Key Items to Get Right When Conducting Randomized Controlled Trials of Social Programs

February 2016

This publication was produced by the Evidence-Based Policy team of the Laura and John Arnold Foundation (now Arnold Ventures).

We welcome comments and suggestions on this document (jbaron@arnoldventures.org).

Purpose

This is a checklist of key items to get right when conducting a randomized controlled trial (RCT) to evaluate a social program or practice.

The checklist is designed to be a practical resource for researchers and sponsors of research. It describes items that are critical to the success of an RCT in producing valid findings about a social program's effectiveness. This document is limited to key items, and does not address all contingencies that may affect a study's success.[1]

Items in this checklist are categorized according to the following phases of an RCT:

1. Planning the study;
2. Carrying out random assignment;
3. Measuring outcomes for the study sample; and
4. Analyzing the study results.

1. Key items to get right in planning the study

☐ **Choose (i) the program to be evaluated, (ii) target population for the study, and (iii) key outcomes to be measured. These should include, wherever possible, ultimate outcomes of policy importance.**

As illustrative examples:

- An RCT of a pregnancy prevention program preferably should measure outcomes such as pregnancies or births, and not just intermediate outcomes such as condom use.
- An RCT of a remedial reading program preferably should measure outcomes such as reading comprehension, and not just participants' ability to sound out words.
- An RCT of a video surveillance and gunshot sensor system in high-crime areas should preferably measure outcomes such as violent crime rates, and not just gunshots detected.

Measuring ultimate outcomes is important because, in many cases, intermediate outcomes are not a reliable predictor of the ultimate outcomes of policy importance.

☐ **Decide whether the study should randomly assign individuals (*e.g.,* students), or clusters of individuals (*e.g.,* schools, communities).**

Random assignment of individuals is usually the most efficient and least expensive approach. However, it may be necessary to randomly assign clusters rather than, or in addition to, individuals in situations such as the following:

(a) The program may have sizable "spillover" effects on nonparticipants.

For example, a handwashing campaign to prevent childhood disease (*e.g.,* diarrhea, parasites) in developing countries could well reduce disease not only among children who wash their hands,

but also among their siblings and peers (due to the presence of fewer carriers). In such a case, an RCT would likely need to randomly assign whole communities to treatment and control groups to accurately determine the program's effect. An RCT that only randomizes individual children within a community to treatment and control groups will underestimate the program's effect to the extent the program reduces disease among control-group children.

– or –

(b) The program is delivered at a cluster level – such as a school-wide reform program for low-performing schools, or a policing strategy implemented in crime hotspots.

To evaluate such programs, an RCT will generally need to randomly assign a sufficient sample of clusters (*e.g.,* schools, crime hotspots) to treatment and control groups in order to create two such groups that are similar in all key respects except that the treatment clusters receive the program while the control clusters do not.[2]

☐ **Conduct a statistical ("power") analysis to estimate the minimum number of individuals or clusters to randomize in order to determine whether the program has a meaningful effect.**

The purpose of conducting a power analysis is to determine how many individuals or clusters must be randomized in order to have confidence that the study will detect meaningful effects of the program, should they exist. The analysis will require making a judgment about the minimum effect size you seek to detect, based on such factors as (i) the smallest effect that would cause policy officials to view the program as worthwhile and meriting expansion, considering its cost and other factors; and (ii) the program's likely effect size based on prior studies. It is important that the power analysis take into account key features of the study design, such as whether individuals or clusters will be randomly assigned, as described in the endnote[3] of this document. The endnote also includes links to helpful online tools for conducting a power analysis.

☐ **Engage the staff of the program being evaluated ("program staff") as partners in the RCT, helping them to understand that the study is important, ethical, and practical.**

Partnering with program staff is usually essential, for several reasons. First, their hands-on understanding of the program's operations and target population can be of great help in planning key aspects of the study, such as (i) how to embed random assignment in program operations in a way that is effective and minimally intrusive; and (ii) how to successfully recruit individuals to enroll in the study. In addition, you will often need program staff to help with (or at least not undermine) key elements of study implementation, such as (i) carrying out random assignment; and (ii) ensuring control group members do not participate in the program. Strategies for enlisting the support of program staff are beyond the scope of this document, but are addressed by two excellent resources cited in the endnote.[4]

☐ **Develop and publicly register a detailed plan of how the study outcomes will be analyzed (*i.e.,* "pre-specify" the analysis).**

Pre-specification of the analysis, prior to random assignment, is important for the following reason. If an RCT estimates a program's effect on multiple outcomes and/or multiple subgroups, it is likely to find statistically-significant effects that are merely the result of chance – *i.e.*, false-positives. This is because, for each estimated effect, the test for statistical significance has about a 1 in 20 chance of

producing a false-positive. Pre-specification prevents researchers from estimating a large number of effects, and then selectively citing the few that are statistically significant as evidence of a program's effectiveness (when they could be false-positives). It also prevents researchers from trying many different statistical methods – *e.g.,* to adjust for missing outcome data – and then selecting the one that produces their hoped-for findings (when most other methods yield a different conclusion).

Thus, studies should pre-specify:

(a) **One or a few primary hypotheses about the program's effects that the study will test**, including the primary outcome(s) and subgroups (if any) to be examined and the hypothesized direction of the effect(s).

If the study will estimate more than one effect in testing the primary hypotheses, the study should specify an appropriate statistical method it will use to adjust for estimating multiple effects.

(b) **Secondary hypotheses that the study will also test but give less evidentiary weight**, recognizing that effects found on secondary outcomes and subgroups could be the result of chance, as discussed above.

(c) **Statistical methods that the study will use to estimate the effects on the specified outcomes and subgroups.**

These include, for example, the methods that the study will use to estimate the treatment's effects for the full sample, as well as for any subgroups; to test the statistical significance of these effects; to adjust for missing data (*e.g.,* due to sample attrition); and to adjust for pre-program differences between the treatment and control groups. The description of these methods should have sufficient detail to allow an independent party to replicate the findings if they had access to the relevant data.

Prior to study inception, the analysis plan should be registered at an appropriate website, such as Open Science Framework, clinicaltrials.gov, American Economic Association RCT registry, or Institute of Education Sciences Registry of RCTs.[5] (The Laura and John Arnold Foundation requires all research grantees to register their plans on the Open Science Framework, at a minimum.)

2. Key items to get right in the random assignment process

☐ **Take steps to protect the integrity of the random assignment process, including:**

(a) Have someone without a vested interest in the study outcome conduct random assignment; or, if it is carried out by program staff, provide training and protocols to ensure it is done faithfully.

In many studies it will be necessary to have program staff, typically those who enroll individuals in the program, carry out random assignment. In such cases, it is vital to:

- Train staff on how to conduct random assignment, preferably using a web-based system that precludes an individual being assigned twice;

- Provide scripts, protocols, and training to staff on how to explain the study – including random assignment – to prospective sample members, and how to inform control-group members of their status;
- Train staff on how to obtain individuals' consent to participate in the study (if such consent is required), and to do so prior to random assignment, as discussed below;
- Ensure that staff cannot interfere with random assignment – for example, by being able to anticipate whether the next assignment will be treatment or control, and to change the order in which individuals are assigned; and
- Train staff on how to prevent control group members from participating in the program and explain that such participation would make the program appear less effective.

(b) Use an assignment method that is truly random – *e.g.,* computer-generated random numbers, rather than the first digit of an individual's social security number (which correlates with the individual's place of birth).[6]

(c) If Institutional Review Board or other rules make it necessary to obtain individuals' consent to participate in the study, obtain such consent prior to random assignment (or at least prior to sample members' learning whether they are in the treatment or control group).

If obtained afterward, sample members' knowledge of their group assignment might affect their decision on whether to consent, thus undermining the equivalence of the treatment and control groups.

(d) If clusters (*e.g.,* classrooms) are randomized, ensure that the placement of individuals *(e.g.,* students) within these clusters is unaffected by whether the cluster is in the treatment or control condition.

For example, if a study randomizes classrooms in order to evaluate a new classroom curriculum, it should ensure that the school leadership is not (for example) disproportionately placing its higher-need students in the treatment classrooms, thus undermining the equivalence of the treatment and control groups. Possible steps to prevent this include: (i) placing students in classrooms *prior to* randomly assigning classrooms; or (ii) randomly assigning students to classrooms (in addition to randomly assigning classrooms to treatment and control groups).

☐ **Obtain the following data on each sample member randomly assigned (or member of a cluster that is randomly assigned):**

(a) Information needed to track the sample member over time, in order to measure and analyze his or her outcomes, including:

- A unique personal identifier (*e.g.,* name, social security number, and/or birthdate);
- The date on which the individual was randomly assigned;
- Whether the individual was assigned to the treatment or control group;
- The random assignment ratio applied to the individual (*e.g.,* 50:50 chance of being assigned to the treatment versus control group, or 60:40 chance); and,
- If study outcomes will be measured with a follow-up survey, the individual's contact information as well as that of friends or family who will know how to reach the individual if his or her contact information changes.

(b) Where feasible, pre-program measures of the outcomes that the program seeks to affect (*e.g.,* in a job training RCT, earnings for two years prior to random assignment), as well as any other descriptive data you wish to obtain on the sample.

Such pre-program data are very useful in (i) describing the population being studied, so that readers of the study will know to whom its findings apply (*e.g.,* workers whose average earnings are $27,000 per year, and whose average age is 35); (ii) confirming whether random assignment successfully created treatment and control groups that are highly similar in key pre-program characteristics; (iii) enabling the study to detect smaller program effects with a given sample size (because such data can later be used in the study analysis to control for variation in outcomes among sample members that is not due to the program); and (iv) enabling analyses of the program's effects on subgroups (*e.g.,* workers with pre-program earnings below the poverty level, workers who are female). Such data should be obtained prior to the time of random assignment, since data collected after that point could reflect the effects of the program.

☐ **Monitor what services the control group members are receiving, and take steps to prevent – or at least minimize instances of – their participation in the program.**

Monitoring what services the control group receives is important, first, for interpreting the study's results. Specifically, the study will produce an estimate of the program's effect *compared to* the control condition. So, to understand that estimate, it is important to know what the control condition is (*e.g.*, usual services available in the community, an alternative program, or no services at all).

Second, such monitoring can help you identify any cases where control group members are participating in the program (as "cross-overs") or otherwise being affected by it ("contaminated"), so that you can work with program staff to correct or minimize such cases, and/or adjust for such cases when later analyzing study outcomes. Such monitoring – coupled with engagement and training of program staff (as discussed above) to help maintain distinct treatment and control groups – can be an RCT's most effective safeguard against cross-over and contamination problems.

3. Key items to get right in measuring outcomes for the study sample

☐ **If the study will use a survey or test to measure outcomes (such as antisocial behavior or academic achievement), make sure the validity of the survey or test is well established.**

In other words, the survey or test should be backed by evidence that it accurately measures individuals' true outcomes. Validation of a survey used to measure delinquency, for example, might consist of evidence that it correlates closely with other measures, such administrative data on school suspensions or criminal arrests. In some RCTs, researchers may also find it useful to administer surveys or tests that do not yet have such evidence of validity, but we suggest doing so only as a supplement to the use of valid, well-established measures.

☐ **If the study asks sample members to self-report outcomes, corroborate their reports, wherever possible, with objective measures.**

For instance, an RCT of a substance-abuse prevention program should, wherever possible, corroborate sample members' self-reported substance use with other measures, such as saliva tests for smoking or urine tests for drugs, since individuals tend to under-report negative behaviors. This may lead to inaccurate study results if the treatment and control groups under-report to a different degree.

Similarly, an RCT of a bullying awareness and prevention program should, wherever possible, corroborate self-reported bullying victimization with independent measures, such as playground observations by an impartial observer. The main reason is that the program, by increasing youth awareness of what constitutes bullying, may cause an increase in self-reported bullying victimization in the treatment group even when the true level of bullying is constant or decreasing.

- **Where appropriate, keep members of the study team who collect outcome data unaware ("blinded") as to who is in the treatment versus control group.**

  Blinding is important when the study measures outcomes using observations, surveys, tests, or other instruments that are not fully structured, possibly allowing the person doing the measuring room for subjective judgment. Blinding protects against the possibility that any biases or hopes the measurer may have (*e.g.,* that the program will be found effective) could influence his or her outcome measurements, consciously or unconsciously. Blinding would be important, for example, in a study that measures preschoolers' cognitive or social skills through observations of the children's play behavior.

- **Make every effort to obtain outcome data for the vast majority of sample members originally randomized (*i.e.*, minimize sample "attrition").**

  As a general guideline, the study should obtain outcome data for at least 80 percent of the sample members originally randomized, including members assigned to the treatment group who did not participate in or complete the program. Furthermore, the study should aim for approximately the same follow-up rate in the treatment versus control group.

  Maximizing the follow-up rate (*i.e.,* minimizing attrition) is important because those individuals lost to follow-up have essentially self-selected themselves out of the sample and may have done so for different reasons in the treatment versus control group, thus undermining the equivalence of the two groups and causing inaccurate estimates of the treatment's effect. This is especially likely to be true if the attrition rate differs between the two groups. To understand why, consider that, in many studies, individuals lost to follow-up tend to be more at risk (*e.g.,* less motivated, more transient) than individuals retained in the sample. Thus, a higher attrition rate in the treatment group versus the control group (for example) could leave the treatment group with fewer at-risk individuals than the control group, potentially leading to superior treatment-group outcomes even if the program is ineffective.

  A detailed discussion of how to minimize attrition is beyond the scope of this document, but the endnote contains brief practical advice on the topic.[7]

- **To the extent possible, make sure that outcome data are collected in the same way, and at the same time, from treatment and control group members.**

  For instance, an RCT of a program to prevent recidivism among ex-offenders should measure recidivism rates for treatment and control group members using the same method (*e.g.,* the same

database of arrest records), and over the same time period measured from the point of random assignment. This is to ensure that any difference in outcomes between the two groups is the result of the treatment and not simply differences in how or when outcomes were measured. Crime rates often vary substantially over the course of a year (higher in summer, lower in winter), so in this example, any difference between the treatment and control group in the dates of follow-up could produce inaccurate estimates of the program's effects.

- **Wherever possible, measure whether the program's effects endure long enough to constitute meaningful improvement in participants' lives (*e.g.,* a year, or hopefully longer).**

  This is important because initial program effects often diminish quickly after the program ends, as demonstrated in RCTs in diverse areas such as early childhood education, substance abuse prevention, and job training. In most cases, it is the longer-term effect, rather than the immediate effect, that is of greatest policy and practical significance. (However, in some program areas, such as domestic violence prevention, the immediate effect may also be of great importance.)

4. Key items to get right in the analysis of study results

- **In estimating the program's effects, keep sample members in the original group to which they were randomly assigned. This even applies to:**

  (a) **Treatment group members who failed to participate in or complete the program** (*i.e.,* "no-shows" or "partial completers"); and

  (b) Control group members who may have participated in or benefitted from the program (*i.e.,* "cross-overs," or "contaminated" members of the control group).

  Retaining these individuals in their original group is called an "intention-to-treat" analysis. Such an analysis is important for study validity because these individuals likely have different characteristics than other members of their assigned group (*e.g.,* no-shows may be less motivated than other treatment group members), so removing them from their assigned group could well undermine the equivalence of the treatment and control groups and lead to inaccurate study results.

  Some RCTs also seek to measure, as a main study goal, the program's effects on those sample members who actually participated in the program (i.e., excluding no-shows in the treatment group and/or including cross-overs from the control group). These are called "treatment-on-treated" effects, and in many cases a valid estimate of these effects can be derived from the intention-to-treat effects.[8] In such cases, treatment-on-treated effects should be analyzed and reported in addition to (not as a substitute for) the intention-to-treat effects.

- **Estimate and report the program's effects on all outcomes and subgroups pre-specified in the analysis plan, following the procedures set out in the plan.[9] Of particular note, the study should:**

  (a) **Estimate and report whether each effect is statistically significant, or close to significant**, so as to help readers gauge how likely it is to be a true effect as opposed to a result of chance;

(b) **Estimate and report the size of each effect, preferably in real-world terms that convey the practical importance of the effect** – *e.g.,* an improvement in reading comprehension of a half grade-level, or a reduction in the teen pregnancy rate from 15% in the control group to 8% in the treatment group;

(c) **Use estimation methods that account for the different random assignment ratios faced by different sample members** (*e.g.*, 50:50 chance of being assigned to the treatment versus control group, or 60:40 chance), in studies where these ratios vary across the sample; and

(d) **Where possible, in estimating the program's effects, adjust for any differences between the treatment and control groups in their pre-program characteristics**, using an appropriate statistical method. Such adjustment should follow the approach pre-specified in the analysis plan.

☐ **Be sure that the above tests for statistical significance of the above effects take into account key features of the study design, such as:**

- Whether individuals or clusters were randomly assigned (as discussed earlier); and
- Whether the sample was sorted into groups ("stratified") prior to randomization, with random assignment taking place within each group (*e.g.,* in a delinquency prevention RCT, older youth versus younger youth).

☐ **Wherever possible, keep members of the study team who conduct the above analyses blinded as to who is in the treatment versus control group.**

In some cases, this could be accomplished by simply asking a colleague, before the analysis begins, (i) to re-label the treatment group's outcome data as the X (or Y) data set, and the control group's outcome data as the Y (or X) data set, and (ii) not to tell the study team which data set is treatment versus control. Doing so is important because certain analytical decisions – such as identifying implausible data values – cannot be pre-specified. Blinding helps prevent the study team from consciously or unconsciously making such decisions in a way that leads to their hoped-for result.[10]

☐ **Develop and report a table showing pre-program characteristics of the treatment and control groups, as well as statistical tests examining whether the two groups were similar.**

This will enable readers to gauge whether random assignment was indeed successful in creating two highly-similar groups, and whether the study's analysis appropriately adjusted for any pre-program differences (hopefully minor) between the two groups.

☐ **Develop and report an analysis of whether sample attrition created differences between the treatment and control groups that reduce the reliability of the study findings.**

This might include a table showing the pre-program characteristics of the treatment and control group members who remained in the sample after attrition, as well as statistical tests examining whether these two groups of remaining members were similar.

References

[1] Some useful, more comprehensive guides to the design, implementation, and analysis of RCTs include: Larry L. Orr, *Social Experimentation: Evaluating Public Programs With Experimental Methods*, Sage Publications, Inc., 1999; Esther Duflo, Rachel Glennerster, and Michael Kremer, "Using Randomization in Development Economics Research: A Toolkit," Abdul Latif Jameel Poverty Action Lab (J-PAL), December 12, 2006; and "E9 Statistical Principles for Clinical Trials," Food and Drug Administration, U.S. Department of Health and Human Services, September 1998.

[2] Evaluation of a program delivered at the cluster level (such as a schoolwide reform program, or hotspots policing strategy) generally requires that clusters be randomly assigned, for two reasons. First, the nature of these programs often precludes individual random assignment. To evaluate a strategy of increased police patrols in crime hotspots, for instance, it would not be possible to randomly assign one city resident to receive increased patrols and his or her next-door neighbor to receive fewer patrols, since they both live on the same street.

Second, in these cases, random assignment of whole clusters is often needed to isolate the effect of the program from other factors. For example, in a study of a schoolwide reform program, schools in the sample will differ in (i) whether they are implementing the reform program, and (ii) other school characteristics, such as teacher quality and school facilities. Therefore, if the study randomly assigns individual students to one school that has implemented the reform program versus another school that has not, the study will not be able to isolate the effect of the reform program from the effect of other school characteristics, such as teacher quality. Such a study therefore will likely need to randomly assign a sizable sample of schools to treatment and control groups, to ensure that the two groups are equivalent not only in student characteristics but also in school characteristics.

[3] It is important that the power analysis take into account key features of the study design, such as: (i) whether individuals or clusters will be randomly assigned; (ii) whether the sample will be sorted into groups ("stratified") prior to randomization, with random assignment taking place within each group (*e.g.,* in a delinquency prevention RCT, older youth versus younger youth); and (iii) whether, in analyzing study results, statistical methods, such as analysis of covariance, will be used to control for some of the variation among sample members in the study's targeted outcomes.

Online, open-access software programs for conducting power analyses include: Steve Raudenbush et. al., *Optimal Design Software for Multi-level and Longitudinal Research* (Version 3.01), 2011; and Nianbo Dong and Rebecca Maynard, *PowerUp! A Tool for Assisting Study Designs and Statistical Power Analysis*, 2015.

Other useful resources on power analysis include: Peter Z. Schochet, "Statistical Power for Random Assignment Evaluations of Education Programs," Mathematica Policy Research, June 22, 2005; Larry Hedges and Christopher Rhoads, "Statistical Power Analysis in Education Research," National Center for Special Education Research, Institute of Education Sciences, U.S. Department of Education, NCSER 2010-3006, 2009; and Brendon McConnell and Marcos Vera-Hernandez, "Going Beyond Simple Sample Size Calculations: a Practitioner's Guide," Institute for Fiscal Studies, Working Paper W15/17, June 2015.

[4] Larry L. Orr, *Social Experimentation: Evaluating Public Programs With Experimental Methods*, op. cit., no. 1, pp. 141-153; and Judith M. Gueron, "The Politics of Random Assignment: Implementing Studies and Impacting Policy," MDRC, 1999.

[5] In some cases, it may be necessary to revise the analysis plan after the study gets underway. As an illustrative example, an education RCT that pre-specified high school graduation as a primary outcome may need to switch measures (*e.g.*, to college enrollment) if it turns out that 95% of control group members are graduating high school, leaving little room for a difference in graduation rates between the treatment and control group. In such cases, the revision to the analysis plan should be made prior to any analysis of the program's effects, and reported in a transparent manner on the website at which the study is registered.

[6] Preferably, the sequence of computer-generated assignments to the treatment versus control group should be recorded, so that the researchers can verify whether program staff faithfully adhered to that sequence in allocating sample members to the two groups.

[7] The following advice for minimizing sample attrition is excerpted from Sarah A. Avellar and Timothy Silman, "What Isn't There Matters: Attrition and Randomized Controlled Trials," Office of Planning, Research and Evaluation, Administration for Children and Families, U.S. Department of Health and Human Services, Report #2014-49, August 2014.

"Overall attrition can be minimized by using multiple methods to contact sample members at follow-up; contact can be

made in person, by phone, via email, or online. Researchers can also continually track response rates to monitor the success of data collection efforts, adjusting their strategies and increasing their effort if response rates are low. Minimizing differential attrition is more important than minimizing overall attrition. Researchers should collect data at equal rates from everyone in the sample (both program and control group members). They also should recognize that more effort may be required to reach people in one group or the other. For example, it may be easier to contact program group members because of their connection with the program."

[8] Larry L. Orr, *Social Experimentation: Evaluating Public Programs With Experimental Methods*, op. cit., no. 1, p. 62 and 210. Howard S. Bloom, "Accounting for No-Shows in Experimental Evaluation Designs," *Evaluation Review*, vol. 8, April 1984, pp. 225-246. Joshua D. Angrist, Guido W. Imbens, and Donald B. Rubin, "Identification of Causal Effects Using Instrumental Variables," *Journal of the American Statistical Association*, vol. 91, no. 434, 1996, pp. 444-455.

[9] An authoritative and widely-used resource on the reporting of RCTs is the CONSORT (CONsolidated Standards of Reporting Trials) 2010 guideline.

[10] Robert MacCoun and Saul Perlmutter, "Blind Analysis: Hide Results to Seek the Truth," *Nature*, vol. 526, issue 7572, October 7, 2015.