10-2016

Randomized Controlled Trials

Robert Boruch
University of Pennsylvania, robertb@gse.upenn.edu

Rui Yang

Jordan Hyatt

Herb Turner III

Recommended Citation

Boruch, R., Yang, R., Hyatt, J., & Turner, H. (2016). Randomized Controlled Trials. *CRESP Working Paper and Briefing Series,* Retrieved from http://repository.upenn.edu/gse_pubs/396

Randomized Controlled Trials

Abstract
This paper covers the topic of randomized controlled trials in social, educational, criminological, health, and other human service sectors. It is studded with illustrations from developed and developing countries. We address the basic ideas that underlie trials in different ways, and cover contemporary definitions and vernacular, some history, and idea of cumulating evidence from such trials including recent work on replication and meta-analyses. Standards for evidence and reporting are considered. We attend to statistical matters and also recognize important non-statistical matters that must be taken into account in designing and executing such trials. Cluster randomized, place randomized, and other designs for such trials are covered on account of their increasing importance.

Disciplines
Education | Educational Assessment, Evaluation, and Research | Other Education

Title: Randomized Controlled Trials

Authors and Affiliations: Robert Boruch (university of Pennsylvania), Rui Yang (American Institutes for Research) , Jordan Hyatt (Drexel University) , Herb Turner III (Analytica, Inc.)

Date: October 2016

Prepared for: University Commons, and for Bent Greve (Ed) *Handbook of Social Policy Evaluation*. Cheltenham UK: Edward Elgar Publishing

0. Introduction

This paper covers the topic of randomized controlled trials in social, educational, criminological, health, and other human service sectors. It is studded with illustrations from developed and developing countries. We address the basic ideas that underlie trials in different ways, and cover contemporary definitions and vernacular, some history, and idea of cumulating evidence from such trials including recent work on replication and meta-analyses. Standards for evidence and reporting are considered. We attend to statistical matters and also recognize important non-statistical matters that must be taken into account in designing and executing such trials. Cluster randomized, place randomized, and other designs for such trials are covered on account of their increasing importance.

1.Definitions, Vernacular, and Rationale

A randomized controlled trial is a study in which people, entities, or places are randomly allocated to one or more interventions. One of the interventions may be a control condition that receives no special treatment, and which is then construed as the counterfactual.

The aim of a randomized controlled trial is to identify causal relationships through (a) a fair comparison of the different interventions in estimating their effects and (b) a legitimate statistical statement of one's confidence in the results of the comparison. Item (a) means that, at the outset of the trial, there are no *systematic* differences between the groups being compared, on account of the random allocation. In statistical language, there will be no bias in estimating the mean differences in the outcomes from each arm of the trial, if the trial is carried out properly. Item (b) means that chance differences--normal variation in behavior of people or organizations--are taken into account. This is accomplished through formal tests of statistical hypotheses or through the estimation of statistical confidence intervals.

Vernacular

A randomized controlled trial may also be called a "randomized experiment." We use this phrase interchangeably with "randomized controlled trial" in what follows. When entities such as schools or service agencies or hospitals are randomly allocated to different interventions, the study is usually called a "cluster randomized trial." If geographic regions, such as city neighborhoods or crime hot spots, are allocated randomly to different programs, the study is often designated as a "place randomized trial." The phrase "group randomized trials" is at times used to characterize such studies in psychological research.

"Quasi-experiments" and "observational studies" aim to estimate effects of interventions, but they do not include the randomization features of a trial. In these kinds of studies, the researchers do not have complete control over the conditions to which the experimental units are exposed. That is, neither they nor other agencies can randomly allocate the units to different interventions. See the chapter by Hanjoerg Gaus and Christoph Mueller in this volume and references below on non-randomized trials.

The phrase "natural experiment" is used in evaluation contexts at times. This may imply studies in which the allocation of people or entities is haphazard. It may mean that the allocation is based on an arguably random process such as birth dates. Or it may mean that two groups got formed in some unspecified way and that they might then be compared with regard to outcomes. There is no agreed upon technical definition of the phrase. Consequently, the evaluator must be wary.

Rationale, Put Simply

Riecken and Boruch (1974), among others, note that the biggest advantage of the randomized experiment, especially as compared to quasi-experiment, is that an experimental study can yield unbiased estimates of relative effects and is therefore a strong basis for drawing causal inference. Put in other words, the properly run randomized experiment assures internal validity of the findings.

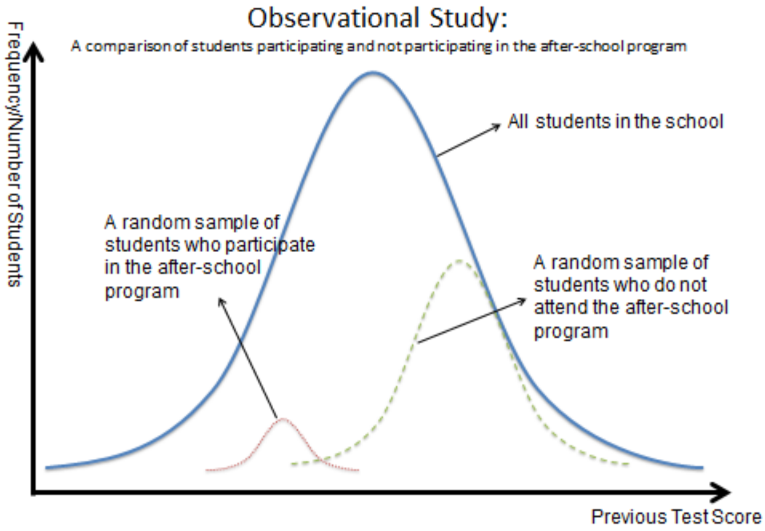
Campbell and Stanley (1963), for instance, discussed eight threats to internal validity in studies that are intended to estimate relative differences in effects of interventions but do not entail randomized assignment. These threats include history, maturation, testing, instrumentation, statistical regression to the mean, selection, mortality, and interaction of maturation and selection. Compared to other statistical designs for estimating the relative effects of interventions, randomized designs are less vulnerable to the challenges imposed by such threats.

The advantages of experiments versus quasi-experimental studies and observational studies can be illustrated graphically, in terms of statistical models, and in the context of empirical examples. These are covered next.

Rationale: Graphical Portrayal

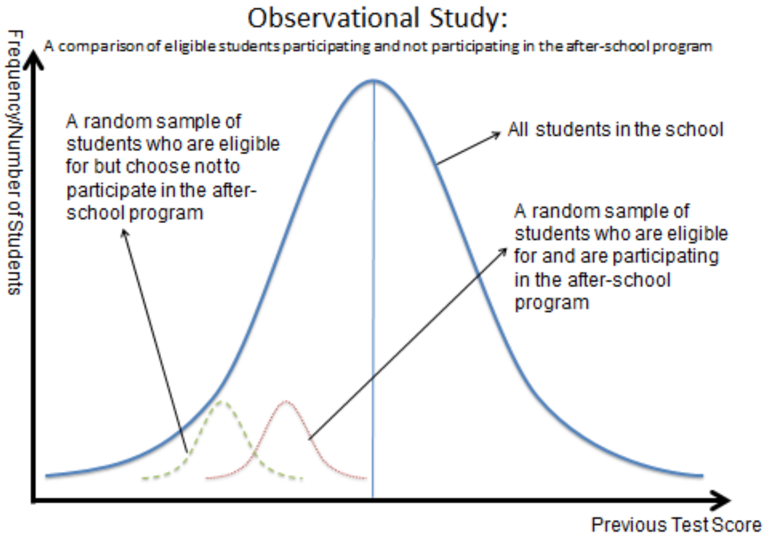
Suppose a research team wants to estimate the effects of an after-school program on students' test scores. In an observational study, without knowing (or requiring) any more information, the researchers might draw a random sample of students who participated in the program and a sample of students who did not, and then compare the scores from the two samples. As shown in Figure 1 below, the researchers may then conclude that students attending the program have a lower average score. An important variable omitted here, among the many possible contenders, is the students' previous scores. It is possible that only students who have scored poorly in the previous test entered the program . The unknown factor may bias the results and, therefore, the comparison is not fair.

Figure 1: *Examining the Effects of an After-school Program—Observational Study 1.*



In order to address that weakness, researchers may modify their original design and draw samples only from students who were eligible for the after-school program and compare those that could have, but did not participate, to those students who completed the program. As illustrated in Figure 2, however, the sample of students who participated in the after-school program had higher post-intervention test scores. This cannot solve all potential selection biases; only screening characteristics have been controlled for, i.e. the previous test scores. Another important variable that might be ignored in this scenario (among many) is student's baseline motivation to improve. It is possible that, students who are more motivated are more likely to enter the after-school program. Thus, it might be their motivation itself, rather than the intervention, that helps them to improve the scores.

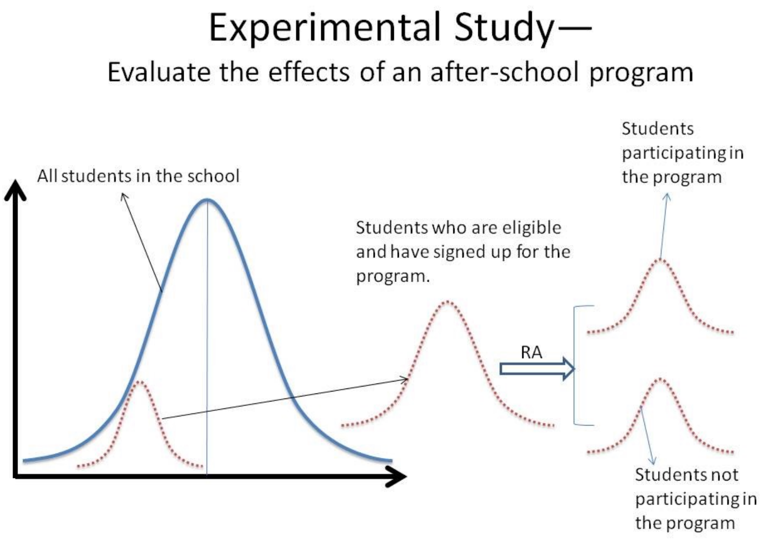
Figure 2: *Evaluating the Effects of an After-school Program—Observational Study 2.*


Observational Study:
A comparison of eligible students participating and not participating in the after-school program
Frequency/Number of Students
A random sample of students who are eligible for but choose not to participate in the after-school program
All students in the school
A random sample of students who are eligible for and are participating in the after-school program
Previous Test Score

Numerous variables- some known and many more unknown- may influence the estimate of interest. Observational studies often cannot accommodate this because the right variables must be identified, must be measured in the right way, and must be entered into analyses in the right functional form.

Figure 3, portrays how researchers can design an experiment to get an unbiased estimate of the effects of the after-school program. After identifying students who are eligible and willing to participate in the after-school program, the researchers randomly assign (RA) into two groups, one receiving the intervention and the other not. Although this requires a significant beforehand effort by the researchers- and cooperation by the school-, the resulting analysis provides a fair comparison.

Figure 3: *Evaluating the Effects of an After-school Program—Experimental Study.*



Experimental Study—
Evaluate the effects of an after-school program

All students in the school
Students who are eligible and have signed up for the program.
RA
Students participating in the program
Students not participating in the program

Rationale: Statistical Model Portrayal

Using the same school intervention example, let participation in the after-school program denoted by *x*, where *x*=1 indicates the child's' participation in the program and *x*=0 indicates no participation. The outcome for each unit is a test score denoted by *y*. If *y* is affected only by x, then fitting the model

$y=\beta x+\varepsilon$, will yield the estimate of treatment effect is $\beta = r_{yx}\frac{s_y}{s_x}$, where r_{yx} is the correlation between y and x, and s_y and s_x are the standard deviations for y and x, respectively. All causes of the outcomes, apart from the intervention, are embodied in the error term which is uncorrelated with the receipt of the intervention on account of random assignment.

In a non-randomized study, the y's may be systematically affected not only by the intervention's presence and absence, but also by other variables that characterize how students wind up in the program or not, such as motivation or parental resources or their pre-intervention test scores . In particular, each student's previous test score (x_1) might be influential. If so, the analyst then may employ this as a control variable in the model to account for some variance that is unexplained in the simpler model. So the model becomes $y=\beta x+\beta_1 x_1+\varepsilon$, and the estimate of effect equals

$\beta = \frac{r_{yx} - r_{yx_1} * r_{xx_1}}{1 - r^2_{xx_1}} \frac{s_y}{s_x}$. A researcher can add as many control variables as is practically possible, though the availability of data will limit the scope of variables that can be viable included in the model. If the influential variable is ignored, the resulting estimate of effect will be inflated. That is, positive correlation between x and x_1 will decrease the size of β, thus shrinking the effect size estimate.

In quasi-experiments and other non-randomized studies that aim to estimate effects, researchers try their best to include as many relevant control variables as possible. The big challenge is that there are often desirable variables that cannot be measured (for example, home resources, in the example above) and others that are relevant but unknown and therefore unobserved. Consequently, the bias of the estimated relative effect of the interventions cannot be fully eliminated. In a randomized trial, since the interventions are randomly assigned, the receipt of the intervention will be independent of all other observed and unobserved variables. In this way, even if some important control variables are omitted, researcher can remain confident about the accuracy of the parameter estimate for the difference in outcomes of the interventions (*i.e.* the effect of a particular intervention or treatment relative to a control condition).

Rationale: Empirical Portrayal

Consider an example drawn from criminological research in the United States. In the 1970's, a New Jersey prison started a program to prevent at-risk juveniles from committing crime by letting the prison inmates present lectures to the juveniles about the inmates' lessons and experiences. The program, known as "Scared Straight," was judged to have achieved a 94% success rate in some non-randomized studies (Finckenauer, 1982) and it received favorable media attention throughout the country (U.S. House, 1979). Scared Straight expanded to other states and six other nations.

A randomized controlled trial subsequently conducted by Finckenauer (1982), however, found no effects on the participants' criminal behavior. Petrosino, Turpin-Petrosino and Buehler (2002) later undertook a systematic review of *all* randomized tests of the Scared Straight programs. This resulted in finding that the intervention usually had *no* discernible effect and, at times, significantly increased offending among juveniles relative to doing nothing at all. The inability to identify and use relevant control variables in the earlier quasi-experiments led to inflated estimates of a positive effect for the program.

This example shows how misleading a non-experimental study could be and how unexpected negative impacts might be revealed in a fair comparison. It is worth noting, however, that despite the rigorous empirical findings from the 1980s, programs following the Scared Straight model continue to be employed in some American jurisdictions and has spawned a television series detailing the experience of kids in such programs.

2.Historical Development of Controlled Trials and Contemporary Illustrations

Mounting comparative trials, which were not necessarily randomized, dates at least to about 600 B.C. when Daniel of Judah compared the health effects of vegetarian diet with a Babylonian diet over a 10-day period (Jadad, 1998, Stolberg, Norman and Trop, 2004). More than 2,000 years later, after experimental philosophy and the idea of intentional comparison became popular after the scientific revolution of the 1400's, James Jurin compared the mortality rates of naturally occurring smallpox with that of cases occurring as a result of inoculation (Meldrum, 2000). Around the same time in the Chinese Qing Dynasty, a branch of scholarship known as Kao Zheng ushered a new epistemology valuing empirical measurement rather than theoretical interpretation of ancient Confucian texts (Boruch and Rui, 2009).

Randomized experiments first appeared in psychological research in laboratories, undertaken by Charles Sanders Peirce.   Later, randomized trials were used in agriculture due to Jerzy Neyman and Ronald Fisher (Neyman, 1923).   Dodd (1934) is seen as a founder for his study of the effects of a hygiene program when compared with some untreated control groups.   Around the same time, Liao Shicheng undertook social experiments in Shanghai as an advocate of transforming education studies into science. Shicheng conducted a randomized experiment to examine the effects of an education intervention, called the "Dalton Plan", on students' achievement by using a value-added model (廖世承, 1925).

An important and thorough examination of the history of the first appearance of randomized trials in the social and education sectors is given by Forsetlund et al (2007). These scholars, from Norway and the UK, trace the earliest verifiable experience to 1928 studies at Purdue University on the effectiveness of counseling. Dehue (2001) covered the contemporary history of trials in the social sector.

The first large scale trial in medicine in the UK was undertaken in 1948. It concerned the effect of Streptomycin treatment on pulmonary tuberculosis. One of the authors, Austin Bradford Hill, is often given major credit for the modern medical randomized trial (Hill, 1952). Since then, the methodology of randomized controlled trial has been increasingly accepted in the medical arena. Evans, Thornton, and Chalmers' (2006, 2008) summary, published in English, Chinese, and other languages, provided a concise history of efforts to use such designs in handling dozens of illnesses. The Cochrane Collaboration, discussed below, has records on over 150,000 trials and promotes the idea known as "evidence based medicine" (Stolberg, Norman & Trop, 2004).

In the social science sector, the acceptance of randomized trials has been slower than that in medicine. The Head Start Program, launched in 1965 by the United States Department of Health and Human Services, is one of the longest-running programs to address systemic poverty in the US. With a budget of $8.1 billion in 2011, it has periodically included randomized trials in the early childhood sectors to evaluate its effectiveness. In the 1990s and 2000s, experimental studies were devoted to study the impacts of Head Start (St. Pierre, 1990).

By the 20st century, randomized controlled trials had become the "gold standard" of evidence in medical research. In social sciences, this approach to estimating effects and making comparisons has also increased in importance, though at a slower pace. In the U.S., the Institute of Education Sciences (IES), for example, has sponsored numerous trials involving random allocation of individuals, teachers, or entire schools to different interventions. From 2002 to 2009, for instance, over 100 sizeable trials were mounted (Boruch, Weisburd, and Berk, 2010). This a marked increase from previous periods of time, though still not reaching the levels seen in the health fields.

The increased prevalence of RCTs in the social sectors has resulted in a number of milestones. Each designed to isolate the causal effects of promising interventions on prescribed outcomes. The following examples are from studies in low, middle and high income countries, and from different human sectors.

- The PROGRESA Trials in Mexico, later called Oportunidades, aimed to understand whether conditional cash transfer payments made to the mothers of children in poor villages would lead to increased enrollment in schools, fewer drop-outs from schools, and better school attendance. Over 300 villages were randomly allocated to the intervention and nearly 200 villages were randomly allocated to a control condition from a common pool of poor villages to learn that such cash transfers did indeed result in children staying in schools. See Behrman, Parker, and Todd (2010) and Parker and Teruel (2005), for example.

- Trials in low income countries have been mounted to generate unbiased estimates of the relative effects of interventions on outcomes and to assure valid statistical statements of one's confidence in results. Bruns, Filmer, and Patrinos (2011), for instance, reviewed evidence from numerous cluster randomized trials on education interventions and economic interventions in developing countries.

- In the US, the Tennessee class size trial, called STAR, was a remarkable precedent. It involved the random allocation of 300 kindergarten classrooms in nearly 80 schools to engage in small class sizes versus usual class sizes coupled with a full time teacher aide. The results support the idea that, in a heterogeneous society such as in Tennessee, small class sizes lead to better achievement of children. The initial results are given in Finn and Achilles (1980). More recent work is given in Konstantopoulos and Sun (2012) and long term economic effects on earnings are given by Chetty et al (2011).

- Numerous place randomized trials have been undertaken to understand whether intensive police patrol strategies in high crime neighborhoods prevent crime, and lead to no migration of criminals to more peaceful neighborhoods (Boruch, Weisburd and Berk, 2010). Positive results have helped to justify policing strategy targeting crime "hot spots" so as to reduce criminal activity. Piquero and Weisburd (2010) provide illustrations and technical details on randomized trials in the police, prison, rehabilitation, and crime control areas.

Such studies are merely illustrative. The rapidly increasing number of trials conducted in a variety of sectors has engendered a need to keep track of them and to summarize their results periodically. It has led to the creation of new journals and organizations that focus heavily on the use and products of the approach. It is to these topics that we proceed next.

3.Compendiums on Randomized Trials and Summaries of the Results, and Journals and Organizations Concerned with Randomized Trials.

Compendiums and Systematic Reviews

The international Campbell Collaboration (http://campbellcollaboration.org) works to summarize the results from numerous randomized controlled trials and high quality quasi-experiments so as to assure that society need not depend on a single study. This Collaboration covers education, crime and justice, and social services. The Campbell Collaboration's older sibling, the Cochrane Collaboration (http://cochrane.org), has a similar mission but confines attention to the health sector, *i.e.,* judging the quality of evidence and to synthesize results from randomized trials.

In the US, the *What Works Clearinghouse* (http://www.whatworks.ed.gov) was created by the federal government to identify education interventions, and to determine whether dependable evidence on their effectiveness exists. High priority is given to evidence from randomized controlled trials. For more details on these and other efforts to screen the quality of evidence on studies of the effectiveness of interventions, see Boruch and Rui (2008, 2009). Most such efforts depend heavily on rigorous randomized trials, and on meta-analyses of their results.

Relevant Journals

Peer reviewed journals that specialize in reporting on randomized controlled trials and, at times, advances in related statistical methods are readily accessible. See for instance: *Journal of Research on Educational Effectiveness* (ISSN 1934-5739); *Trials* (http://www.trialsjournal.com); *Clinical Trials* (http://ctj.sagepub.org); and *Journal of Experimental Criminology.* Of course, many other scientific journals also publish reports on experiments periodically; some of these are cited in the reference list of this chapter. Advances in statistical methods that pertain to trials appear periodically in journals such as *Statistical Sciences, Journal of the Royal Statistical Society, Journal of the American Statistical Association,* and others. The *Randomized Social Experiments eJournal* covers work done mainly by economists who do research in education and other sectors; see http://www.ssrn.com/update/ern/ern_random-social-experiments.html .

Organizations

A variety of organizations have developed the capacity to design and implement randomized trials in developed and developing countries. The organizations which are involved in international work include (but are certainly not limited to): the Poverty Action Laboratory at the Massachusetts Institute of Technology, American Institutes for Research (http://www.air.org); Mathematica Policy Research (http://www.mathematica-mpr.com); MDRC (http://www.mdrc.org); RAND Corporation (http://www.rand.org); and the Urban Institute (http://www.urban.org). Organizations such as the National Bureau of Economic Research (http://www.nber.org) also produce reports on experiments in their discipline.

People from these organizations often collaborate with researchers at universities in designing trials and analyzing statistical results. And, of course, university faculty members undertake randomized trials when the resources are ample, and provide education about the approach. In the UK, University of York, Cambridge and Oxford, for instance, routinely provide courses on the topic. Many institutions in the US do so; the University of Pennsylvania, Stanford, Northwestern University, and Harvard University are among them.

Most such organizations provide free access to recent reports on trials on their websites. If they do not, the agency that sponsored the trial will usually provide results on its website.

Funding Organizations

The statistical design of a trial, its implementation, and analysis of results require resources apart from knowledgeable staff and willingness of the participants to engage in a trial. In recent years in the U.S., for example, funding for such trials in education, welfare, food and nutrition, prevention of crime, disease, and substance abuse, and rehabilitation has come from every relevant federal government agency. Government agencies in Mexico, Colombia, India, the UK, Germany, some countries of Africa, and others have contributed funds to such studies. And multi-national funders of programs also invest in funding controlled trials on those investments, in the interest of "accountability," *e.g.* Bruns, Filmer, and Patrinos (2011). Reports of the statistical results of randomized controlled trials typically acknowledge the funding source and readers may then wish to read these to learn more.

4.Contexts of Randomized Trials

Evaluation Policy Context

To understand the importance of randomized controlled trials, consider a broader context of evaluation policy. Boruch (1997, p.22) and others frame the core questions for evaluation research as follows:

> What is the severity, scope, and nature of the problem? How do we know?
>
> What programs, projects or practices are being implemented to reduce the problem? How do we know?
>
> What are the effects? How do we know?
>
> What are the relative cost effectiveness measures, and how do we know?

The first questions address the definition of the issue that the intervention is seeking to resolve. The rationale is that nobody can be confident about designing solutions to a problem before understanding the problem itself. Mixed-methods, a combination of quantitative and qualitative methods, are often used in this step. Surveys, for example, are sent out to samples of the population affected by the problem, and interviews and focus groups are conducted to collect in-depth information about the scope of the problem, potential challenges for study implementation, and data that may inform the design of the intervention itself.

The second group of questions focuses on the implementation of the intervention, program, or project. It engenders further questions such as: "Does the program/project follow the plan?," "What modifications have been made to the implementation?," and "Do the members of the intervention's delivering group meet expectations?" These inquiries help the researcher understand the context of the study- and move beyond the idealized version of the intervention.   Sometimes researchers use this opportunity to measure mediating variables[1], factors unrelated to the intervention that may impact outcomes, to learn why and how well the intervention has been implemented.   Monitoring processes can help one understand if the intervention was indeed deployed, and uncovering what might be the key elements that likely contributed to its success or its failure, including if the curriculum did not achieve its goal or if the teachers did not follow the prescribed steps.

The third class of inquiry shifts the attention from the process of implementing an intervention to estimating an intervention's effect relative to a counterfactual. This phase of inquiry is usually where the randomized controlled trial can offer meaningful answers. The answer to this question indicates how the program's target individuals would have performed in the absence of the program, which of course is the focus here. More detailed is provided later.

The final set of questions addresses the economic implications of the results of a trial. For example, some programs are effective but the costs may outweigh the benefits. Under those circumstances, the program may be terminated because the investors cannot afford the sustained costs or support the work force required. Cost-effectiveness analysis is a complex task, as the cost of a program can be difficult to quantify even if one has good estimates of effectiveness. The direct costs and short-term benefits are sometimes easy to measure. These include staffing, materials and data collection expenses,

[1] A mediating variable is a variable which accounts for the relationship between a predictor and an outcome. Mediating variables usually explain how and why certain events have effects on certain criteria.

as well as the fiscal measurement of the change in behaviors such as the monetary benefits of reduction in criminal recidivism or increase in educational attainment or reduced dependence on welfare. The indirect costs and long-term benefits often require complex models that perforce are more speculative especially if one considers the replicability or sustainability of the intervention.   Levin and McEwan (2001) provide technical guidance and illustrations, including coverage of some of the experiments cited in this chapter.

Social Context

A variety of conditions may prevent or limit the use of randomized trials in a given social setting.   Coyle et al. (1991, p. 183) present five circumstances that justify the selection of a non-randomized approach: (1) decision makers' tolerance of ambiguity in estimating the effect of the new program; (2) the assumption that competing explanations of the program's effect are negligible; (3) the political l or legal requirement that all individuals eligible for the intervention must be involved in the program; (4) the preference for a non-randomized trial in meeting standards of ethical propriety; and (5) the explicitness of theory-based or data-based predictions of effectiveness.

Putting these conditions in a different way, Boruch (1997) declared the following. If human rights will be violated, then do not do a randomized trial or design the trial so as to recognize those rights.   If a non-randomized trial will suffice, given the tolerance for ambiguity and the ability to make a forecast about how people would have behaved in the absence of the intervention, then don't use a trial. If the results of the trial will not be used, there is no point to doing the trial. Nor is the any point to mounting a trial if there is no uncertainty about the value of the proposed regimen.

5. Resources for Trial Design and Analysis and of Results

The earliest works on the statistical underpinnings of randomized controlled trials are readily accessible, *e.g.* Fisher (1935) and Neyman (1925). Technical details on the statistical design of randomized trials are given in Box et al (2005) and Kirk (2012), among others. Murray (1998) covers group randomized trials in psychological research, and related statistical design matters for hierarchical studies are covered by Raudenbush and Bryk (2002) mainly in education. Hayes and Moulton (2009) cover the technical aspects for cluster randomized trials in health and medicine.

Statistical aspects of randomized trials are important. But so too are the managerial , social, and political –institutional aspects. The contents of the Mosteller-Boruch (2001) volume for instance include papers on each, in the contexts of social welfare, education, and social services. Gueron and Rolston's (2013) history covers the last 40 years of controlled trials in the welfare and manpower training sectors in the U.S., giving special attention to the political-institutional issues and how they were resolved.

Simple Trials

A straight-forward way of assessing the statistical dependability of an estimated treatment effect is a simple t test (for continuous outcome measure) or proportion test (for dichotomous outcome measure). When the randomization is properly carried out with an eligible target group there are no systematic differences between the treatment group and control group on key variables, and these tests will often suffice in simple experimental designs.

A robust way of testing treatment effects, especially in small trials, is to use a randomization test (Edgington and Onghena, 2007; Bookmeyer and Chen, 1998). The basic idea of the test is to compute results for all possible combinations of the observational units allocated to either the treatment group or the control group, pretending they have not been assigned a group membership yet. The actual difference in the outcomes of each combination and probability of the actual difference occurring is then compared to the empirical distribution of all possible outcomes. One feature of the randomization test lies in the fact that it has no underlying model assumptions (Small et al., 2008).

Assuring Balance

Imbalance in the groups that are randomly composed may differ on account of chance despite the fact that there will be no systematic difference between the groups. Before examining the treatment effects, researchers test baseline equivalence between groups on some key independent variables using simple t tests. A second involves fitting a logistic regression model with the group membership being the dependent variable and the variable we want to test for equivalence being the independent variable. A significant relationship between the predictor and the outcome signals the lack of baseline equivalence. Variables that are not baseline equivalent in the analysis sample need to be included in sophisticated analysis models to adjust the estimate of the intervention's effects.

Researchers can use a variety of specific study designs to reduce the chance of an "unlucky" randomization before the intervention is implemented. In a matched-pair design, for instance, participants are paired according to their similarities of some key factors. Then randomization is conducted *within* each pair to allocate one participant to intervention group and the other to the control group. It ensures the two groups will be more similar to one another than a complete randomized design. Imai et al. (2009) showed that if the pair-wise matching was done properly, it could enhance the precision in estimates of the effects of interventions. One downside of a matched pair design is that it results in a loss of statistical degrees of freedom and therefore reduces the statistical power of the analysis. The trial's designers need to identify the trade-offs before making a decision about which design fits the context the best. Morgan and Rubin (2016) present more advanced methods that entail re-randomization when the initial randomization is unsatisfactory and provide explicit statistical criteria for balance that guide the approach.

Hierarchical and Complex Trials

As mentioned earlier, randomized trials can be classified according to the level at which the randomization is conducted. Since cluster randomized trials are now widely used in the social sector, some basic technical issues related to CRTs are outlined here. Donner and Klar (2000), Hayes and Moulton (2009) and Raudenbush and Bryk (2002) cover statistical methods for hierarchical arrangements in controlled trials and the latter consider non-randomized observational studies.

For example, consider an intervention for which the randomization happens at the school level and the intervention is to increase student's academic achievements. To make it simple, suppose one has only have two variables, apart from the assignment to intervention or control groups, so as to increases the precision of analysis. The student's test score (Y) as the response variable, the student's gender (G= 0 if male and 1 is female) is the student-level control, and the school type (S=0 is public school and 1 is private school) is the school-level control. Earlier techniques for analyzing cluster data either aggregated all information to the school level and treat schools as the units for analyses, or they disaggregated all the information to the student level. The problem with the first approach is the loss of all the lower-level information. The concern with the latter method lies in the fact that there will be local dependence for students in the same school, therefore violating the independence assumption of the error distribution.

Using hierarchical linear modeling resolves the issue. As can be seen below, the student-level model is fit first, then the coefficients at the lower level are modeled on variables at the higher level.

Level-1 Model (Student-Level):

$$Y_{\mathrm{ij}} = \beta_{0j} + \beta_{1j} G_{ij} + r_{ij},\ r_{ij} \sim N(0,\ \sigma^2),$$

Where Y_{ij} is the test score for student i in school j, G_{ij} is the gender for student i in school j, β_{0j} is the average achievement for male students in school j, β_{1j} is the effect of gender (the difference between being a female and being a male) in school j, and r_{ij} is the random error term with a normal distribution.

Level-2 Model (School-level):

$$\beta_{0j} = \gamma_{00} + \gamma_{01} S_j + \mu_{0j}$$
$$\beta_{1j} = \gamma_{10} + \gamma_{11} S_j + \mu_{1j,}$$
$$(\mu_{0j}, \mu_{1j})^T \sim \mathrm{N}[(0,0)^T, \begin{pmatrix} \tau_{00} & \tau_{01} \\ \tau_{01} & \tau_{11} \end{pmatrix}]$$

Where S_j is the school type for school j, γ_{00} and γ_{10} are the average public school achievement and gender effect, γ_{01} is the effect of school type on achievement, γ_{11} is the interaction effect of school type on student gender, and μ_{0j} and μ_{1j} are the school-level errors for intercept and for gender effects.

Combining the two levels together, we have:

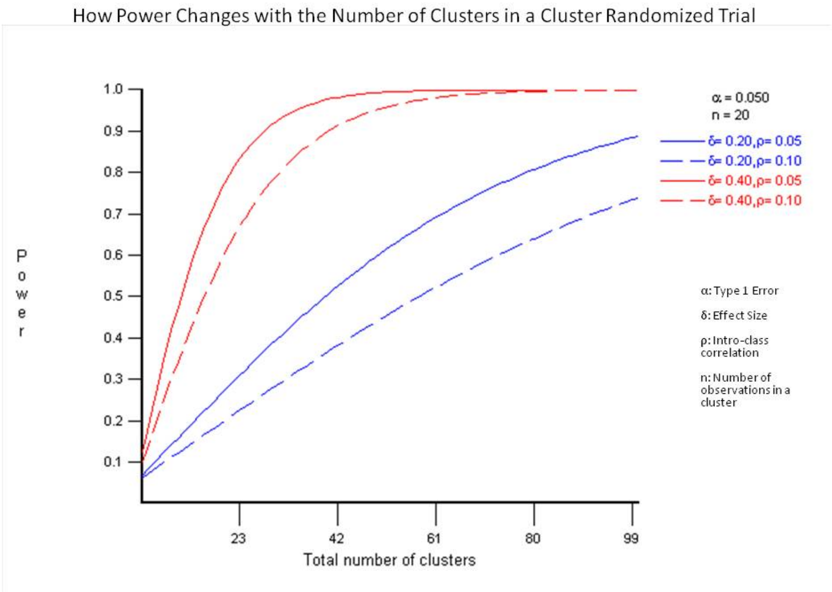
$$Y_{\mathrm{ij}} = \gamma_{00} + \gamma_{01} S_j + \gamma_{10} G_{ij} + \gamma_{11} S_j G_{ij} + \mu_{0j} + \mu_{1j} G_{ij} + r_{ij}$$

The model described above is a full model, which is also called an intercepts-and-slopes-as-outcomes model. μ_{0j} and μ_{1j} are the random effects of schools. If we treat the school effects as fixed, there will be no μ_{0j} and μ_{1j}, and the model is the same as a two-way ANOVA model with interactions. If no control variables are included in the model, the model is called an unconditional model which can provide us information about the variance component (how much of the total variance is between-school variance and how much is within-school variance). If the student-level variable is absent, the model becomes a conditional means model. If no school-level variable is available, the model becomes a random-coefficients model. Detailed discussion about those models is beyond the scope of this chapter. Of the statistical software for hierarchical modeling, SAS statistical software a mixed procedure which looks at the model from the variance-covariance structure of the error terms, and HLM software (Raudenbush et al 2004) deals with it using the logic as described above.

The design of a good randomized trial requires attention to statistical power analysis so as to assure that the mean effects of the tested intervention are detectable. In cluster randomized trials, important work has been done by Spybrook et al. (2007) and Dong and Maynard (2013). Basically, in a two-level scenario, it is the number of places, rather than the number of individuals within places, that is crucial for the calculation of power. In the school intervention example above, we the unconditional model provides the variance component information—the intra-class correlation (ICC)—as $\frac{\tau_{00}}{\tau_{00}+\sigma^2}$, i.e. the percentage of variance at the school level relative to the total variance in outcomes. Knowing the ICC, one can compute the design effect, DEFF= $1+ICC(n-1)$, where n is the average number of observations within a cluster. DEFF indicates the loss of efficiency engendered by using cluster sampling instead of random sampling of units across clusters. For example, in the school intervention case, one may need 200 independent students to achieve a power of 0.80. Assuming that the ICC=0.2, and n=30 in a study in which independent schools are clusters, DEFF=6.8. The sample size required will be 200*6.8=1360 students; therefore, one needs to sample at least 45 schools.

A larger ICC indicates the clusters are more different than similar (or individuals in the same cluster are more alike), resulting in the need to sample more clusters. Figure 4 below is a graph obtained by using the Optimal Design software (Dong and Maynard 2013). It illustrates statistical power (verticle axis) changes when the total number of clusters (horizontal axis) changes, assuming a two-level scenario of the kind described , with a P=.05 type 1 error and an average 20 observations in each cluster.

Figure 4: The Relationship of Power with the Number of Clusters in a Two-level Cluster Randomized Trial.



Similar statistical power analysis can be done using software that is dependable and accessible on the WT Grant Foundation's web site. Any such work depends on what the trial's designer and substantive experts expect about the size of effects that are important, notably minimally detectable effect sizes. Empirical evidence on expected effect sizes is given in Hill, Bloom, Black, & Lipsey (2008), for instance.

6.Controversies, Disagreements, and Contentions

For evaluators who desire to produce fair and defensible estimates of the relative effects of interventions, a randomized controlled trial has a real benefit, provided, of course, that it is executed well. At the most basic level, the use of randomization minimizes the possibility of systematic differences between groups at the outset of the study. On the other hand, passive observational studies are often easier to mount than randomized trials. RCTs are very specific to the context in which they are implemented, which limits the generalizability (external validity) of the findings. Although replications of the trial in different settings and the use of meta-analytic techniques can provide some measure of generalizability, this approach is not appropriate for all types of studies.

There are, of course, well thought out approaches to analyzing non-randomized trials and observational studies that attempt to estimate the effects of interventions. Rosenbaum's (2010, 2002) books are very informative on this account. The methodological advances can produce interesting, useful, and defensible results. This is provided that one is willing to make assumptions that hinge on the statistical models that are presumed to underlie the structure of the data. For instance, in making a fair between-group comparison in a non-randomized study so as to estimate an intervention's effects, one must usually assume that one has the right statistical model (functional form), with the right variables in the model (no important ones having been omitted), and that these variables are measured in the right way.

Debates wax and wane about whether, when, and how these statistical models may suffice, and the role of controlled trials in the modeling context. Imbens' (2010) work in response to work by Heckman and Urzua (2010) and Deaton's (2009) are instructive. An experiment does not rely on any such assumptions, providing the design and protocol are well-implemented.

7.Realities of Randomized Controlled Trials

The design of any randomized trial must suit the setting in which it will be conducted and each trial is unique in some respects, even when interventions have been proved successful based on trials in other locations. Reconnaissance prior to the trial in a new setting is then essential for evaluators. This implies basic questions that must be answered. Will the units of random assignment, for instance, be individuals or institutions? Will the experiment's environment be stable enough to assure that the interventions can be deployed and the experiment carried out?

Ethics must be considered. Strategies for tailoring the trial's design to satisfy ethical standards, to assure that people's rights are recognized, are outlined by Boruch et al (2012). Politics and legislatures play a role at times in demanding (or not demanding or forbidding) dependable evidence based on randomized trials; Gueron and Rolston (2013) give history in the welfare sector in the U.S. And a theory or logic about how the intervention is supposed to work is essential prior to the start of any trials. The statistician, focused only on the analytical framework, does not need to know all of this. These are the concerns of the evaluation specialist, the project manager and the study's overall director. But the statistician *must* be able to tailor the experiment's design to recognize realities and to fit the design to the setting.

8.Access to Data from Randomized Trials for Reanalysis and Replication

At times, the data produced in randomized trials are made available through the internet so as to permit independent secondary analyses of the data. These analyses may be in the interest of confirming or disconfirming original analyses. Or, they may be in the interest of uncovering new results: doing different kinds of analyses or deeper analyses of certain subgroups in the study.

For instance, data from Mexico's PROGRESA/Oportunidades trial on conditional cash transfers, have been used by economists and education researchers. The data are at:

http://evaluacion.oportunidades.gob.mx:8010/en/index.php. Chetty et al (2011) used accessible data on the Tennessee class size experiments to and linked these records over 20+ years to understand the effect of the smaller class sizes on wages.

By making data available, the research team provides the opportunity for others to duplicate and verify their work, as well as to extend the analyses, using new method or sets of variables that not part of the initial research plan. Publically-sponsored research studies must often make their data and results available as a condition for the use of government funding. Data from trials or other empirical studies, such as surveys sponsored by the United States Department of Education's Institute for Education Sciences, are made available often, as are data from experiments sponsored by the US National Institute of Justice, National Institutes of Mental Health, and others. Making data available is costly and complex, however, and not all data can be made available for secondary analysis. Nonetheless, the Institute for Education Sciences (US) have demonstrated initiative in assuring that applicants for large scale grants for experiments and related studies include a data sharing plan in their applications. See http://ies.ed.gov/datasharing_policy.asp. These rules to access government contracts, as opposed to grants, are demanding. Nonetheless, one can apply for access to data sets under licensing agreements. See http://ies.d.gov/pubsearch/licenses/asp for a description of data sets available from randomized trials. Data from research in the social sciences, economic and other fields are also housed at the Interuniversity Consortium for Political and Social Research (ICPSR): https://www.icpsr.umich.edu/icpsrweb/landing.jsp). The conditions vary for access and use of these data.

Independent replication of studies is important in the sciences so as to enhance understanding beyond reanalysis of a particular data set from a particular trial. Exact and uniform replication of a randomized trial done in one place at one point in time may not be possible at another place and time, of course.

People in each site differ systematically. Times change. The local culture and context may require adaption rather than exact replication. In the arena of social, health and crime prevention, Flay et al (2005), for instance, provides standards of evidence in prevention research. Valentine et al (2011) lay out different ways social scientists and practitioners can think about replication standards.

9. Acknowledgements

Work on this and related topics has been supported by National Science Foundation Grant DRL-1337237. We are grateful to Xinyin Chen, to graduate students in Penn's ED 871 Seminar, and to an anonymous reviewer for comments on earlier drafts of this chapter.

10.References

Behrman, J. R., Parker, S. W., and Todd, P.E. (2010) Do Conditional Cash Transfers for Schooling Generate Lasting Benefits? A Five Year Follow-up of PROGRESA/Oportunidades. *Journal of Human Resources. 46*(1), 93-122.

Boruch, R. (1997). *Randomized Experiments for Planning and Evaluation: A Practical Guide*. Thousand Oaks, CA: Sage.

Boruch R., Merlino, J., & Porter, A. (2012) Where Teachers Are Replaceable Widgets, Education Suffers. Education Week, 31(27), 20-21. Retrieved from http://www.edweek.org/ew/articles/2012/04/04/27porter.h31.html?qs=churn

Boruch, R. and Rui, N. (2008) From Randomized Controlled Trials to Evidence Grading Schemes*. Journal of Evidence Based Medicine, 1,* 41-49.

Boruch, R., Weisburd, D. and Berk, R. (2010) Place Randomized Trials. In A. Piquero and D. Weisburd (Eds) *Handbook of Quantitative Criminology*. New York: Springer, pp.481-502.

Boruch, R. and 芮宁 (2009) 从 RCT 到证据分级评价系统：循证实践在社会科学领域的发展现状。中国循证医学杂志， 9(1)，12-18.

Boruch, Robert, Cecil, Joe, Turner, Herb, Victor, Timothy, and Hyatt, Jordan (2012) Resolving Ethical Issues in Randomised Controlled Trials. In Erica Bowen and Sarah Brown (Eds) *Perspectives on Evaluating Criminal Justice and Corrections: Advances in Program Evaluation Volume 13.* Bingley UK: Emerald Group Publishing, pp 95-128.

Box, G. E., Hunter, J. S. and Hunter, W. G. (2005). Statistics for Experiments: Design, Innovation, and Discovery, 2nd Edition. Hoboken, NJ: Wiley

Brookmeyer, R., & Chen, Y. Q. (1998). Person-time analysis of paired community intervention trials when the number of communities is small. *Statistics in Medicine, 17*, 2121-2132.

Bruns, B., Filmer, D., and Patrinos, H. (2011) *Making Schools Work: New Evidence on Accountability Reforms.* Washington DC: World Bank.

Campbell, D. T., and Stanley, J. *Experimental and quasi-experimental designs for research*. Chicago: Rand McNally and Company, 1963.

Chetty, Raj, Friedman, John, Hilger, Nathaniel, Saez, Emmanuel, Schazenbach, Diane, and Yagan, Danny (2011) How Does Your Kindergarten Classroom Experience Affect Your Earnings? Evidence from Project STAR. *Quarterly Journal of Economics*. *26* (4), 1593-1660.

Coyle, S., Boruch, R., and Turner, C. (Eds.). (1991). *Evaluating AIDS prevention programs*. Washington, DC: National Academy of Sciences Press.

Deaton, A., (2009) Instruments of Development: Randomization in the Tropics, and the Search for the Elusive Keys to Economic Development. *NBER Working Paper* #14690.

Dehue, Trudy (2001) Establishing the Experimenting Society: The Historical Origins of Social Experimentation according to Controlled Designs. *American Journal of Psychology, 114* (2) 283-302.

Dodd, S. C. (1934). *A Controlled Experiment on Rural Hygiene in Syria*. Beirut: Publ. Fac. Arts Sci., Am. Univ. Beirut (Soc. Sci. Ser. No. 7). 128 pp.

Dong, N. and Maynard, R. A. (2013). PowerUp! A Tool for Calculating Minimum Detectable Effect Sizes and Sample Size Requirements for Experimental and Quasi-experimental Designs. *Journal of Research on Educational Effectiveness*. 6 (1) doi: 10.1080/19345747.2012.673143

Donner, A. and Klar, N. (2000) *Design and Analysis of Cluster Randomized Trials in Health Research.* London: Arnold.

Edgington, E. S., and Onghena, P. (2007) *Randomization Tests.* (Fourth Edition). London and New York: Chapman and Hall/CRC.

Evans, I., Thornton, H., and Chalmers, I. (2006) *Testing Treatments: Better Research for Better Health Care.* London: The British Library.

Evans, I., Thornton, H., and Chalmers, I. (2008). 验证治疗措施的公平—高质量研究促进高质量卫生保健。中国循证医学杂志编辑部。

Finckenauer, J.O. (1982). *Scared Straight and the Panacea Phenomenon*. Englewood Cliffs, NJ: Prentice Hall.

Finn, J. and Achilles, C. (1990) Answers and Questions about Class Size: A Statewide Experiment. *American educational Research Journal. 27,* 557-577.

Fisher, R. A. (1935) *The Design of Experiments*. Oxford, England: Oliver & Boyd.

Flay B. R., Biglan, A., Boruch, R., Castro, F. G., Gottfredson, D., Kellam, S., Moscicki, E., Schinke, S., Valentine, J. C. and Ji, P. (2005) Standards of Evidence: Criteria for Efficacy, Effectiveness and Dissemination. *Prevention Science, 6(3)*, 151-175.

Forsetlund, Louise, Chalmers, Iain, and Bjorndahl, Arild (2007) When was Random Assignment First Used to Generate Comparison Groups in Experiments to Assess the Effects of Social Interventions? *Econ. Innov.New Techn,* Volume *16* (5), 371-384.

Gueron, J. and Rolston, H. (2013) *Fighting for Reliable Evidence*. New York: Russell Sage Foundation.

Hayes, R. and Moulton, L. (2009) *Cluster Randomized Trials*. Boca Raton Florida: Chapman and Hall/CRC/Taylor and Francis Group.

Heckman, J., and S. Urzua., (2009) “Comparing IV With Structural Models: What Simple IV Can and Cannot Identify,”NBER Working Paper, # 14706.

Hill, A. B. (1952). The Clinical Trial. *New England Journal of Medicine, 247*, 113 –119.

Hill, C. J., Bloom, H. S., Black, A. R. and Lipsey, M. W. (2008), Empirical Benchmarks for Interpreting Effect Sizes in Research. *Child Development Perspectives, 2*, 172–177.

Imai, K., King, G. and NALL, C. (2009). The essential role of pair matching in cluster-randomized experiments, with application to the Mexican universal health insurance evaluation. *Statistical Sciences*. *24* 29–53.

Imbens, G. (2010) Better LATE than Nothing. *Journal of Economic Literature 48,* 399-423.

Jadad, A. R. (1998) *Randomised Controlled Trials: A User's Guide*. London, England: BMJ Books.

Kirk , R. E. (2012). *Experimental design: Procedures for the Behavioral Sciences (4th ed.).* Thousand Oaks, CA: Sage Publications, Inc.

Konstantopoulos, S. and Sun, M. (2012) Is the Persistence of Teacher Effects in Early Grades Larger for Lower Performing Students? *American Journal of Education. 118*(3), 309-339.

Levin, Henry M. and McEwann, Patrick, J. (2001) *Cost-Effectiveness Analysis: Methods and Applications.* (2nd Edition). Thousand Oaks California and London England: Sage Publications.

Liao, Shicheng (廖世承). (1925). 东大附中道尔顿制实验报告。商务印书馆。

Meldrum ML (2000). A brief history of the randomized controlled trial: From oranges and lemons to the gold standard. *Hematol Oncol Clin North Am* 14 (4): 745–60

Morgan, Kari and Rubin, Donald (2016) Rerandomization to Balance Tiers of Covariates. *Journal of the American Statistical Association. 110* (512), 1412-1421.

Mosteller, F. and Boruch, R. (2002) (Eds) *Evidence Matters.* Washington DC: Brookings Institution Press.

Murray, D. M. (1998) *Design and Analysis of Group Randomized Trials.* Oxford: Oxford University Press.

Neyman, J. (1923) On the Application of Probability Theory to Agricultural Experiments. Essay on Principles. Section 9. *Statistical Science 5(4):* 465–472.

Parker, S. and Teruel, G. (2005) Randomization and Social Program Evaluation: The Case of Progresa. *Annals of the American Academy of Political and Social Science,* Volume 599 (May), 199-219.

Petrosino, A., Turpin-Petrosino, C., and Buehler, J. (2002). "Scared Straight" and other juvenile awareness programs for preventing juvenile delinquency (Cochrane Review). In: *The Cochrane Library*, Issue 2. Oxford: Update Software.

Piquero, A. R. and Weisburd, D. (Eds) (2010) *Handbook of Quantitative Criminology.* New York, Dordrect, Heidelberg, and London: Springer.

Raudenbush, S.W. (1997). Statistical analysis and optimal design for cluster randomized trials, *Psychological Methods 2(2*): 173-185.

Raudenbush, S. W. and Bryk, A. (2002). *Hierarchical Linear Models: Applications and Data Analysis Methods*. 2nd ed.. Thousand Oaks, CA: Sage; 2002.

Raudenbush, S.W., and Liu, X.F. (2000). Statistical power and optimal design for multisite randomized trials, *Psychological Methods 5(2*): 199-213.

Raudenbush, S., Bryk, A., Cheong, Y., Congdon, R., and duToit, M. (2004) *HLM6: Hierarchical Linear and Nonlinear Modeling.* Lincolnwood Illinois: Scientific Software International.

Riecken, H. W. and Boruch, R. (Eds) (1974). *Social Experimentation.* Academic Press, New York.

Rosenbaum, Paul (2002) *Observational Studies.* (Second Edition). New York: Springer.

Rosenbaum, Paul (2010) *Design of Observational Studies*. New York, Dordrecht, Heidelberg, London: Springer.

Shadish, W. R., Cook, T. D., and Campbell, D. T. (2002). *Experimental and Quasi-Experimental Designs for Generalized Causal Inference*. Boston: Houghton-Mifflin.

Small, D. S., Ten Have, T. R., and Rosenbaum, P. R. (2008). Randomization inference in a group-randomized trial of treatments for depression: covariate adjustment, noncompliance, and quantile effects. *Journal of the American Statistical Association 103*, 271–279.

Soydan, Haluk and Palinkas, Lawrence (2014*) Evidence Based Practice in Social Work: Development of a New Professional Culture*. London and New York: Routledge/Taylor and Francis Group.

Stolberg, H. O., Norman, G., & Trop, I. (2004). Randomized Controlled Trials. *The American Journal of Roentgenology, 183(6)*, 1539-1544.

St. Pierre R., et al. (1990). *National Evaluation of the Comprehensive Child Development Program: Study Design*. Cambridge, MA: Abt Associates Inc.

U.S. House (1979). Committee on Education and Labor. *Oversight on Scared Straight: Hearings before the House Subcommittee on Human Resources.* 96th Congress, 1st Sess. 4 June. Washington, DC: Government Printing Office.

Valentine, J. C. and others (2011). Replication in Prevention Science. *Prevention Science. 12,* 103-117.