5

Randomized Controlled Trials

A Gold Standard or Gold Plated?[1]

Leonard Bickman

Stephanie M. Reich

Randomized controlled or clinical trials (RCTs) have been taking on increasing importance, especially outside of the medical field.[1] The number of RCTs is increasing as well as the number of areas in which they are conducted (Bloom, 2008; Boruch, Weisburd, Turner, Karpyn, & Littell, 2009). Moreover, these designs are being recommended and privileged over other designs by prestigious research organizations (e.g., Shavelson & Towne, 2002). In addition, several U.S. federal agencies deemed the RCT as the gold standard that should be used not only in considering the funding of research and evaluation but also in initiating and terminating programs (Brass, Nunez-Neto, & Williams, 2006). However, over the last several years, there has been considerable debate about whether RCTs should be considered the ultimate standard design (Cook & Payne, 2002; Maxwell, 2004).

While most would argue that the RCT is a powerful research design, many debate whether it should be labeled as the *gold standard* for research trying to determine causality. Dissenters of this design as the model for research cite issues of appropriateness, ethics, feasibility, and cost, arguing that other methods can answer causal questions equally well. Most claim that RCTs are more appropriate for medical and basic science investigations,

---

i AUTHORS' NOTE: We are not the first to use metals to describe standards. Both Rossi (1987) and Berk (2005) have used similar terms.

where such procedures as blinding or masking of experimental conditions are possible, and not for the bulk of behavioral science research. For instance, Scriven (2005) argues that numerous other designs can infer causation for far less money and in less time than the RCT. He lists cross-sectional designs with triangulation of information, regression discontinuity designs, and interrupted time series designs as just a few equally, if not more, credible designs. Others note that some research areas are not amenable to RCTs, such as safe-sex interventions for HIV/AIDS prevention (e.g., Van de Ven & Aggleton, 1999); broad-spectrum programs (e.g., Cook & Campbell, 1979); and, comically, the efficacy of parachutes (Smith & Pell, 2003).

The debate over RCTs as the gold standard for research is highly relevant to the evaluation of social and educational programs that are designed to improve the quality of lives of others. At the heart of this dispute is whether RCT is the only credible design for testing causal relationship. That is, are the findings believable, trustworthy, convincing, and reliable, and do specific designs, such as the RCT, truly yield more credible findings? To address these issues, this chapter will first briefly focus on the issue of credibility and its subjective nature. Then we will consider how credible RCTs are for the study of cause-and-effect relationships. We will describe some of the important limitations of RCTs and briefly consider alternative designs.

Credibility

Credibility is a highly subjective term. The quality of evidence cannot be determined without knowing which questions were asked and why, what evidence was gathered to answer these questions, who asked the questions and gathered the evidence, how the evidence was gathered and analyzed, and under which conditions the evaluation was undertaken. In addition to these foundational issues, credibility is also influenced by the depth and breadth of the study as well as whether the findings are based on a single study or conclusions drawn about the effectiveness of sets or types of interventions. In other words, much more goes into the judgment of the credibility of research than just the design. While we will discuss some of the broader issues related to credibility, we will concentrate mostly on how the design of the study affects its credibility, since this is the focus of the current debate over the whether RCT should be the standard for research and evaluation.

One aim of this chapter is to discuss elements of RCTs that increase or threaten credibility. This is a daunting task, since we believe that credibility is highly subjective. Credibility is a perceived quality and does not reside in an object, person, or piece of information. Thus what is called credible will be different for different people and under different situations. For assessing credibility in

evaluation, we argue that there needs to be a consensus among persons recognized as experts in what they label *credible*. In this chapter, we will describe the elements of an RCT that make it more or less credible to determine causal relations.

It is important to note that the RCT's designation as the gold standard is based on theoretical or potential characteristics of RCTs. It is also the case that many of the limitations attributed to RCTs are more potential problems than proven actual limitations. The chapter will describe some of the issues that arise when implementing an RCT that affect its credibility.

Causation, Epistemology, and Credibility

An important component of credibility is what is defined as knowledge and which methods are used to obtain, discover, or create this knowledge. For instance, in comparing qualitative and quantitative methods, the ontological, epistemological, and theoretical issues of what are "data," how they are analyzed, and what interpretations are drawn from them are viewed quite differently (Brannan, 2005). Although there have been some efforts to critique the quality, trustworthiness, and transferability of qualitative research (e.g., Rolfe, 2006)—and, similarly, the quality and internal, external, statistical conclusion—and construct validity of quantitative research (Shadish, Cook, & Campbell, 2002), epistemological disagreements exist on both sides. Even those who utilize mixed methods acknowledge that the epistemological arguments remain (Howe, 2004; Mertens & Hesse-Biber, 2013; Onwuegbuzie & Leech, 2005).

Although epistemology is an important component of credibility, defining it is beyond the scope of this chapter. Discussing what qualifies as evidence or "truth" would take far more space than this book allows. Therefore, for the purpose of this chapter, we will focus solely on the credibility of studies utilizing quantitative methods. In particular, our discussion will examine claims that the *cause* resulted in the *effect* and whether RCTs should be the standard in quantitative program evaluation and other applied research approaches. While we limit this discussion to post-positivistic quantitative methods, we recognize that no single method can be used to address every question and that different methodologies may yield different answers.

The Cost of Making a Wrong Decision about Causality

We know that there are costs in making a wrong decision. To call a program effective when it is not means that valuable resources may be wasted and the search for other means to solve the problem will be hindered. Some programs may not only be ineffective but also harmful. In such cases, the costs of a wrong

decision would be higher. On the other hand, falsely labeling a program as ineffective would mean that clients who would have benefited from the intervention would not have that benefit. Some examples illustrate the point. Imagine the identification of a drug as effective when it is not. Patients taking this drug would not have their condition improve and, in some cases, would continue to deteriorate. Further, if producing the drug takes hundreds of millions of dollars, the financial cost of this mistake is high. Conversely, if the drug were ineffective but costs only pennies to manufacture, there would be little waste of financial resources. While those taking the medication would not benefit from the treatment, the low cost would not inhibit the development of other medications. Thus it is more likely that patients would receive effective medication in the future (assuming pharmaceutical research continues), although not as likely as if the drug was accurately labeled as ineffective. While mislabeling a drug as effective is problematic, this problem is exacerbated when the cost is high. The ramifications of the drug example above are somewhat straightforward, yet for most interventions, the situation is not as clear. Most studies do not compute the costs of the intervention, let alone the cost of a wrong conclusion. Thus credible research is needed to minimize wrong decisions and the subsequent human and financial costs. Later, we will describe some actual medical decisions that were based on the use of a weak design.

The RCT as the Gold Standard of Evidence

RCTs are often thought of as providing more credible evidence than other research designs. This sentiment was demonstrated by the U.S. Secretary of Education's change in funding priorities in 2003 to give highest priority to RCTs.[2] While we agree that an RCT is a powerful research design; it is not immune to threats to validity. In fact, RCTs are just as vulnerable to threats to construct and statistical conclusion validity as other designs. Below, we critique the validity and subsequent credibility of RCTs, with special attention to internal validity—that is, the cause-effect relationship.

Why RCTs Are High in Internal Validity

RCTs have high internal validity because each participant has an equal chance of being assigned to each group, which reduces the likelihood of a systematic bias in the selection and assignment of participants, the main threat to internal validity. Thus random assignment rules out numerous rival hypotheses for the effects found that are based on initial participant differences. We will not describe all the threats to internal validity, as practically any book on program

evaluation or research methods will list them. Instead, we will focus on the more visible threats and those that have not been widely discussed in relation to experimental designs.

More Visible Threats to Internal Validity That RCTs Share

While RCTs minimize some of the threats to internal validity, they are not immune from them all. In fact, the majority of threats to drawing valid causal inferences remain. Some of these threats are well known (such as experimenter effects, allegiance, history, and attrition) and others are less often acknowledged. First we will address the more often-cited threats.

EXPERIMENTER EFFECTS

RCTs can be influenced by the behavior of the experimenter/evaluator. The most commonly acknowledged behaviors are experimenter expectancies in which the experimenter knows who is getting the intended treatment or *cause*, and this either influences how the *effect* is recorded or how the participant is treated, thus influencing the observed effect. Using blinding or masking most often controls for this threat. The most common example of this is double-blinded drug studies in which both the participants and researcher do not know who is receiving the experimental drug or placebo. It is rare to have single-blinded let alone double-blinded RCTs in the social sciences (Kumar & Oakley-Browne, 2001). Moreover, although hundreds of studies have been conducted on experimenter expectancy effects (Rosenthal, 1976) and the broader term, *demand characteristics* (Rosnow, 2002), it is not known how strong these effects are in the real world and under which conditions they will produce a significant bias in the results. A similar caution is expressed about the so-called *Hawthorne effect*, for which we believe there is scant supporting evidence (Adair, Sharpe, & Huynh, 1989).

ALLEGIANCE EFFECTS

Similar to experimenter expectancies is the bias caused by allegiance. Allegiance bias occurs when the experimenter develops or supports the treatment (intended cause), and this influences the observed or reported effect. For instance, in the area of clinical psychology, Luborsky and colleagues (1999) compared 29 different individual therapies and found a correlation of 0.85 between differential investigator allegiances and differential treatment outcomes. Even if participants are randomly assigned to conditions, the experimenters' expectations and behaviors can systematically bias results. Making sure that persons other than the originator of the program conduct the evaluation can control for the allegiance effect.

LOCAL HISTORY

The validity of RCTs can be threatened when one group is exposed to external events that can affect the outcome, but not the other group. In these situations, it does not matter how people were assigned, since a systematic bias is introduced to one group of participants and not the other. While it may be impossible for a researcher to prevent such an occurrence, the possibility of such an event requires the monitoring of both experimental and control conditions.

ATTRITION

One of the strengths of random assignment is its ability to minimize the effects of systematic initial differences among participants. However, this equivalence may be threatened when attrition occurs, especially if the attrition differentially affects one group and not the other. When attrition happens more in one condition than another, or if different types of attrition occur, then the experimental design is transformed into a nonequivalent comparison group design and the benefits of random assignment may be lost (Lipsey & Cordray, 2000). Systematic attrition—that is, attrition that is not at random—can have grave effects on the accuracy of causal inferences (Shadish, Hu, Glaser, Kownacki, & Wong, 1998), especially if the reasons for dropping out are inaccessible and how participants vary is unknown (Graham & Donaldson, 1993).

There are several approaches to diagnosing attrition (Foster & Bickman, 2000), but one of the more popular approaches to correcting the attrition problem is propensity scoring (e.g., Leon et al., 2006; Shadish, Luellen, & Clark, 2006; VanderWeele, 2006). Propensity scoring may also be useful even if differential attrition did not appear to occur. In these situations, propensity scores help account for differences that are not obvious (Rosenbaum & Rubin, 1983). However, the use of statistical equating techniques is not without controversy (Steiner & Cook, 2013). Some question whether it accomplishes its goal while others argue about how it should be done (Baser, 2006; West & Thoemmes, 2008). Moreover, there are important innovative approaches to strengthening the validity of quasi-experiments that do not use statistical adjustment but focus on designing more internally valid research approaches (Cook & Wong, 2008).

Less Well-Recognized Threats to Internal Validity

When random assignment is feasible and well implemented, it can be an effective method for untangling the effects of systematic participant differences from program effects (Boruch, 1998; Keppel, 1991). However, random assignment alone is not sufficient to ensure high-quality evaluations with credible findings. As noted above, numerous other design, measurement, and implementation

issues must be considered in order to make causal inferences. Some of these, which are described above, are commonly noted in the literature. Other threats to internal validity are not often acknowledged. Thus we would like to draw your attention to a few.

SUBJECT PREFERENCE BEFORE RANDOM ASSIGNMENT

Since properly implemented random assignment ensures an equal chance of assignment to each group, it ignores individual preferences, including individual decision making preferences that might exist (McCall & Green, 2004; McCall, Ryan, & Plemons, 2003). For instance, drawing a sample from only those who are willing to be randomly assigned may not produce equivalent groups if participants have preferences for one condition over another. Suppose 80 people want condition A and 20 prefer condition B. Of the 50 people randomly assigned to condition A, the chances are that 40 (or 80%) will end up in their preferred condition. Of the 50 people assigned to condition B, 10 (or 20%) will be in their preferred condition.

Willingness to be randomly assigned and preferences for a treatment condition are clearly not the same. Faddis, Ahrens-Gray, and Klein (2000) experienced this problem in their effort to compare school-based and home-based Head Start services. In this study, researchers found that many families who were randomly assigned to home-based childcare programs rather than Head Start (school-based) centers never enrolled their children and when they did, they were more likely to attrite. Thus the families that enrolled and remained in the evaluation may have been systematically different by condition from those families that did not complete enrollment. It would also appear from the attrition patterns of this study that some families prefer center-based Head Start to home-based Head Start services. These preferences may have affected how beneficial each type of program was to families and the conclusions drawn from the comparison.

Similar results have been found in mental health services research (Corrigan & Salzer, 2003; Macias et al., 2005). For instance, Chilvers and colleagues (2001) found that patients who chose counseling did better than those who received the services because of randomization. Similarly, in a review of the influence of patient and physician preferences in medical research, King and colleagues (2005) found evidence that preferences can influence outcomes; however, this effect was minimized in larger trials.

In order to address the effect of preference in RCTs, some have advocated for a patient preference RCT/comprehensive cohort design or two-stage randomized design (Brewin & Bradley, 1989; Howard & Thornicroft, 2006). In the first two designs, some of the people with clear treatment preferences are enrolled in their desired treatment while the rest (those with and without

strong preferences) are randomly assigned to a treatment condition. In the two-stage randomized design, all participants are randomized into two groups. The first group is able to select their treatment and the second group is randomized to the treatment condition. These variations of the RCT allow evaluators to estimate the degree to which a preference systematically biases results as well as assess how representative the randomized sample is to the general population (those with preferences and without). While this approach may work, it is not feasible in most conditions, since there are rarely the abundant resources (i.e., participants and money) needed to apply these designs.

UNMASKED ASSIGNMENT

Theoretically, RCTs should not have any selection bias because of the random assignment of participants to conditions. Unfortunately, sometimes the random assignment process breaks down. The potential breakdown may be suspected when the person doing the random assignment knows which condition will be used next. This is described as an unmasked trial. Berger and Weinstein (2004) found several instances of this problem in major clinical trials, and Greenhouse (2003) warns that finding significant baseline differences with an unmasked trial is clear evidence of the manipulation of random assignment.

SMALLER SAMPLE SIZE THAN EXPECTED

The reduction of selection bias due to random assignment is based on the notion that with a large enough sample, potentially biasing participant characteristics will be distributed evenly across groups. However, many studies do not include sample sizes large enough to truly protect against systematic bias. As Keppel (1991) warns,

> we never run sufficiently large numbers of subjects in our experiments to qualify for the statistician's definition of the "long run." In practice, we are operating in the "short run," meaning that we have no guarantee that our groups will be equivalent with regards to differences in environmental features or the abilities of subjects. (p. 15)

In small sample studies, the threat of selection bias is much greater, regardless of random assignment of participants to conditions.

ONE-TIME RANDOM ASSIGNMENT

If we took the same people and reassigned them to groups randomly, there would inevitably be differences in group means obtained, as "any observed

comparisons between the outcome for the experimental group and the outcome for the control group confounds possible treatment effects with random variation in group composition" (Berk, 2005, p. 422). Thus findings from an RCT only provide information on a specific configuration of people and do not necessarily apply if the people were regrouped. Thus all single studies should be judged with caution. This is a reminder that in science all findings are provisional—regardless of the method used.

INACCURATE UNIT OF ASSIGNMENT

Some evaluation and research questions utilize different units of an organization for random assignment but another level for data analyses. If the data analysis level is used as the level of assignment instead of the organization, then the benefits of the RCT are minimized. For example, a professional training program for teachers that should increase student achievement uses individual student achievement scores for the analysis, which ignores nonindependence of students in the same class. Instead, the proper unit of analyses should be the class average, not the individual student (Boruch, Weisburd, & Berk, 2010). Large threats to validity are introduced when a typical RCT, rather than cluster RCT (described later in the chapter), is used (Raudenbush, 2008). Further, when interventions are at the population level, meta-analyses have found that a bias toward identifying more treatment effects with RCT than observational time series designs exist (Concat, Shah, & Horwitz, 2000; Sanson-Fisher, Bonevski, Green, & D'Este, 2007).

Threats to Between-Group Differences—Even in RCTs

There are other threats to the internal validity of randomized experiments found in almost any textbook on research design that seem plausible but appear to be without any empirical support, at least none we could find.

RESENTFUL DEMORALIZATION

When the control group or a weaker treatment group knows that they are not getting the treatment, they may become demoralized and not perform as well. Thus differences at posttest may be due to decreased performance of the control group and not enhanced performance of the treatment group. Unfortunately, it is difficult to demonstrate that the control group deteriorated while the treatment group held its own rather than a real effect occurring. Blinding or masking helps protect against this. Moreover, we do not know if

this artifact is rare or the magnitude of its effect. Thus this may be a potential threat with little practical consequences.

COMPENSATORY RIVALRY—JOHN HENRY EFFECT

Knowledge of whether you are in the treatment or control group can also have the opposite effect of demoralization. Instead of losing motivation, the control group members may become even more motivated to perform, just to "show them." John Henry may have been a real person, but his relevance to research design is based on legend. John Henry was a "steel-driving man" who competed against a steam machine to drive spikes into the railroad crossties. Although he won the race, he died soon after. Because of his knowledge of the machine's advantage, John Henry did not become demoralized but became more motivated to beat the machine. In a similar fashion, a control group's members might compete more strongly against a more heavily resourced experimental condition to demonstrate their prowess.

Lack of External Validity in RCTs

While this chapter focuses on the credibility of RCTs in drawing causal inferences, we must note the most often-cited criticism of this design is its reduced external validity. Clearly, if the focus of an evaluation is whether a program is effective, then understanding for whom and under what conditions it is effective is important as well. Unfortunately, the nature of volunteerism and the evaluator's ability only to randomly assign those willing to be in any group can limit the generalizability of findings. This has led many to question whether causal inferences are enough of a benefit when RCTs may create artificial circumstances and findings that may not be generalizable to other settings, people, or times (Cronbach, 1982; Heckman & Smith, 1995; Raudenbush, 2002). Berk (2005) states this strongly by claiming,

> It cannot be overemphasized that unless an experiment can be generalized at least a bit, time and resources have been wasted. One does not really care about the results of a study unless its conclusions can be used to guide future decisions. Generalization is a prerequisite for that guidance. (p. 428)

The validity of the RCT depends upon the investigators' intention to study the effectiveness of a cause in the real world or its efficacy in more of a laboratory context. While Cook and Campbell (1979) described internal validity as the *sine qua non* of research, external validity is essential for research in applied

contexts. It is clear that the health sector values internal validity over external validity (Glasgow et al., 2006), as the research standards of the medical field (e.g., Consolidated Standards of Reporting Trials [CONSORT]) do not deal with external validity. Currently, it is likely that a tightly designed study that is high in internal validity is more likely to be published and have a grant application approved even if it is low in external validity.

Lack of generalizability is a common concern for medical research, as participants are often selected to be younger, healthier (lacking comorbid diseases than the one being studied), and, for women, not pregnant (Van Spall, Toren, Kiss, & Fowler, 2007). Such selection of participants identifies treatments that may not be effective to types of people most likely to need treatment (Apisarnthanarax et al., 2013; Bhatt & Cavender, 2013; Rothwell, 2005; Van Spall et al., 2007). For instance, people over the age of 65 account for 53% of new cancer diagnoses but are only 33% of the participants in clinical trials for cancer treatments (Leaf, 2013). Leaf's opinion piece about the generalizability of RCTs in the *New York Times* illustrated concern among the general public about RCTs, given the over 100 commentaries that followed this publication.

Judging external validity is often more difficult than assessing internal validity. How are we to know if the results will apply in the future in some other context? Predicting the future is difficult under any circumstances, but it is especially difficult when we do not know what factors moderate the treatment. We agree that it is too limiting to focus solely on internal validity. The best way to assure external validity is to conduct the study in a setting that is as close as possible to the one that the program would operate in if it were adopted and to include the people that would typically use that setting or be affected by the problem the intervention hopes to ameliorate. We see the recent emphasis on research transportability and bridging science and practice to be a step toward valuing external validity.

APPROPRIATENESS OF THE COMPARISON

Another aspect of external validity beyond volunteerism is the use of the appropriate comparison. Some RCTs are well implemented but conceptually flawed due to the use of an inappropriate comparison group. When this occurs, the internal validity may be high but the utility of the study (i.e., external validity) is low. For instance, in a review of pharmaceutical-funded versus independent investigator-implemented RCTs, Lexchin, Bero, Djulbegovic, and Clark (2003) found that new drugs were typically compared to placebo conditions rather than another medication. As such, large effects were found yet the drugs' performance in comparison to current treatments was unexplored. Here, the RCTs were well executed but were not much of a contribution to science or human well-being. Thus external validity is essential for the credibility if the findings are going to be applied.

Other Issues Raised with RCTs

PRIVILEGING CERTAIN TYPES OF RESEARCH

The belief that RCTs are indeed the gold standard of research implies that other studies employing other designs are weaker and thus less credible. The implications of this are that research areas amenable to the conduct of RCTs are by their very nature more credible. This introduces an unintended side effect of championing certain topic areas as having more credible results when randomized designs are easier to conduct in those areas. For example, it is much easier to conduct a randomized double-blind study of a psychotropic drug than to evaluate a type of psychotherapy. This implies that drug studies are more credible, on the average, than psychotherapy studies. In a similar fashion, total coverage or mandated programs cannot be feasibly evaluated using an RCT. In the former, everyone is receiving the treatment and thus none can be randomly assigned because it would be illegal to withhold the benefit. This does not mean that these programs are immune to study, only that they cannot be studied using an RCT. However, their findings can still be credible.

A similar issue of privileging research occurs when the use of RCTs promotes more investigation into a low-priority area, resulting in resources being spent in less beneficial areas. For instance, in the area of HIV interventions, before truly potent antiretroviral therapy was discovered, 25 RCTs found spuriously significant effects for a variety of treatments of approved, controversial, and contraindicated medications. This resulted in what Ioannidis (2006) calls a *domino effect*, when "one research finding being accepted leads to other findings becoming seemingly more credible as well. This creates webs of information and practices to which we assign considerable credibility, while they may all be false and useless" (p. e36).

ETHICAL ISSUES

There are ethical issues in using the RCT because of the possibility that effective treatments may be denied to some. There have been many discussions of the ethics of design (Boruch et al., 2009; Fisher & Anushko, 2008; Sieber, 2009), and we will briefly summarize them here. First, if it is known with some degree of certainty that one treatment is better than another, then one must question why the study is to be conducted. It is only when we do not know the relative effectiveness that an RCT is called for. Second, in almost all cases, a treatment group is compared to another treatment and not to a no-treatment condition. This use of an active control is important for methodological as well as ethical issues. Some conditions are especially appropriate, from an ethical point of view, for using an RCT. When there are more in need than there are

treatments, it seems especially fair to distribute the treatment by lot or randomly, thus giving everyone an equal chance of obtaining the experimental treatment.

Other times, a RCT is not appropriate if the outcome of a comparison group is known with certainty. For instance, in designing bulletproof fabrics, researchers draped the cloth over pigs and fired a high-caliber weapon at the fabric. In this study, there was no need to shoot at naked pigs, since the outcome of the control condition was well established (Boruch, 2005a). Further difficulties can arise and bias results when the clinical staff does not believe that the effectiveness of the treatment is unknown. Why select a treatment to test if it was not likely that it was effective? Another bias that may be introduced is when the clinician believes that the treatment will work better with a particular type of client. This is one reason why the investigator should maintain strict control over the random assignment process.

Sometimes the random assignment of participants would be unethical, making an RCT inappropriate. For instance, the attempted Children's Environmental Exposure Research Study of the effects of pesticides on babies in Florida was halted due to national response to the questionable ethics of such a study to be carried out by the Environmental Protection Agency (Johnson, 2005). Just because an RCT is possible does not mean it should be conducted. Smith and Pell (2003) make this point well in their satirical article, "Parachute Use to Prevent Death and Major Trauma Related to Gravitational Challenge: Systematic Review of Randomized Controlled Trials."

MULTILEVEL/CLUSTER DESIGNS

In addition to ethical issues, feasibility factors into whether the RCT is the best design to use. Randomized experiments where the unit of assignment is not an individual but a group (such as a school or classroom) offer a special challenge because of covariation due to nesting of units within other units. This research is most often found in educational research, where the treatment may be introduced at the class, school, or even district level. In such cases, there is a strong consensus that the appropriate analysis is at the unit of random assignment. Thus if schools are randomly assigned, then the school should be the unit of analysis. The major drawback to this position is the drastic reduction of degrees of freedom. In the not-too-distant past, researchers randomly assigning an intervention to, for example, eight schools with 500 students each would analyze the data as if there were 4,000 participants (8 x 500) instead of eight. By not considering nesting, the analyst is not taking into account the intercorrelation coefficient (ICC) between students and classes within a school. This ICC can seriously affect the statistical power of the design.

For instance, Varnell, Murray, Janega, and Blitstein (2004) reviewed 60 group- or cluster-randomized trials (GRTs) published in the *American Journal of Public Health and Preventive Medicine* from 1998 through 2002. The authors found that only nine (15.0%) GRTs reported evidence of using appropriate methods for sample size estimation. Of 59 articles in the analytic review, 27 (45.8%) reported at least one inappropriate analysis and 12 (20.3%) reported only inappropriate analyses. Nineteen studies (32.2%) reported analyses at an individual or subgroup level, ignoring group or included group as a fixed effect. Thus interclass correlations were largely ignored. In an attempt to deal with this problem, there are now CONSORT standards (described later) that can be used in evaluating the quality of cluster designs (Campbell, Elbourne, & Altman, 2004).

As noted earlier, using the correct level of analysis has important implications for the design of RCTs. Now, instead of counting students, we need to count schools. While there is not a one-to-one loss in statistical power (i.e., one student is equivalent to one school), it will typically take close to 40 schools to detect a small to medium effect size between two conditions. While this is a difficult requirement, it is more feasible in education than in other areas. Fortunately, there are school districts that have many schools within them. However, this is not the case for areas outside of education. In the area of mental health, randomizing at the mental health center level is difficult. We (LB) conducted an RCT using 40 different mental health sites (Bickman, Kelley, Breda, De Andrade, & Riemer, 2011). If we had not been collaborating with the country's largest provider of mental health services for children, we do not believe the study could have been conducted. Moreover, the expense of conducting these multisite studies is very high when compared to single-site research.

Another concern with cluster RCT designs is determining how to measure outcomes (the effect) and who should be viewed as the participant needing consent. Take, for example, the NEXUS study, a cluster RCT of the effectiveness of a physician educational intervention to reduce the use of lumbar radiographs (Eccles et al., 2001). Although the intervention targeted general practitioners and randomization occurred at the clinic level, no doctors were consented. The "effect" was the number of radiographs ordered per 1000 patients, but no patients were informed that their medical chart was a component of this study (Weijer et al., 2011). Thus additional ethical concerns arise when RCT is used at a group level.

REPORTING OF RCTS

It is beneficial to have standards of quality of RCTs, such as the CONSORT standards (described later) used primarily in the medical field, but it is important

to inquire the degree to which these standards are actually followed in journals. This question has been addressed primarily in several medical fields, with uneven reporting of participant characteristics and attrition by group (Lu, Yao, Gu, & Shen, 2013; Post, de Beer, & Guyatt, 2013). Many medical journals now mandate CONSORT reporting, but this is not the standard for many behavioral science publications that disseminate RCT research and evaluation. Such reporting is important in determining the credibility of the findings and determining their generalizability to other people, places, and times.

Do RCTs Have Different Outcomes from Other Designs?

While RCTs are often called the *gold standard* of research, one must question whether these designs yield different results from quasi-experimental designs. A very visible example of an RCT producing findings that were different from those of a nonrandomized design is the research on hormone replacement therapy for women. Previous nonrandom trials indicated positive effects of the therapy, while a randomized trial found negative effects (Shumaker et al., 2003). However, a more detailed examination suggests that differences in outcomes could be explained by differences in the samples studied (Hernan et al., 2008). In a meta-analysis comparison of psychotherapy studies using RCTs and quasi-experiments, Shadish and Ragsdale (1996) concluded that under some circumstances, a well-conducted quasi-experiment could produce adequate estimations of the results obtained from a randomized study; however, they concluded that randomized experiments are still the gold standard.

Since then, other studies have been completed comparing experimental and quasi-experimental designs. These studies have not produced consistent findings. Some research has found different outcomes favoring randomized experiments (e.g., Glazerman, Levy, & Myers, 2003), while others found that quasi-experiments produced outcomes of unknown accuracy (e.g., Rosenbaum, 2002). However, all the previous studies shared a flaw that made the results even less certain. All of them confound assignment method with other study variables. Shadish, Clark, and Steiner (2008) used an innovative doubly randomized preference trial procedure to untangle these confounds by first randomly assigning students to either a random assignment condition or a self-selection procedure. The authors found that both the random assignment condition and self-selection condition produced similar results after adjusting the self-selection procedure with propensity scores. However, the authors caution that these results may not generalize, since they were conducted in a college laboratory using college students as participants, and that the results appear to be sensitive to how missing data in the predictors were handled. The reader is referred to

commentaries (Hill, 2008; Little, Long, & Lin, 2008; Rubin, 2008) and a rejoinder by the authors for a more in-depth discussion of comparisons between randomized and nonrandomized designs.

Existing statistical approaches to nonexperimental data appear insufficient to compensate for biases that may arise when the pattern of missing data cannot be properly modeled, such as when there are no standards for treatment, when affected populations have limited access to treatment, or when there are high rates of treatment dropout.

Additionally, due to great heterogeneity among people and their lifestyles, typical RCT may not be able to detect effects for certain people under different conditions. Recent pharmaceutical studies have failed to identify why some people respond well to medications while others do not or how to account for the aggregation of data, which may mask impacts, especially when there is insufficient statistical power to detect effects due to participant heterogeneity (Button et al., 2013). This has led some to argue for the use of "data-intensive mega-trials," with sample sizes of tens of thousands of participants for high-profile (i.e., "blockbuster") medications (Ioannidis, 2013). In considering population variability and sample size needs, there is evidence that registries and administrative data can provide insights into the impact of interventions better than RCTs alone (Joynt, Orav, & Jha, 2013). For instance, in the field of pediatric oncology, the use of a registry has helped identify effective treatment protocols for more and less commonly occurring pediatric cancers (Steele, Wellemeyer, Hansen, Reaman, & Ross, 2006). The registry, in conjunction with smaller RCTs, has greatly increased the survival rates of childhood cancer, even though the incidence of pediatric cancers is increasing (National Cancer Institute, 2008). Cardiology has recently followed suit with the creation of the National Cardiovascular Data Registry (NCDR), which is currently targeting clinical practices and outcomes of cardio catheters nationwide (Dehmer et al., 2012).

Approaches to Judging the Credibility of RCTs

There have been several approaches to evaluating the quality of an RCT. Probably the most widespread is the CONSORT. Around 1995, two efforts to improve the quality of reports of RCTs led to the publication of the CONSORT statement. These standards were developed for medical clinical trials but can be used with some modification in any RCT. The CONSORT statement consists of a checklist and flow diagram for reporting a RCT. It was designed for use in writing, reviewing, or evaluating reports of simple two-group parallel RCTs. The standards apply to the reporting of an RCT but may be considered a proxy of the actual conduct of the study. This assumes that the published article accurately describes the methods. Huwiler-Müntener, Juni, Junker, and Egger

(2002) found that the methodological quality of published articles as rated by reviewers was associated with the reported quality indicated by a subset of the CONSORT standards. Soares and colleagues (2004) compared the published quality of RCTs performed by the Radiation Therapy Oncology Group to the actual protocol used in each study. The authors found that the published version of the article underestimated the quality of the protocol used. Unfortunately, the authors only compared the absolute level of quality and not the correlation between quality of the reports and quality of the protocol. The key aspects of the checklist that relate specifically to RCTs are summarized in Table 5.1.

There have been several studies in quite a few medical specialties that have examined whether research published in their journals has improved since the release of the standards. Kane, Wang, and Garrard (2007) examined RCTs published in two medical journals before and after the CONSORT guidelines were evaluated; one journal used the CONSORT statement (*Journal of the American Medical Association* [*JAMA*]) and one did not (*New England Journal*

Table 5.1 Key Aspects of the Checklist That Relate Specifically to RCTs

Section and Topic	*Descriptor*
Randomization	
a. Sequence generation	a. Method used to generate the random allocation sequence, including details of any restrictions (e.g., blocking, stratification)
b. Allocation concealment	b. Method used to implement the random allocation sequence (e.g., numbered containers or central telephone), clarifying whether the sequence was concealed until interventions were assigned
c. Implementation	c. Who generated the allocation sequence, who enrolled participants, and who assigned participants to their groups?
Blinding (masking)	Whether or not participants, those administering the interventions, and those assessing the outcomes were blinded to group assignment. If done, how the success of blinding was evaluated.

SOURCE: Based on Moher, Schulz, and Altman (2001).

of Medicine [*NEJM*]). The results indicated that reporting improved in both journals, but JAMA showed significantly more improvement in all aspects of RCT reporting. Several other studies found similar results (Moher, Jones, & Lepage, 2001; Plint et al., 2006).

The publication of the CONSORT standards has had a greater effect on research in medicine than on research in the behavioral sciences. For instance, Spring, Pagoto, Knatterud, Kozak, and Hedeker (2007) examined the analytic quality of RCTs published in two leading behavioral journals and two medical journals. One of the criteria used was *intention to treat* (ITT), where the analysis includes all participants kept in the assigned group, regardless of whether they experienced the condition in that group. Not only did more reports in medical journals (48%) state that they were going to use ITT than in behavioral journals (24%) but more also used it correctly in the medical journals (57%) than in behavioral journals (34%). Moreover, the articles in the top psychology journals were less likely than those in medical journals to describe a primary outcome, give a reason for estimating study size, describe the denominators that were used in the analysis of the primary outcomes, and account for missing data in analyses.

In the area of social and behavioral research, the Society for Prevention Research (SPR) has established broader standards that provide criteria with which to judge the credibility of evidence for efficacy, effectiveness, and dissemination (Flay et al., 2005). These broader concerns require criteria dealing with the intervention, measures, analysis, and other aspects of research. We will focus on those standards related to determining the credibility of causal statements. These are

> *Standard 3:* The design must allow for the strongest possible causal statements;
>
> *Standard 3.b:* Assignment to conditions needs to minimize the statistical bias in the estimate of the relative effects of the intervention and allow for a legitimate statistical statement of confidence in the results; and
>
> *Standard 3.b.i:* For generating statistically unbiased estimates of the effects of most kinds of preventive interventions, random assignment is essential. (p. 157)

SPR supports the use of nonrandomized designs when necessary—for example, for total coverage programs or when ethical considerations do not allow such a design. The alternatives suggested are the interrupted time series and regression-discontinuity designs. The requirements of these latter two designs severely limit their use. A third design, matched case-control design, is viewed as acceptable only when there is a pretest demonstration of group equivalence. To be credible, this necessitates demonstrating equivalence by using sufficiently powered tests on several baselines or pretests of multiple outcomes and the inclusion of major covariates. Steiner and colleagues (2010) have

shown that it is critical to identify the appropriate covariates if bias is going to be statistically reduced. The key is to provide convincing evidence that the lack of a random assignment process did not result in a correlation between unmeasured variables and condition.

Other Approaches to Establishing Causality

There are other research designs found to be scientifically acceptable in other disciplines that do not involve experiments, let alone RCTs. These disciplines include geology, astronomy, engineering, medical forensics, and medical laboratory testing. In some cases, it is impossible to conduct experiments, such as in astronomy. Like RCTs, approaches in these fields rely on observational methods. These fields have extremely precise predictions, exacting measurement, and exceptionally large numbers of replications in common. While it is important to note that other observational methods are scientifically acceptable ways to establish causality, it is equally important to understand that they are credible only under conditions that rarely, if ever, occur in the social and behavioral sciences.

An approach that is more suitable to the social sciences is known as *program theory*, *theory-driven method*, or *pattern-matching method* (Bickman & Peterson, 1990; Chen & Rossi, 1983; Donaldson, 2003; Scriven, 2005). These are nonexperimental, observational methods that use complex predictions to support a causal hypothesis. In some ways, they are similar to astronomical research, without the precision. This approach is not in opposition to RCTs and has been used in RCTs and quasi-experiments as a way of understanding what is occurring in the black box of the program. We fully realize that RCTs can directly answer only a very limited number of questions, and we must depend on other approaches to fill the gaps. Further, the use of RCTs can be strengthened by the inclusion of qualitative methods, which may help with the interpretation and credibility of research and evaluations findings. For instance, the New Hope anti-poverty intervention used of post-hoc qualitative case studies to disentangle observed quantitative impacts of the RCT and to create a follow-up survey (Gibson & Duncan, 2005).

While both experimental and quasi-experimental designs have numerous threats to the validity of conclusions drawn, the RCT, when implemented well, controls for more threats than nonexperimental designs (in the social sciences). As such, well-executed RCTs are more credible in determining causal relationships. However, the argument has been made that they are inherently less feasible, more costly, and more difficult to implement. Having conducted many RCTs and quasi-experimental designs, we do not agree.

Feasibility

RCTs have been criticized as being difficult to conduct and lacking feasibility. We would argue that the existence of so many RCTs belies that criticism. In a report to Congress, the Congressional Research Service cited the number of RCTs as of 2002 as 250,000 in medicine and 11,000 in all the social sciences combined (Brass et al., 2006). While the number is 20 times more in medicine, 11,000 is still a significant number of RCTs. Two other examples are the Cochrane Collaboration (www.cochrane.org), which has over 350,000 RCTs registered, and the Campbell Collaboration (www.campbellcollaboration.org; Petrosino, Boruch, Rounding, McDonald, & Chalmers, 2000), which contains over 13,000 RCTs in the social sciences (Boruch, 2005b). While there are some specific conditions in which RCTs could not or should not be implemented, these appear to be rare rather than the modal situation.

Practical Issues in the Conduct of RCTs

Currently, both authors are conducting RCTs in such areas as mental health, education, and parenting. In addition to these studies, the senior author has conducted over 20 large-scale RCTs in his career in several areas, representing over $20 million in external funding. Having worked with numerous types of designs, we have not found RCTs to be more difficult to implement than quasi-experimental designs. In fact, the design is usually not the most troublesome aspect of field experimentation.

LIMITED TREATMENT RESOURCES

The *Congressional Research Service Report* (Brass et al., 2006) suggests that RCTs take longer to conduct, are more expensive, and, in general, are more difficult to implement than nonrandomized designs. This is clearly true when RCTs are compared to preexperimental designs such as the simple post-only or pre–post designs. As noted earlier, the fair comparison is with designs that include a control group.

There are some conditions that are optimal for an RCT. If there are more people who want a service than can be provided for, then the fairest way to determine who receives the service is usually by lot or random assignment. Often, the service provider might insist that severity should serve as the criteria for admissions. It is possible to argue that within every measure of severity, there is a band of uncertainty within which random assignment could be done. However, this limits the power of the study as well as the external validity. In such cases, the regression-discontinuity design may be more appropriate.

NONEQUIVALENT CONTROL GROUP

Finding a comparison group can often be more difficult than randomly assigning participants to conditions. Assuming that such a group could be found, there is still the difficulty of convincing the control group organization to collect data when there is no benefit to the organization or clients for participating in the study. In addition, more assumptions must be made and more analyses must be conducted to assess a pretest of group differences and, when possible, propensity for group assignment.

COST IS COST

As mentioned earlier, cost is often raised as a consideration in implementing an RCT. The cost of including a comparison group should be the same whether random assignment is used or not. However, in many quasi-experiments, the control group is not in the same location as the treatment group, often necessitating increased expenses either due to travel or the staffing of a remote site. If the experimenter is very confident of the randomization process (e.g., the sample is very large), then it is possible to do a posttest-only design with the assumption that the treatment and control groups were equivalent before the study started. This would cut data collection costs in half. Hak, Wei, Grobbee, and Nichol (2004) found only one empirical study that looked at the cost effectiveness of different study designs testing the same hypothesis. However, it was not applicable to this discussion of RCTs because it compared a case-control design to a cohort design. We do not have empirical evidence that RCTs are more expensive.

RANDOM ASSIGNMENT

Negotiating for the use of an RCT, where it is ethical and legal, is not as difficult as some may make it appear. For instance, we have implemented random assignment in a study in which parents called to obtain mental health services for their children. While there was initially some resistance by one staff person, the random assignment apparently posed no problem to parents, since 84% agreed to participate in the study (Bickman, Summerfelt, & Noser, 1997).

One of the issues in implementing an RCT is the danger of *crossover* in which those assigned to the treatment group end up in the control group (usually because the organization did not provide the promised services) or when control group participants are exposed to the treatment (also known as *contamination* and *diffusion*). In some situations, physical separation of the participants reduces the probability of this problem, but in the above example, the

parents were all in the same community. Moreover, the service organization could not legally or ethically refuse to treat children. The latter issue was dealt with by asking parents, before they became clients, if they were willing to participate in the random assignment, with the incentive that their child would not be put on the waiting list if he or she was selected for treatment. All of the system-of-care clients received care in that system. In the control group, 6% of the cases received services from the system of care at some point in the study. Thus crossover was not a problem in this particular study (Bickman, Summerfelt, & Noser, 1997).

Crossover problems may occur in educational experiments that are implemented at the school level and take more than a school year to conduct. In these cases, it is not unusual for some students to transfer between schools. In one of our studies, 1.4% of the over 1,000 students changed to schools that had a different experimental condition in the first year, and less than 1% changed schools in the second year.

The issue of crossover analysis can be dealt with in a conservative fashion by using an ITT approach in the analysis. In this case, the analysis is conducted using the original assignment, regardless of the condition the participant experienced. It is conservative because it will water down any potential treatment effects but it maintains the advantages of the random assignment. A discussion of the ITT analysis may be found in Nich and Carroll (2002). Subsequent analyses of the sample can be conducted to look at how those who actually received treatment responded. This type of Treatment on the Treated (TOT) analysis can compare outcomes that are based on receiving treatment versus other artifacts (Dobie & Fryer, 2011; Morris & Gennetian, 2003).

Resistance to random assignment may be a problem, but in our experience, we have not found it to be a significant issue. However, in an experiment we conducted on pediatrician diagnosis and treatment of attention-deficit/hyperactivity disorder (ADHD), we found that the pediatricians took an extraordinarily long time to commit to participate. We think there was a conflict between their values as scientists and not wanting to be subjects in a study. Still, this would have probably occurred whether participants were randomly assigned or not.

Conclusion: So What Counts as Credible Evidence?

While RCTs may be prone to numerous threats to validity, they are nonetheless one of the most credible designs available to researchers. We have described many of the problems of RCTs, both in implementation and in concept. However, we still view them to be a credible choice for quantitative research. They are not really a "gold standard" in the sense of being perfect, but to paraphrase what Winston Churchill said about democracy, we conclude, "For

determining causality, in many but not all circumstances, the randomized design is the worst form of design except all the others that have been tried."

This chapter has explored the RCT as the gold standard for credible research. As noted, credibility of research is determined by assessing whether the findings are believable, trustworthy, convincing, and reliable. Specifically, judging credibility necessitates information about the research questions asked: what evidence was gathered to answer these questions, who asked the questions and gathered the evidence, how the evidence was gathered and analyzed, and under which conditions the evaluation was undertaken. In addition to these foundational issues, credibility is also influenced by the depth and breadth of the study as well as whether the findings are based on a single study or multiple studies. For assessing credibility in evaluation, we argue that there also needs to be a consensus among persons recognized as experts on what they label credible. While credibility is affected by what is viewed as knowledge or truth, this chapter is limited to discussing only post-positivistic quantitative methods. As such, issues of credibility are influenced by statistical conclusion, internal, construct, and external validity.

The RCT is as vulnerable to threats to statistical conclusion validity and construct validity as other methods. However, it is protected against one of the main threats to internal validity: selection bias. Even with this protection, there are several other well-recognized threats as well as less commonly acknowledged threats to internal validity. As described in the chapter, some of the well-recognized threats include experimenter effects; allegiance effects; local history; and attrition, especially differential attrition. Less-familiar threats include participant preferences prior to randomization, unmasked assignment, small sample size, and one-time random assignment.

When considering issues of external validity, RCTs may create an artificial situation in which the findings are not very generalizable. In such cases, credibility of the application of the evaluation is reduced. When conducting any RCT, it is important to use an appropriate comparison and to be sure that group random trials are used when making comparisons across settings or in situations where interclass covariation will influence results. While RCTs may still be prone to numerous threats to validity, this chapter has argued that they are still one of the most credible designs available to researchers and evaluators.

Notes

1. We will use the term *RCT*, known as a randomized clinical or control trial, to represent all randomized experiments, not just clinical trials.
2. While certain quasi-experimental designs were included in this priority, randomized designs were preferred when possible.

References

Adair, J. G., Sharpe, D., & Huynh, C. (1989). Hawthorne control procedures in educational experiments: A reconsideration of their use and effectiveness. *Review of Educational Research, 59*(2), 215–228.

Apisarnthanarax. S., Swisher-McClure, S., Chiu, W., Kimple R. J., Harris, S. L., Morris, D. E., & Tepper, J. E. (2013). Applicability of randomized trials in radiation oncology to standard clinical practice. *Cancer, 119*(16), 3092–3099. DOI:10.1002/cncr.28149

Baser, O. (2006). Too much ado about propensity score models? Comparing methods of propensity score matching. *Value in Health, 9*(6), 377–385.

Berger, V. W., & Weinstein, S. (2004). Ensuring the comparability of comparison groups: Is randomization enough? *Controlled Clinical Trials, 25*(5), 515–524.

Berk, R. A. (2005). Randomized experiments as the bronze standard. *Journal of Experimental Criminology, 1,* 417–433.

Bhatt, D. L., & Cavender, M. A. (2013). Are all clinical trial sites created equal? *Journal of the American College of Cardiology, 61*(5), 580–581.

Bickman, L., Kelley, S., Breda, C., De Andrade, A., & Riemer, M. (2011). Effects of routine feedback to clinicians on youth mental health outcomes: A randomized cluster design. *Psychiatric Services, 62*(12), 1423–1429.

Bickman, L., & Peterson, K. (1990). Using program theory to describe and measure program quality. In L. Bickman (Ed.), Advances in program theory. *New Directions for Evaluation, 47,* 61–72.

Bickman, L., Summerfelt, W. T., & Noser, K. (1997). Comparative outcomes of emotionally disturbed children and adolescents in a system of services and usual care. *Psychiatric Services, 48,* 1543–1548.

Bloom, H. S. (2008). The core analytics of randomized experiments for social research. In P. Alasuutari, J. Brannen, & L. Bickman (Eds.), *Handbook of social research methods* (pp. 115–133). London, England: SAGE.

Boruch, R. F. (1998). Randomized controlled experiments for evaluation and planning. In L. Bickman & D. Rog (Eds.), *Handbook of applied social research methods* (pp. 161–191). Thousand Oaks, CA: SAGE.

Boruch, R. F. (2005a). Comments on the papers by Rawlings and Duflo-Kremer. In G. K. Pitman, O. N. Feinstein, & G. K. Ingram (Eds.), *Evaluating development effectiveness* (pp. 232–239). New Brunswick, Canada: Transaction Publishers.

Boruch, R. F. (Ed.). (2005b, May). Place randomized trials: Experimental tests of public policy [Special issue]. *Annals of the American Academy of Political and Social Sciences, 599.*

Boruch, R. F. (2007). Encouraging the flight of error: Ethical standards, evidence standards, and randomized trials. *New Directions for Evaluation, 113*, 55–73.

Boruch, R. F., Weisburd, D., & Berk, R. (2010). Place randomized trials. In A. R. Piquero & D. Weisburd (Eds.), *Handbook of quantitative criminology* (pp. 481–502). New York, NY: Springer.

Boruch, R. F., Weisburd, D., Turner, H., Karpyn, A., & Littell, J. (2009). Randomized controlled trials for evaluation and planning. In L. Bickman & D. Rog (Eds.),

Handbook of applied social research methods (2nd ed., pp. 147–181). Thousand Oaks, CA: SAGE.

Brannan, J. (2005). Mixing methods: The entry of qualitative and quantitative approaches into the research process. *International Journal of Social Research Methodology, 8*(3), 173–184.

Brass, C. T., Nunez-Neto, B., & Williams, E. D. (2006, March 6). *Congress and program evaluation: An overview of randomized controlled trials (RCTs) and related issues.* Washington, DC: Congressional Research Service, Library of Congress. Retrieved May 18, 2007, from http://assets.opencrs.com/rpts/RL33301_20060307.pdf

Brewin, C. R., & Bradley, C. (1989). Patient preferences and randomised clinical trials. *British Medical Journal, 299,* 313–315.

Button, K. S., Ioannidis, J. P., Mokrysz, C., Nosek, B. A., Flint, J., Robinson, E. S., & Munafo, M. R. (2013). Power failure: Why small sample size undermines the reliability of neuroscience. *National Review of Neuroscience, 14*(5), 365–376. DOI: 10.1038/nrn3475.

Campbell, M. K., Elbourne, D. R., & Altman, D. G. (2004). CONSORT statement: Extension to cluster randomised trials. *British Medical Journal, 328,* 702–708.

Chen, H. T., & Rossi, P. H. (1983). Evaluating with sense: The theory-driven approach. *Evaluation Review, 7,* 283–302.

Chilvers, C., Dewey, M., Fielding, K., Gretton, V., Miller, P., Palmer, B., . . . Harrison, G. (2001). Antidepressant drugs and generic counselling for treatment of major depression in primary care: Randomised trial with patient preference arms. *British Medical Journal, 322,* 772–775.

Concat, J., Shah, N., & Horwitz, R. I. (2000). Randomized controlled trials, observational studies and the hierarchy of research design. *New England Journal of Medicine, 342*, 1887–1892.

Cook, T. D., & Campbell, D. T. (1979). *Quasi-experimentation: Design and analysis issues for field settings.* Boston, MA: Houghton Mifflin.

Cook, T. D., & Payne, M. R. (2002). Objecting to the objections to using random assignment in educational studies. In F. Mosteller & R. Boruch (Eds.), *Evidence matters: Randomized trials in education research* (pp. 150–178). Washington, DC: Brookings Institution.

Cook, T. D., & Wong, V. C. (2008). Better quasi-experimental practice. In P. Alasuutari, J. Brannen, & L. Bickman (Eds.), *Handbook of social research methods.* London, England: SAGE.

Corrigan, P., & Salzer, M. (2003). The conflict between random assignment and treatment preference: Implications for internal validity. *Evaluation and Program Planning, 26,* 109–121.

Cronbach, L. J. (1982). *Designing evaluation and social programs.* San Francisco, CA: Jossey-Bass.

Dehmer, G. J., Weaver, D., Roe, M. T., Milford-Beland, S., Fitzgerald, S., Hermann, A., . . . Brindis, R. G. (2012). A contemporary view of diagnostic cardiac catheterization and percutaneous coronary intervention in the United States: A report from the CathPCI Registry of the National Cardiovascular Data Registry,

2010 through June 2011. *Journal of the American College of Cardiology, 60*(20), 2017–2031. DOI: 10.1016/j.jacc.2012.08.966

Dobie, W., & Fryer, R. G. (2011). Are high-quality schools enough to increase achievement among the poor? Evidence from the Harlem Children's Zone. *American Economic Journal: Applied Economics, 30,* 158–187.

Donaldson, S. I. (2003). Theory-driven program evaluation. In S. I. Donaldson & M. Scriven (Eds.), *Evaluating social programs and problems: Visions for the new millennium.* (pp. 109–141). Mahwah, NJ: Erlbaum.

Eccles, M., Steen, N., Grimshaw, J., Thomas, L., McNamee, P., Soutter, J., . . . Bond, S. (2001). Effect of audit and feedback, and reminder messages on primary-care radiology referrals: A randomised trial. *Lancet, 357*(9266), 1406–1409.

Faddis, B., Ahrens-Gray, P., & Klein, E. (2000). *Evaluation of Head Start family child care demonstration* (Final Report). Washington DC: Commissioner's Office of Research and Evaluation.

Fisher, C. B., & Anushko, A. E. (2008). Research ethics in social science. In P. Alasuutari, J. Brannen, & L. Bickman (Eds.), *Handbook of social research methods* (pp. 95–110). London, England: SAGE.

Flay, B. R., Biglan, A., Boruch, R. F., Castro, F. G., Gottfredson, D., Kellam, S., . . . Ji, P. (2005). Standards of evidence: Criteria for efficacy, effectiveness, and dissemination. *Prevention Science, 6*(3), 151–175.

Foster, E. M., & Bickman, L. (2000). Refining the costs analyses of the Fort Bragg evaluation: The impact of cost offset and cost shifting. *Mental Health Services Research, 2*(1), 13–25.

Gibson, C., & Duncan, G. (2005). Qualitative/quantitative synergies in a random-assignment program evaluation. In T. Weisner (Ed.), *Mixed methods in the study of child and family life* (pp. 283–303). Chicago, IL: University of Chicago Press.

Glasgow, R. E., Green, L. W., Klesges, L. M., Abrams, D. B., Fisher, E. B., Goldstein, M. G., . . . Orleans, T. (2006). External validity: We need to do more. *Annals of Behavioral Medicine, 31*(2), 105–108.

Glazerman, S., Levy, D. M., & Myers, D. (2003). Nonexperimental versus experimental estimates of earnings impacts. *The Annals of the American Academy of Political and Social Science, 589,* 63–93.

Graham, J., & Donaldson, S. (1993). Evaluating interventions with differential attrition: The importance of nonresponse mechanisms and use of follow-up data. *Journal of Applied Psychology, 78*(1), 119–128.

Greenhouse, S. W. (2003). The growth and future of biostatistics: A view from the 1980s. *Statistics in Medicine, 22,* 3323–3335.

Hak, E., Wei, F., Grobbee, D. E., & Nichol, K. L. (2004). A nested case-control study of influenza vaccination was a cost-effective alternative to a full cohort analysis. *Journal of Clinical Epidemiology, 57,* 875–880.

Heckman, J. J., & Smith, J. A. (1995). Assessing the case for social experiments. *Journal of Economic Perspective, 9*(2), 85–110.

Hernan, M., Alonso, A., Logan, R. Grodstein, F., Michels, K., Stamfer, M., . . . Robins, J. M. (2008). Observational studies analyzed like randomized experiments: An application

to postmenopausal hormone therapy and coronary heart disease. *Epidemiology*, *19*(6), 766–779.

Hill, J. (2008). Comment. *Journal of the American Statistical Association, 103*(484), 1346–1350.

Howard, L., & Thornicroft, G. (2006). Patient preference randomised controlled trials in mental health research. *British Journal of Psychiatry, 188,* 303–304.

Howe, K. (2004). A critique of experimentalism. *Qualitative Inquiry, 10*(1), 42–61.

Huwiler-Müntener, K., Juni, P., Junker, C., & Egger, M. (2002). Quality of reporting of randomized trials as a measure of methodologic quality. *JAMA, 287,* 2801–2804.

Ioannidis, J. (2006). Evolution and translation of research findings: From bench to where? *PLoS Clinical Trials, 1*(7), e36.

Ioannidis, J. (2013). Mega-trials for blockbusters. *JAMA, 309*(3), 239–240.

Johnson, S. L. (2005). *Children's environmental exposure research study.* Washington, DC: U.S. Environmental Protection Agency. Retrieved May 30, 2007, from http://www.epa.gov/cheers.

Joynt, K. E., Orav, E. J., & Jha, A. K. (2013). Mortality rates for Medicare beneficiaries admitted to critical access and non-critical access hospitals, 2002–2010. *JAMA, 309*(13), 1379–1387. DOI:10.1001/jama.2013.2366

Kane, R., Wang, J., & Garrard, J. (2007). Reporting in randomized clinical trials improved after adoption of the CONSORT statement. *Journal of Clinical Epidemiology, 60*(3), 241–249.

Keppel, G. (1991). *Design and analysis: A researcher's handbook.* Upper Saddle River, NJ: Prentice Hall.

King, M., Nazareth, I., Lampe, F., Bower, P., Chandler, M., Morou, M., . . . Lai, R. (2005). Impact of participant and physician intervention preferences on randomized trials: A systematic review. *Journal of the American Medical Association, 293*(9), 1089–1099.

Kumar, S., & Oakley-Browne, M. (2001). Problems with ensuring a double blind. *Journal of Clinical Psychiatry, 62*(4), 295–296.

Leaf, C. (2013, June 13). Do clinical trials work? *New York Times,* p. SR1.

Leon, A. C., Mallinckrodt, C. H., Chuang-Stein, C., Archibald, D. G., Archer, G. E., & Chartier, K. (2006). Attrition in randomized controlled clinical trials: Methodological issues in psychopharmacology. *Biological Psychiatry, 59*(11), 1001–1005.

Lexchin, J., Bero, L., Djulbegovic, B., & Clark, O. (2003). Pharmaceutical industry sponsorship and research outcome and quality: Systematic review. *British Medical Journal, 326,* 1167–1170.

Lipsey, M., & Cordray, D. (2000). Evaluation methods for social intervention. *Annual Review of Psychology, 51,* 345–375.

Little, R. J., Long, Q., & Lin, X. (2008). Comment. *Journal of the American Statistical Association, 103*(484), 1344–1346.

Lu, Y., Yao, Q., Gu,J., & Shen, C. (2013). Methodological reporting of randomized clinical trials in respiratory research in 2010. *Respiratory Care*, published ahead of print, January 9, 2013. DOI:10.4187/respcare.01877

Luborsky, L., Diguer, L., Seligman, D. A., Rosenthal, R., Krause, E. D., Johnson, S., . . . Schweizer, E. (1999). The researcher's own therapy allegiances:

A "wild card" in comparison treatment efficacy. *Clinical Psychology: Science and Practice, 6,* 95–106.

Macias, C., Barreira, P., Hargreaves, W., Bickman, L., Fisher, W., & Aronson, E. (2005). Impact of referral source and study applicants' preference for randomly assigned service on research enrollment, service engagement, and evaluative outcomes. *American Journal of Psychiatry, 162*(4), 781–787.

Maxwell, J. (2004). Causal explanation, qualitative research, and scientific inquiry in education. *Educational Researcher, 33*(2), 3–11.

McCall, R., & Green, B. (2004). Beyond methodological gold standards of behavioral research: Considerations for practice and policy. *Social Policy Report, 18*(2), 3–12.

McCall, R., Ryan, C., & Plemons, B. (2003). Some lessons learned on evaluating community-based, two-generation service programs: The case of the Comprehensive Child Development Program. *Journal of Applied Developmental Psychology, 24*(2), 125–141.

Mertens, D., & Hesse-Biber, S. (2013). Mixed methods and credibility of evidence in evaluation. *New Directions for Evaluation, 138*, 5–13. DOI: 10.1002/ev.20053

Moher, D., Jones, A., & Lepage, L. (2001). Use of the CONSORT statement and quality of reports of randomized trials: A comparative before-and-after evaluation. *JAMA, 285,* 1992–1995.

Moher, D., Schulz, K. F., & Altman, D. G. (2001). The CONSORT statement: Revised recommendations for improving the quality of reports of parallel-group randomized trials. The CONSORT Group. *JAMA, 285,* 1987–1991.

Morris, P., & Gennetian, L. (2003). Identifying the effects of income on children's development using experimental data. *Journal of Marriage and Family, 65*(3), 716–729.

National Cancer Institute. (2008). *Childhood Cancers*. Retrieved on March 30, 2014, from http://www.cancer.gov/cancertopics/factsheet/Sites-Types/childhood

Nich, C., & Carroll, K. M. (2002). Intention to treat meets missing data: Implications of alternate strategies for analyzing clinical trials data. *Drug and Alcohol Dependence, 68*(2), 121–130.

Onwuegbuzie, A., & Leech, N. (2005). On becoming a pragmatic researcher: The importance of combining quantitative and qualitative research methodologies. *International Journal of Social Research Methodology, 8*(5), 375–387.

Petrosino, A., Boruch, R. F., Rounding, C., McDonald, S., & Chalmers, I. (2000). The Campbell Collaboration social, psychological, educational and criminological trials register (C2-SPECTR) to facilitate the preparation and maintenance of systematic reviews of social and educational interventions. *Evaluation and Research in Education, 14*(3), 206–219.

Plint, A. C., Moher, D., Morrison, A., Schulz, K., Altman, D. G., Hill, C., & Gaboury, I. (2006). Does the CONSORT checklist improve the quality of reports of randomised controlled trials? A systematic review. *Medical Journal of Australia, 185*(5), 263–267.

Post, P. N., de Beer, H., & Guyatt, G. H. (2013). How to generalize efficacy results of randomized trials: Recommendations based on a systematic review of possible approaches. *Journal of Evaluation in Clinical Practice, 19*, 638–643. DOI:10.1111/j.1365–2753.2012.01888.x

Raudenbush, S. W. (2002, February 6). *Identifying scientifically-based research in education.* Paper presented at the Working Group Conference in Washington, DC. Retrieved May 30, 2007, from http://www.ssicentral.com/hlm/techdocs/ScientificallyBasedResearchSeminar.pdf

Raudenbush, S. W. (2008). Many small groups. In J. de Leeuw & E. Meijer (Eds.), *Handbook of multilevel analysis* (pp. 207–236). New York, NY: Springer.

Rolfe, G. (2006). Validity, trustworthiness, and rigour: Quality and the idea of qualitative research. *Journal of Advanced Nursing, 53*(3), 304–310.

Rosenbaum, P. R. (2002). *Observational studies* (2nd ed.). New York, NY: Springer-Verlag.

Rosenbaum, P. R., & Rubin, D. B. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika, 70*(1), 41–55.

Rosenthal, R. (1976). *Experimenter effects in behavioral research* (enlarged ed.). New York, NY: Irvington.

Rosnow, R. L. (2002). The nature and role of demand characteristics in scientific inquiry. *Prevention & Treatment, 5,* 37.

Rossi, P. (1987). The iron law of evaluation and other metallic rules. *Research in Social Problems and Public Policy, 4,* 3–20.

Rothwell, P. M. (2005). External validity of randomised controlled trials: 'To whom do the results of this trial apply?" *Lancet, 365*(9453), 82–93.

Rubin, D. B. (2008). Comment: The design and analysis of gold standard randomized experiments. *Journal of the American Statistical Association, 103*(484), 1350–1353.

Sanson-Fisher, R. W., Bonevski, B., Green, L. W., & D'Este, C. (2007). Limitations of the randomized controlled trial in evaluating population-based health interventions. *American Journal of Preventative Medicine, 33*(2), 155–161. DOI: 10.1016/j.amepre.2007.04.007

Scriven, M. (2005, December). *Can we infer causation from cross-sectional data?* Paper presented at the School-Level Data Symposium, Washington, DC. Retrieved May 28, 2007, from http://www7.nationalacademies.org/bota/School-Level%20Data_Michael%20Scriven-Paper.pdf.

Shadish, W. R., Clark, M. H., & Steiner, P. M. (2008). Can nonrandomized experiments yield accurate answers? A randomized experiment comparing random to nonrandom assignment. *Journal of the American Statistical Association, 103*(484), 1334–1356.

Shadish, W. R., Cook, T., & Campbell, D. (2002). *Experimental and quasi-experimental designs for generalized causal inference.* Boston, MA: Houghton Mifflin.

Shadish, W. R., Hu, X., Glaser, R. R., Kownacki, R. J., & Wong, T. (1998). A method for exploring the effects of attrition in randomized experiments with dichotomous outcomes. *Psychological Methods, 3,* 3–22.

Shadish, W. R., Luellen, J. K., & Clark, M. H. (2006). Propensity scores and quasi-experiments: A testimony to the practical side of Lee Sechrest. In R. R. Bootzin & P. E. McKnight (Eds.), *Strengthening research methodology: Psychological measurement and evaluation* (pp. 143–157). Washington, DC: American Psychological Association.

Shadish, W. R., & Ragsdale, K. (1996). Random versus nonrandom assignment in controlled experiments: Do you get the same answer? *Journal of Consulting and Clinical Psychology, 64,* 1290–1305.

Shavelson, R. J., & Towne, L. (Eds.). (2002). *Scientific research in education* (National Research Council. Committee on Scientific Principles for Educational Research). Washington, DC: National Academy Press.

Shumaker, S. A., Legault, C., Rapp, S. R., Thal, L., Wallace, R. B., Ockene, J. K., . . . Wactawski-Wende, J. (2003). Estrogen plus progestin and the incidence of dementia and mild cognitive impairment in postmenopausal women: The Women's Health Initiative memory study: A randomized controlled trial. *Journal of the American Medical Association, 289,* 2651–2662.

Sieber, J. E. (2009). Planning ethically responsible research. In L. Bickman & D. Rog (Eds.), *Handbook of applied social research methods* (2nd ed., pp. 106–141). Thousand Oaks, CA: SAGE.

Smith, G., & Pell, J. (2003). Parachute use to prevent death and major trauma related to gravitational challenge: Systematic review of randomised controlled trials. *British Medical Journal, 327,* 1459–1461.

Soares, H. P., Daniels, S., Kumar, A., Clarke, M., Scott, C., Swann, S., & Djulbegovi, B. (2004). Bad reporting does not mean bad methods for randomised trials: Observational study of randomised controlled trials performed by the Radiation Therapy Oncology Group. *British Medical Journal, 328,* 22–24.

Spring, B., Pagoto, S., Knatterud, G., Kozak, A., & Hedeker, D. (2007). Examination of the analytic quality of behavioral health randomized clinical trials. *Journal of Clinical Psychology, 63*(1), 53–71.

Steele, J. R., Wellemeyer, A., Hansen, M., Reaman, G. H., & Ross, J. A. (2006). Childhood cancer research network: A North American childhood cancer research network. *Cancer Epidemiology Biomarkers & Prevention, 15*, 1241–1242.

Steiner, P. M., & Cook, D. L. (2013). Matching and propensity scores. In T. D. Little (Ed.), *The Oxford handbook of quantitative methods* (vol. 1, pp. 237–259). New York, NY: Oxford University Press.

Steiner, P. M., Cook, T. D., Shadish, W. R., & Clark, M. H. (2010). The importance of covariate selection in controlling for selection bias in observational studies. *Psychological Methods, 15,* 250–267.

Van de Ven, P., & Aggleton, P. (1999). What constitutes evidence in HIV/AIDS education? *Health Education Research, 14*(4), 461–471.

Van Spall, H., Toren, A., Kiss, A., & Fowler, R. (2007). Eligibility criteria of randomized controlled trials published in high-impact general medical journals: A systematic sampling review. *JAMA, 297*(11), 1233–1240. DOI:10.1001/jama.297.11.1233

VanderWeele, T. (2006). The use of propensity score methods in psychiatric research. *International Journal of Methods in Psychiatric Research, 15*(2), 95–103.

Varnell, S. P., Murray, D. M., Janega, J. B., & Blitstein, J. L. (2004). Design and analysis of group-randomized trials: A review of recent practices. *American Journal of Public Health, 94*(3), 393–399.

Weijer, C. Grimshaw, J. M., Taljaard, M., Binik, A., Boruch, R., Brehaut, J. C., . . . Zwarenstein, M. (2011). Ethical issues posed by cluster randomized trials in health research. *Trials, 12*, 100–111. DOI:10.1186/1745–6215–12–100

West, S. G., & Thoemmes, F. (2008). Equating groups. In P. Alasuutari, J. Brannen, & L. Bickman (Eds.), *Handbook of social research methods* (pp. 414–430). London, England: SAGE.