Green Lab SOP

Donald P. Green, Winston Lin, and Alexander Coppock

Version 0.2: June 9, 2015

Contents

This standard operating procedure describes the default practices of the experimental lab group lead by Donald P. Green at Columbia University. This guide is meant as a stopgap, not a replacement for rigorous project-specific pre-analysis plans (PAPs). In particular, many experiments encounter problems in implementation or unforseen logistical complications and so veer into territory undescribed by the PAP.

This is a living document. If ever you encounter an experimental situation not covered herewithin, please email or submit an issue request on GitHub. Additionally, when referencing this document, please be sure to note the version.

When randomization doesn't go according to plan

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. In these situations, we will use randomization inference, excluding the disallowed randomizations.

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 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 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.)

Outliers

Except as specified in the PAP, we will not delete or edit outlying values merely because they are large. 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.

Reliance on Permutation Methods

For significance tests and p-values, we will either report Studentized permutation tests (Chung and Romano 2013) or use permutation methods to check the accuracy of asymptotic approximations.

For an example of the former, see below under "Attrition." For an example of the latter, 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.

Attrition

We will routinely perform three types of checks for asymmetrical attrition:[1]

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[2] 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] We will use at least 10,000 randomizations, and our seed for the random number generator will be 1234567.
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 (e.g., Gerber and Green (2012, 107)). As in check #2, we will implement the test as a Studentized permutation test, using at least 10,000 randomizations and the seed 1234567.

In checks #2 and #3, p-values below 0.10 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 worst-case bounds,[4] 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.

[1] 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 "Covariate adjustment."

[2] Wooldridge (2010, 62).

[3] Note that in the case of a two-arm trial, the Studentized permutation test does not compare the estimated treatment effect with its empirical distribution, but instead compares a heteroskedasticity-robust t-statistic with its empirical distribution. For background and motivation, see Romano (2009) and Chung and Romano (2013).

[4] Aronow et al. (2015).

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. (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.

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.

When some subjects receive "partial treatment"

(Gerber and Green 2012, 164–65) discuss several approaches for estimating treatment effects when some subjects receive a full dose of the intended intervention and others receive only part of it. Unless variation in treatment dosage is randomized, we will follow the approach where "the researcher simply considers all partially treated subjects as fully treated" (Gerber and Green 2012, 165).

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 classified as 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 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 0.10 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 of the hypothesis that all coefficients on the covariates are zero. A p-value below 0.10 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.

Covariate adjustment

Default estimation methods

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.)

Choice of covariates

Ordinarily our choice of covariates for adjustment will have been specified in the PAP. 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.[5]

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,

[5]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.

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.[6] 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.[7] However, if observed imbalances are allowed to influence the choice of covariates,[8] 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 "Robustness checks" below.

Missing covariate values

Observations with missing covariate values will be included in the analysis 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. 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.

Robustness Checks

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, and if alternative estimates differ substantially, we will note the lack of robustness and be appropriately restrained in our conclusions.

[6] 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."

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

[8] 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.

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),[9] 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.[10]

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.[11]

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.

Pre-specified analyses that don't appear in the published paper

We will make all such analyses available in a document that will be referenced in the paper.

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 a recognizable caller ID (e.g., "Vote '98 Headquarters") would count as contact.

[9] 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 (2014).

[10] Freedman (2008); Firth and Bennett (1998). Lin gave an informal discussion in a blog comment.

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

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.

References

nocite: |

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 Ke 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. http://www.amazon.com/Field-Experiments-Design-Analysis-Interpretation/dp/0393979954.

Greenberg, D., and B. 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.

Judkins, D. R., and K. E. Porter. 2014. "Robustness of Ordinary Least Squares in Randomized Clinical Trials."

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.

Nickerson, David W. 2005. "Scalable protocols offer efficient design for field experiments." *Political Analysis* 13 (3): 233–52. doi:10.1093/pan/mpi015.

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.

Wooldridge, Jeffrey M. 2010. *Econometric Analysis of Cross Section and Panel Data*. MIT Press. https://books.google.com/books/about/Econometric/_Analysis/_of/_Cross/_Section/_an.html?id=yov6AQAAQBAJ/&pgis=1.