Key Items To Get Right When Conducting a Randomized Controlled Trial in Education

December 2005

This publication was produced by the Coalition for Evidence-Based Policy, in partnership with the What Works Clearinghouse (a joint venture of the American Institutes for Research and the Campbell Collaboration). It was produced under a contract with the U.S. Education Department's Institute of Education Sciences (Contract #ED-02-CO-0022). The views expressed herein do not necessarily reflect the views of the Institute of Education Sciences.

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

Purpose

This is a checklist of key items to get right when conducting a randomized controlled trial to evaluate an educational program or practice ("intervention").

It is intended as a practical resource for researchers and sponsors of research, describing items that are often critical to the success of a randomized controlled trial. A significant departure from any one of these items, we believe, may well lead to erroneous conclusions about whether the intervention is effective in the sample and setting where it is tested. This document limits itself to key items, and does not try to address all contingencies that may affect the study's success.

This checklist includes the What Works Clearinghouse's criteria for reviewing studies, as well as some additional items. The checklist items that correspond with a What Works Clearinghouse criterion are marked with an endnote that describes the correspondence.

This checklist is organized into the following four sections:

1. Key items to get right in planning the study;
2. Key items to get right in the random assignment process;
3. Key items to get right in measuring outcomes for the study sample; and
4. Key items to get right in the analysis of study results.

1. Key items to get right in planning the study

☐ **Decide on (i) the specific intervention to be evaluated, and (ii) the key outcomes to be measured. These should include, wherever possible, the ultimate outcomes the intervention seeks to affect.**

For example, a study of a third-grade remedial reading program should, to the extent possible, evaluate the program's effect on ultimate outcomes such as reading comprehension, and not just surrogate outcomes such as word attack or word identification skills. Similarly, a study of a middle-school substance-abuse prevention program should, wherever possible, evaluate the program's effect on ultimate outcomes such as initiation of drug, alcohol, or tobacco use, and not just surrogate outcomes such as attitudes toward drugs. The reason is that improvements in surrogate outcomes (e.g., word attack/identification skills, attitudes toward drugs) may not always translate into improvements in the ultimate outcomes of interest (reading proficiency, reduction in drug use).

☐ **Decide whether the study should randomly assign individuals (e.g., students), or groups (e.g., classrooms or schools), to determine the intervention's effect.**

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

(a) **The intervention may have sizeable "spillover" effects on individuals other than those who receive it.**

For example, if there is good reason to believe that a school-based substance-abuse prevention program may produce sizeable reductions in drug use not only among the students in the program, but also among their peers within the school (through peer-influence), it will probably be necessary to randomly assign whole schools to intervention and control groups to determine the program's effect. A study that only randomizes individual students within a school to intervention versus control groups will underestimate the program's effect to the extent the program reduces drug use among both intervention and control-group students in the school.

For interventions where this spillover effect is likely to be small, however, random assignment within a school – of individual students and/or of classrooms and teachers - may still be a viable approach, and the more cost-effective one.

(b) **The intervention is delivered to groups such as classrooms or schools (e.g., a classroom curriculum or schoolwide reform program), and you want to distinguish the effect of the intervention from the effect of other group characteristics (e.g., quality of the classroom teacher).**

For example, in a study of a new classroom curriculum, classrooms in the sample will usually differ in two ways: (i) whether they use the new curriculum or not, and (ii) who is teaching the class. Therefore, if the study (for example) randomly assigns individual students to two classrooms that use the curriculum versus two classrooms that don't, the study will not be able to distinguish the effect of the curriculum from the effect of other classroom characteristics, such as the quality of the teacher. Such a study will therefore probably need to randomly assign whole classrooms and teachers (a sufficient sample of each) to intervention and control groups, to ensure that the two groups are equivalent not only in student characteristics but also in classroom and teacher characteristics.

For similar reasons, a study of a schoolwide reform program will probably need to randomly assign whole schools to intervention and control groups, to ensure that the two groups are equivalent not only in student characteristics but also school characteristics (e.g., teacher quality, average class size).

☐ **Conduct a statistical analysis to estimate the minimum number of individuals and/or groups to randomize in order to determine whether the intervention has a meaningful effect.**[1]

The purpose of such an analysis – known as a "power" analysis – is to ensure that enough individuals or groups are randomized to be confident that the study will detect meaningful effects of the intervention should they exist. The analysis will require you to make a judgment about the minimum effect size you are seeking to detect – a judgment that you might base on such factors as (i) what previous studies suggest as the intervention's likely effect size, and (ii) what effect size would justify the intervention's cost, or make adoption of the intervention attractive to schools seeking a gain in student achievement or other outcomes of a certain magnitude. Useful resources for conducting a power analysis are referenced in the endnote.[2]

It is important that the power analysis take into account key features of the study design, including:

(a) **Whether individuals (e.g., students) and/or groups (e.g., classrooms or schools) will be randomly assigned.**

(b) **Whether the sample will be sorted into groups prior to randomization** – such as high-achievers, average-achievers, and low-achievers – with the random assignment taking place within each group. (Such sorting is known as "stratification" or "blocking.")

(c) **Whether the study intends its estimates of the intervention's effect (i) to apply only to the sites – e.g., schools – in the study, or (ii) to be generalizable to a larger population** – e.g., to all schools participating in a federal program. (The two approaches are known respectively as "fixed-effects" versus "random-effects" models.)

(d) **Whether, in analyzing study outcomes, statistical methods (e.g., multivariate regression analysis) will be used to increase the study's ability to detect meaningful effects of the intervention.**

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

☐ **Take steps to protect the integrity of the random assignment process,[3] including:**

(a) **Have someone without a vested interest in the study outcome conduct the random assignment** of sample members to the intervention and control groups.

(b) **Use a random assignment method that is truly random** (e.g., computer-generated random numbers, rather than first letter of the last name).

(c) **If groups (e.g., classrooms) are randomized, ensure that the placement of individuals within these groups (e.g., placement of students in classrooms) is unaffected by whether the group is in the intervention or control condition.**

Thus, if a study randomizes classrooms and teachers in order to evaluate a new classroom curriculum, it should ensure that the school is not (for example) disproportionately placing its advanced students in the intervention classrooms – an occurrence which would undermine the equivalence of the intervention and control groups. Possible steps to prevent such an occurrence include: (i) placing students in classrooms *prior to* randomly assigning classrooms to the intervention and control groups; (ii) keeping the school officials who assign students to classrooms unaware of which classrooms are in the intervention and control groups; or (iii) *randomly* assigning students to classrooms (in addition to randomly assigning classrooms to intervention and control groups).

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

(a) **Obtain the data needed to accurately track the sample member over the course of the study, and analyze his or her outcomes, including –**

- a unique personal identifier (e.g. name and/or social security number),
- the date on which he or she was randomly assigned,
- whether he or she was assigned to the intervention or control group, and
- the random assignment ratio applied to him or her (e.g., 50:50 chance of being assigned to the intervention versus control group, or 60:40 chance, etc.).

(b) **Obtain pre-intervention measures of the outcomes that the intervention seeks to affect (e.g., reading or math skills)**, as well as any other descriptive data you wish to obtain on the sample. Such pre-intervention measures are extremely useful in (i) confirming whether the intervention and control groups are similar in their key characteristics prior to the intervention, and (ii) increasing the study's ability to detect meaningful effects of the intervention should they exist (via statistical methods for analyzing outcomes such as multivariate regression analysis). If at all possible, such data should be obtained just prior to the time of random assignment, since (i) data collected after that point could reflect the effects of the intervention; and (ii) it is often more difficult to locate all sample members after that point and/or to obtain their cooperation in providing the data.

☐ **Monitor what services the control group members may be receiving, and take steps to prevent their "crossing over" to the intervention group, or otherwise benefiting from the intervention (i.e., being "contaminated").**

Monitoring what services the control group members receive is important for two reasons. First, the study will produce an estimate of the intervention's effect *compared to* the control condition, and so it is important to know what that control condition is (e.g., the school's usual services, an alternative intervention, or no services at all). Second, such monitoring can help you identify and minimize any cross-overs by, or contamination of, control group members – occurrences which can undermine the study's ability to determine the effect of the intervention compared to the control condition.

More generally, you can often prevent cross-overs by continually monitoring the composition of both the intervention and control groups, to make sure that, to the maximum extent possible, they include only those sample members originally assigned to them through the random assignment process.

You can often minimize contamination of the control group by engaging school officials, teachers, program providers, and other study participants as partners in the study; giving them an understanding of the importance of having distinct intervention and control conditions; and enlisting their help in maintaining that distinction (e.g., asking teachers in the intervention group receiving a new classroom curriculum not to share curriculum materials or teaching strategies with the teachers in the control group for the duration of the study).

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

☐ **In measuring outcomes, use tests or other instruments whose ability to accurately measure outcomes is well-established.**[4]

Specifically, the main tests or other instruments that the study uses to measure outcomes should be backed by evidence that they are (i) "reliable" (i.e., yield similar responses in re-tests or with different raters), and (ii) "valid" (i.e., accurately measure the true outcomes that the intervention seeks to affect). To measure academic skills, we suggest you use instruments whose reliability and validity are widely accepted.

- **When study sample members are asked to "self-report" outcomes, corroborate their reports, to the extent possible, with independent and/or objective measures.**

  For instance, studies of substance-abuse or violence-prevention programs should, wherever possible, corroborate sample members' self-reported substance use or violent behavior with other measures, such as saliva tests for smoking, drug tests, school disciplinary records, official arrest data, and third-party observations. Such corroboration need not necessarily be applied to the full sample, just a subsample of sufficient size. The reason for corroborating self-reports is that people tend to under-report such undesirable behaviors. This may lead to inaccurate study results, to the extent the intervention and control groups under-report by different amounts.

- **If outcomes are measured with tests or other instruments that allow the testers room for subjective judgment, keep them unaware of who is in the intervention and control groups.**

  Such "blinding" of testers helps protect against the possibility that any bias they may have (e.g., as proponents of the intervention) could influence their outcome measurements. Blinding would be appropriate, for example, in a study that measures the word identification skills of first graders through individually-administered tests, or a study that measures the incidence of hitting on the playground through playground observations. In both these cases, the testers' biases might consciously or unconsciously affect their measurements. Blinding may be less important if outcomes are measured in ways that do not allow the tester room for subjective judgment, such as standardized written tests or fully-structured interviews.

- **Seek outcome data for all sample members originally randomized, including intervention group members who fail to participate in or complete the intervention.**

  This is important because those intervention group members who fail to participate in or complete the intervention may well differ from other sample members in their motivation level and other unmeasured characteristics, and it is not possible to identify their counterparts in the control group with the same characteristics and remove them from the sample. Therefore, failure to collect the outcome data of these "no-show" intervention group members would undermine the equivalence of the intervention and control groups.

  If, as we suggest, your study seeks outcome data for all sample members originally randomized, it will produce estimates of the intervention's effect on those intended to be treated. These are known as "intention-to-treat" estimates. Under some conditions, it may be possible to use such estimates to then infer the intervention's effect on those who actually participated, using a "no-show" adjustment.[5]

- **Make sure that outcome data are collected in the same way, and at the same time, from intervention and control group members.**

  For example, a study of a reading intervention should administer the same test of reading skills to intervention and control group members, at the same time (measured from the point of random assignment) and in comparable settings. This is to ensure that any difference in outcomes between the intervention and control groups is the result of the intervention and not simply differences in how or when outcomes were measured.

- **Make every effort to obtain outcome data for the vast majority[6] of the sample members originally randomized.[7]**

  Maximizing sample retention (also known as "minimizing attrition") is important because those individuals whose outcome data cannot be obtained may differ from other sample members in important ways (e.g., motivation, educational achievement level, demographics). Thus, their attrition from the study may well undermine the equivalence of the intervention and control groups. This is especially likely to be true if the amount of attrition differs between the intervention and control groups, but may also be true even if these amounts are the same.

  Some studies choose to obtain outcome data for only a subset of the study sample, in order to reduce the cost of the study or for other reasons. For such studies, it is important that the subset be defined in ways that preserve the equivalence of the intervention and control groups – e.g., by selecting random subsamples of each group, or choosing a subset that is defined by pre-intervention characteristics such as gender or pre-intervention test scores. It is also important to maximize sample retention in this subset, for the same reasons as those discussed in the preceding paragraph.

- **Wherever possible, collect data on long-term outcomes of the intervention (e.g., a year after the intervention has ended, preferably longer), to determine whether its effects are sustained.**

  This is important because the effect of many interventions diminishes substantially within 2-3 years after the intervention ends, as demonstrated in randomized controlled trials in diverse areas such as early reading, school-based substance-abuse prevention, and prevention of childhood depression. In most cases, it is the longer-term effect, rather than the immediate effect, that is of greatest practical and policy significance. (However, for some interventions, such as the use of weapons detectors to prevent school violence, the immediate effect may also be of great importance.)

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

- **Make sure all those originally randomized to the intervention and control groups are retained in their group when analyzing study results[8] – even:**

  (a) **Intervention group members who fail to participate in or complete the intervention** (retaining them in the intervention group is consistent with an "intention-to-treat" approach, discussed earlier); and

  (b) **Control group members who may have participated in or benefited from the intervention** (i.e., "cross-overs," or "contaminated" members of the control group, as discussed earlier).

  Retaining these individuals in their original group is needed to preserve the initial equivalence of the intervention and control groups over the course of the study, as discussed earlier.[9]

- **Estimate and report the intervention's effect on all outcomes measured, including (i) the statistical significance of each effect; and (ii) the magnitude of each significant effect, in easily-understood terms.**

Specifically, the study should test the statistical significance of each effect at conventional levels (usually the .05 level in a two-tailed test), to determine whether the effect is likely to be the result of the intervention as opposed to chance. And the study should attempt to estimate and report the magnitude of each significant effect in "real-world" terms that convey its practical importance, such as an improvement in reading comprehension of a half grade-level, or a reduction in the percentage of students using illicit drugs from 20 to 14 percent. (In some cases, you may wish to present additional estimates of the magnitude in terms such as "standardized effect sizes" or "odds ratios.")

If the study used a different random assignment ratio for different sample members (e.g., in year 1 it assigned 50 students each to the intervention and control groups, and in year 2 it assigned 50 and 25 respectively to the two groups), the study's estimates of the intervention's overall effect should account for this through a "weighted" analysis.[10]

If, as discussed earlier, you obtained data on the pre-intervention characteristics of sample members (e.g., their pre-intervention reading or math scores), you should estimate the intervention's effects using statistical methods such as multivariate regression analysis that will increase the study's ability to detect such effects should they exist.

- **Be sure that the above tests for statistical significance take into account key features of the study design,[11] including:**

  (a) **Whether individuals (e.g., students) or groups (e.g., classrooms or schools) were randomly assigned.**

  (b) **Whether the sample was sorted into groups prior to randomization** (as discussed earlier).

  (c) **Whether the study intends its estimates of the intervention's effect to apply only to the sites (e.g., schools) in the study, or to be generalizable to a larger population** (as discussed earlier).

- **Conduct and report statistical tests of whether the intervention and control group were similar in key characteristics prior to the intervention.[12]**

  If the study has randomized a sufficiently large sample, these tests are likely to confirm that intervention and control groups were similar in the vast majority of characteristics, allowing that by chance there might have been some minor differences. If differences greater than this are found, (i) you should review whether the random assignment process might have been violated, and (ii) it is particularly important to use statistical methods such as multivariate regression analysis (as discussed above) to adjust for pre-intervention characteristics of the study sample.

- **Conduct and report an analysis of whether sample attrition created differences between the intervention and control groups.**

  For example, this might be an analysis of whether the intervention and control group members who remain in the sample after attrition are equivalent in their pre-intervention characteristics (e.g., their educational achievement level prior to the intervention, and demographic traits).

References

[1] The corresponding What Works Clearinghouse criterion, used to rate a study's statistical analysis, is: *Precision of Estimate: Is the sample large enough for sufficiently precise estimates of effects?*

[2] Resources that may be helpful in conducting power analyses include: Peter Z. Schochet, *Statistical Power for Random Assignment Evaluations of Education Programs*, prepared for the U.S. Education Department's Institute of Education Sciences, June 22, 2005; Steve Raudenbush et. al., *Optimal Design Software for Group Randomized Trials*, and a user's manual for the software; the William T. Grant Foundation's free consulting service in the design of group-randomized trials; and Howard Bloom, *Randomizing Groups to Evaluate Place-Based Programs*, prepared for a conference of the Society for Research on Adolescence, March 2, 2004.

[3] The corresponding What Works Clearinghouse criterion, used to determine whether a study meets its evidence standards, is: *Randomization: Were participants placed into groups randomly?*

[4] The corresponding What Works Clearinghouse criterion, used to rate a study's outcome measures, is: *Reliability: Is there evidence that the scores on the outcome measure were acceptably reliable?*

[5] Useful resources that discuss the "no-show" adjustment, and how and when it can be used, include Larry L. Orr, *Social Experimentation: Evaluating Public Programs With Experimental Methods*, Sage Publications, Inc., 1999, p. 62; and Howard S. Bloom, "Accounting for No-Shows in Experimental Evaluation Designs," *Evaluation Review*, vol. 8, April 1984, pp. 225-246.

[6] For example, the Institute of Education Sciences publication *Identifying and Implementing Educational Practices Supported By Rigorous Evidence: A User-Friendly Guide* advises that, "A general guideline is that [a] study should lose track of fewer than 25 percent of individuals originally randomized – the fewer, the better." Similarly, the Office of Management and Budget's guidance on *What Constitutes Strong Evidence of a Program's Effectiveness* advises that, "As a general guideline, [a] study should obtain outcome data for at least 80 percent of the individuals (or other groups) originally randomized."

[7] The corresponding What Works Clearinghouse criteria, used to determine whether a study meets its evidence standards, are: (i) *Differential Attrition: Is there a differential attrition problem that is not accounted for in the analysis?* and (ii) *Overall Attrition: is there a severe overall attrition problem that is not accounted for in the analysis?*

[8] The corresponding What Works Clearinghouse criterion, used to rate a study's statistical analysis, is: *Statistical Assumptions: Are statistical assumptions necessary for analysis met?*

[9] As noted in section 3, under some conditions, it may be possible to obtain estimates of the intervention's effect on those who actually received it (as opposed to "no-shows") using the "no-show" adjustment discussed in endnote 5. Also, a variation on this technique can sometimes be used to adjust for "cross-overs" – see Larry L. Orr, *Social Experimentation: Evaluating Public Programs With Experimental Methods*, op. cit., no. 5, p. 210.

[10] A helpful resource for conducting such an analysis is Larry L. Orr, *Social Experimentation: Evaluating Public Programs With Experimental Methods*, op. cit., no. 5, p. 213.

[11] The corresponding What Works Clearinghouse criterion, used to rate a study's statistical analysis, is: *Statistical Assumptions: Are statistical assumptions necessary for analysis met?*

[12] The corresponding What Works Clearinghouse criterion, used to determine whether a study meets its evidence standards, is: *Baseline Equivalence: Were the groups comparable at baseline, or was incomparability addressed by the study authors and reflected in the effect size estimate?*