Hierarchy of Study Designs for Evaluating the Effectiveness of a STEM Education Project or Practice

Published in 2007

This publication was produced by the Coalition for Evidence-Based Policy, with funding support from the William T. Grant Foundation, Edna McConnell Clark Foundation, and Jerry Lee Foundation.

We welcome comments and suggestions on this document (jbaron@coalition4evidence.org).

Hierarchy of Study Designs For Evaluating the Effectiveness of a STEM Education Project or Practice

This document contains a narrative overview of the hierarchy, followed by a one-page graphic summary.

Purpose of the Hierarchy:

To help agency/program officials assess which study designs are capable of producing scientifically-valid evidence on the effectiveness of a STEM education project or practice ("intervention" [1]).

More specifically, the hierarchy –

- Encompasses study designs whose purpose is to estimate an intervention's effect on educational outcomes, such as student math/science achievement, or Ph.D. completion. (These are sometimes called "impact" studies.) The hierarchy does not apply to other types of studies that serve other purposes (e.g., implementation studies, longitudinal cohort studies).[2]

- Recognizes that many designs, including less rigorous impact studies,[3] can play a valuable role in an overall research agenda. It is not meant to imply that rigorous impact studies are appropriate for all interventions, or the only designs that produce useful knowledge.[4]

- Is intended as a statement of general principles, and does not try to address all contingencies that may affect a study's ability to produce valid results.

Basis for the Hierarchy:

- **It is based on the best scientific evidence about which study designs are most likely to produce valid estimates of an intervention's true effect** – evidence that spans a range of fields such as education, welfare/employment, criminology, psychology, and medicine.[5] This evidence shows that many common study designs often produce erroneous conclusions, and can lead to practices that are ineffective or harmful.

- **It is broadly consistent with the standards of evidence used by federal agencies and other authoritative organizations across a number of policy areas and contexts, including** –

  - Department of Education[6]
  - Department of Justice, Office of Justice Programs[7]
  - Department of HHS, Substance Abuse and Mental Health Services Administration[8]
  - Food and Drug Administration[9]
  - Helping America's Youth (a White House initiative)[10]
  - Office of Management and Budget[11]
  - American Psychological Association[12]
  - National Academy of Sciences, Institute of Medicine[13]
  - Society for Prevention Research.[14]

  Consistent with the hierarchy below, these various standards all recognize well-designed randomized controlled trials, where feasible, as the strongest design for evaluating an intervention's effectiveness, and most recognize high quality comparison-group studies as the best alternative when a randomized controlled trial is not feasible.

HIERARCHY OF STUDY DESIGNS

I. A well-designed randomized controlled trial, where feasible, is generally the strongest study design for evaluating an intervention's effectiveness.

A. Definition: Randomized controlled trials measure an intervention's effect by randomly assigning individuals (or groups of individuals) to an intervention group or a control group. Randomized controlled trials are sometimes called "experimental" study designs.

For example, suppose one wishes to evaluate, in a randomized controlled trial, whether providing struggling math students in third grade with supplemental one-on-one tutoring is more effective than simply providing them with the school's existing math program. The study would randomly assign a sufficiently large number of third-grade students to either an intervention group, which receives the supplemental tutoring, or to a control group, which only receives the school's existing math program. The study would then measure the math achievement of both groups over time. The difference in math achievement between the two groups would represent the effect of the supplemental tutoring compared to the school's existing program.

B. The unique advantage of random assignment: It enables you to assess whether the intervention itself, as opposed to other factors, causes the observed outcomes.

Specifically, the process of randomly assigning a sufficiently large number of individuals into either an intervention group or a control group ensures, to a high degree of confidence, that there are no systematic differences between the groups in any characteristics (observed and unobserved) except one – namely, the intervention group participates in the intervention, and the control group does not. Therefore, assuming the randomized controlled trial is properly carried out, the resulting difference in outcomes between the two groups can confidently be attributed to the intervention and not to other factors.

By contrast, nonrandomized studies by their nature can never be entirely confident that they are comparing intervention participants to non-participants who are equivalent in observed *and unobserved* characteristics (e.g., motivation). Thus, these studies cannot rule out the possibility that such characteristics, rather than the intervention itself, are causing an observed difference in outcomes between the two groups.

C. Random assignment alone does not ensure that a trial is *well-designed* and thus likely produce valid results; other key features well-designed trials include the following[15]:

- Adequate sample size;
- Random assignment of groups (e.g., classrooms) instead of, or in addition to, individuals when needed to determine the intervention's effect;
- Few or no systematic differences between the intervention and control groups prior to the intervention;
- Outcome data is obtained for the vast majority of sample members originally randomized (i.e., there is low sample "attrition");
- Few or no control group members "cross over" to the intervention group after randomization, or otherwise benefit from the intervention (i.e., are "contaminated");
- An analysis of study outcomes that is based on all sample members originally randomized, including those who fail to participate in the intervention (i.e., "intention-to-treat" analysis);

- Outcome measures that are highly correlated with the true outcomes that the intervention seeks to affect (i.e., "valid" outcome measures) – preferably well-established tests, and/or objective, real-world measures (e.g., percent of students graduating with a STEM degree);
- Where appropriate, evaluators who are unaware of which sample members are in the intervention group versus the control group (i.e., "blinded" evaluators);
- Preferably long-term follow-up;
- Appropriate tests for statistical significance (in group-randomized trials, "hierarchical" tests that are based both on the number of groups and the number of individuals in each group);
- Preferably, evaluation of the intervention in more than one site and/or population – preferably schools/institutions and populations where the intervention would typically be implemented.

II. ***Well-matched*** **comparison-group studies can be a second-best alternative when a randomized controlled trial is not feasible.**

A. Definition: A "comparison-group study" compares outcomes for intervention participants with outcomes for a comparison group chosen through methods other than randomization.[16]

Comparison-group studies are sometimes called "quasi-experimental" studies.

For example, a comparison-group study might compare students participating in an intervention with students in neighboring schools who have similar demographic characteristics (e.g., age, sex, race, socioeconomic status) and educational achievement levels.

B. Among comparison-group studies, those in which the intervention and comparison groups are *very closely matched* in key characteristics are most likely to produce valid results.

The evidence suggests that, in most cases, such well-matched comparison-group studies seem to yield correct overall conclusions about whether an intervention is effective, ineffective, or harmful. However, their estimates of the size of the intervention's impact are still often inaccurate, possibly resulting in misleading conclusions about the intervention's policy or practical significance. As an illustrative example, a well-matched comparison-group study might find that a class-size reduction program raises test scores by 40 percentile points – or, alternatively, by 5 percentile points – when its true effect is 20 percentile points.

C. A full discussion of matching is beyond the scope of this paper,[17] but a key principle is that the two groups should be closely matched on characteristics that may predict their outcomes.

More specifically, in an educational study, it is generally important that the two groups be matched on characteristics that are often correlated with educational outcomes – characteristics such as students' educational achievement prior to the intervention, demographics (e.g., age, sex, race, poverty level), geographic location, time period in which they are studied, and methods used to collect their outcome data.

In addition, the study should preferably choose the intervention and matched comparison groups "prospectively" – i.e., *before* the intervention is administered. This is because if the intervention and comparison groups are chosen by the evaluator *after* the intervention is administered ("retrospectively"), the evaluator may consciously or unconsciously select the two groups so as to generate his or her desired results. Furthermore, it is often difficult or impossible for the reader of the study to determine whether the evaluator did so.

III. Other common study designs – including pre-post studies, and comparison-group studies without careful matching – can be useful in generating hypotheses about what works, but often produce erroneous conclusions.

A. A pre-post study examines whether participants in an intervention improve or become worse off during the course of the intervention, and then attributes any such improvement or deterioration to the intervention.

B. The problem with a pre-post study is that, without reference to a comparison group, it cannot answer whether participants' improvement or deterioration would have occurred anyway, even without the intervention. This often leads to erroneous conclusions about the effectiveness of the intervention. Such studies should therefore not be relied upon to inform policy decisions, but may still be useful in generating hypotheses about what works that merit confirmation in more rigorous studies (e.g., randomized controlled trials or well-matched comparison-group studies).

C. Likewise, comparison-group studies without close matching often produce erroneous conclusions, because of differences between the two groups that differentially affect their outcomes.

This is true even when statistical techniques (such as regression adjustment) are used to correct for observed differences between the two groups. Therefore, such studies – like pre-post studies – should not be relied upon to inform policy decisions, but may still be useful in hypothesis-generation.

D. Despite their limitations, these less rigorous designs can play a key role in a larger research agenda. One research strategy, for example, is to sponsor or conduct low-cost, less rigorous studies of a wide range of interventions, to identify areas where additional research, using more rigorous methods, is warranted.

References

[1] In this document, the term "intervention" refers generically to any project, practice, or strategy funded by a federal STEM education program. As an illustrative example, it might refer to (i) a specific science curriculum developed and implemented in a single school district under a government grant; or (ii) a model science curriculum implemented by many school districts with funding from several federal programs; or (iii) an enhancement to such a model curriculum that is being tested against the original model.

[2] There are many important study designs to which this hierarchy does not apply – designs that address questions other than an intervention's effect on educational outcomes. Examples include implementation studies (designed to answer questions such as "To what extent is the intervention being implemented as intended, and how does the experience of intervention participants differ from non-participants?") and longitudinal cohort studies (designed to answer questions such as "What early risk factors are associated with students' failure to reach eighth-grade proficiency in mathematics?").

These studies can often be valuable complements to impact studies. As illustrative examples –

- If an implementation study finds that an intervention is not even being implemented correctly, it may mean that a rigorous impact study is premature since it would likely find no effect of the intervention.

- If a longitudinal cohort study finds that a student's inability to master basic arithmetic in first grade closely predicts his or her failure to reach eighth-grade proficiency in math and science, it may suggest the need to focus resources on rigorous impact studies to identify effective early math interventions.

[3] In this document the term "rigorous" (as in "rigorous studies") refers to a study's ability to produce valid estimates an intervention's true effect.

[4] A full discussion of where the more rigorous designs are appropriate and feasible, and where they are not, is beyond the scope of this paper, but as a general matter –

- They tend to be more feasible and cost-effective when used to evaluate well-defined interventions funded by federal STEM educations programs, rather than the programs themselves (e.g., a specific teacher professional development model funded by the federal Mathematics and Science Partnerships (MSP) program, rather than the MSP program itself).

- It makes sense to rigorously evaluate an intervention only where doing so is likely to yield important knowledge about "what works" that justifies the cost of the study. As illustrative examples: (i) a rigorous design might be the appropriate tool to evaluate a promising, well-defined classroom curriculum which, if shown effective, could then be replicated on a larger scale; (ii) a rigorous design might not yet be appropriate for a classroom curriculum that is in the early stages of development and whose key elements are still being refined.

For a useful discussion of when rigorous study designs, such as randomized controlled trials, are possible and when they are not, see Office of Management and Budget, *What Constitutes Strong Evidence of Program Effectiveness*, http://www.whitehouse.gov/omb/part/2004_program_eval.pdf, 2004, pp. 8-13.

[5] For a useful overview of the scientific evidence on which studies designs are most likely to produce valid estimates of an intervention's effect, see Office of Management and Budget, *What Constitutes Strong Evidence of Program Effectiveness*, op. cit., no. 4., pp. 4-8.

The following are citations to the relevant literature in education, welfare/employment, and other areas of social policy. Howard S. Bloom, Charles Michalopoulos, and Carolyn J. Hill, "Using Experiments to Assess Nonexperimental Comparison-Groups Methods for Measuring Program Effects," in *Learning More From Social Experiments: Evolving Analytic Approaches*, Russell Sage Foundation, 2005, pp. 173-235. James J. Heckman et. al., "Characterizing Selection Bias Using Experimental Data," *Econometrica*, vol. 66, no. 5, September 1998, pp. 1017-1098. Daniel Friedlander and Philip K. Robins, "Evaluating Program Evaluations: New Evidence on Commonly Used Nonexperimental Methods," *American Economic Review*, vol. 85, no. 4, September 1995, pp. 923-937. Thomas Fraker and Rebecca Maynard, "The Adequacy of Comparison Group Designs for Evaluations of Employment-Related Programs," *Journal of Human Resources*, vol. 22, no. 2, spring 1987, pp. 194-227. Robert J. LaLonde, "Evaluating the Econometric Evaluations of Training Programs With Experimental Data," *American Economic Review*, vol. 176, no. 4, September 1986, pp. 604-620. Roberto Agodini and Mark Dynarski, "Are Experiments the Only Option? A Look at Dropout Prevention Programs," *Review of Economics and Statistics*, vol. 86, no. 1, 2004, pp. 180-194. Elizabeth Ty Wilde and Rob Hollister, "How Close Is Close Enough? Testing Nonexperimental Estimates of Impact against Experimental Estimates of Impact with Education Test Scores as Outcomes," Institute for Research on Poverty Discussion paper, no. 1242-02, 2002, at http://www.ssc.wisc.edu/irp/.

This literature is systematically reviewed in Steve Glazerman, Dan M. Levy, and David Myers, "Nonexperimental Replications of Social Experiments: A Systematic Review," Mathematica Policy Research discussion paper, no. 8813-300, September 2002. The portion of this review addressing labor market interventions is published in "Nonexperimental versus Experimental Estimates of Earnings Impact," *The American Annals of Political and Social Science*, vol. 589, September 2003, pp. 63-93.

[6] U.S. Department of Education, "Scientifically-Based Evaluation Methods: Notice of Final Priority," *Federal Register*, vol. 70, no. 15, January 25, 2005, pp. 3586-3589. U.S. Education Department, Institute of Education Sciences, *What Works Clearinghouse Study Review Standards*, February 2006, http://www.w-w-c.org/reviewprocess/study_standards_final.pdf.

[7] U.S. Department of Justice, Office of Juvenile Justice and Delinquency Prevention, *Model Programs Guide*, at http://www.dsgonline.com/mpg2.5/ratings.htm; U.S. Department of Justice, Office of Justice Programs, *What Works Repository*, December 2004.

[8] U.S. Department of Health and Human Services, Substance Abuse and Mental Health Services Administration, "Changes to the National Registry of Evidence-Based Programs and Practices," *Federal Register*, vol. 71, no. 49, March 14, 2006, pp. 13132-13155.

[9] The Food and Drug Administration's standard for assessing the effectiveness of pharmaceutical drugs and medical devices, at 21 C.F.R. §314.12.

[10] Helping America's Youth initiative, program rating criteria at http://www.helpingamericasyouth.gov/programtool-how.htm.

[11] Office of Management and Budget, *What Constitutes Strong Evidence of Program Effectiveness*, op. cit., no. 4.

[12] American Psychological Association, "Criteria for Evaluating Treatment Guidelines," *American Psychologist*, vol. 57, no. 12, December 2002, pp. 1052-1059.

[13] Institute of Medicine (IOM), *Knowing What Works in Health Care: A Roadmap for the Nation*, The National Academies Press, 2008

[14] Society for Prevention Research, *Standards of Evidence: Criteria for Efficacy, Effectiveness and Dissemination*, April 12, 2004, at http://www.preventionresearch.org/sofetext.php.

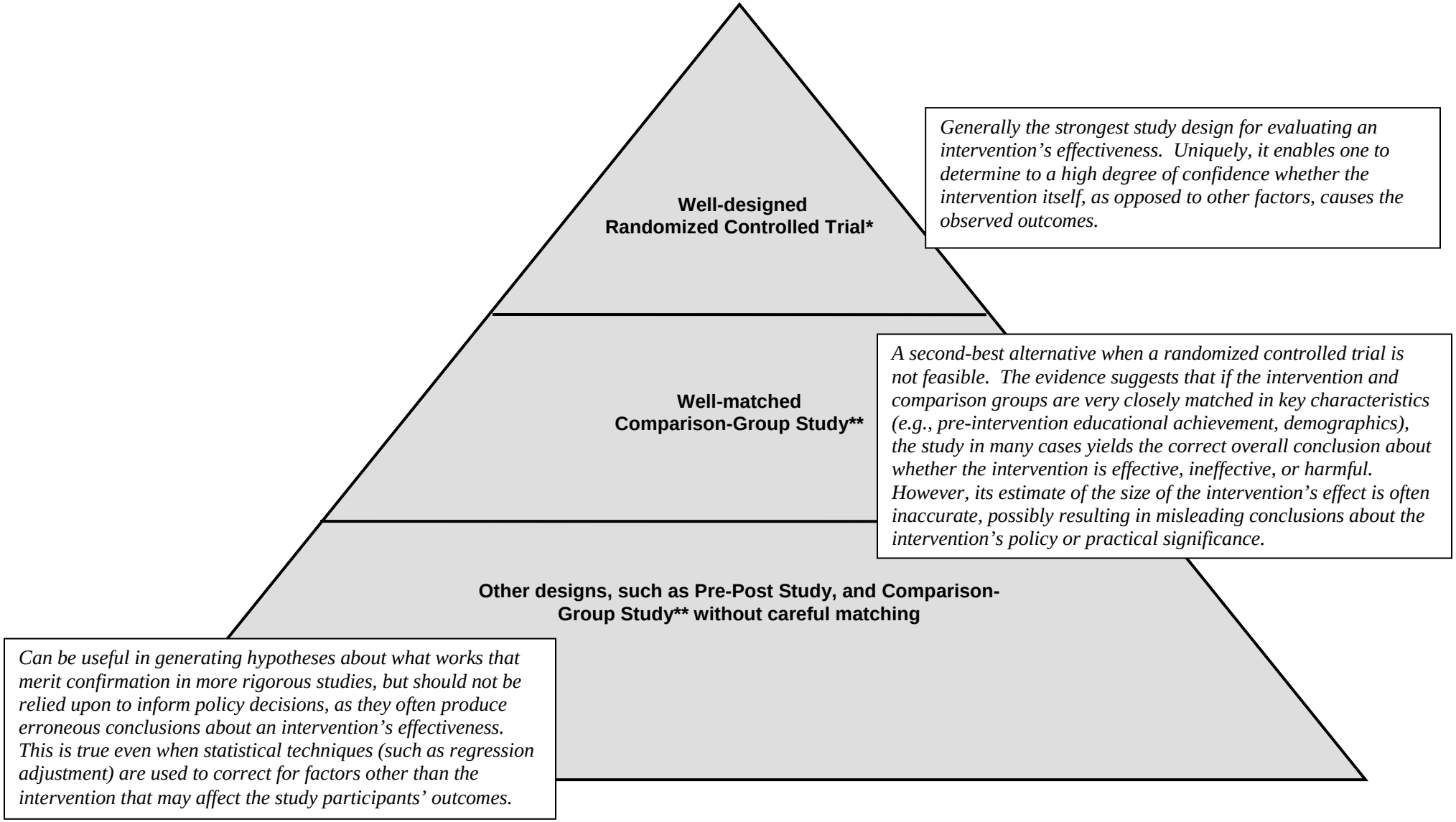
[15] The following guides contain a more complete discussion of what constitutes a well-designed randomized controlled trial. U.S. Department of Education, Institute of Education Sciences, *Identifying and Implementing Educational Practices Supported By Rigorous Evidence: A User-Friendly Guide*, produced by the Coalition for Evidence-Based Policy, December 2003, http://www.ed.gov/rschstat/research/pubs/rigorousevid/index.html. *Key Items to Get Right When Conducting a Randomized Controlled Trial In Education*, produced by the Coalition for Evidence-Based Policy in partnership with the Institute of Education Sciences' What Works Clearinghouse, December 2005, http://www.whatworkshelpdesk.ed.gov/guide_RCT.pdf.

[16] "Comparison-group studies" encompass a wide range of designs that compare intervention participants to nonparticipants selected through means other than randomization. Such designs include (among many others) regression-discontinuity studies and econometric studies.

[17] The following guides contain a more complete discussion of what constitutes a well-matched comparison-group study. U.S. Department of Education, Institute of Education Sciences, *Identifying and Implementing Educational Practices Supported By Rigorous Evidence: A User-Friendly Guide*, op. cit., no. 15. Coalition for Evidence-Based Policy, *Which Study Designs Can Produce Rigorous Evidence of Program Effectiveness? A Brief Overview*, January 2006, http://www.evidencebasedpolicy.org/docs/RCTs_first_then_match_c-g_studies-FINAL.pdf.

Graphic Summary:

Hierarchy of Study Designs For Evaluating the Effectiveness of a STEM Educational Intervention



* A randomized controlled trial is sometimes called an "experimental" study.

** A comparison-group study is sometimes called a "quasi-experimental" study.