Standard operating procedures for Don Green's lab at Columbia

Winston Lin, Donald P. Green, and Alexander Coppock

Version 1.05: June 7, 2016

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 (forthcoming).

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 Özler, Anselm Rink, Cyrus Samii, Michael Schwam-Baird, Uri Simonsohn, Amber Spry, Dane Thorley, Anna Wilke, and José 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.

Standard errors, confidence intervals, and significance tests

Bell–McCaffrey adjustment for standard errors and confidence intervals

Many scholars agree that uncertainty due to sampling variability should be communicated not solely via null hypothesis significance testing, but via methods that focus on the magnitude of the estimated effect and the

[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.

range of magnitudes compatible with the data, such as confidence intervals (Moher et al. 2010; Vazire 2016; Greenland et al. 2016; Wasserstein and Lazar forthcoming). Some journals require or encourage reporting of confidence intervals (CIs), while in others it is more typical to report estimated standard errors (SEs) along with point estimates. SEs are of interest not so much for their own sake as for enabling the construction of CIs and tests (Efron 1986; Pustejovsky and Tipton 2016). Thus, if SEs are reported without explicit CIs, they should ideally be accompanied by sufficient information for CI construction, such as the degrees of freedom for the distribution of the t-statistic.

Our default method for SEs, degrees of freedom, and CIs is the Bell–McCaffrey (BM) adjustment (Bell and McCaffrey 2002). Imbens and Kolesár (forthcoming) give a very helpful discussion of the procedure and "recommend that empirical researchers should, as a matter of routine, use the BM confidence intervals."[3] Similar to Welch's unequal-variances t-test, the BM adjustment combines a bias-reduced version of the "robust" or "cluster-robust" SE with a Satterthwaite approximation for the degrees of freedom of the t-distribution. PAPs are (as always) free to deviate from this default approach. Our primary goal here is to obtain approximately valid CIs for average treatment effects (where "valid" means that the actual coverage probability equals or exceeds the nominal coverage). A researcher preparing a PAP for a specific experiment may have reason to believe, based on statistical theory and/or simulations, that a simpler method (e.g., using a critical value from the standard normal distribution instead of the t-distribution when the sample is sufficiently large, or using the classical OLS SE when subjects are randomly assigned to two equal-sized groups) will achieve the same goal, that a different approach is needed due to complexities in the experimental design or the point estimation procedure, or that a different approach will yield more precision.

Below, we give example R code (a slight adaptation of Michal Kolesár's `BMlmSE()` function) to perform the BM adjustment given a fitted OLS regression.[4]

Example R code for Bell–McCaffrey adjustment

In the following R code, the function `BMlmSE()` takes these arguments:

- `model` (required): An object returned by `lm()`, representing a fitted OLS regression.
- `clustervar` (optional): A factor whose levels correspond to the clusters. If `clustervar` is supplied, the function uses Bell and McCaffrey (2002)'s bias-reduced cluster-robust variance estimator. If `clustervar` is left unspecified (meaning that the user does not want clustered SEs), the function uses the HC2 robust variance estimator.
- `ell` (optional): A vector specifying a linear combination of coefficients to compute an SE and degrees of freedom for. `ell` must have the same length as the vector of regression coefficients. For example, if `model` is an object returned by `lm(outcome ~ treatment + covariate)`, then `ell = c(0, 1, 0)` specifies that we are only interested in the coefficient on `treatment`. If `ell` is left unspecified, the function computes SEs and degrees of freedom for all coefficients. (The adjusted degrees of freedom will generally be different for each coefficient, unlike the classical OLS degrees of freedom.)
- `IK` (optional): A logical value. If `clustervar` is unspecified, `IK` has no effect on the results. If `clustervar` is supplied, `IK` determines whether the degrees of freedom are computed using Imbens and Kolesár (forthcoming)'s method (if `IK` is `TRUE` or unspecified) or Bell and McCaffrey (2002)'s method (if `IK` is `FALSE`). Our SOP uses Bell and McCaffrey (2002)'s method.

The function returns a list with these components:

[3] Cameron and Miller (2015), Pustejovsky and Tipton (2016), and Young (2016) also give helpful discussions and simulation evidence supporting similar adjustments.

[4] The BM adjustment has been extended to weighted least squares (WLS) regression (McCaffrey, Bell, and Botts 2001; Pustejovsky and Tipton 2016). James Pustejovsky's `clubSandwich` R package, which implements this extension (among other things), is in active development at the time of this writing. For now, our SOP adopts the BM adjustment for OLS regression as our default and shows how to use OLS with treatment–block interactions instead of WLS to analyze block-randomized experiments with treatment probabilities that vary by block.

- `Vhat`: The estimated covariance matrix.
- `dof`: The adjusted degrees of freedom. (If `ell` is specified, `dof` is a scalar. If `ell` is unspecified, `dof` is a named vector with one element for each regression coefficient.)
- `adj.se`: The SEs derived from `Vhat` *with* Imbens and Kolesár (forthcoming)'s additional adjustment to "convert the dof adjustment into a procedure that only adjusts the standard errors." (If `ell` is specified, `adj.se` is a scalar. If `ell` is unspecified, `adj.se` is a named vector with one element for each regression coefficient.) This is *not* our default method. `adj.se` can be multiplied by 1.96 to form the margin of error for a 95% CI. However, multiplying `adj.se` by 1.645 is *not* generally a valid way to form the margin of error for a 90% CI.
- `se`: The SEs derived from `Vhat` *without* Imbens and Kolesár (forthcoming)'s additional adjustment to "convert the dof adjustment into a procedure that only adjusts the standard errors." (If `ell` is specified, `se` is a scalar. If `ell` is unspecified, `se` is a named vector with one element for each regression coefficient.) This *is* our default method. To form the margin of error for a CI at any confidence level, multiply `se` by the appropriate critical value from the t-distribution with `dof` degrees of freedom. We give example R code for a 95% CI later in this section. If we are not reporting CIs explicitly, we will report `se` together with `dof`.[5]
- `se.Stata`: The cluster-robust SE computed by the formula that Stata has used (see Imbens and Kolesár forthcoming). (If `clustervar` is unspecified, `se.Stata` is `NA`.) This is *not* our default method.

```
## This is a slightly modified version of Michal Kolesar's BM_StandardErrors.R
## The original file was downloaded from:
##   https://github.com/kolesarm/Robust-Small-Sample-Standard-Errors
##   (Latest commit 2a80873 on Aug 26, 2015)
## We made minor changes for efficiency, as well as two changes that affect the
## code's functionality:
## 1) MatSqrtInverse() now stops with an error message if its argument is not of
##    full rank:
##    "Bell-McCaffrey SE undefined. This happens, e.g., when a dummy regressor is 1
##     for one cluster and 0 otherwise."
## 2) The list returned by BMlmSE() now includes "se".

# The MIT License (MIT)
#
# Copyright (c) 2015 Michal Kolesar
#
# Permission is hereby granted, free of charge, to any person obtaining a copy
# of this software and associated documentation files (the "Software"), to deal
# in the Software without restriction, including without limitation the rights
# to use, copy, modify, merge, publish, distribute, sublicense, and/or sell
# copies of the Software, and to permit persons to whom the Software is
# furnished to do so, subject to the following conditions:
#
# The above copyright notice and this permission notice shall be included in all
# copies or substantial portions of the Software.
#
# THE SOFTWARE IS PROVIDED "AS IS", WITHOUT WARRANTY OF ANY KIND, EXPRESS OR
# IMPLIED, INCLUDING BUT NOT LIMITED TO THE WARRANTIES OF MERCHANTABILITY,
# FITNESS FOR A PARTICULAR PURPOSE AND NONINFRINGEMENT. IN NO EVENT SHALL THE
```

[5] Exceptions can be made when `dof` is 500 or greater. Critical values from the t- and normal distributions are then approximately equal, so it is acceptable for a table to report only SEs and note that the corresponding degrees of freedom were always greater than or equal to 500.

```r
# AUTHORS OR COPYRIGHT HOLDERS BE LIABLE FOR ANY CLAIM, DAMAGES OR OTHER
# LIABILITY, WHETHER IN AN ACTION OF CONTRACT, TORT OR OTHERWISE, ARISING FROM,
# OUT OF OR IN CONNECTION WITH THE SOFTWARE OR THE USE OR OTHER DEALINGS IN THE
# SOFTWARE.

## Compute Bell-McCaffrey Standard Errors
library(sandwich)
library(Matrix)

message1 <- paste0(
    'Bell-McCaffrey SE undefined. This happens, e.g., when a dummy regressor is 1 ',
    'for one cluster and 0 otherwise.'
)

MatSqrtInverse <- function(A) {
    ##  Compute the inverse square root of a matrix
    if (rankMatrix(A) < NROW(A)) stop(message1)
    ei <- eigen(A, symmetric = TRUE)
    d2 <- 1/sqrt(ei$values)
    ## diag(d2) is d2 x d2 identity if d2 is scalar, instead we want 1x1 matrix
    ei$vectors %*% (if (length(d2)==1) d2 else diag(d2)) %*% t(ei$vectors)
}

BMlmSE <- function(model, clustervar=NULL, ell=NULL, IK=TRUE) {
    X <- model.matrix(model)
    sum.model <- summary.lm(model)
    n <- sum(sum.model$df[1:2])
    K <- model$rank
    XXinv <- sum.model$cov.unscaled # XX^{-1}
    u <- residuals(model)

    df <- function(GG) {                # Compute DoF given G'*Omega*G
        sum(diag(GG))^2 / sum(GG * GG)
    }

    if(is.null(clustervar)) {           # no clustering
        Vhat <- vcovHC(model, type="HC2")
        Vhat.Stata <- Vhat*NA

        M <- diag(n)-X %*% XXinv %*% t(X)       # annihilator matrix
        GOG <- function(ell) {           # G'*Omega*G
            Xtilde <- drop(X %*% XXinv %*% ell / sqrt(diag(M)))
            crossprod(M * Xtilde)
        }
    } else {
        if(!is.factor(clustervar)) stop("'clustervar' must be a factor")

        ## Stata
        S <- length(levels(clustervar)) # number clusters
        uj <- apply(u*X, 2, function(x) tapply(x, clustervar, sum))
        Vhat.Stata <- S/(S-1) * (n-1)/(n-K) * sandwich(model, meat = crossprod(uj)/n)

        ## LZ2
```

```
    tXs <- function(s) {
        Xs <- X[clustervar==s, , drop=FALSE]
        MatSqrtInverse(diag(NROW(Xs))-Xs%*% XXinv %*% t(Xs)) %*% Xs
    }
    tX <- lapply(levels(clustervar), tXs) # list of matrices

    tu <- split(u, clustervar)
    tutX <- sapply(seq_along(tu),function(i) crossprod(tu[[i]],tX[[i]]))
    Vhat <- sandwich(model, meat = tcrossprod(tutX)/n)

    ## DOF adjustment
    tHs <- function(s) {
        Xs <- X[clustervar==s, , drop=FALSE]
        index <- which(clustervar==s)
        ss <- outer(rep(0,n),index)     # n x ns matrix of 0
        ss[cbind(index,1:length(index))] <- 1
        ss-X %*% XXinv %*% t(Xs)
    }
    tH <- lapply(levels(clustervar), tHs) # list of matrices

    Moulton <- function() {
        ## Moulton estimates
        ns <- tapply(u, clustervar,length)
        ssr <- sum(u^2)
        rho <- max((sum(sapply(seq_along(tu), function(i)
            sum(tu[[i]] %o% tu[[i]])))-ssr) / (sum(ns^2)-n), 0)
        c(sig.eps=max(ssr/n - rho, 0), rho=rho)
    }

    GOG <- function(ell) {
        G <- sapply(seq_along(tX),
                    function(i)  tH[[i]] %*% tX[[i]] %*% XXinv %*% ell)
        GG <- crossprod(G)

        if (IK==TRUE) {            # IK method
            Gsums <- apply(G, 2, function(x) tapply(x, clustervar, sum)) # Z'*G
            GG <- Moulton()[1]*GG + Moulton()[2]*crossprod(Gsums)
        }
        GG
    }
}

if (!is.null(ell)) {
    se <- drop(sqrt(crossprod(ell,Vhat) %*% ell))
    dof <- df(GOG(ell))
    se.Stata <- drop(sqrt(crossprod(ell,Vhat.Stata) %*% ell))
} else {
    se <- sqrt(diag(Vhat))
    dof <- sapply(seq(K), function(k) df(GOG(diag(K)[,k])))
    se.Stata <- sqrt(diag(Vhat.Stata))
}
names(dof) <- names(se)
```

```
    return(list(vcov=Vhat, dof=dof, adj.se=se*qt(0.975,df=dof)/qnorm(0.975),
                se=se,
                se.Stata=se.Stata))
}
```

Here is an example of R code using `BMlmSE()` to construct a 95% CI for the average treatment effect (ATE).

```
## For this example, assume:
## - We have an experiment with simple or complete randomization of subjects.
##   10 subjects are assigned to treatment and 5 to control.
## - Our PAP specified that we would estimate ATE using an OLS regression of the
##   outcome on the treatment dummy and a single covariate, but did not specify
##   how we would estimate the SE or construct a CI.

# Generate fake data

set.seed(1234567)
treatment <- c(rep(1, 10), rep(0, 5))
covariate <- rnorm(15)
outcome <- treatment + rnorm(15)

# Use lm() to fit an OLS regression

ols.fit <- lm(outcome ~ treatment + covariate, singular.ok = FALSE)

# Use BMlmSE() to compute the BM SE and degrees of freedom

## The argument ell must have the same length as the vector of regression coefficients.
## In this example, ell = c(0, 1, 0) indicates that we only want to compute
## the Bell-McCaffrey SE and degrees of freedom for the linear combination
##   0 * intercept + 1 * (coef on treatment) + 0 * (coef on covariate),
## i.e., the coefficient on treatment.

bm <- BMlmSE(model = ols.fit, ell = c(0, 1, 0))

# Construct a 95% confidence interval for the average treatment effect

point.estimate <- coef(ols.fit)['treatment']

critical.value <- qt(0.975, df = bm$dof)  # Critical value for 95% CI
margin.of.error <- critical.value * bm$se

ci = c(point.estimate - margin.of.error, point.estimate + margin.of.error)
names(ci) = c('lower','upper')

point.estimate  # Estimated average treatment effect
bm$se           # HC2 robust SE
bm$dof          # Bell-McCaffrey degrees of freedom
ci              # Bell-McCaffrey confidence interval
```

Avoiding regression models that do not allow the BM adjustment

The BM adjustment's SE estimators are undefined for some regression models:

- The HC2 robust SE (returned by `BMlmSE()` when `clustervar` is unspecified) is undefined if any observation has a leverage of 1, which happens when the regressors include a dummy variable that equals 1 for that observation and 0 for all others (or vice versa), and in other scenarios where the values of the regressors are such that the OLS solution will always fit that observation perfectly, no matter what the outcome values are.[6] In this situation, `BMlmSE()` returns values of `NaN` (Not a Number).
- The BM cluster-robust SE (returned by `BMlmSE()` when `clustervar` is supplied) is undefined when the matrix $I_i - H_{ii}$ (Bell and McCaffrey 2002, 7) for some cluster i is not of full rank, which happens, for example, when the regressors include a variable that would be constant if one cluster were dropped (e.g., a dummy variable for one cluster or for a subset of one cluster). In this situation, `BMlmSE()` prints the following error message: "Bell-McCaffrey SE undefined. This happens, e.g., when a dummy regressor is 1 for one cluster and 0 otherwise."[7]

We think these situations can usually be prevented by specifying parsimonious regression models and avoiding dummy covariates that identify one or two subjects or clusters. Also, they can be detected by attempting the BM adjustment with mock outcome data before treatment effects are estimated (since the leverages and the $I_i - H_{ii}$ matrices depend only on the regressors, not the outcome). If the PAP has specified a regression model that makes the BM adjustment undefined, we will use mock outcome data to help determine which covariates need to be dropped to make the BM adjustment well-defined.

Use of clustered standard errors

Our default approach is to use clustered SEs if and only if a regression includes multiple observations for some randomization units, and in that case, to cluster the SEs at the level of the randomization unit (not at any higher level). For example:

- When individual subjects are randomly assigned to treatment, the randomization unit is the subject. In the typical case where a regression includes only one observation per subject, we will not use clustered SEs. If a regression includes multiple observations per subject, we will use SEs clustered at the subject level.
- When households are randomly assigned to treatment, the randomization unit is the household. If a regression includes only one observation per household, we will not use clustered SEs. If a regression includes multiple observations per household (e.g., one observation per person, with more than one person in some households), we will use SEs clustered at the household level (not at any higher level such as county or state).

In cases where we use clustered SEs, our default SE estimator is Bell and McCaffrey (2002)'s bias-reduced cluster-robust SE. The SE and associated degrees of freedom can be computed using the `BMlmSE()` function above, with `IK = FALSE` and `clustervar` set to a factor variable whose values identify the clusters. (For example, if `hhid` is a variable that uniquely identifies households, then one can supply the argument `clustervar = as.factor(hhid)` to `BMlmSE()` to compute robust SEs clustered at the household level.)

Taking block randomization into account in SEs and CIs

Additional issues arise in block-randomized experiments:

[6] For example, if the sample includes only one libertarian entomologist, and the regressors include a dummy for "libertarian" and a dummy for "libertarian non-entomologist," then the libertarian entomologist's observation has a leverage of 1.

[7] Pustejovsky and Tipton (2016) provide a bias-reduced cluster-robust SE estimator that is always well-defined and is implemented in the `clubSandwich` R package. We have not yet used `clubSandwich`, but this seems a promising development.

1. If treatment assignment probabilities vary by block, simple comparisons of the treatment and control groups' average outcomes do not yield unbiased or consistent estimates of treatment effects. Some methods to address this problem are discussed in Gerber and Green (2012) (sections 3.6.1 and 4.5).

2. Regardless of whether treatment assignment probabilities vary by block, SEs and CIs may be too conservative if they do not take the blocking into account (Bruhn and McKenzie 2009).[8]

Weighted regression (Gerber and Green 2012, 116–17) addresses issue #1 but does not by itself address issue #2. Taking a weighted average of block-specific treatment effect estimates and estimating its SE (Gerber and Green 2012, 77) is a possible way to address both issues. It is not obvious how to calculate the degrees of freedom for the distribution of the resulting t-statistic (an issue that can matter in small samples). However, OLS regression with treatment–block interactions (Schochet 2010, sec. 4.2.3) can reproduce both the weighted average point estimate and its estimated SE. Let B be the number of blocks, and let X_b be a dummy for block b, for $b = 1, \ldots, B-1$ (omitting the dummy for block B to avoid collinearity). To compute a weighted average of the block-specific treatment–control differences in means, with weight w_b on block b and $\sum_{b=1}^{B} w_b = 1$, one can use the estimated coefficient on the treatment dummy T in an OLS regression of the outcome Y on T, $X_1, \ldots, X_{B-1}$, and $T \cdot (X_1 - w_1), \ldots, T \cdot (X_{B-1} - w_{B-1})$. In simulations, we have found that the HC2 robust SE for the coefficient on T in this regression is equivalent to the SE estimator in Gerber and Green (2012) (p. 77).

Our default methods for analyzing block-randomized experiments rely on OLS regression, using the Bell–McCaffrey adjustment for SEs, degrees of freedom, and CIs. Whether we include block dummies and treatment–block interactions in the regression will depend on sample sizes and other aspects of the experimental design, as described below. We do not attempt to cover every possible design here, but focus on some commonly used ones. The subsections that follow assume there are only two treatment arms (i.e., either one treatment group and one control group, or two treatment groups without an additional control group). We plan to cover designs with three or more treatment arms in a future version of the SOP.

As always, PAPs are free to deviate from the default methods described below. They may also use these defaults as a starting point and add other covariates to the regression model.

In what follows, let $\overline{X}_b$ denote the sample mean of X_b, for $b = 1, \ldots, B-1$. That is, $\overline{X}_b = N_b/N$, where N is the total number of subjects in the experimental sample and N_b is the number of subjects in block b.

Block randomization of individuals, with treatment assignment probabilities constant across blocks

If each block has at least 5 individuals assigned to each treatment arm,[9] we will estimate the average treatment effect by running an OLS regression of Y on T, $X_1, \ldots, X_{B-1}$, and $T \cdot (X_1 - \overline{X}_1), \ldots, T \cdot (X_{B-1} - \overline{X}_{B-1})$.

If some block has fewer than 5 individuals assigned to some treatment arm:

- If the blocks were formed by coarsening a continuous variable X, we will estimate the ATE by running an OLS regression of Y on T, X, and $T \cdot (X - \overline{X})$, where $\overline{X}$ is the sample mean of X.
- Otherwise, we will not account for blocking in our SEs and CIs unless the PAP specified a method for doing so.

[8] Blocking tends to improve precision if the blocking criterion is correlated with the outcome. Thus, SEs that ignore the effects of blocking tend to be conservative. However, if the blocking criterion is only weakly correlated with the outcome, then this conservatism is only mild, and accounting for blocking may reduce power by reducing degrees of freedom (Imbens 2011). Thus, if blocking is done for logistical reasons and not for precision improvement, a researcher preparing a PAP may decide to deviate from the SOP and ignore the blocking in the analysis.

[9] There are two main reasons for requiring a minimum number of individuals per treatment arm per block here: (1) If some block has fewer than 2 observations in some arm, the usual SE estimator for a weighted average of block-specific differences in means (Gerber and Green 2012, 77) is undefined and so is the HC2 robust SE for the equivalent regression coefficient. We use a minimum of 5 to allow for some attrition. (2) If attrition reduces the number of observations to 1 for some block, then including that block's dummy is equivalent to dropping the observation. E.g., in a matched pair design, if pair dummies are included in the regression, then whenever one subject in a pair has missing outcome data, the other subject's observation is effectively dropped. Dropping such observations can lead to bias (Gerber and Green 2012, 241–43).

Block randomization of individuals, with varying treatment assignment probabilities

If each block has at least 5 individuals assigned to each treatment arm:

1. Except in the extreme case described in item #2 below, we will use the same method as in the preceding subsection: estimate the average treatment effect by running an OLS regression of Y on T, $X_1, \ldots, X_{B-1}$, and $T \cdot (X_1 - \overline{X}_1), \ldots, T \cdot (X_{B-1} - \overline{X}_{B-1})$.

2. An alternative approach, sometimes called the least-squares dummy variable (LSDV) method, is to run an OLS regression of the outcome on treatment and block dummies (without interactions). 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, 256; Angrist and Pischke 2009, 75; Gerber and Green 2012, 119). In contrast, the estimand for the method in item #1 above is the unweighted ATE, which gives block j a total weight proportional to N_j. In the extreme case where there is at least one block j such that

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

   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.[10] See the section "Analysis of block randomized experiments with treatment probabilities that vary by block" under "Using covariates in analysis" for discussion of additional issues that arise when the LSDV approach is used.

If some block has fewer than 5 individuals assigned to some treatment arm, then, in the extreme case described in item #2 above, we will again use the LSDV method, but otherwise our default methods are:

- If each block has a different treatment probability, use the same method as in item #1 above.
- Otherwise, reduce the number of dummy variables in the regression by replacing block dummies with treatment probability dummies. Let $C < B$ be the number of possible treatment probabilities; let $Z_1, \ldots, Z_{C-1}$ be a set of dummy variables for all but one treatment probability; and let $\overline{Z}_c$ be the sample mean of Z_c, for $c = 1, \ldots, C-1$. If the blocks were formed by coarsening a continuous variable X, estimate the average treatment effect by running an OLS regression of Y on T, $Z_1, \ldots, Z_{C-1}$, X, $T \cdot (Z_1 - \overline{Z}_1), \ldots, T \cdot (Z_{C-1} - \overline{Z}_{C-1})$, and $T \cdot (X - \overline{X})$, where $\overline{X}$ is the sample mean of X. If the blocks were not formed by coarsening a continuous variable, use the same approach but omit the terms X and $T \cdot (X - \overline{X})$.

Block-randomized treatment/placebo designs

In a treatment/placebo design (Gerber and Green 2012, 161–64), the planned analysis compares treatment group compliers with placebo group compliers. Block randomization (if used) can ensure that the block composition of the entire treatment group matches that of the entire placebo group, but cannot ensure the same degree of balance between treatment group compliers and placebo group compliers. In fact, it is possible for a block to contain one or more compliers in the treatment group but none in the placebo group, or vice versa. We therefore do not use block dummies in our default methods for analyzing block-randomized treatment/placebo experiments.

If the treatment assignment probability is constant across blocks:

[10] 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 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 in the ATE than in the LSDV estimand. In this situation, we would use LSDV.

- If the blocks were formed by coarsening a continuous variable X, we will estimate the ATE by running an OLS regression of Y on T, X, and $T \cdot (X - \overline{X})$, where $\overline{X}$ is the mean of X among all compliers in the treatment and placebo groups.
- Otherwise, we will not account for blocking in our SEs and CIs unless the PAP specified a method for doing so.

If the treatment assignment probabilities vary across blocks, we will use the same method as in the final bullet of the above section on "Block randomization of individuals, with varying treatment assignment probabilities," except that $\overline{Z}_c$ and $\overline{X}$ should be defined as the means of Z_c and X among compliers only.

Block randomization of clusters

In a cluster-randomized experiment, if the number of individuals varies across clusters, researchers should decide whether their estimand gives equal weight to each individual or to each cluster (Green and Vavreck 2008; Imbens 2013). If the estimand gives equal weight to each cluster (or if the clusters are all of the same size, in which case the two estimands are equivalent), it is straightforward to adapt the methods from the sections above on "Block randomization of individuals": Using cluster-level averages as the observations, run an OLS regression with one observation per cluster, and substitute "cluster" for "individual" or "subject" in the descriptions of the methods (e.g., change "If each block has at least 5 individuals assigned to each treatment arm" to "If each block has at least 5 clusters assigned to each treatment arm").

If the cluster sizes vary and the estimand gives equal weight to each individual, our default methods use OLS regression with one observation per individual (and, as noted earlier, our default SE estimator in this case is the Bell and McCaffrey (2002) bias-reduced cluster-robust SE):

1. If each block has at least 20 clusters assigned to each treatment arm,[11] or if each block has a different treatment probability, estimate the average treatment effect by running an OLS regression of Y on T, $X_1, \ldots, X_{B-1}$, and $T \cdot (X_1 - \overline{X}_1), \ldots, T \cdot (X_{B-1} - \overline{X}_{B-1})$.
2. If the conditions for item #1 are not met:

- If the treatment assignment probability is constant across blocks: If the blocks were formed by coarsening a continuous variable X, we will estimate the ATE by running an OLS regression of Y on T, X, and $T \cdot (X - \overline{X})$, where $\overline{X}$ is the sample mean of X.[12] If the blocks were not formed by coarsening a continuous variable, we will not account for blocking in our SEs and CIs unless the PAP specified a method for doing so. (If no regression model is specified in the PAP, we will regress Y on T.)
- If the treatment assignment probability varies across blocks: If the blocks were formed by coarsening a continuous variable X, we will estimate the ATE by running an OLS regression of Y on T, $Z_1, \ldots, Z_{C-1}$, X, $T \cdot (Z_1 - \overline{Z}_1), \ldots, T \cdot (Z_{C-1} - \overline{Z}_{C-1})$, and $T \cdot (X - \overline{X})$ (where, as before, C is the the number of possible treatment probabilities, $Z_1, \ldots, Z_{C-1}$ are a set of dummy variables for all but one treatment probability, and overlines denote sample means). If the blocks were not formed by coarsening a continuous variable, we will use the same approach but omit the terms X and $T \cdot (X - \overline{X})$.

[11] There are three main reasons for requiring a minimum number of clusters per treatment arm per block here. Two of them are analogous to those given in a footnote in the section on "Block randomization of individuals, with treatment assignment probabilities constant across blocks." Additionally, when cluster sizes vary within blocks and the number of clusters per block is small, controlling for blocks may increase bias (Middleton and Aronow 2015).

[12] In some cases, precision may be improved by estimating a regression model that allows the between-cluster relationship between X and Y to differ from the within-cluster relationship (Raudenbush 1997, 181–82; Klar and Darlington 2004). We have not adopted such a regression as our default method, but it may be worth considering for some PAPs.

One-tailed or two-tailed test?

We will report two-tailed significance tests unless the PAP specifies a one-tailed test or some other approach.[13]

Studentized permutation test

For significance tests and p-values, our default method is a Studentized permutation test (Chung and Romano 2013; Janssen 1997; Romano 2009, sec. 3) that compares the t-statistic for the average treatment effect (i.e., the estimated ATE divided by its estimated SE) with its empirical distribution under random reassignments of treatment that follow the same randomization scheme as the actual experiment.[14] We will use 10,000 randomizations and the random number seed 1234567. 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, 33).

Here is an example of R code for a Studentized permutation test.

```
## For this example, assume:
## - We have an experiment with complete randomization of 200 subjects.
##   50 subjects are assigned to treatment and 150 to control.
## - Our PAP specified that we would estimate ATE using an OLS regression of the
##   outcome on the treatment dummy and a single covariate.
## - Either the PAP specified that we would use the HC2 SE, or
##   it did not specify an SE estimator so we are following the SOP and using HC2.

library(randomizr)
library(sandwich)

set.seed(1234567)

n.rands <- 10000  # no. of randomizations for permutation test
N <- 200          # no. of subjects
N.treated <- 50   # no. of subjects assigned to treatment

## Generate "observed" data for this example

treated <- c(rep(1, N.treated), rep(0, N - N.treated))
covariate <- rnorm(N)
outcome <- 0.1 * treated + rnorm(N)

## Run the pre-specified regression with the observed data.
## Save the observed t-statistic for ATE.

fit <- lm(outcome ~ treated + covariate, singular.ok = FALSE)

est.ate <- coef(fit)['treated']  # Estimated average treatment effect
se <- sqrt( vcovHC(fit, type = 'HC2')['treated','treated'] )  # HC2 robust SE
```

[13] 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 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) for a sympathetic view and Hurlbert and Lombardi (2012) for a critical view of asymmetric tests.

[14] This test differs from the classic Fisher–Pitman permutation test in that the test statistic is the t-statistic, not the estimated ATE itself.

```
t.observed <- est.ate / se

# Permutation test

t.sim <- rep(NA, n.rands)

for (i in 1:n.rands) {
  ## Simulate random assignment of treatment

  treated.sim <- complete_ra(N, m = N.treated)

  ## Run the pre-specified regression with the simulated data.
  ## Save the simulated t-statistic for ATE.

  fit.sim <- lm(outcome ~ treated.sim + covariate, singular.ok = FALSE)
  est.ate.sim <- coef(fit.sim)['treated.sim']
  se.sim <- sqrt( vcovHC(fit.sim, type = 'HC2')['treated.sim','treated.sim'] )

  t.sim[i] <- est.ate.sim / se.sim
}

## "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.sim <= t.observed)
p.right <- mean(t.sim >= t.observed)

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

We recommend attempting the permutation test with mock outcome data and actual covariate data before analyzing the actual outcome data. The mock permutation test may reveal that on some randomizations, the t-statistic cannot be computed because the regressors are collinear or because the HC2 or BM SE is undefined (see the section above on "Avoiding regression models that do not allow the BM adjustment"). In such cases, covariates should be dropped from the model until the mock permutation test runs without errors.

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 an experiment with random assignment of individuals. Let N be the number of subjects, and 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 (2012) 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), singular.ok = FALSE)

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

```
##
## t test of coefficients:
##
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  1.787658   0.088533 20.1921  < 2e-16 ***
## Z            0.272489   0.135330  2.0135  0.04544 *
## X1_c         0.704749   0.066646 10.5745  < 2e-16 ***
## X2_c         2.959756   0.180087 16.4351  < 2e-16 ***
## Z:X1_c      -0.083130   0.107644 -0.7723  0.44090
## Z:X2_c       0.225950   0.270757  0.8345  0.40502
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

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

```
##
## t test of coefficients:
##
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  2.03830    0.18976 10.7414   <2e-16 ***
## Z           -0.26248    0.27692 -0.9479   0.3444
## ---
## 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.[15]

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.

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.[16] 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.[17] However, if observed imbalances are allowed to influence the choice of covariates,[18] the following guidelines should be observed:

[15] 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.

[16] 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."

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

[18] 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.

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.[19] 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

# 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:
```

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

```r
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), singular.ok = FALSE)

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

```
##
## t test of coefficients:
##
##                 Estimate Std. Error t value  Pr(>|t|)
## (Intercept)     1.524330   0.187873  8.1136 5.737e-14 ***
## Z               0.197747   0.271709  0.7278    0.4676
## X1_c            0.098306   0.170398  0.5769    0.5647
## X2_c            0.526768   0.384638  1.3695    0.1724
## X2_missing_c   -0.075837   0.471082 -0.1610    0.8723
## Z:X1_c         -0.067737   0.249650 -0.2713    0.7864
## Z:X2_c         -0.506846   0.549334 -0.9227    0.3573
## Z:X2_missing_c -0.118223   0.684370 -0.1727    0.8630
## ---
## 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),[20] we will provide supplementary

[20]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) and Judkins and Porter (2016).

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.[21]

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, 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

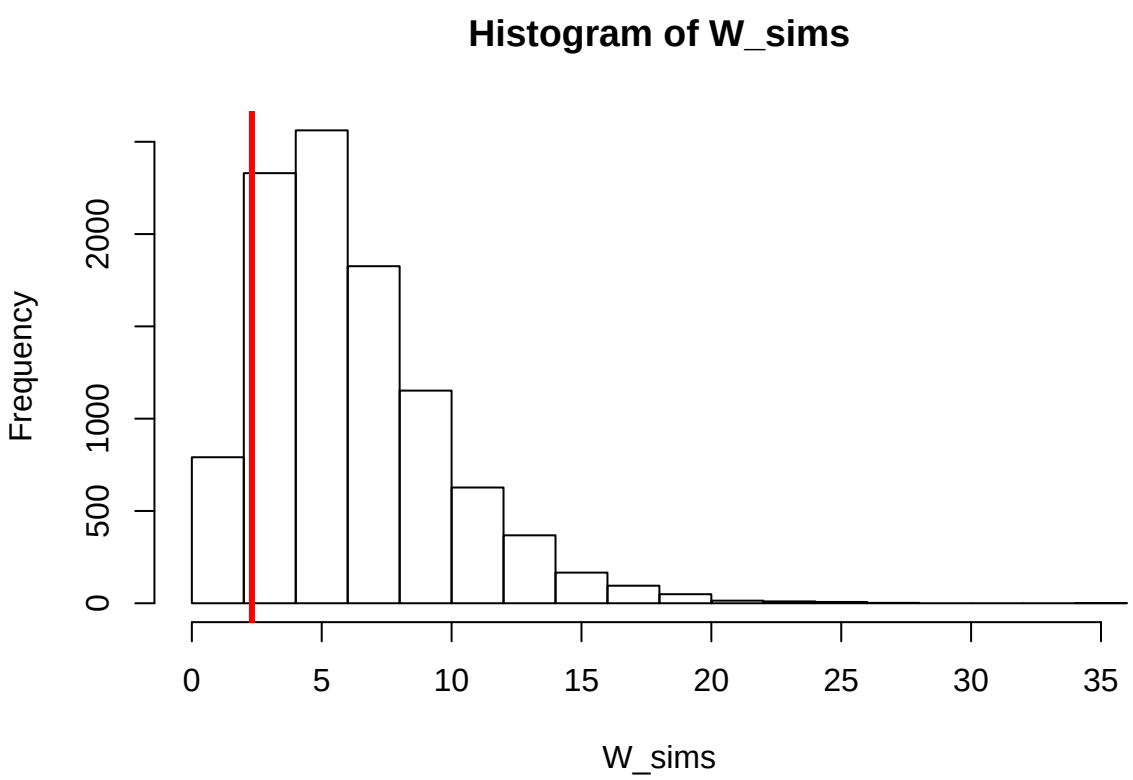
```
suppressMessages({
  library(randomizr)
  library(sandwich)
})

# Generate Covariates

set.seed(1234567)

N <- 1000

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
```

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

```r
fit <- lm(Z ~ gender + age + lincome + party + education, singular.ok = FALSE)

# 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, singular.ok = FALSE)

  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

The section "Taking block randomization into account in SEs and CIs" describes our default methods for estimating treatment effects in block randomized experiments and explains that we would use the least-squares dummy variable (LSDV) method in certain extreme cases.

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 in block j a weight proportional to $1 - P_j$, and give each control group observation in that block a weight proportional to P_j, where P_j is the probability of assignment to treatment. (This results in a weighted treatment group mean and weighted control group mean that give block j a total weight proportional to $N_j P_j (1 - P_j)$, where N_j is the number of subjects in block j. Thus, the weighting is the same as in the LSDV estimand.)

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

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 $N_jP_j(1-P_j)$, the same as in 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

See also the subsection "Block-randomized treatment-placebo designs" under "Taking block randomization into account in SEs and CIs".

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, 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, 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, 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, 226–27). (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.)

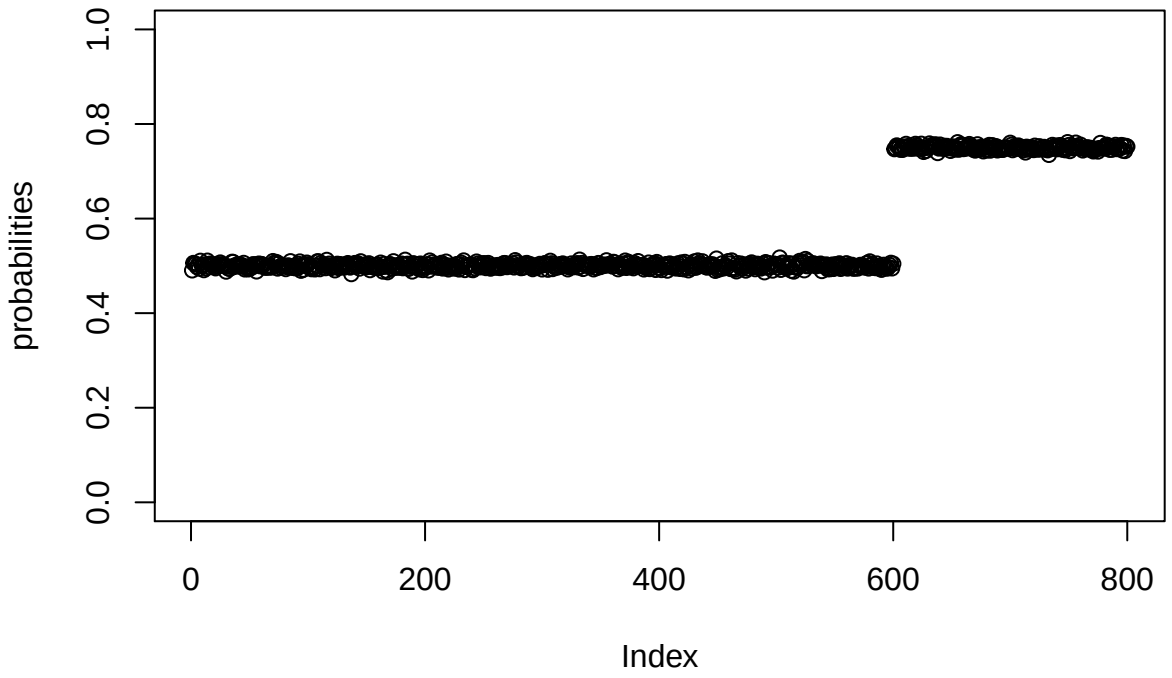
Example: Fundraising Experiment

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.

```
# 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.[22]

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".)

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%.

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

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."

Bell, Robert M., and Daniel F. McCaffrey. 2002. "Bias Reduction in Standard Errors for Linear Regression with Multi-Stage Samples." *Survey Methodology* 28 (2): 169–81. http://www.statcan.gc.ca/pub/12-001-x/2002002/article/9058-eng.pdf.

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.

Cameron, A. Colin, and Douglas L. Miller. 2015. "A Practitioner's Guide to Cluster-Robust Inference." *Journal of Human Resources* 50 (2): 317–72. doi:10.3368/jhr.50.2.317.

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.

Efron, B. 1986. "Discussion: Jackknife, Bootstrap and Other Resampling Methods in Regression Analysis." *Annals of Statistics* 14 (4): 1301–4. http://www.jstor.org/stable/2241457.

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 S., 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.

Green, Donald P., and Lynn Vavreck. 2008. "Analysis of Cluster-Randomized Experiments: A Comparison of Alternative Estimation Approaches." *Political Analysis* 16 (2): 138–52. http://www.jstor.org/stable/25791925.

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.

Greenland, Sander, Stephen J. Senn, Kenneth J. Rothman, John B. Carlin, Charles Poole, Steven N. Goodman, and Douglas G. Altman. 2016. "Statistical Tests, P-values, Confidence Intervals, and Power: A Guide to Misinterpretations." *European Journal of Epidemiology* 31 (4): 337–50. doi:10.1007/s10654-016-0149-3.

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.

Hurlbert, Stuart H., and Celia M. Lombardi. 2012. "Lopsided Reasoning on Lopsided Tests and Multiple Comparisons." *Australia and New Zealand Journal of Statistics* 54 (1): 23–42. doi:10.1111/j.1467-842X.2012.00652.x.

Imbens, Guido W. 2011. "Experimental Design for Unit and Cluster Randomid Trials." https://web.archive.org/web/20160329102918/http://cyrussamii.com/wp-content/uploads/2011/06/Imbens_June_8_paper.pdf.

———. 2013. "Design and Analysis of Randomized Experiments: The Costs and Benefits of Stratification [slides]." http://ps-experiments.ucr.edu/conferences/imbens.pdf.

Imbens, Guido W., and Michal Kolesár. forthcoming. "Robust Standard Errors in Small Samples: Some Practical Advice." *Review of Economics and Statistics*. https://www.princeton.edu/~mkolesar/papers/small-robust.pdf.

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. 2016. "Robustness of Ordinary Least Squares in Randomized Clinical Trials." *Statistics in Medicine* 35 (11): 1763–73. doi:10.1002/sim.6839.

Klar, Neil, and Gerarda Darlington. 2004. "Methods for Modelling Change in Cluster Randomization Trials." *Statistics in Medicine* 23 (15): 2341–57. doi:10.1002/sim.1858.

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." *Annals of Applied Statistics* 7 (1): 295–318. doi:10.1214/12-AOAS583.

Lin, Winston, and Donald P. Green. forthcoming. "Standard Operating Procedures: A Safety Net for Pre-Analysis Plans." *PS: Political Science and Politics*. http://www.stat.berkeley.edu/~winston/sop-safety-net.pdf.

McCaffrey, Daniel F., Robert M. Bell, and Carsten H. Botts. 2001. "Generalizations of Biased Reduced Linearization." https://web.archive.org/web/20151110163144/http://www.amstat.org/sections/SRMS/Proceedings/y2001/Proceed/00264.pdf.

Middleton, Joel A., and Peter M. Aronow. 2015. "Unbiased Estimation of the Average Treatment Effect in Cluster-Randomized Experiments." *Statistics, Politics and Policy* 6 (1–2): 39–75. doi:10.1515/spp-2013-0002.

Moher, David, Sally Hopewell, Kenneth F. Schulz, Victor Montori, Peter C. Gøtzsche, P. J. Devereaux, Diana Elbourne, Matthias Egger, and Douglas G. Altman. 2010. "CONSORT 2010 Explanation and Elaboration: Updated Guidelines for Reporting Parallel Group Randomised Trials." *BMJ* 340: c869. doi:10.1136/bmj.c869.

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.

Pustejovsky, James E., and Elizabeth Tipton. 2016. "Small Sample Methods for Cluster-Robust Variance Estimation and Hypothesis Testing in Fixed Effects Models." http://blogs.cuit.columbia.edu/let2119/files/2016/01/ClusterRobustTesting_FE_models.pdf.

Raudenbush, Stephen W. 1997. "Statistical Analysis and Optimal Design for Cluster Randomized Trials." *Psychological Methods* 2 (2): 173–85. doi:10.1037/1082-989X.2.2.173.

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. doi:10.1007/978-1-4419-1213-8.

Schochet, Peter Z. 2010. "Is Regression Adjustment Supported by the Neyman Model for Causal Inference?" *Journal of Statistical Planning and Inference* 140 (1): 246–59. doi:10.1016/j.jspi.2009.07.008.

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

Vazire, Simine. 2016. "Editorial." *Social Psychological and Personality Science* 7 (1): 3–7. doi:10.1177/1948550615603955.

Wasserstein, Ronald L., and Nicole A. Lazar. forthcoming. "The ASA's Statement on P-Values: Context, Process, and Purpose." *American Statistician*. doi:10.1080/00031305.2016.1154108.

Wooldridge, Jeffrey M. 2010. *Econometric Analysis of Cross Section and Panel Data*. 2nd ed. MIT Press. http://site.ebrary.com/lib/columbia/detail.action?docID=10453042.

Young, Alwyn. 2016. "Improved, Nearly Exact, Statistical Inference with Robust and Clustered Covariance Matrices Using Effective Degrees of Freedom Corrections." http://personal.lse.ac.uk/YoungA/Improved.pdf.