Standard operating procedures for Don Green's lab at Columbia

Winston Lin, Donald P. Green, and Alexander Coppock

Version 1.04: Dec. 22, 2015

Contents

This standard operating procedures (SOP) document describes the default practices of the experimental research group led by Donald P. Green at Columbia University. These defaults apply to analytic decisions that have not been made explicit in pre-analysis plans (PAPs). They are not meant to override decisions that are laid out in PAPs. For more discussion of the motivations for SOP, see Lin and Green (2015).

This is a living document. To suggest changes or additions, please feel free to e-mail us[1] or submit an issue on GitHub. Also, when referencing this document, please be sure to note the version and date.

Many thanks to Peter Aronow, Susanne Baltes, Jake Bowers, Al Fang, Macartan Humphreys, Berk Ozler, Anselm Rink, Michael Schwam-Baird, Uri Simonsohn, Amber Spry, Dane Thorley, Anna Wilke, and Jose Zubizarreta for helpful comments and discussions.

Preliminary checks on data preparation and study implementation

After collection of the raw data, all data processing steps leading up to creation of the files for analysis will be performed by computer code. We will make these computer programs publicly available.

The following types of checks should be performed and documented before analyses are publicly disseminated or submitted for publication:[2]

- Verifying that treatment occurred. Supportive evidence can include receipts for expenses associated with treatment, spot checks by more than one observer in the field, geotagged photographs, and/or manipulation checks (statistical evidence that random assignment created a treatment contrast in the intended direction—see, e.g., Gerber and Green (2012), Box 10.3, p. 335).
- Verifying that outcome data were gathered (e.g., via receipts, spot checks, and/or having a second team member independently gather data for a sample of cases).
- Verifying that there were at least two team members involved in all raw data collection efforts (including, but not limited to, the gathering of data from survey organizations and government agencies).
- Verifying that the computer programs for data processing correctly implement the intended logic and that, when applied to the raw data, these programs reproduce the data used in the analysis.
- Examining the distributions and missingness rates of all variables used in the analysis and their precursors in the raw data files. These variables should be checked both for plausibility and for consistency across the stages of data processing. The checks should normally include not only univariate summaries, but also cross-tabulations of related variables.
- Other checks described under "Covariate imbalance and the detection of administrative errors" and "Verifying that randomization was implemented as planned".

Any unresolved data preparation or processing issues will be disclosed in a research report or supplementary materials.

Hypothesis testing

One-tailed or two-tailed test?

We will report two-tailed hypothesis tests unless the PAP specifies a one-tailed test or some other approach.[3]

[1] winston.lin@columbia.edu (Lin), dpg2110@columbia.edu (Green), and a.coppock@columbia.edu (Coppock).

[2] If any of these checks require access to confidential data, we will obtain Institutional Review Board approval for the team members performing the checks.

[3] Olken (2015) (p. 70) discusses one other approach: "An interesting hybrid alternative would be to pre-specify asymmetric tests: for example, to reject the null if the result was in the bottom 1 percent of the distribution or in the top 4 percent, or the

Use of permutation methods

For significance tests and p-values, we will either (1) report Studentized permutation tests or (2) use permutation methods to check the accuracy of asymptotic approximations.

For an example of (1), see below. The Studentized permutation t-test (Janssen (1997)) compares a heteroskedasticity-robust t-statistic (instead of the estimated average treatment effect itself) with its empirical distribution under random reassignments of treatment.[4] Unless otherwise specified in the PAP, our Studentized permutation t-tests will use the t-statistic based on the HC0 or HC1 robust standard error estimator,[5] and p-values for two-tailed tests will be computed according to the following convention: "In general, if you want a two-sided P-value, compute both one-sided P-values, double the smaller one, and take the minimum of this value and 1" (Rosenbaum (2010), p. 33).

For an example of (2), see Lin (2013, 309–13), where a simulation permuting the treatment indicator is used to check the validity of confidence intervals based on robust standard errors.

For all permutation methods, we will use at least 10,000 randomizations and the random number seed 1234567.

Example: Studentized Permutation Test

```
suppressMessages({
  library(randomizr)
  library(sandwich)
  library(lmtest)
})
```

```
## Warning: package 'randomizr' was built under R version 3.1.3
```

```
set.seed(1234567)

N <- 200

# Create potential outcomes
Y0 <- rnorm(N)
Y1 <- Y0 + 0.25

# Conduct random assignment
Z <- complete_ra(N, m = 50)
Y_obs <- Y1*Z + Y0*(1-Z)

# Conduct Estimation
fit <- lm(Y_obs ~ Z)

# Obtain Heteroskedasticity-robust t-statistic
```

bottom 0.5 and the top 4.5 percent, and so on. These asymmetric tests would gain much of the statistical power from one-sided tests, but still be set up statistically to reject the null in the presence of very large negative results." See also Tukey (1993) (p. 276).

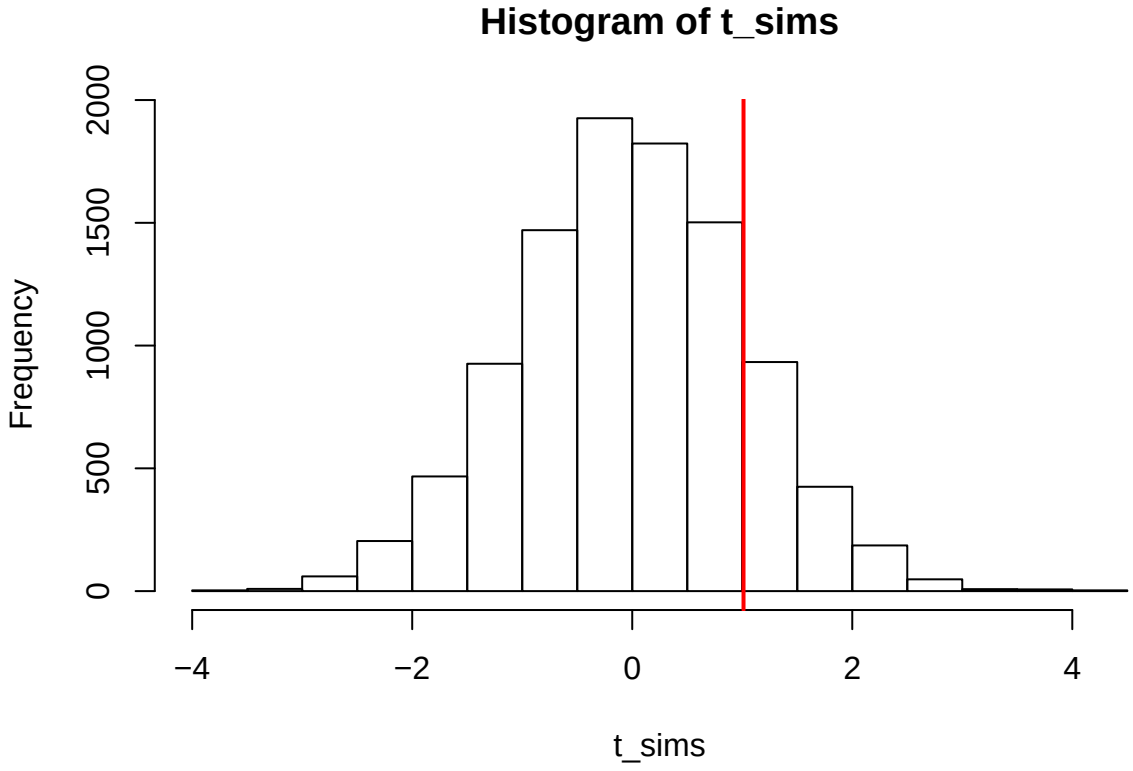
[4] See also Romano (2009) and Chung and Romano (2013).

[5] Since HC1 is just HC0 multiplied by a constant, the choice between these two estimators has no effect on the permutation test. MacKinnon (2013) (pp. 456-458) found in simulations that the power of Studentized bootstrap tests "decreases monotonically from HC1 to HC2, HC3, and finally HC4. ... The best HCCME [heteroskedasticity-consistent covariance matrix estimator] for asymptotic inference may not be the best one for bootstrap inference." We conjecture that, similarly, Studentized permutation tests tend to have more power with HC0 or HC1 than with HC2, HC3, or HC4.

```r
t_obs <- coeftest(fit, vcov = vcovHC(fit, type = 'HC0'))['Z','t value']

# Conduct simulation
sims <- 10000
t_sims <- rep(NA, sims)

for(i in 1:sims){
  Z_sim <- complete_ra(N, m = 50)
  fit_sim <- lm(Y_obs ~ Z_sim)
  t_sims[i] <- coeftest(fit_sim, vcov = vcovHC(fit_sim, type = 'HC0'))['Z_sim','t value']
}

# "In general, if you want a two-sided P-value, compute both one-sided P-values,
#  double the smaller one, and take the minimum of this value and 1."
#    Rosenbaum (2010), Design of Observational Studies, p. 33, note 2
# (Other options exist, but this is our default.)

p.left  <- mean(t_sims <= t_obs)
p.right <- mean(t_sims >= t_obs)

p <- min(2 * min(p.left, p.right), 1)
p
```

```
## [1] 0.3172
```

```r
hist(t_sims)
abline(v = t_obs, lwd=2, col="red")
```



Using covariates in analysis

Default methods for estimating average treatment effects

Estimation methods for the primary analysis will normally have been specified in the PAP. For reference in what follows, here we describe our default methods for a unit-randomized experiment with N subjects. Let $M < N$ denote the largest integer such that at least M subjects are assigned to each arm.

- If $M \geq 20$, we use least squares regression of Y on T, X, and T * X, where Y is the outcome, T is the treatment indicator, and X is a set of one or more mean-centered covariates (see "Choice of covariates" below for guidelines on the choice and number of covariates). The coefficient on T estimates the average effect of assignment to treatment. See Lin (2012a) for an informal description of this estimator.
- If $M < 20 \leq N$, we use least squares regression of Y on T and X.
- If $N < 20$, we use either difference-in-differences or difference-in-means. (Section 4.1 in Gerber and Green discusses the efficiency comparison between these two estimators. Again, the choice will typically be specified in the PAP.)

Example: Default Covariate Adjustment Procedure

```
suppressMessages({
  library(randomizr)
  library(sandwich)
  library(lmtest)
})

N <- 200

# Make some covariates
X1 <- rnorm(N)
X2 <- rbinom(N, size = 1, prob = 0.5)

# Make some potential outcomes
Y0 <- .6*X1 + 3*X2 + rnorm(N)
Y1 <- Y0 + .4

# Conduct a random assignment and reveal outcomes
Z <- complete_ra(N, m= 100)
Y_obs <- Y1*Z + Y0*(1-Z)

# Mean-center the covariates
X1_c <- X1 - mean(X1)
X2_c <- X2 - mean(X2)

# Conduct Estimation
fit_adj <- lm(Y_obs ~ Z + Z*(X1_c + X2_c))

# Robust Standard Errors
coeftest(fit_adj, vcov = vcovHC(fit_adj, type = "HC2"))
```

```
##
```

```
## t test of coefficients:
##
##              Estimate Std. Error t value  Pr(>|t|)
## (Intercept)  1.548750   0.097417 15.8981 < 2.2e-16 ***
## Z            0.158787   0.141514  1.1221   0.26322
## X1_c         0.773147   0.118270  6.5371  5.41e-10 ***
## X2_c         2.813019   0.198191 14.1935 < 2.2e-16 ***
## Z:X1_c      -0.286591   0.146341 -1.9584   0.05162 .
## Z:X2_c       0.210276   0.282875  0.7434   0.45817
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
# Compare to unadjusted model
fit_unadj <- lm(Y_obs ~ Z)
coeftest(fit_unadj, vcov = vcovHC(fit_unadj, type = "HC2"))
```

```
##
## t test of coefficients:
##
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  1.665825   0.177443   9.388   <2e-16 ***
## Z           -0.069557   0.258543  -0.269   0.7882
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Choice of covariates for regression adjustment

Ordinarily our choice of covariates for adjustment will have been specified in the PAP. For voter turnout experiments, the SOP section "Issues specific to voter turnout experiments" gives a default set of covariates in case the PAP fails to specify the choice.

With M and N as defined above, we will include no more than $M/20$ covariates in regressions with treatment-covariate interactions, and no more than $N/20$ covariates in regressions without such interactions.[6]

If PAP has failed to specify the choice of covariates, if the experiment is not a voter turnout study, and if the number of available baseline covariates (excluding higher powers, other transformations, and interactions between covariates) is 10 or fewer and does not exceed the limits above, we will include all the covariates in our regressions.

In general, covariates should be measured before randomization. To make any exceptions to this rule, we need to have a convincing argument that either (1) the variable is a measure of pre-randomization conditions, and treatment assignment had no effect on measurement error, or (2) although the variable is wholly or partly a measure of post-randomization conditions, it could not have been affected by treatment assignment. (Rainfall on Election Day would probably satisfy #2.)

Occasionally a new source of data on baseline characteristics becomes available after random assignment (e.g., when political campaigns join forces and merge their datasets). To decide which (if any) variables derived from the new data source should be included as covariates, we will consult a "blind jury" of collaborators or colleagues. The jury should not see treatment effect estimates or any information that might suggest whether inclusion of a covariate would make the estimated effects bigger or smaller. Instead, they should be asked which covariates they would have included if the new data source had been available before the PAP was registered.

[6] The purpose of this rule of thumb is to make it unlikely that adjustment leads to substantially worse precision or appreciable finite-sample bias. If time allows, simulations (using baseline data or prior studies) could provide additional guidance during the development of a PAP.

Covariates should generally be chosen on the basis of their expected ability to help predict outcomes, regardless of whether they appear well-balanced or imbalanced across treatment arms.[7] But there may be occasions when the covariate list specified in the PAP omitted a potentially important covariate (due to either an oversight or the need to keep the list short when N is small) with a nontrivial imbalance. Protection against ex post bias (conditional on the observed imbalance) is then a legitimate concern.[8] However, if observed imbalances are allowed to influence the choice of covariates,[9] the following guidelines should be observed:

1. If possible, the balance checks and decisions about adjustment should be finalized before we see unblinded outcome data.
2. The *direction* of the observed imbalance (e.g., whether the treatment group or the control group appears more advantaged at baseline) should not be allowed to influence decisions about adjustment. We will either pre-specify criteria that depend on the size of the imbalance but not its direction, or consult a "blind jury" that will not see the direction of imbalance or any other information that suggests how the adjustment would affect the point estimates.
3. The estimator specified in the PAP will always be reported and labeled as such, even if alternative estimates are also reported. See also "Unadjusted estimates, alternative regression specifications, and nonlinear models" below.

Missing covariate values

Observations with missing covariate values will be included in the regressions that estimate average treatment effects, as long as the outcome measure and treatment assignment are non-missing. Ordinarily, methods for handling missing values will have been specified in the PAP. If not, we will use the following approach:

1. If no more than 10% of the covariate's values are missing, recode the missing values to the overall mean. (Do not use arm-specific means.)
2. If more than 10% of the covariate's values are missing, include a missingness dummy as an additional covariate and recode the missing values to an arbitrary constant, such as 0.[10] If the missingness dummies lead us to exceed the M / 20 or N / 20 maximum number of covariates (see above under "Choice of covariates"), revert to the mean-imputation method above.

Example: Recoding Missing Covariates

```
suppressMessages({
  library(randomizr)
  library(sandwich)
  library(lmtest)
})

N <- 200
```

[7] As Bruhn and McKenzie (2009, 226) emphasize, "greater power is achieved by always adjusting for a covariate that is highly correlated with the outcome of interest, regardless of its distribution between groups."

[8] See Lin (2012b; 2013, 308) and references therein for discussion of this point.

[9] Commonly used standard error estimators assume that we would adjust for the same set of covariates regardless of which units were assigned to which treatment arm. Letting observed imbalances influence the choice of covariates violates this assumption. In the scenario studied by Permutt (1990), the result is that the significance test for the treatment effect has a true Type I error probability that is lower than the nominal level—i.e., the test is conservative.

[10] This method is described in Gerber and Green (2012), p. 241.

```r
# Make some covariates
X1 <- rnorm(N)
X2 <- rbinom(N, size = 1, prob = 0.5)

# Make some potential outcomes
Y0 <- .6*X1 + 3*X2 + rnorm(N)
Y1 <- Y0 + .4

# Conduct a random assignment and reveal outcomes
Z <- complete_ra(N, m= 100)
Y_obs <- Y1*Z + Y0*(1-Z)

# Some covariate values are missing:
X1_obs <- X1
X2_obs <- X2

X1_obs[sample(1:N, size = 10)] <- NA
X2_obs[sample(1:N, size = 50)] <- NA

# Less than 10% of X1_obs is missing, so:
X1_obs[is.na(X1_obs)] <- mean(X1_obs, na.rm = TRUE)

# More than 10% of X2_obs is missing, so:
X2_missing <- is.na(X2_obs)
X2_obs[X2_missing] <- 0

# Mean-center the covariates
X1_obs_c <- X1_obs - mean(X1_obs)
X2_obs_c <- X2_obs - mean(X2_obs)
X2_missing_c <- X2_missing - mean(X2_missing)

# Conduct Estimation
fit_adj <- lm(Y_obs ~ Z + Z*(X1_c + X2_c + X2_missing_c))

# Robust Standard Errors
coeftest(fit_adj, vcov = vcovHC(fit_adj, type = "HC2"))
```

```
##
## t test of coefficients:
##
##                 Estimate Std. Error t value  Pr(>|t|)
## (Intercept)     1.614419   0.183535  8.7962 8.091e-16 ***
## Z               0.626398   0.263746  2.3750   0.01853 *
## X1_c            0.233659   0.181516  1.2873   0.19955
## X2_c           -0.039053   0.371192 -0.1052   0.91632
## X2_missing_c   -0.119343   0.378039 -0.3157   0.75258
## Z:X1_c         -0.180135   0.250866 -0.7181   0.47360
## Z:X2_c         -0.346390   0.530964 -0.6524   0.51494
## Z:X2_missing_c  0.735601   0.529865  1.3883   0.16666
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Unadjusted estimates, alternative regression specifications, and nonlinear models

Our primary analysis will be based on a pre-specified covariate-adjusted estimator (unless $N < 20$), but we will also report unadjusted estimates as a robustness check. Results from alternative regression specifications may also be reported as specified in the PAP, or as allowed under "Choice of covariates" above, or as requested by referees. We will make clear to readers which estimator was pre-specified as primary.

For binary or count-data outcomes, some referees prefer estimates based on nonlinear models such as logit, probit, or Poisson regression. Although we disagree with this preference (the robustness of least squares adjustment in RCTs is supported by both theory and simulation evidence),[11] we will provide supplementary estimates derived from nonlinear models (using marginal effects calculations) if requested by referees. We prefer logits to probits because adjustment based on the probit MLE is not misspecification-robust.[12]

Covariate imbalance and the detection of administrative errors

We will perform a statistical test to judge whether observed covariate imbalances are larger than would normally be expected from chance alone. In an experiment with a binary treatment and a constant probability of assignment to treatment, the test involves a regression of the treatment indicator on the covariates and calculation of a heteroskedasticity-robust Wald statistic for the hypothesis that all the coefficients on the covariates are zero (Wooldridge (2010), p. 62). The covariates to be included in the regression should be specified in the PAP. (For voter turnout experiments, the SOP section "Issues specific to voter turnout experiments" gives a default set of covariates in case the PAP fails to specify the choice.) If the experiment is block-randomized with treatment probabilities that vary by block, we will also include dummy variables for the varying treatment probabilities in the regression, and we will test the hypothesis that all coefficients on the covariates, excluding the treatment probability dummies, are zero.

We will use a permutation test (randomization inference) to calculate the p-value associated with the Wald statistic.

In an experiment with multiple treatments, we will perform an analogous test using multinomial logistic regression of treatment on covariates.

A p-value of 0.01 or lower should prompt a thorough review of the random assignment procedure and any possible data-handling mistakes. If the review finds no errors, we will report the imbalance test, proceed on the assumption that the imbalance is due to chance, and report estimates with and without covariate adjustment.

Example: Permutation Test of Covariate Balance

```
suppressMessages({
  library(randomizr)
  library(sandwich)
})

# Generate Covariates

set.seed(1234567)

N <- 1000
```

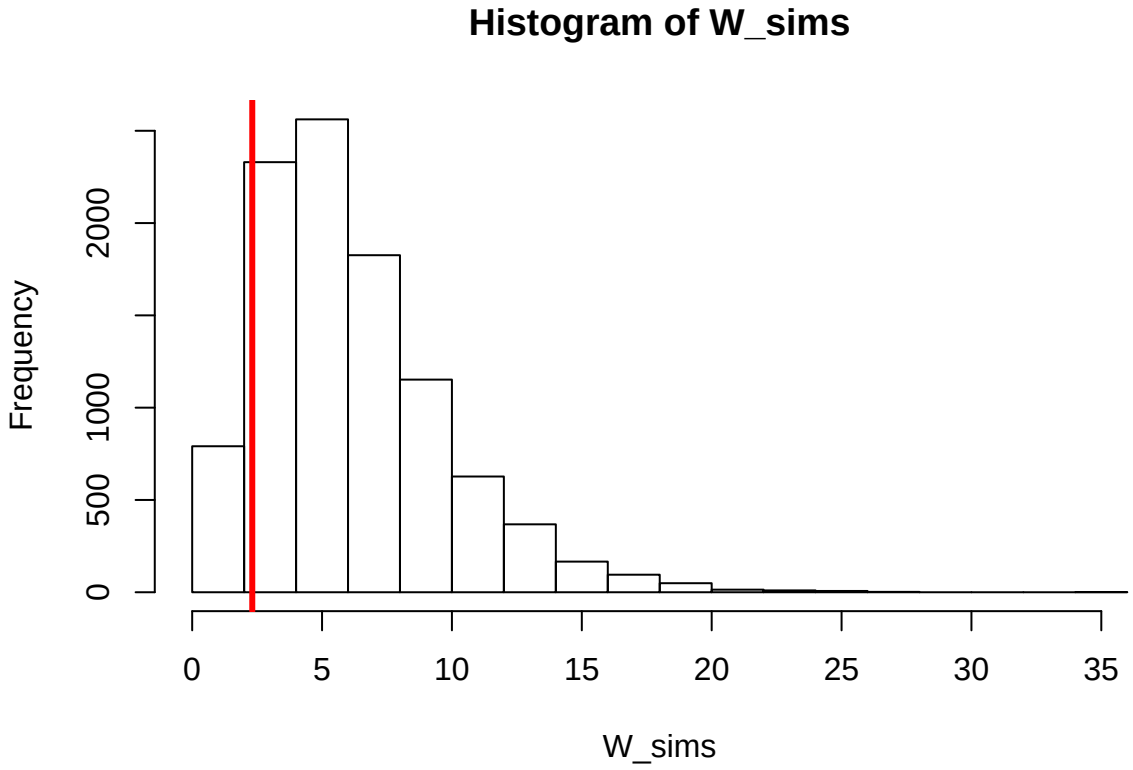
[11] For asymptotic theory, see Lin (2013), where all the results are applicable to both discrete and continuous outcomes. For simulations, see Humphreys, Sanchez de la Sierra, and van der Windt (2013) or Judkins and Porter (forthcoming).

[12] Freedman (2008); Firth and Bennett (1998). Lin gave an informal discussion in a comment on the *Mostly Harmless Econometrics* blog.

```r
gender <- sample(c("M", "F"), N, replace=TRUE)
age <- sample(18:65, N, replace = TRUE)
lincome <- rnorm(N, 10, 3)
party <- sample(c("D", "R", "I"), N, prob=c(.45, .35,.2), replace=TRUE)
education <- sample(10:20, N, replace=TRUE)

# Conduct Random Assignment
Z <- complete_ra(N, 500)

# Regress treatment on covariates
fit <- lm(Z ~ gender + age + lincome + party + education)

# Obtain observed heteroskedasticity-robust Wald statistic
# See Wooldridge (2010), p. 62
# Null hypothesis is that the slope coefficients are all zero, i.e.
#  R beta = 0
#  where beta is the 7 x 1 vector of coefficients, including the intercept
#  and R is the 6 x 7 matrix with all elements zero except
#   R[1,2] = R[2,3] = R[3,4] = R[4,5] = R[5,6] = R[6,7] = 1

Rbeta.hat <- coef(fit)[-1]
RVR <- vcovHC(fit, type <- 'HC0')[-1,-1]
W_obs <- as.numeric(Rbeta.hat %*% solve(RVR, Rbeta.hat))  # Wooldridge, equation (4.13)

# Compare to permutation distribution of W

sims <- 10000
W_sims <- numeric(sims)

for(i in 1:sims){
  Z_sim <- complete_ra(N, 500)
  fit_sim <- lm(Z_sim ~ gender + age + lincome + party + education)

  Rbeta.hat <- coef(fit_sim)[-1]
  RVR <- vcovHC(fit_sim, type <- 'HC0')[-1,-1]
  W_sims[i] <- as.numeric(Rbeta.hat %*% solve(RVR, Rbeta.hat))
}

# Obtain p-value
p <- mean(W_sims >= W_obs)
p
```

```
## [1] 0.8903
```

```r
hist(W_sims)
abline(v = W_obs, lwd=3, col="red")
```



Analysis of block randomized experiments with treatment probabilities that vary by block

Estimating treatment effects

When treatment assignment probabilities vary by block, we will use weighted least squares regression with inverse probability-of-assignment weights (IPW) to estimate the average treatment effect (ATE), except in extreme cases (defined below). See Gerber and Green (2012), section 4.5 for discussion of the IPW method.

An alternative approach, sometimes called the least-squares dummy variable (LSDV) method, is to run an OLS regression of the outcome on treatment, block indicators, and covariates. The estimand for the LSDV method is a weighted ATE, giving each block j a total weight proportional to $N_j P_j (1 - P_j)$, where N_j is the number of subjects in the block and P_j is their probability of assignment to treatment (Angrist (1998), p. 256; Angrist and Pischke (2009), p. 75; Gerber and Green (2012), p. 119). In contrast, the estimand for the IPW method is the unweighted ATE, which gives block j a total weight proportional to N_j. We will use IPW except in the extreme case where there is at least one block j such that

$$\frac{N_j}{\sum_j N_j} > 20 \frac{N_j P_j (1 - P_j)}{\sum_j N_j P_j (1 - P_j)}.$$

In that case, we will use LSDV and we will explain to readers that we pre-specified a weighted ATE as our estimand in an attempt to improve precision. For background, see Angrist and Pischke (2009) (p. 76, footnote 23) and Gerber and Green (2012) (pp. 119-120).

For example, imagine a block-randomized experiment with two blocks: 500 of 1,000 subjects in block 1 are assigned to treatment, while 5 of 100,000 in block 2 are assigned to treatment. The IPW estimand (the unweighted ATE) places a total weight of $N_2/(N_1+N_2) = 99.01\%$ on the second, relatively uninformative block, while the LSDV estimand gives the same block a total weight of $N_2 P_2 (1-P_2)/[N_1 P_1 (1-P_1) + N_2 P_2 (1-P_2)] =$

1.96%. Thus, block 2 is weighted over 50 times more heavily by IPW than by LSDV. In this situation, we would use LSDV instead of IPW.

Presentation of statistics describing covariate balance

For the formal test for covariate balance, see the section above on "Covariate imbalance and the detection of administrative errors".

For tables or figures describing covariate balance, we will follow the advice in Gerber and Green (2012), pp. 120-121, with the following amendment to their footnote 19: If LSDV is the primary treatment effect estimator, then instead of using the weights in their equation (4.12), give each treatment group observation a weight proportional to $1 - P_j$ and each control group observation a weight proportional to P_j. (This results in a weighted treatment group mean and weighted control group mean that give block j a total weight proportional to $N_j P_j(1 - P_j)$, which matches the LSDV estimand.)

Presentation of statistics describing overall baseline characteristics of the study sample (not separated by treatment arm)

If the primary treatment effect estimator is IPW, summary statistics on overall characteristics of the sample should be unweighted, since the IPW estimand is the unweighted ATE. If LSDV is the primary treatment effect estimator, the summary statistics should give each observation a weight proportional to $P_j(1 - P_j)$, since this gives block j a total weight proportional to $N_j P_j(1 - P_j)$, which matches the LSDV estimand.

Sample exclusions and the coding of outcome variables

In general, we will avoid coding outcomes in ways that cause subjects to be excluded from the estimation of average treatment effects. Such exclusions are especially problematic when treatment assignment may affect the chances that a sample member will be excluded. For example, for an outcome such as the amount donated to a political campaign, we will not exclude subjects with outcome values of zero from our analyses.

Exceptions include instances where sample exclusions do not threaten the symmetry between the randomly assigned treatment arms. For example, in an experiment with a treatment/placebo design, we will report analyses that exclude noncompliers if checks #1-4 in the section on "Treatment/placebo designs" yield satisfactory results.

In a voter turnout experiment, we will exclude subjects who voted before treatment began, since their outcomes could not have been affected by treatment (see the section on "Issues specific to voter turnout experiments"). In other cases where data collected after random assignment identifies a subgroup of subjects whose outcomes were determined before treatment began, we will exclude that subgroup if (1) there was no plausible way for outcomes to be affected by treatment assignment before the actual treatment began and (2) empirical investigations analogous to checks #2-4 in the section on "Treatment/placebo designs" yield no evidence that this sample exclusion creates noncomparability between the treatment arms.

The section on "Issues specific to survey or laboratory experiments" discusses other examples.

Noncompliance

In experiments that encounter noncompliance with assigned treatments, our analysis will include a test of the hypothesis that the average intent-to-treat effect is zero.

Estimating treatment effects when some subjects receive "partial treatment"

Gerber and Green (2012) (pp. 164-165) discuss several approaches for estimating treatment effects when some treatment group members receive a full dose of the intended intervention and others receive only part of it. If there is no noncompliance in the control group, we will follow the approach where "the researcher simply considers all partially treated subjects as fully treated" (Gerber and Green 2012, 165) (thus adopting the most expansive definition of treatment in order to make the instrumental variables exclusion restriction plausible), unless either the variation in treatment dosage is randomized or the PAP specifies both the dosage measure and the method for analyzing the dose-response relationship.

Treatment/placebo designs

In a treatment/placebo design, subjects are randomly assigned to be encouraged to receive either the treatment or a placebo (Gerber and Green 2012, 161–64). Those who actually receive the treatment or placebo are revealed to be compliers. The intended analysis compares the outcomes of treatment group compliers vs. placebo group compliers. However, if the encouragement efforts differ between the two arms, the two groups of compliers may not be comparable. To evaluate their comparability, we will perform the following checks:

1. Implementation: Were the treatment and placebo administered by the same personnel? Were these personnel blinded to subjects' random assignments until after compliance status was determined, or if not, were the treatment and placebo administered symmetrically in their timing, place, and manner?
2. Comparison of compliance rates across arms: We will perform a two-tailed unequal-variances t-test of the hypothesis that treatment assignment does not affect the compliance rate.
3. Comparison of compliers' baseline characteristics across arms: Using compliers only, we will estimate a linear regression of the treatment group indicator on baseline covariates and perform a heteroskedasticity-robust F-test (Wooldridge (2010), p. 62) of the hypothesis that all coefficients on the covariates are zero.
4. Comparison of noncompliers' outcomes across arms: Using noncompliers only, we will perform a two-tailed unequal-variances t-test of the hypothesis that treatment assignment does not affect the average outcome.

In checks #2-#4, p-values below 0.05 will be considered evidence of noncomparability.

If any of those checks raises a red flag, we will use two-stage least squares to estimate the complier average causal effect, using assignment to the treatment as an instrumental variable predicting actual treatment. In other words, we will analyze the experiment as if it had a conventional treatment/baseline design instead of a treatment/placebo design.

Nickerson's rolling protocol design

In Nickerson's rolling protocol design (Nickerson 2005), researchers create a randomly ordered list of treatment group members (or clusters of treatment group members) and insist that treatment attempts follow this random order. When resources for treatment attempts run out, the bottom portion of the randomly ordered list (i.e., those treatment group members for whom treatment was never attempted) is moved into the control group. To check that this procedure creates comparable groups, we will perform the following checks:

1. Movement of treatment group members into the control group must be based strictly on the random ordering of the list. If, within some section of the list, the personnel administering treatment have nonrandomly chosen to attempt treatment for some subjects but not others, then the entire section and all preceding sections should remain in the treatment group.

2. The decision to stop treatment attempts must be based solely on resources, not on characteristics of the subjects or clusters.

3. Comparison of baseline characteristics: We will estimate a linear regression of the proposed treatment group indicator on baseline covariates and perform a heteroskedasticity-robust F-test (Wooldridge (2010), p. 62) of the hypothesis that all coefficients on the covariates are zero. A p-value below 0.05 will be considered evidence of noncomparability.

If these checks cast doubt on the comparability of treatment and control groups, we will not move any unattempted treatment group members into the control group.

Attrition

"Attrition" here means that outcome data are missing. (When only baseline covariate data are missing, we will still include the observations in the analysis, as explained under "Missing covariate values".) Often, it is unclear theoretically whether missingness threatens the symmetry between treatment and control groups. We will routinely perform three types of checks for asymmetrical attrition:

1. Implementation: Were all treatment arms handled symmetrically as far as the timing and format of data collection and the personnel involved? Did each arm's subjects have the same incentives to participate in follow-up? Were the data collection personnel blind to treatment assignment?

2. Comparison of attrition rates across treatment arms: In a two-arm trial, we will perform a two-tailed unequal-variances t-test of the hypothesis that treatment does not affect the attrition rate. In a multi-arm trial, we will perform a heteroskedasticity-robust F-test (Wooldridge (2010), p. 62) of the hypothesis that none of the treatments affect the attrition rate. In either case, we will implement the test as a Studentized permutation test—i.e., a test that compares the observed t- or F-statistic with its empirical distribution under random reassignments of treatment.

3. Comparison of attrition patterns across treatment arms: Using a linear regression of an attrition indicator on treatment, baseline covariates, and treatment-covariate interactions, we will perform a heteroskedasticity-robust F-test of the hypothesis that all the interaction coefficients are zero. The covariates in this regression will be the same as those used in the covariate balance test (see the section on "Covariate imbalance and the detection of administrative errors"). As in check #2, we will implement the test as a Studentized permutation test.

In checks #2 and #3, p-values below 0.05 will be considered evidence of asymmetrical attrition.

If any of those checks raises a red flag, and if the PAP has not specified methods for addressing attrition bias, we will follow these procedures:

1. Rely on second-round sampling of nonrespondents, combined with extreme value bounds (Aronow et al. 2015) if (a) the project has adequate resources and (b) it is plausible to assume that potential outcomes are invariant to whether they are observed in the initial sample or the follow-up sample. If either (a) or (b) is not met, go to step 2.

2. Consult a disinterested "jury" of colleagues to decide whether the monotonicity assumption for trimming bounds (Lee 2009; Gerber and Green 2012, 227) is plausible. If so, report estimates of trimming bounds; if not, report estimates of extreme value (Manski-type) bounds (Gerber and Green (2012), pp. 226-227). (If the outcome has unbounded range, report extreme value bounds that assume the largest observed value is the largest possible value.) In either case, also report the analysis that was specified in the PAP.

Outliers

Except as specified in the PAP or as part of a supplemental robustness check, we will not delete or edit outlying values merely because they are very large or very small. However, it is appropriate for outlying values to trigger checks for data integrity, as long as the process and any resulting edits are results-blind and symmetric with respect to treatment arm.

When randomization doesn't go according to plan

Verifying that randomization was implemented as planned

We will have at least two team members check each computer program used to randomly assign treatment, and we will make these programs publicly available. In all such programs, we will use the seed value 1234567 for the random number generator, so that the resulting assignments can be replicated and verified.

See the section "Covariate imbalance and the detection of administrative errors" for a description of the statistical test we will use to judge whether observed covariate imbalances are larger than would normally be expected from chance alone. In addition to reporting the result of this test, we will follow the reporting guidelines in the "Allocation Method" section of Gerber et al. (2014) (Appendix 1, part C).

In the event that these checks reveal any errors, we will report the errors and take them into account in any analyses we report (see below for examples). We will add more specific guidance and examples to our SOP as we learn from our own and/or other researchers' experiences.

Learning of a restricted randomization

Sometimes we may learn or realize ex post that certain randomizations were disallowed. For example, an NGO partner may reveal that they would have canceled the RCT if a particular unit had not been assigned to the treatment group. Or, we may realize that we implicitly did a restricted randomization, since we checked covariate balance prior to implementing the treatment assignment, and if there had been a large enough imbalance, we would have re-randomized.

We will reveal such implicit restrictions in our research reports and articles.

If we can formalize the implicit restriction and reconstruct the set of admissible randomizations, we will analyze the data as suggested in Gerber and Green (2012) (Box 4.5, p. 121): First, if the treatment and control groups are of different sizes, we will use inverse probability-of-assignment weights to estimate the average treatment effect (estimating the weights by simulating a large number of admissible randomizations and tabulating the fraction of randomizations that assign each subject to treatment or control). Second, we will use randomization inference (excluding the disallowed randomizations) to estimate p-values.

If we cannot formalize the implicit restriction, we will keep the pre-specified analysis strategy but will note the issue for readers (e.g., by saying that we checked for covariate balance before implementing treatment assignment but did not have a fixed balance criterion in mind).

Duplicate records in the dataset used for randomization

After treatment has begun, we may learn that there were duplicate records in the dataset that was used to randomly assign subjects. This raises the problems that (1) a subject could be assigned to more than one arm, and (2) subjects with duplicate records had a higher probability of assignment to treatment than subjects with unique records.

How we handle this situation depends on two questions.

Question 1: Were the multiple assignments of duplicate records made simultaneously, or can they be ordered in time?

For example, when applicants for a social program are randomly assigned as their applications are processed, random assignment may continue for months or years, and in unusual cases, a persistent applicant who was originally assigned to the control group may later succeed in getting assigned to treatment under a duplicate record. In that case, the existence and number of duplicate records may be affected by the initial assignment.

If the assignments can be ordered in time, we will treat the initial assignment as the correct one, and any noncompliance with the initial assignment will be handled the same way as for subjects who did not have duplicate records.

If the assignments were made simultaneously, Question 2 should be considered.

Question 2: Is it reasonable to say that if a subject was assigned to more than one arm, one of her assignments "trumps" the other(s)?

For example, in a two-arm trial where the treatment is an attempted phone call and the control condition is simply no attempt (without any active steps to prohibit a phone call), it seems reasonable to decide that treatment trumps control—i.e., assigning a subject with duplicate records to both conditions is like assigning her to treatment. In contrast, in a treatment/placebo design where the treatment and placebo are attempted conversations about two different topics, we would hesitate to assume that treatment trumps placebo. And in a three-arm trial with two active treatments and a control condition, it might be reasonable to assume that one treatment trumps the other if the former includes all of the latter's activities and more, but otherwise we would hesitate to make that assumption.

If the trump assumption can be reasonably made, then in the analysis, we will take the following steps:

1. Deduplicate the records.
2. Use the trump assumption to reclassify any subject who was assigned to more than one arm.
3. Calculate each subject's probabilities of assignment to each arm, where "assignment" means the unique classification from step 2. These probabilities will depend on the number of records for the subject in the original dataset.
4. Use inverse probability-of-assignment weighting (IPW) to estimate treatment effects.

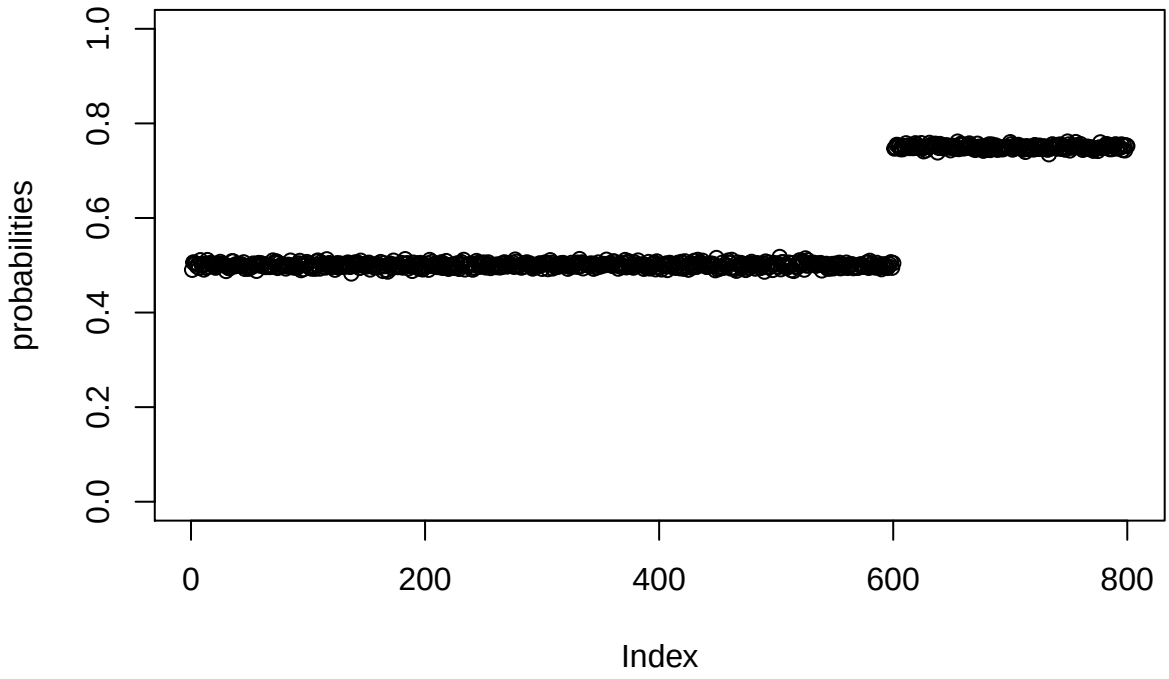
If the trump assumption cannot be reasonably made, then we will replace step 2 with a step that excludes from the analysis any subject who was assigned to more than one arm. We will then check whether steps 3 and 4 still need to be performed. (For example, in a two-arm Bernoulli-randomized trial with intended probabilities of assignment of 2 / 3 to treatment and 1 / 3 to control, a subject with two records has probability 4 / 9 of two assignments to treatment, 4 / 9 of one assignment to treatment and one to control, and 1 / 9 of two assignments to control. Conditional on remaining in the analysis after we exclude subjects who were assigned to both treatment and control, she has probability 4 / 5 of assignment to treatment.)

Example: Fundraising Experiment

Suppose a fundraising experiment randomly assigns 500 of 1,000 names to a treatment that consists of an invitation to contribute to a charitable cause. However, it is later discovered that 600 names appear once and 200 names appear twice. Before the invitations are mailed, duplicate invitations are discarded, so that no one receives more than one invitation.

In this case, the experimental procedure justifies the trump assumption. Names that are assigned once or twice are in treatment, the remainder are in control. It's easy enough in this example to calculate analytic probabilities (0.5 for those who appear once, $1 - (500/1000) \times (499/999) \approx 0.75$ for those who appear twice). However, in some situations, simulating the exact procedure is the best way to determine probabilities (it can also be a good way to check your work!). Here is a short simulation in R that confirms the analytic solution.

```r
# Load randomizr for complete_ra()
library(randomizr)

# Make a list of 1000 names. 200 names appear twice
name_ids <- c(paste0("name_", sprintf("%03d", 1:600)),
              paste0("name_", sprintf("%03d", 601:800)),
              paste0("name_", sprintf("%03d", 601:800)))

# Conduct simulation
sims <- 10000
Z_mat <- matrix(NA, nrow = 800, ncol = sims)
for(i in 1:sims){
  # Conduct assignment among the 1000 names
  Z_1000 <- complete_ra(1000, 500)
  # Check if names were ever assigned
  Z_800 <- as.numeric(tapply(Z_1000, name_ids, sum) > 0)
  # Save output
  Z_mat[,i] <- Z_800
}

# Calculate probabilities of assignment
probabilities <- rowMeans(Z_mat)
plot(probabilities, ylim=c(0,1))
```



The plot confirms the analytic solution. The first 600 names have probability of assignment 0.5, and names 601 through 800 (the duplicates) have probability 0.75.

Other transparency issues

Canceled, stopped, or "failed" RCTs

In extreme circumstances, an RCT may "fail" in the sense that unanticipated problems impose such severe limitations on what we can learn from the study that it becomes unpublishable. Such problems may include a failure to enroll an adequate number of subjects or to implement a meaningful treatment, stakeholder resistance that leads to cancellation of the RCT, or evidence of harm that persuades researchers to stop the RCT early for ethical reasons.[13]

In such cases, we will make publicly available a summary of the design and implementation, the results (if any), and the apparent causes of failure.

Differences between the pre-specified analyses and those that appear in the article

Each published article will reference its PAP. If the article contains analyses that deviate from the PAP, it will make clear that these analyses were not pre-specified. Conversely, if the article omits any pre-specified analyses, it will give a brief description of them and they will be made available in a document that is referenced in the article.

Issues specific to voter turnout experiments

Because our lab frequently evaluates the effects of voter mobilization campaigns, this SOP includes rules designed to impose uniformity across trials.

Coding of voter turnout outcomes often varies across jurisdictions, with some administrative units reporting only whether someone voted and others reporting whether registered voters voted or abstained. We will code turnout as 1 if the subject is coded as having voted and 0 otherwise.

In cases where a post-election list of registered voters no longer includes some members of the treatment and control groups, we will evaluate whether attrition is plausibly independent of treatment assignment using the procedures discussed above. If so, the analysis will focus on just those subjects who have not been removed from the voter registration rolls.

In some instances, voter turnout records include the date on which a ballot is cast. When voter turnout data is date-stamped, our analysis sample will exclude those who voted before treatment began, since their outcomes could not have been affected by treatment.

In canvassing and phone-banking experiments, noncompliance is common. In such cases, contact will be coded broadly to include any form of interaction with the subject that might affect turnout – even a very brief conversation whereby the respondent hangs up after the canvasser introduces himself/herself. Messages left with housemates count as contact. Interactions that do not count as contact include busy signals, no one opening the door, or failure to communicate with the respondent due to language barriers. A phone call from a number with a recognizable caller ID (e.g., "Vote '98 Headquarters") would count as contact.

In instances where canvassing or calling efforts fail to attempt large swaths of the originally targeted treatment group (e.g., a certain group of precincts), an assessment will be made of whether failure-to-attempt was related to the potential outcomes of the subjects. If the scope of the canvassing or calling effort fell short for reasons that seem to have nothing to do with the attributes of the subjects who went unattempted, the subject pool will be partitioned and the analysis restricted to the attempted precincts. (See the section on "Nickerson's rolling protocol design".)

[13] For related discussion, see Greenberg and Barnow (2014).

If the PAP fails to specify the choice of covariates for regression adjustment or for the test of covariate balance, the default set of covariates will include voter turnout in all past elections for which data are available in the voter file, excluding any elections in which turnout rates in the subject pool were below 5%.

Issues specific to survey or laboratory experiments

Do not exclude subjects who discern the purpose of the experiment

Subjects in a lab or survey experiment may indicate in a post-experimental debriefing session that they discerned the hypothesis that we sought to test. We will not exclude these subjects from the analysis. The treatment assignment could have affected whether subjects discerned the hypothesis (see the section on "Sample exclusions and the coding of outcome variables").

Whether to exclude subjects who display inattention in survey experiments

In an online survey experiment, some subjects may be clicking answers arbitrarily to complete the survey quickly. To detect such behavior, researchers sometimes insert "screener" questions (Berinsky, Margolis, and Sances (2014)) to assess whether subjects are paying attention to the content of the survey (e.g., a question that simply directs respondents to check a particular box). In such cases, we will report an analysis excluding the "inattentive" respondents (those who answered the screener questions incorrectly) if (1) there is nothing about the treatment that would cause inattention to be more or less common in one treatment arm and (2) we find no evidence that this sample exclusion creates noncomparability between the treatment arms when checks #2-#4 from the section on "Treatment/placebo designs" are performed in this context (classifying "inattentive" respondents as noncompliers). We will also report results from the full sample as a robustness check.

If either condition (1) or condition (2) above is not satisfied, we will not exclude the "inattentive" respondents.

References

Angrist, Joshua D. 1998. "Estimating the Labor Market Impact of Voluntary Military Service Using Social Security Data on Military Applicants." *Econometrica* 66 (2): 249–88. http://www.jstor.org/stable/2998558.

Angrist, Joshua D., and Jörn-Steffen Pischke. 2009. *Mostly Harmless Econometrics: An Empiricist's Companion*. Princeton University Press.

Aronow, Peter M., Alexander Coppock, Alan S. Gerber, Donald P. Green, and Holger Kern. 2015. "Double Sampling for Missing Outcome Data in Randomized Experiments."

Berinsky, Adam J., Michele F. Margolis, and Michael W. Sances. 2014. "Separating the Shirkers from the Workers? Making Sure Respondents Pay Attention on Self-Administered Surveys." *American Journal of Political Science* 58 (3): 739–53. doi:10.1111/ajps.12081.

Bruhn, Miriam, and David McKenzie. 2009. "In Pursuit of Balance: Randomization in Practice in Development Field Experiments." *American Economic Journal: Applied Economics* 1 (4): 200–232. http://www.jstor.org/stable/25760187.

Chung, EunYi, and Joseph P. Romano. 2013. "Exact and Asymptotically Robust Permutation Tests." *Annals of Statistics* 41 (2): 484–507. doi:10.1214/13-AOS1090.

Firth, D., and K. E. Bennett. 1998. "Robust Models in Probability Sampling." *Journal of the Royal Statistical Society: Series B* 60 (1): 3–21. doi:10.1111/1467-9868.00105.

Freedman, David A. 2008. "Randomization Does Not Justify Logistic Regression." *Statistical Science* 23 (2): 237–49. doi:10.1214/08-STS262.

Gerber, Alan S., and Donald P. Green. 2012. *Field Experiments: Design, Analysis, and Interpretation*. New York: W.W. Norton.

Gerber, Alan, Kevin Arceneaux, Cheryl Boudreau, Conor Dowling, Sunshine Hillygus, Thomas Palfrey, Daniel R. Biggers, and David J. Hendry. 2014. "Reporting Guidelines for Experimental Research: A Report from the Experimental Research Section Standards Committee." *Journal of Experimental Political Science* 1 (01): 81–98. doi:10.1017/xps.2014.11.

Greenberg, David, and Burt S. Barnow. 2014. "Flaws in Evaluations of Social Programs: Illustrations From Randomized Controlled Trials." *Evaluation Review* 38 (5): 359–87. doi:10.1177/0193841X14545782.

Humphreys, Macartan, Raul Sanchez de la Sierra, and Peter van der Windt. 2013. "Fishing, Commitment, and Communication: A Proposal for Comprehensive Nonbinding Research Registration." *Political Analysis* 21 (1): 1–20. doi:10.1093/pan/mps021.

Janssen, Arnold. 1997. "Studentized Permutation Tests for Non-I.I.D. Hypotheses and the Generalized Behrens-Fisher Problem." *Statistics and Probability Letters* 36 (1): 9–21. doi:10.1016/S0167-7152(97)00043-6.

Judkins, David R., and Kristin E. Porter. forthcoming. "Robustness of Ordinary Least Squares in Randomized Clinical Trials." *Statistics in Medicine*.

Lee, David S. 2009. "Training, Wages, and Sample Selection: Estimating Sharp Bounds on Treatment Effects." *Review of Economic Studies* 76 (3): 1071–1102. doi:10.1111/j.1467-937X.2009.00536.x.

Lin, Winston. 2012a. "Regression Adjustment in Randomized Experiments: Is the Cure Really Worse than the Disease? Part I." http://web.archive.org/web/20150505184132/http://blogs.worldbank.org/impactevaluations/node/847.

———. 2012b. "Regression Adjustment in Randomized Experiments: Is the Cure Really Worse than the Disease? Part II." http://web.archive.org/web/20150505184245/http://blogs.worldbank.org/impactevaluations/node/849.

———. 2013. "Agnostic Notes on Regression Adjustments to Experimental Data: Reexamining Freedman's Critique." *The Annals of Applied Statistics* 7 (1): 295–318. doi:10.1214/12-AOAS583.

Lin, Winston, and Donald P. Green. 2015. "Standard Operating Procedures: A Safety Net for Pre-Analysis Plans."

MacKinnon, James G. 2013. "Thirty Years of Heteroskedasticity-Robust Inference." In *Recent Advances and Future Directions in Causality, Prediction, and Specification Analysis: Essays in Honor of Halbert L. White Jr*, edited by Xiaohong Chen and Norman R. Swanson, 437–61. Springer.

Nickerson, David W. 2005. "Scalable Protocols Offer Efficient Design for Field Experiments." *Political Analysis* 13 (3): 233–52. doi:10.1093/pan/mpi015.

Olken, Benjamin A. 2015. "Promises and Perils of Pre-Analysis Plans." *Journal of Economic Perspectives* 29 (3): 61–80. doi:10.1257/jep.29.3.61.

Permutt, Thomas. 1990. "Testing for Imbalance of Covariates in Controlled Experiments." *Statistics in Medicine* 9 (12): 1455–62. doi:10.1002/sim.4780091209.

Romano, Joseph P. 2009. "Discussion of 'Parametric versus Nonparametrics: Two Alternative Methodologies'." *Journal of Nonparametric Statistics*. doi:10.1080/10485250902846900.

Rosenbaum, Paul R. 2010. *Design of Observational Studies*. Springer.

Tukey, John W. 1993. "Tightening the Clinical Trial." *Controlled Clinical Trials* 14 (4): 266–85. doi:10.1016/0197-2456(93)90225-3.

Wooldridge, Jeffrey M. 2010. *Econometric Analysis of Cross Section and Panel Data*. 2nd ed. MIT Press.