02	1.14	1.03	1.08	1.07	.94	.97	1.11	1.12	.94	1.04

Notes: As in Table VII in the paper.

Table A1: Clustering at Authors' Level vs Clustering at Treatment Level
(significance rates in tests of individual treatment effects)

	impact on 12 papers consistently clustering below treatment level				impact on 53 paper average results			
analysis at:	authors' level		treatment level		authors' level		treatment level	
	.01	.05	.01	.05	.01	.05	.01	.05
authors' level	.205	.323			**.216**	**.354**		
treatment level			.220	.359			.221	.364
randomization-t	.97	.95	.37	.64	**.78**	**.87**	.65	.80
bootstrap-t	.90	.91	.32	.59	**.79**	**.84**	.66	.77
jackknife	.92	.86	.44	.62	**.78**	**.83**	.68	.78
randomization-c	.92	.90	.26	.60	.75	.85	.60	.78
bootstrap-c	.91	.89	.91	.87	.88	.93	.88	.93

Notes: Numbers in bold are those reported in Table V of the paper; top row reports average across 12 or 53 papers of the within paper fraction of significant results evaluated using authors' methods (clustering at authors' level or treatment level); values in lower rows are average fraction of significant results evaluated using indicated method divided by the top row. Bootstrap and randomization inference calculated using 10000 samples, with standard errors calculated using authors' methods in the case of -t results.

IV: Impact of Alternative Methods on Summary Results

In this appendix I report the impact on summary results of three deviations from the methods described above and in the paper: (1) clustering at treatment level rather than authors' level; (2) bootstrapping each equation individually, restricting the bootstrap sample to the observations used in the estimating equation alone (rather than bootstrap sampling the entire experimental sample); (3) calculating conditional randomization p-values where possible, so that the p-value of each individual coefficient does not depend upon the null for other treatment coefficients.

As noted in Section II of the paper, in the 12 papers where authors systematically cluster below treatment level (or do not cluster at all) I defer to their decision and re-randomize, bootstrap sample and jackknife at their clustering level, acting as if the treatment units (e.g. sessions or geographical units) are merely nominal. Table A1 examines the impact of clustering the conventional p-value at treatment level and randomizing and bootstrap sampling at the treatment level as well. As usual, the top line of the table reports the average across papers of the fraction of treatment effects that are conventionally significant (to three decimal places), while lower rows report the average across papers of the fraction of effects that are found to be

significant using randomization, bootstrap and jackknife methods divided by the top row (and reported with two decimal places for contrast). Clustering at treatment level raises slightly the fraction of reported treatment effects that are .01 and .05 conventionally significant, but results in a substantial reduction in the relative fraction of significant results found using alternative methods. The alternative methods produce significant coefficients at a rate of around .9 of authors' methods when sampling at the level specified by authors, but (with the exception of the bootstrap-c) yield only .3 to .6 as many significant results when sampling at the treatment level. The right-hand panel of the table shows that these changes would have a large effect on the summary results reported in the paper. While Table V of the paper reports that randomization-t methods on average produce .78 and .87 as many significant results as authors' methods at the .01 and .05 levels, respectively, if treatment level clustering had been imposed on the 12 papers that did not do so, the relative number of significant randomization results in the 53 paper sample would have fallen to .65 and .80 at the two levels.

To keep bootstrap methods similar to those used in my baseline randomization analysis (where I re-randomize treatment across the entire experimental sample to simulate the potential distribution of outcomes), in the results reported in the paper I bootstrap by re-sampling the entire experimental sample. This also facilitates the calculation of the joint distribution of coefficients across equations. However, the bootstrap is usually implemented by only resampling the observations in the estimating equation itself. This may differ from the total experimental sample because of missing data for individual observations or because the regression itself is explicitly restricted to a sub-sample of the experiment based upon conditionals (e.g. covariate values). Table A2 below compares bootstrap coefficient rejection rates found when resampling the entire experimental sample with those found when restricting the resampling to the observations/ clusters present in each individual equation. As shown, bootstrapping at the equation level generally results in slightly lower relative rejection rates, particularly in the case of the bootstrap-t (the results of which are reported in the paper's tables). Since coefficient estimates are not affected by deleting observations that are not in the regression sample, it makes no difference whether I implement the jackknife using the entire sample or the regression sample alone.

Table A2: Relative Coefficient Rejection Rates Found
Using Alternative Forms of Bootstrap Resampling

	resampling the entire experiment				resampling regression specific observations			
	reported coefficients		all coefficients		reported coefficients		all coefficients	
	.01	.05	.01	.05	.01	.05	.01	.05
authors' p-value	**.216**	**.354**	.213	.345	.216	.354	.213	.345
bootstrap-t	**.79**	**.84**	.78	.84	.78	.83	.78	.83
bootstrap-c	.88	.93	.87	.93	.90	.91	.89	.91

Notes: As in Table A1.

As noted in the paper, randomization p-values for individual treatment effects in equations with multiple treatment variables in general depend upon the null assumed for the effects of other treatment measures, as these effects are accounted for in the adjustments to the dependent variable following each re-randomization draw $\mathbf{T}_S$ from $\mathbf{\Omega}$. However, as also noted, it is possible to calculate "conditional" p-values in some cases, where one can consider the universe of potential reallocations of a given treatment measure conditional on holding constant the allocations of other treatment measures. This can, for example, be implemented in cases where there are multiple treatment regimes and no interactions of these regimes with non-treatment covariates. In implementing this procedure, I fully condition, restricting the re-randomization process to the observations in the estimating equation alone, so as to produce results that are as close as possible in spirit to conventional p-values. Thus, the thought experiment for the re-randomization is: "for the participant observations that ended up in this estimating equation, what potential reallocations of treatment variable x, holding constant other treatments, might have occurred given the method in which treatment was allocated (e.g. strata)."

I am able to calculate conditional p-values of this sort for 1294 of the 3254 reported treatment effects in equations with more than one treatment variable and 2235 of the 4950 total treatment effects (including unreported) in equations with more than one treatment variable. However, in some cases the condition that the allocation of other treatment variables be kept constant restricts the number of potential outcomes so much that there are a large number of "ties", generating unusually large p-values when a random number allocates the ties, as in equation (5) in the paper. To avoid this, I drop all cases where the ties account for more than .01

Table A3: Rejection Rates for Individual Treatment Effects Found
Using Alternative Forms of Randomization Inference
(average rejection rates in multi-treatment equations in 25 papers)

	unconditional joint 0 null				conditional individual 0 nulls			
	reported coefficients		all coefficients		reported coefficients		all coefficients	
	.01	.05	.01	.05	.01	.05	.01	.05
authors' p-value	.273	.384	.270	.372	.273	.384	.270	.372
randomization-t	.90	.96	.90	.97	.83	.95	.83	.97
randomization-c	.79	.96	.81	.97	.90	.95	.92	.98

Notes: As in Table A1.

of total randomization draws. This leaves me with 1023 conditional p-values on reported treatment effects and 1849 on all treatment effects in regressions from 25 papers. As shown in Table A3, average relative rejection rates in individual tests at the .05 level are almost identical for conditional p-values for individual nulls as they are for unconditional individual p-values based upon joint nulls. At the .01 level, conditional results produce lower rejection rates in the randomization-t and higher rejection rates in the randomization-c. Since the paper itself reports -t results, use of conditional p-values would make reported results in the paper somewhat less favourable to authors' results.

As noted in the paper, use of alternative randomization inference schemes has a very substantial effect on the change in p-values when conventionally significant results are found to be insignificant. Table A4 below reproduces Table VIII's analysis in the paper of the distribution of randomization p-values for reported results that are found to be significant using authors' methods. As reported in the paper, when an individual treatment effect is statistically significant at the .01 level using authors' methods, only .019 of randomization-t p-values are above .10. The randomization-c finds greater differences, with .073 of randomization-c p-values lying above .10. In the 12 papers where authors systematically clustered below treatment level, when a .01 conventionally significant result is found, the randomization-t p-value is never above .05. However, when both the conventional and randomization test are based upon clustering at treatment level, in the average paper the corresponding randomization-t p-value is greater than .10 almost one-fifth of the time. Turning to the treatment effects in 25 papers where I was able to calculate conditional randomization p-values that do not depend upon the nulls for other

Table A4: Distribution of Randomization P-Values for Individual Treatment Effects that are Conventionally Significant using Alternative Forms of Randomization Inference

	all 53 papers				12 papers				25 papers			
	rand-t		rand-c		authors' clustering		treatment clustering		unconditional p-value		conditional p-value	
	.01	.05	.01	.05	.01	.05	.01	.05	.01	.05	.01	.05
< .01	**.752**	↓	.701	↓	.805	↓	.322	↓	.881	↓	.797	↓
.01 - .05	**.160**	**.853**	.191	.804	.195	.927	.433	.648	.111	.970	.140	.915
.05 - .10	**.068**	**.101**	.034	.101	.000	.018	.058	.134	.003	.020	.015	.027
.10 - .20	**.014**	**.029**	.027	.038	.000	.054	.135	.114	.005	.004	.026	.027
> .20	**.005**	**.017**	.046	.057	.000	.000	.051	.104	.000	.006	.021	.031

Notes: Reported figures are the average across papers of the within paper distribution of randomization p-values when a coefficient is significant at the level specified using authors' methods; (↓) included in the category below; numbers in bold are those reported in Table VIII of the paper; distributions for comparison of authors' vs. treatment clustering and conditional vs. unconditional nulls are based on the randomization-t.

treatment variables in multi-treatment equations, when a conventionally significant result is found the randomization-t based upon the full randomization distribution and assuming a null of zero for all treatment effects finds p-values greater than .10 on average only .005 of the time. In contrast, the conditional randomization p-value that only rerandomizes the treatment associated with a given coefficient, and hence does not depend upon the nulls for other treatment effects, finds a p-value greater than .10 on average .047 of the time. In sum, the results presented in the paper are for those randomization methods which yield the smallest average differences between randomization and conventional rejection rates and the smallest difference between randomization and conventional p-values when a conventional result is statistically significant.

V: Relative Power of Randomization-t and Conventional Robust Methods

This appendix reports the relative power of randomization-t and conventional robust inference in Monte Carlos. I use the data generating processes of the upper panel of Table III in the paper, with observation specific treatment effects which are either fixed (i.e. the same for all participants) or distributed as standard normal or chi^2 variables. I vary the mean of these effects by adding a constant to their distribution and continue to test the null that the mean effect is zero. As in Table III, I conduct 10000 Monte Carlos for each of the three data generating processes for each of three sample sizes (N = 20, 200 and 2000). The results are reported in Table A5 below. The table begins by reporting relative randomization-t to robust inference size and the absolute level of randomization-t size, so as to allow the reader to recall when the tests have correct nominal size and when they have positive size distortions. In the lower panels, I then vary the mean of the data generating process to generate power (i.e. rejection rates) of .10, .25, .50, .75, and .90 in the conventional test of the (false) zero null at the .05 level. The relative power of the randomization-t to that of the conventional test using the robust covariance estimate is reported, as well as the ratio of randomization-t power to size to that found using the conventional test. As was reported in the paper, when both tests have size near nominal value, i.e. in balanced regression designs or large samples, there relative power is identical. In small samples with unbalanced regression design, the power of randomization inference falls below that of the conventional test (which has large size distortions in these cases), but the ratio of power to size is found to be greater (and often much greater) using randomization inference.

Table A5: Relative Power and Size of Randomization-t and Robust Inference

	balanced design			unbalanced design			balanced design			unbalanced design		
	fixed	normal	chi2	fixed	normal	chi2	fixed	normal	chi2	fixed	normal	chi2
	relative randomization-t/robust size						randomization-t size					
20	1.01	.99	1.03	.19	.31	.31	.048	.052	.062	.046	.089	.091
200	1.00	.99	1.01	.76	.80	.79	.048	.052	.055	.051	.051	.065
2000	1.01	.98	1.00	.97	1.00	.97	.049	.048	.045	.052	.052	.052
	relative randomization-t/robust power						relative randomization-t/robust power/size					
	when conventional test has power = .10 at .05 level											
20	1.00	1.00	1.12	NA	NA	NA	.99	1.00	1.09	NA	NA	NA
200	.99	1.01	1.00	.82	.81	.78	.99	1.01	.99	1.07	1.01	.99
2000	1.01	1.00	1.01	.99	.99	.97	1.00	1.02	1.00	1.02	.99	1.00
	when conventional test has power = .25 at .05 level											
20	.99	.99	1.07	.20	NA	NA	.98	1.00	1.04	1.07	NA	NA
200	1.01	1.00	1.01	.86	.84	.81	1.00	1.01	1.00	1.13	1.06	1.03
2000	1.00	1.00	.99	.99	.98	.98	.99	1.02	.99	1.02	.98	1.00
	when conventional test has power = .50 at .05 level											
20	1.00	1.00	1.05	.39	.46	.44	.98	1.00	1.02	2.09	1.48	1.40
200	1.00	1.00	1.00	.90	.91	.88	.99	1.01	.99	1.18	1.14	1.12
2000	1.00	1.00	1.00	.99	.98	.99	.99	1.02	1.00	1.02	.98	1.02
	when conventional test has power = .75 at .05 level											
20	1.00	1.00	1.03	.54	.62	.57	.99	1.00	1.00	2.87	1.99	1.82
200	1.00	1.00	1.00	.94	.95	.94	.99	1.00	.99	1.23	1.19	1.19
2000	1.00	1.00	1.00	1.00	.99	.99	.99	1.01	1.00	1.03	.99	1.02
	when conventional test has power = .90 at .05 level											
20	1.00	1.00	1.01	.69	.77	.72	.98	1.01	.99	3.62	2.46	2.31
200	1.00	1.00	1.00	.97	.97	.97	1.00	1.01	.99	1.27	1.22	1.23
2000	1.00	1.00	1.00	1.00	1.00	1.00	.99	1.02	1.00	1.03	.99	1.02

Notes: NA – not applicable, there is no null for which power of robust inference equals the indicated level, because size already exceeds this value. Reported figures are based upon 10000 Monte Carlo simulations using the data generating processes described in the upper panel of Table III of the paper.

VI: Determinants of Disagreement between Authors' Results and those of the Randomization-t

Tables A6 and A7 analyze the determinants of disagreement between authors' and randomization-t results in the tests of individual treatment effects reported in Section V of the paper. The dependent variable is a 0/1 indicator for disagreement in statistical significance when a treatment effect is significant at the .01 or .05 level using authors' methods (1 = disagreement, randomization-t not significant at that level). The right hand side variables are the maximum cluster or observation leverage, the number of clusters or observations, indicators for a first table or a regression with covariate interactions with treatment, and paper fixed effects. In columns (1) and (5) each cell represents a different regression, with one right hand side variable entered at a time (plus, in the right hand panel, paper fixed effects). Each subsequent column refers to an individual regression, with maximum leverage entered alongside the other variable with reported results (with and without paper fixed effects). The sample in Table A6 is restricted to conventionally significant reported treatment effects, while Table A7 expands the sample to include conventionally significant unreported treatment effects. Standard errors are clustered at the paper level.

Two patterns are apparent in the tables. First, in most specifications maximal cluster/observation leverage is statistically significant, while the coefficients for other right hand side variables generally become insignificant once maximum cluster/observation leverage is included in the regression. Second, the coefficients on other right hand side variables are moved toward zero by the inclusion of maximal leverage in the regression, to the degree that, in one specification or another, they are shrunken by at least 50 percent. In contrast, the coefficients on maximal leverage are hardly changed by the inclusion of the number of observations or dummies for first tables or interactions in the regression. Given the small number of papers in the sample, the statistical significance of maximal leverage should not be taken too seriously. Nevertheless, the two patterns are strongly suggestive that the number of observations and indicators for first tables and covariate interactions, insofar as they explain differences between authors' and randomization-t results, probably do so because of their association with regression design.

Table A6: Determinants of Differences Between Conventional and Randomization-t Significance in Tests of Individual Treatment Effects
(reported treatment effects)

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
	without paper fixed effects				with paper fixed effects			
indicator for disagreement when conventional results are .01 significant (N = 672)								
maximal cl/obs leverage	2.09* (.495)	2.05* (.500)	2.09* (.503)	1.98* (.490)	1.67** (.701)	1.67** (.701)	1.62** (.714)	1.50** (.660)
# of cl/observations	$-6.0e^{-6}$* ($1.7e^{-6}$)	$-2.7e^{-6}$ ($1.5e^{-6}$)			$-2.9e^{-6}$ ($3.7e^{-6}$)	$-2.5e^{-6}$ ($3.4e^{-6}$)		
first table	-.065 (.059)		-.070 (.048)		-.058 (.056)		-.032 (.055)	
covariate interactions	.146 (.093)			.091 (.081)	.167** (.066)			.144** (.069)
indicator for disagreement when conventional results are .05 significant (N = 1234)								
maximal cl/obs leverage	1.66* (.372)	1.65* (.374)	1.63* (.361)	1.66* (.391)	1.55* (.302)	1.55* (.301)	1.53* (.307)	1.47* (.299)
# of cl/observations	$-3.6e^{-6}$** ($1.4e^{-6}$)	$-9.4e^{-7}$ ($1.4e^{-6}$)			$2.4e^{-6}$ ($2.8e^{-6}$)	$2.8e^{-6}$ ($2.9e^{-6}$)		
first table	-.084** (.036)		-.065 (.034)		-.044 (.030)		-.014 (.030)	
covariate interactions	.040 (.062)			-.003 (.055)	.091* (.031)			.063 (.034)

Notes: N = number of observations; cl/obs = cluster or observation; e^{-x} = times 10^{-x}; *, ** = .01 or .05 significant, respectively; standard errors clustered at the paper level. In columns (1) and (5) each regressor is entered individually (i.e. each cell represents a separate regression); other columns report individual regressions with all variables with reported results entered simultaneously.

Table A7: Determinants of Differences Between Conventional and Randomization-t Significance in Tests of Individual Treatment Effects
(reported & unreported treatment effects)

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
	without paper fixed effects				with paper fixed effects			
indicator for disagreement when conventional results are .01 significant (N = 773)								
maximal cl/obs leverage	1.48**	1.44**	1.48**	1.34**	.644	.643	.617	.556
	(.632)	(.632)	(.642)	(.633)	(.712)	(.712)	(.723)	(.666)
# of cl/observations	-6.8e-6**	-4.1e-6**			-3.0e-6	-2.8e-6		
	(2.6e-6)	(2.0e-6)			(3.8e-6)	(3.6e-6)		
first table	-.050		-.050		-.044		-.033	
	(.055)		(.045)		(.048)		(.050)	
covariate interactions	.167			.110	.154*			.143**
	(.086)			(.078)	(.056)			(.060)
indicator for disagreement when conventional results are .05 significant (N = 1457)								
maximal cl/obs leverage	1.39*	1.38*	1.38*	1.37*	.951**	.953**	.926**	.895**
	(.373)	(.374)	(.370)	(.391)	(.380)	(.381)	(.381)	(.359)
# of cl/observations	-4.1e-6**	-1.6e-6			2.3e-6	2.6e-6		
	(1.8e-6)	(1.4e-6)			(2.7e-6)	(2.8e-6)		
first table	-.069**		-.056		-.045		-.027	
	(.031)		(.029)		(.026)		(.028)	
covariate interactions	.062			.016	.091*			.074**
	(.056)			(.050)	(.027)			(.029)

Notes: as in Table A6.

VII: Papers in the Experimental Sample

The following are the papers in the experimental sample. The acronym at the beginning of each reference is the code used to identify the paper in the public use do-files.

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

(AKL) 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.

(ABHOT) 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.

(ALO) 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.

(AL) 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.

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

(ABS) 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.

(BBLP) 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

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

(BL) 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.

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

(CILS) 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.

(CC1) Camera, Gabriele and Marco Casari. 2014. "The Coordination Value of Monetary Exchange: Experimental Evidence." *American Economic Journal: Microeconomics* 6 (1): 290–314.

(CMS) 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.

(CC2) Chen, Roy and Yan Chen. 2011. "The Potential of Social Identity for Equilibrium Selection." *American Economic Review* 101 (6): 2562–2589.

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

(CHKL) 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.

(CGTTTV) 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.

(DDK) 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.

(DKR) 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.

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

(D) 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.

(DR) 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.

(DR2) 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.

(ER) 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.

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

(FG) 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.

(FJP) 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.

(FPPR) 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.

(FL) 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.

(GJKM) Giné, Xavier, Pamela Jakiela, Dean Karlan, and Jonathan Morduch. 2010. "Microfinance Games." *American Economic Journal: Applied Economics* 2 (3): 60-95.

(GRS) 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.

(GKB) 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.

(GMR) 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.

(GGY) 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.

(GKN) 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.

(HS) 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.

(IZ) 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.

(KL) 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.

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

(KMP) 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.

(LLLPR) 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.

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

(LMW) 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.

(MSV) 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.

(MMW) 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.

(MMW2) 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.

(OT) 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.

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

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

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

(VDR) 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.

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