<u>Random Controlled Trials (RCTs)</u>

Well-designed RCTs are considered the gold standard for measuring an intervention's impact across many diverse fields of human inquiry, such as education, welfare and employment, medicine, and psychology.[i] This is based on persuasive evidence that (i) they are superior to other methods in estimating an intervention's true effect; and (ii) the most common study designs – including "pre-post" studies and "comparison-group" (or "quasi-experimental") studies without careful matching – often produce erroneous conclusions. The following discussion elaborates, and also suggests an alternative when RCTs are not feasible.

A. <u>Definition</u>: RCTs are studies that measure an intervention's effect by randomly assigning individuals (or groups of individuals) to an intervention group or a control group.

For example, suppose that a school district wants to rigorously evaluate whether a new teacher professional development curriculum is more effective than the district's existing curriculum. The district might undertake an RCT which randomly assigns teachers to either an intervention group, which receives the new curriculum, or to a control group, which uses the existing curriculum. The RCT would then measure outcomes – such as teacher content knowledge or test scores of their students – for both groups over a period of time. The difference in outcomes between the two groups would represent the effect of the new curriculum compared to the existing curriculum.

B. <u>The unique advantage of random assignment</u>: 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 RCT 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.

C. <u>Evidence supporting RCTs</u>: There is persuasive evidence that –

(i) Well-designed RCTs are superior to other study designs in estimating an intervention's true effect; and

(ii) Well-matched comparison-group designs may be a good alternative when an RCT is not feasible.

Specifically:

- **"Pre-post" study designs often produce erroneous results.**

  Definition: 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.

  The problem with this type of study is that, without reference to a control group, it cannot answer whether the participants' improvement or deterioration would have occurred anyway, even without the intervention. This often leads to erroneous conclusions about the effectiveness of the intervention.

> **Example.** A pre-post study of Even Start – a federal program designed to improve the literacy of disadvantaged families – found that the children in the program made substantial improvements in school readiness during the course of the program (e.g., an increase in their national percentile ranking on the Picture Peabody Vocabulary Test from the 9th to the 19th percentile). However, an RCT of Even Start carried out by the same researchers found that the children in the *control* group improved by approximately the same amount over the same time period. Thus, the program had no *net* impact on the children's school readiness. If the researchers had only carried out the pre-post study, and not the RCT, their results would have suggested erroneously that Even Start is highly effective in increasing school readiness.[ii]

- **The most common "comparison group" study designs (also known as "quasi-experimental" designs) also lead to erroneous conclusions in many cases.**

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

  For example, a comparison-group study of a new teacher professional development curriculum might compare outcomes for teachers who receive the new curriculum to outcomes for a group of teachers in a neighboring school who do not receive the new curriculum.

  In education and other areas, a number of "design replication" studies have been carried out to examine whether and under what circumstances comparison-group studies can replicate the results of RCTs. These investigations have shown that most comparison-group studies in education and other areas of social policy produce inaccurate estimates of an intervention's effects. This is because of differences between the intervention and comparison groups that differentially affect their outcomes.[iii]

- **However, *well-matched* comparison-group studies can produce valuable knowledge, and may be a good alternative when an RCT is not feasible.**

  Specifically, the design replication studies noted above generally support the value of comparison-group studies in which the comparison group is *very* *closely matched* with the intervention group –e.g., in student test scores prior to the intervention, demographic characteristics, time period in which the two groups are studied, and methods used to collect their outcome data. Among comparison-group studies, these well-matched studies are the most likely to generate valid conclusions about an intervention's effectiveness. However, their estimates of the magnitude of an intervention's effect are often inaccurate, and in some instances they still produce erroneous overall conclusions about whether the intervention is effective, ineffective, or harmful.

This body of evidence therefore suggests that well-matched comparison-group studies can establish *possible* evidence of an intervention's effectiveness, thereby generating good hypotheses that merit confirmation in RCTs. And in cases where RCTs are not feasible or not yet available, such well-matched studies may serve as a second-best alternative.

D. RCTs may not be feasible in some cases – e.g., due to study participants' concerns about random assignment.

For example, in the MSP program, some schools and/or teachers may have concerns about randomly assigning some teachers to a control group that will not participate in the MSP project. We believe there are often effective strategies that you can use to address and overcome their concerns (discussed immediately below); however, if these are unsuccessful, you may wish to solicit well-matched comparison-group studies as a second-best alternative. If you do, we suggest you keep in mind that very careful matching of the intervention and comparison group – particularly on student test scores prior to the program – increases the chances that the study will produce valid estimates of an MSP project's effect.

E. You may be able to overcome schools' and teachers' concerns about random assignment through steps such as the following:

- In cases where an MSP project cannot enroll all eligible teachers due to budget or capacity limitations, you can make a strong case that random assignment – i.e., a lottery – is a fair way to determine which teachers will participate.

- You can offer control-group teachers participation in the MSP project after a one-year or two-year delay, if the project proves to be effective.

- You can offer control-group teachers an alternative program of professional development. The RCT would then be evaluating the effectiveness of the MSP project *compared to* that of the other program.

[i] See, for example, Office of Management and Budget, *Program Assessment Rating Tool (PART) Guidance for FY 2006 Budget*, p. 24, http://www.whitehouse.gov/omb/part/2006_part_guidance.pdf; the Food and Drug Administration's standard for assessing the effectiveness of pharmaceutical drugs and medical devices, at 21 C.F.R. §314.12; "The Urgent Need to Improve Health Care Quality," Consensus statement of the Institute of Medicine National Roundtable on Health Care Quality, *Journal of the American Medical Association*, vol. 280, no. 11, September 16, 1998, p. 1003; and *Standards of Evidence: Criteria for Efficacy, Effectiveness and Dissemination*, Society for Prevention Research, April 12, 2004, at http://www.preventionresearch.org/sofetext.php.

[ii] Robert G. St. Pierre et. al., "Improving Family Literacy: Findings From the National Even Start Evaluation," Abt Associates, September 1996.

[iii] Howard S. Bloom et. al., "Can Nonexperimental Comparison Group Methods Match the Findings from a Random Assignment Evaluation of Mandatory Welfare-to-Work Programs?" MDRC Working Paper on Research Methodology, June 2002, at http://www.mdrc.org/ResearchMethodologyPprs.htm. 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," Mathematica Policy Research, Inc., August 2001, at http://www.mathematica-mpr.com/PDFs/redirect.asp?strSite=experonly.pdf. 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.