Evaluating Methods for Estimating Program Effects
Charles S. Reichardt
American Journal of Evaluation 2011 32: 246 originally published online 22 February 2011
DOI: 10.1177/1098214011398954

The online version of this article can be found at:
http://aje.sagepub.com/content/32/2/246
Evaluating Methods for Estimating Program Effects

Charles S. Reichardt

Abstract
I define a treatment effect in terms of a comparison of outcomes and provide a typology of all possible comparisons that can be used to estimate treatment effects, including comparisons that are relatively unknown in both the literature and practice. I then assess the relative merit, worth, and value of all possible comparisons based on the criteria of bias, precision, generalizability, ease of implementation, and cost. Which comparison is best depends on how these criteria are weighted and on the constraints of the specific research setting. I hope readers come to recognize a wider range of comparisons than they had previously, appreciate the value of considering all possible comparisons, and see how my typology of comparisons provides the basis for making fair appraisals of the relative strengths and weaknesses of different types of comparisons in the presence of the contingencies that are most likely to arise in practice.

Keywords
program effects, randomized experiments, quasi-experiments, bias

What are the best methods for estimating the effects of a social program? That question lies at the heart of some of the longest running and most heated debates in the field of evaluation, as any seasoned reader of the American Journal of Evaluation (AJE) can attest. Twenty years ago, for example, the pages of AJE prominently featured a lively dispute over qualitative versus quantitative methods that was sparked by Lee Sechrest’s (1991) presidential address to the American Evaluation Association (Fetterman, 1992; Lincoln & Guba, 1992, 1994; Sechrest, 1992; Sechrest, Babcock, & Smith, 1993). And the volume of AJE appearing in 2010 contained commentary on much the same qualitative versus quantitative debate as it was reinvigorated by the Department of Education’s policy (drafted in 2003 and finalized in 2005) giving priority in funding to randomized experiments over other methods (Cook, Scriven, Coryn, & Evergreen, 2010; Gargani, 2010). That much the same dispute can span decades may suggest to some evaluators that progress has not been made in our shared understanding of evaluation methods. But knowledge about evaluation methods does accumulate, even if only slowly and tentatively. Indeed, knowledge has accumulated sufficiently over the past two decades that it is useful to try to summarize what evaluators have learned about
the relative merit, worth, and value of methods for estimating program effects. The present article is my attempt to provide such a summary.

Debates about the relative merit, worth, and value of different methods for estimating effects sometimes founder because commentators are unclear about the definition of an effect. To avoid this potential source of confusion, I provide an explicit (and widely accepted) definition of an effect and defend it against its critics. I then assess the merit, worth, and value of different methods for estimating effects based on that definition. I would not complain if critics reject my definition of an effect (and thereby reach different conclusions about the relative merit, worth, and value of different methods for estimating effects) as long as those critics provide an explicit and defensible alternative definition of an effect. I am more than willing to consider alternative definitions. But I am hesitant to evaluate methods for estimating effects without being clear about the criteria for evaluation which includes being clear about how an effect is defined. I hope to convince readers of the importance of grounding the evaluation of methods for estimating effects on an explicit definition of an effect.

I define an effect in terms of a comparison and I provide a typology of comparisons. The typology covers all possible comparisons that can be used to estimate effects including comparisons that are not widely recognized either in the literature or in practice. Although I do not rely heavily on the distinction between qualitative and quantitative methods, each comparison in the typology can be used by either a qualitative or quantitative researcher. And whether a researcher approaches an evaluation from either a qualitative or quantitative perspective, whenever a comparison is used to estimate an effect, that comparison is located somewhere in the typology. I hope readers will come to recognize a wider range of comparisons than they had previously, appreciate the value of considering all the different types of comparisons in the typology, and see how the typology provides a unified way of addressing intrinsic sources of bias in effect estimates.

Comparisons for estimating the effect of a program can be assessed on a variety of criteria including bias, precision, generalizability, ease of implementation, and cost. Which comparison is best depends on how the various criteria are weighted under a particular set of circumstances. I describe the strengths and weaknesses of the comparisons in the typology on a range of criteria recognizing that an advantage on one criterion may be offset by a disadvantage on another and that a potential strength on one criterion can turn into weakness, or vice versa, depending on the specific circumstances. In some debates about methods, proponents tend to overvalue the attributes of favored comparisons while undervaluing the attributes of disfavored comparisons because generalizations about strengths and weaknesses are drawn without adequate regard for the contingencies that often arise in practice. In contrast, I try to emphasize rather than ignore contingencies. For example, because of the substantial misunderstandings about benefits and drawbacks of random assignment (RA) that exist in the literature, I take pains to explain exactly what randomization does and does not accomplish both under ideal conditions and in the presence of the likely contingencies of attrition and noncompliance to treatment assignment. Although addressing contingencies increases the complexity of the discussion, the generalizations I draw are likely to be more defensible than otherwise. Plenty of room remains for healthy and vigorous debate about methods based on differences of opinion about the relative weights to place on criteria such as bias, precision, generalizability, ease of implementation, and cost. But healthy debate is not furthered by misrepresentations due to faulty generalizations. I hope readers will learn how to make fair appraisals of the relative strengths and weaknesses of different types of comparisons in the presence of the contingencies that are most likely to arise in practice.

Although I cite novel designs and new statistical methods, my focus is not on reviewing new methods as much as on understanding old ones. This is because the greatest methodological advances in estimating program effects in the past 25 years have come not from the introduction of new methods as much as from a deeper understanding and appreciation of the methods that have long been available.
The Causal Question Being Asked
A variety of questions can be asked about cause and effect. Two particularly important questions are:

Question #1: What is the effect of a given cause?
Question #2: What is the cause of a given effect?

Question #1 is asked when evaluators estimate the effects of a social program. Question #2 is asked when police detectives investigate the causes of a person’s death, pathologists investigate the causes of an illness, epidemiologists investigate the causes of an epidemic, sociologists investigate the causes of poverty, or evaluators investigate why a program failed to be implemented as planned. Both questions are important in policy making, but it is important to distinguish between the two because the best methods for answering one question may differ from the best methods for answering the other, or for answering other questions about causality. Therefore, when engaged in methodological debates about causality, it is important to be clear which causal question is being asked, otherwise disputants risk talking past each other.

In the present article, I am concerned with answering only Question #1. Other causal and noncausal questions besides Question #1 are certainly important in evaluation. But there is enough meat on the bone of Question #1 to occupy my attention here. And Question #1 rather than Question #2 is usually the question most directly asked when an evaluator estimates the effects of a program. But in spite of the critical differences between these two questions, evaluators sometimes confuse them, as I will show shortly.

Defining an Effect
Question #1 is a question about effects. In determining the best methods for answering Question #1, it is useful (if not necessary) to define what is meant by an effect. I define an effect in the following fashion. The effect of treatment X compared to treatment Y is the difference between (a) what would have happened at time 2 if treatment X had been implemented at time 1, and (b) what would have happened at time 2 if treatment Y had been implemented at time 1 instead of treatment X, but (c) everything else at time 1 had been the same (Reichardt, 2006). In this definition, the cause of the effect is the difference between treatment X and treatment Y.

A special case of the preceding definition of an effect arises when (a) treatment Y is equivalent to what would have happened if treatment X had not been implemented and (b) there is no ambiguity about what would have happened if treatment X had not been implemented. Under these two conditions, the definition of an effect simplifies to the following. The effect of treatment X is the difference between (a) what would have happened at time 2 if treatment X had been implemented at time 1, and (b) what would have happened at time 2 if treatment X had not been implemented at time 1 but (c) everything else at time 1 had been the same. In this special case of the definition, the cause of the effect is treatment X. Of course, using the special case of the definition risks ambiguity because there may be more than one way for treatment X not to have been implemented. For example, not implementing a novel treatment might mean, in different circumstances, that either no treatment, a standard treatment, or a placebo treatment would have been implemented instead. And even for a treatment as simple as taking a couple of aspirin versus not taking a couple of aspirin, it may matter what was done while not taking the aspirin (e.g., if a person drank a glass of water when taking the aspirin it may matter if the person would have still drank a glass of water if he or she had not taken the aspirin). In general, it is best to be explicit about what is entailed in treatment Y and to use the general form of the definition rather than the special case. Another advantage of the general form of the definition is that it allows an effect to be defined in instances when treatment Y is not the absence of treatment X. For example, the general form of the definition defines an effect when
treatments X and Y are two alternative, innovative treatments neither of which can be conceptualized as the absence of the other. And circumstances often arise where it is useful to assess the effect of the difference between two conceptually distinct treatments and not just the effect of a difference between a treatment and the absence of that treatment.

The definition of an effect given above relies on a comparison between what happens after X is implemented and what happens after Y is implemented. The comparison in the definition is called a counterfactual comparison (and the definition of an effect is therefore called a counterfactual definition) because the comparison is impossible (contrary to fact) to obtain in practice. The comparison is impossible to obtain in practice because it is impossible both to implement treatment X and to implement treatment Y instead of treatment X, with everything else remaining the same. Scriven (2008, 2009; Cook et al., 2010, p. 108) characterizes the “counterfactual approach” as not being the “correct analysis of causation” because of the inability of counterfactual definitions to cope with the problem of overdetermination. The problem of overdetermination arises when two sufficient causes are present and both produce the same effect. For example, consider a felon who is executed by a firing squad of five shooters who shoot simultaneously. In this case, it is impossible to say that any one of the shooters killed the felon because the felon would have died if (counterfactually) any one of the shooters had not fired. So according to a counterfactual definition, none of the shooters can be said to have killed the felon, even though it is true that the felon was killed by the shooters. Nor can all five of the shooters be said unambiguously to be the cause of death because the felon would have died even if only one of the shooters had fired rather than all five. Hence simple counterfactual definitions of causes lead to ambiguities in the face of overdetermination.

But note that the problem of overdetermination as described above arises when assessing the cause of the felon’s death which is answering Question #2 while the present discussion is interested only in Question #1. It is correct that overdetermination causes problems for counterfactual definitions, but only for counterfactual definitions of causes and not of effects. That is, overdetermination is a problem when asking what caused the felon’s death, which is Question #2. But overdetermination is not a problem when assessing the effects of different treatments, which is Question #1 and which is the only focus of the present article. Overdetermination is not a problem for Question #1 because questions about the effect of a given cause can be answered unambiguously as long as both treatment X and treatment Y are specified unambiguously. If treatment X consists of all five shooters firing their weapons while treatment Y consists of none of the five shooters firing, then the effect of all five shooters firing (as compared to having none of the five shooters firing) is the death of the felon. In contrast, if treatment X consists of all five shooters firing while treatment Y consists of only four of the shooters firing, which is a difference of one shooter firing in the setting where the other four shooters all fire, then the effect of that one shooter firing in the context of four other shooters firing is unambiguously not the death of the felon. Alternatively, the effect of one shooter firing in the context of no other shooters firing would be the death of the felon. In this way, the counterfactual definition of an effect makes it clear, as well it should, that context matters when assessing the effects of a given cause. The point is that overdetermination does not introduce ambiguity about the effects of a given cause as long as both treatment X and treatment Y are specified unambiguously, as I have advised be done.

So Scriven is wrong to imply that overdetermination is a problem for the proposed counterfactual definition of an effect. Scriven’s error arises because he confuses Question #2 with Question #1. Answering Question #2 requires a definition of the cause of a given effect while answering Question #1 requires a definition of the effect of a given cause. A substantial amount of philosophical discussion has been devoted to defining causality with some important contributions made by Scriven (1968, 1975). But when philosophers consider the analysis of causality, they almost exclusively focus on defining causes rather than effects. And providing a completely adequate definition of the cause of a given effect has proven to be extremely difficult (Brand, 1976, 1979). The philosophical
literature has long revealed that simplistic counterfactual definitions of the cause of a given effect are inadequate because of overdetermination, as illustrated by the firing squad example. But for the present article, I am interested in defining only the effect for a given cause and not the cause of a given effect. And, to reiterate, the counterfactual definition of the effect of a given cause (unlike simplistic counterfactual definitions of the cause of a given effect) is not invalidated by the presence of overdetermination. As long as you are careful to define both treatment X and treatment Y, overdetermination is not a problem for the counterfactual definition of an effect. Overdetermination is only a problem to the extent one leaves treatment Y ambiguous, as I have warned against doing (also see Gargani, 2010).

Except for very sporadic dissent (e.g., Dawid, 2000), the counterfactual definition of an effect is widely accepted among statisticians where it is an integral and explicit part of what has come to be called the Rubin model of causality (Holland, 1986; Rubin, 2005). Some discussions of methods for estimating effects in the social sciences explicitly base their presentations on a counterfactual definition (Reichardt, 2006; Reichardt & Mark, 1998; Shadish, Cook, & Campbell, 2002), but most do not provide an explicit definition of an effect. Nonetheless, the logic and methods of experimentation and quasi-experimentation (as presented, e.g., in Campbell & Stanley, 1966; Cook & Campbell, 1979; Shadish et al., 2002) are widely accepted across the social sciences and are completely compatible with a counterfactual definition of an effect.

I am unaware of any adequate alternatives to the counterfactual definition of an effect. Attempts at providing alternatives often founder because of circularity: they define the connection between cause and effect by evoking the concept of causality which is what they are supposed to be defining rather than assuming (cf. Mohr, 1999). Nonetheless a useful alternative definition might already exist or be crafted in the future which supports different methods for estimating effects than are supported by the counterfactual definition. If so, discussions of methods for estimating effects should make clear the definition of an effect being assumed so as to avoid misunderstandings. In what follows, I assume the counterfactual definition of an effect.

Types of Comparisons
As noted above, the counterfactual definition of an effect is based on a counterfactual comparison between what would have happened if treatment X had been implemented and what would have happened if treatment Y had been implemented instead of treatment X, but everything else had been the same. Also as noted before, this counterfactual comparison is impossible to obtain in practice because it is impossible both to implement treatment X and to implement treatment Y instead of treatment X, with everything else being the same. Any comparison of outcomes that can be implemented in practice will differ from the counterfactual comparison. As Reichardt (2006) notes, a treatment effect can be estimated using any of four types of practical comparisons in place of the counterfactual comparison. The four types of practical comparisons are distinguished by the nature of the units (either recipients, settings, times, or outcome variables) assigned to receive different treatment conditions and then compared. That is, a researcher can compare what happens after treatment X is implemented to what happens after treatment Y is implemented by assigning the two treatments to different recipients, different settings, different times, or different outcome variables. I describe each of these four types of practical comparisons below. I also distinguish among three ways of assigning units to treatment conditions: random assignment (RA), assignment based on a quantitative variable which I label quantitative assignment (QA), and assignment that is neither random nor based on a quantitative variable which I label nonquantitative assignment (non-QA).

Each of the three ways of assigning units to treatment conditions can be applied to each of the four types of units. Crossing the three ways of assigning units to treatments with the four types of units creates the three-by-four typology of comparisons which is presented graphically in Table 1.
Comparisons in which units are assigned at random are called randomized experiments. As Table 1 reveals, there are four types of randomized experiments. There are also four types of QA comparisons and four types of non-QA comparisons. Each of the 12 types of comparisons is described next.

### Comparisons Between Recipients

In a comparison between recipients, different recipients (e.g., people) receive different treatments. That is, one group of recipients is given treatment X while another group of recipients is given treatment Y. Assigning recipients to different treatments at random produces an RA comparison between recipients, which is one type of randomized experiment. A QA comparison between recipients arises when recipients are assigned to different treatments based on a cutoff score on a quantitative measurement of the recipients. For example, recipients might be assessed on a quantitative measure of need or merit and assigned to receive one of the treatments if their scores on need or merit exceed a specified cutoff value, and otherwise assigned to the other treatment. In Campbellian nomenclature, a QA comparison between recipients is called a regression-discontinuity design (Shadish et al., 2002). Assigning recipients to treatments neither randomly nor quantitatively produces a non-QA comparison between recipients. For example, a non-QA comparison is created when recipients either choose their own treatment condition or are assigned to treatment conditions by administrators based on qualitatively assessed criteria. In Campbellian nomenclature, a non-QA comparison between recipients is called a nonequivalent comparison group design.

### Comparisons Between Settings

In a comparison between settings, different settings receive different treatments. For example, the effect of adding a traffic light to intersections could be assessed by adding traffic lights to some intersections and not adding traffic lights to other intersections, and comparing subsequent accidents at both sets of intersections. Assigning intersections to different treatments at random would produce an RA comparison between settings, which is one type of randomized experiment. Assigning intersections to treatment conditions based on a cutoff score on a quantitative assessment of the intersections would produce a QA comparison between settings. For example, a set of intersections could be selected, the number of traffic accidents at each intersection during the preceding year could be

---

**Table 1. A Typology of Comparisons**

<table>
<thead>
<tr>
<th>Assignment to Treatments</th>
<th>Nonrandom Assignment</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Units of Assignment</strong></td>
<td><strong>Random Assignment</strong> (RA)</td>
</tr>
<tr>
<td>Recipients</td>
<td>RA comparison between recipients</td>
</tr>
<tr>
<td>Settings</td>
<td>RA comparison between settings</td>
</tr>
<tr>
<td>Times</td>
<td>RA comparison between times</td>
</tr>
<tr>
<td>Outcome variables</td>
<td>RA comparison between outcome variables</td>
</tr>
</tbody>
</table>
assessed, and those intersections with more accidents than a specified cutoff value could be assigned
to receive a traffic light. Assigning settings to treatments neither at random nor quantitatively
produces a non-QA comparison between settings. For example, political considerations, which were
not measured quantitatively, might be used to determine which intersections receive traffic lights.³

Comparisons Between Times

In a comparison between times, different times receive different treatments. For example, the effect
of a medication for attention deficit disorder could be assessed for a given individual by having the
individual take the medication on some days and not on other days. Differences in attentiveness
across the two sets of days could then be used to estimate the effect of the treatment. Assigning times
(e.g., days) to different treatment conditions at random would produce an RA comparison between
times, which is one type of randomized experiment. Assigning times to treatment conditions based
on a cutoff score on a quantitative assessment of times would produce a QA comparison between
times. In Campbellian nomenclature, if chronological time is the variable used for the quantitative
assessment of times, the QA comparison that results is called an interrupted time-series (ITS) design
(Marcantonio & Cook, 1994). Assigning treatment conditions to times neither at random nor
quantitatively produces a non-QA comparison between times. An example is a simple before–after
comparison where outcomes immediately before an intervention is introduced are compared to
outcomes immediately after the intervention is introduced.

Comparisons Between Outcome Variables

In a comparison between outcome variables, different outcome variables receive different treat-
ments. For example, imagine an educational television program intended to teach children to pro-
nounce the letters of the alphabet. Half the letters are taught, one per week, during the first
13 weeks the show is broadcast. At the end of the 13th week, children who watched the show are
tested on their ability to pronounce all the letters in the alphabet, where the pronunciation of each
letter is a different outcome variable. To estimate the effects of the program, the scores on one set
of outcomes variables (e.g., the pronunciations of the letters presented on the show) are compared to
the scores on the other set of outcome variables (e.g., the pronunciations of the letters not presented
on the show). Assigning outcome variables (e.g., letters of the alphabet) to treatment conditions
(e.g., to be either taught or not taught on the television show) at random produces an RA comparison
between outcome variables, which is one type of randomized experiment. In the context of labora-
tory experimentation, such comparisons are called within-subject designs. Assigning outcome vari-
ables to treatment conditions based on a cutoff score on a quantitative assessment of the outcome
variables produces a QA comparison between outcome variables. For example, the letters of the
alphabet could be ordered based on their frequency of appearance in the English language and the
most frequent letters could be taught on the television show during its first season. Assigning out-
come variables to treatment conditions neither at random nor quantitatively produces a non-QA
comparison between outcomes variables. For example, the letters of the alphabet presented on the
TV program might have been chosen because they were the first letters in the names of common
animals and the producers of the television show wanted to use animals in the program materials.

Combinations of Comparisons

In addition to comparisons between either recipients, settings, times, or outcome variables, combi-
nations of comparisons are also possible. For example, it is possible to compare what happens after
treatment X is implemented to what happens after treatment Y is implemented using a comparison
between different recipients and between different times. An ITS design that includes a control
series of observations from a group of recipients who do not receive the treatment is an example
of such a combination of comparisons (also see Coryn, Schroter, & Hanssen, 2009). Scriven (2003) provides another example of a combination design:

“If, on twelve occasions, we switch a single class from lectures to the new model of highly interactive learning, running pre and post tests on [all] such occasions, and we find that the learning gains in the interactive mode are always very close to triple those in the lecture mode, albeit on different content, and such changes do not occur spontaneously, we have proven beyond reasonable doubt, for this class at least, that the interactive approach represents an improvement of great educational importance, despite the absence of a matched control group.”

If the times at which the teacher introduces either lectures or interactive learning are assigned at random and if the different contents for the lessons are assigned to these two treatment conditions at random, the estimate of the treatment effect would be based on both an RA comparison between times and an RA comparison between outcome variables. Without RA in either way, the estimate of the treatment effect would be based on both a non-QA comparison between times and a non-QA comparison between outcome variables.

Alternatively, a researcher could assign 20 classrooms to each receive 10 lessons taught by lecture and 10 lessons taught by interactive learning with both the lesson materials that are assigned to the two treatment conditions and the order in which the lesson materials are received varied randomly. In such a design, the estimate of the treatment effect would again be based on an RA comparison between both outcome variables and times, with each classroom of students serving as its own control.

Eckert (2000) provides an additional example of a combination comparison. He assessed the effects of a learning module by comparing performance before the module was taught to performance after the module was taught using different test questions on the pretreatment and post-treatment assessments, where the different test questions were assigned to the two assessments at random. The estimate of the treatment effect in this design was based on an RA comparison between outcome variables and a non-QA comparison between times. Also see Jagannathan, Camasso, and Killingsworth (2004) for an example of an evaluation that used estimates of program effects derived from multiple comparisons.

For purposes of classification, a comparison which is a combination of comparisons spans multiple cells in Table 1. As a result, a comparison which is a combination of comparisons both benefits and suffers from the properties of the comparisons in all the applicable cells in Table 1. Nonetheless, whether a comparison is a simple comparison or a combination of comparisons, it falls under the rubric of Table 1. That is, any comparison used to estimate a program effect entails a comparison of recipients, settings, times and/or outcome variables, and can be classified as involving either RA, QA, or non-QA comparisons.

Not all of the 12 simple comparisons in Table 1 are widely recognized. In particular, not all four types of randomized experiments are well recognized. However, to have a well-stocked arsenal of designs at one’s intellectual disposal and to fully understand the strengths and weaknesses of different types of comparisons, an evaluator needs to appreciate the design possibilities represented by all the 12 cells in Table 1. In addition, anyone leveling criticisms against a type of comparison (such as randomized experiments) should make sure their criticisms apply to all forms in which that comparison can be implemented, or else restrict the domain of their criticism appropriately.

Is a Comparison Necessary?
If an evaluator estimates a treatment effect by comparing what happened after treatment X was implemented to what happened after treatment Y was implemented, the evaluator must use one or more of the comparisons described above. But an evaluator is not logically required to estimate
a treatment effect by comparing what happened after treatment X was implemented to what happened after treatment Y was implemented. Indeed, various methods such as theory-based approaches in which outputs are estimated based only on theory and the strength of inputs (Cook, 2002) and Scriven’s (1976; Mohr, 1999) modus operandi method have been proposed that do not require comparisons.

Evaluators who do not accept the counterfactual definition of an effect may be particularly drawn to approaches that do not rely on comparisons. For example, Scriven (1976, 2005, 2008, 2009; also see Mohr, 1999) explicitly rejects the counterfactual definition of an effect (although he does not provide an alternative definition) and argues that effects can be assessed by “direct observation” by which he seems to mean without use of a comparison. But at least some of the examples Scriven uses to illustrate the direct observation of effects, in fact, rely on comparisons. For example, Scriven (2005, p. 8) suggests a driver can directly observe “that pressing the brake pedal causes your car to slow.” But observing your car slow down relies on a comparison between how fast your car was traveling before you pressed the brake and how fast it was traveling after you pressed the brake, which is a comparison between times. In fact, most observations of this type are based on a sophisticated, if unconscious, ITS comparison and not just a simple before–after comparison.

So although it is possible to go about estimating program effects without either an explicit or implicit comparison, evaluators nonetheless very often base their estimates of effects on comparisons even when the estimates are derived from case studies in which the comparison being drawn is not immediately obvious. For example, Scriven (2008, p. 22) provides the following description of a case study of the evaluation of aid given to farmers in East Africa: “...after determining that a substantial improvement in welfare has followed the arrival of aid, and has been sustained for a few years, we check for the presence of more than a dozen other possible causes of this observed subsequent increase in welfare...” which shows how the potential presence of a treatment effect was determined using a comparison drawn over time between what happened before aid arrived and what happened after aid arrived. And Scriven (2008) approvingly cites Patton’s (2008) case study of the effect of a stealth campaign to influence a Supreme Court decision where, again, an effect was estimated using a comparison drawn over time between what happened before the campaign was implemented and what happened afterward. In addition, even though it is possible to estimate effects without a comparison, decision makers quite often request that estimates be derived from a comparison of empirically observed outcomes.

The bottom line is that methods for estimating effects that rely on empirical comparisons of outcomes have held, and likely will continue to hold, a place of prominence in evaluation. I suspect the emphasis on empirical comparisons of outcomes arises, at least in part, because of the acceptance, either implicitly or explicitly, of a counterfactual definition of an effect. In any case, methods of estimating effects that do not rely on a comparison will not be the focus of attention in my presentation. The modus operandi method (and other noncomparative methods) can be highly useful add-ons to comparison methods. But my presentation does not consider noncomparative methods by themselves. My purpose is to assess only empirical comparisons of outcomes for estimating program effects (either with or without noncomparative add-ons). A relative assessment of methods for deriving estimates of program effects using empirical comparisons of outcomes versus methods for deriving estimates of program effects without using empirical comparisons of outcomes will have to wait for another day. For those undertaking such an assessment, I suspect conclusions would depend greatly on how an effect is defined, so providing an explicit definition would be useful.

**Evaluating Comparisons**

When comparisons of outcomes are used to estimate treatment effects, alternative explanations for differences in outcomes must be taken into account lest the estimate be incorrect. That estimating
treatment effects requires ruling out alternative explanations was first immortalized in Campbell and Stanley (1966) and later updated in both Cook and Campbell (1979) and Shadish et al. (2002). Scriven (2008) labels the process of ruling out alternative explanations the general elimination method (GEM) and Reichardt (2000) provides a typology of the mechanisms by which alternative explanations can be eliminated when estimating treatment effects.

The need to rule out alternatives explanations is no different when estimating treatment effects than when drawing any conclusion. The way to warrant any claim to knowledge is by identifying alternatives and assessing their implications. If you have been diligent in coming up with alternatives and in assessing their implications, and if the implications of one explanation hold true much more than the implications of any of the alternatives, you are then justified in tentatively accepting that explanation as true. Such a path to knowledge is the path of eliminating alternative explanations. As Scriven (2008) makes clear, the GEM is the only method available for warranting claims to knowledge whether those claims are derived from quantitative, qualitative, or any other method of investigation.

Which is the best comparison for estimating treatment effects depends on the degree to which alternative explanations can be ruled out. The sections that follow discuss several of the most important alternative explanations (as well as other factors) to consider in evaluating comparisons for estimating treatment effects.

Selection Differences

Selection differences are differences between the units (e.g., recipients, settings, times, or outcome variables) in the different treatment conditions. In particular, selection differences are the differences between the units in the treatment groups that would arise even if the units did not receive different treatments. For example, consider a comparison where one group of recipients receives treatment X and another group receives treatment Y. Differences in outcomes between these two groups could be due not just to treatment effects but also to the effects of initial differences between the individuals in the two groups—differences that are present regardless of the effects of the treatments. The initial differences between the two groups of recipients are selection differences.

By definition, the counterfactual comparison that defines an effect is free of selection differences. In contrast, all practical comparisons have selection differences. It is simply impossible to obtain a comparison between what happened after treatment X was implemented and what happened after treatment Y, instead of treatment X, was implemented without having selection differences. However, the four types of comparisons (i.e., comparisons between recipients, settings, times, and outcome variables) each has a different primary source of selection difference (Reichardt, 2006). In comparisons between recipients, selection differences are primarily differences between the recipients in the different treatment conditions. In comparisons between settings, selection differences are primarily differences between the settings in the different treatment conditions. In comparisons between times, selection differences are primarily differences between the times in the different treatment conditions. (In Campbellian nomenclature, selection differences in comparisons between times include differences due to maturation, testing, instrumentation, history, mortality, and statistical regression.) In comparisons between outcome variables, selection differences are primarily differences between the outcome variables in the different treatment conditions. When a treatment effect is estimated using a combination of comparisons, at least one type of selection difference is always present and multiple types of selection differences are usually present.

Note that in Campbellian nomenclature, selection differences are defined as differences only between the recipients in comparisons between recipients. I define the notion of selection differences more broadly (than in the Campbellian tradition) to include differences between the recipients, settings, times, or outcome variables in each of the different types of comparisons. I use the same
Selection differences can produce outcome differences between the treatment groups that either masquerade as treatment effects or mask treatment effects. The effects that selection differences have on outcome differences may or may not be substantial in any given comparison. But because selection differences are always present in any practical comparison, their effects must be taken into account (if in no other way than by showing their effects are insubstantial) if estimates of treatment effects are to be credible. The remainder of the present section is concerned with the means of taking selection differences into account.

In what follows, I conceptualize selection differences as being composed of two components—random differences and systematic differences. Random selection differences diminish to zero as the sample size (i.e., the number of units) increases. Random selection differences introduce no bias in the estimate of a treatment effect. In contrast, the size of systematic selection differences does not necessarily change as the sample size increases. In addition, systematic selection differences can introduce bias in the estimate of a treatment effect. I also distinguish instances where attrition is and is not present, where attrition means data is missing from some units in the study. Each of these cases will be addressed separately below. The same conclusions hold for all types of units. That is, the same conclusions apply whether a comparison is drawn between recipients, settings, times, or outcome variables.

Removing Bias in the Absence of Attrition

When there is no attrition, RA removes systematic selection differences. This means RA comparisons are free from bias due to selection differences, in the absence of attrition. The statistical analyses of data from randomized experiments without attrition need do nothing to remove bias due to selection differences. The bias is removed by the very process of RA itself. There is no such automatic removal of bias from selection differences in either QA or non-QA comparisons. By their very nature, QA comparisons create groups that differ systematically on the quantitative measurement used for treatment assignment and on related variables, and hence create systematic selection differences. Selection differences in non-QA comparisons also tend to be systematic.

In QA comparisons, the systematic effects of section differences can be taken into account by modeling the relationship between the outcome scores and the QA variable. This relationship must be modeled as it would exist in the absence of a treatment effect. If the relationship would be curvilinear in the absence of a treatment effect, the analyst must model the correct curvilinear shape. There can be a good deal of uncertainty about how to model this relationship properly. To the extent there is uncertainty, there will be uncertainty about whether the systematic effects of selection differences have been taken into account properly (Reichardt & Henry, 2012). Ways of minimizing bias due to curvilinearity is a topic of research, especially in economics where the regression-discontinuity design has received a resurgence of attention (Imbens & Lemieux, 2008). Because the analyst does not need to perform such modeling to remove bias with RA, this source of uncertainty is not present with RA comparisons.

In non-QA comparisons, selection differences can be taken into account either by (a) measuring the effects of selection differences and subtracting those measured effects from the observed outcome differences between the groups or (b) by comparing units from the different treatment groups after they have been matched (either physically or statistically) to take account of the presence of selection differences. Consider each of these two options in turn.

In its simplest form, option (a) entails demonstrating that the effects of selection differences are sufficiently small that they can be ignored (Reichardt, 2000). Eckert (2000) provides an example that...
uses a before–after comparison to estimate the effects of a training program. Because of the short
time period between the before and after measurements, the very focused nature of the program, and
the program’s relatively large effects, Eckert (2000) is able to argue persuasively that any effects of
selection differences due to history, maturation, testing, instrumentation, or mortality are likely too
small to account for the observed differences in outcomes. If the effects of selection differences are
not small enough to be ignored, an additional comparison is usually required on which the effects of
selection differences are measured and then subtracted from the observed group difference on the
outcome scores. This additional comparison must assess the effect of selection differences on the
same scale of measurement as used to assess outcomes otherwise the analyst will be subtracting
apples from oranges. Unfortunately, evaluators usually cannot be confident they can obtain an addi-
tional comparison that accurately measures the effects of selection differences.

The matching of units required in option (b) can be pursued in either of two ways—units can be
matched on the determinants of the outcome scores or on the determinants of group selection
(Reichardt, 1979). In either case, the determinants of outcomes scores or group selection must be
accurately measured for each unit. This can be a difficult and uncertain task. The analyst may not
know much about the determinants of either outcome scores or group selection and even if the deter-
mnants are known, they may not be easy to measure. If any of the determinants are omitted or if any
of the determinants are measured imperfectly, bias due to selection differences can remain. Struc-
tural equation modeling (SEM) is often used to match units on the determinants of the outcome
scores because SEM allows researchers to model measurement error in the determinants of the out-
come scores. Propensity scores provide a means of matching units on the determinants of group
selection, have been the subject of extensive research in the field of statistics over the past few of
decades (Rosenbaum & Rubin, 1983; Rubin, 2005), and are being widely used in practice (Leow,
Marcus, Zanutto, & Boruch, 2004; Peck, 2007; Smith, Brandon, Lawton, & Krohn-Ching, 2010).
Whichever approach to option (b) is used, taking account of the systematic effects of selection dif-
f erences in non-QA comparisons requires additional data and additional assumptions compared to
RA comparisons and the validity of the additional assumptions will often be highly uncertain. Tak-
ing account of selection differences in non-QA comparisons using option (b) is also usually far more
uncertain than taking account of selection differences in QA comparisons. One way to characterize
the additional uncertainty is the following. The analysis of data from a QA comparison is akin to a
propensity score approach to matching. As noted above, failure to include all the determinants of
group selection or failure to measure the determinants perfectly can lead to bias in a matching anal-
ysis. In QA comparisons, an available quantitative measurement is the sole determinant of group
selection so the determinant of selection is perfectly known and perfectly measured. In contrast,
in non-QA comparisons the determinants of group selection are imperfectly known and imperfectly
measured.

The bottom line is the following. Whether option (a) or (b) is used, far greater uncertainty usually
exists in non-QA comparisons, than in either RA or QA comparisons, about whether estimates of
treatment effects are reasonably free from the systematic effects of selection differences. In general,
the best strategy with a non-QA comparison is to implement the treatment so it has a large effect
relative to variability among the units and design the study so the effects of selection differences are
small, because such actions tend to reduce the effects of uncertainty about how well biases due to
selection differences have been addressed.

Taking Account of Random Selection Differences

Even though RA removes the systematic effects of selection differences, random effects of
selection differences are still present in RA comparisons. The same holds for the other two types of
comparisons (QA and non-QA). That is, whether or not the systematic effects of selection
The means of taking account of the effects of random selection differences are the same in the three types of comparisons (RA, QA, and non-QA). Classic statistical procedures (including statistical significance tests and confidence intervals) were designed expressly to assess and take account of the effects of random selection differences. Statistical procedures require some knowledge of the dependence among observations and, if parametric procedures are to be used, some knowledge of the shape of the distribution of observations, but the necessary assumptions do not differ across the different types of comparisons whether RA, QA, or non-QA. And in any case, appropriate statistical procedures can usually be implemented equally well across the different types of comparisons whether RA, QA, or non-QA. So the ease or difficulty of taking account of the random effects of selection differences provides little basis for choosing among these three types of comparisons.

But although statistical models are equally applicable to RA, QA, and non-QA comparisons, the different types of comparisons usually require different numbers of units if the statistical procedures are to have the same degree of power and precision. For example, even under ideal conditions, QA comparisons generally require more than twice as many units to have the same power and precision as RA comparisons (Cappelleri, Darlington, & Trochim, 1994; Goldberger, 1972). Non-QA comparisons also tend to require more units than RA comparisons but circumstances vary too much to be able to usefully quantify the difference. Because RA removes systematic selection differences in the absence of attrition and because the effects of random selection differences diminish with sample size, a researcher using an RA comparison can make the effects of selection differences as small as desired, in the absence of attrition, simply by increasing the sample size. The same cannot be said about QA and non-QA comparisons because of uncertainty about taking the systematic effects of selection differences into account. Increasing the sample size in QA and non-QA comparisons decreases the random effects of selection differences, but uncertainty about the systematic effects of selection differences usually remains even as sample size increases, so researchers using QA or non-QA comparisons cannot be confident the effects of selection differences decrease toward 0 as the sample size increases, even in the absence of attrition.

The simplest statistical models assume independence among the outcomes from different units. However, dependencies often arise when units are times, outcome variables, and recipients who are grouped within clusters (such as students within classrooms). Statistical methods that address dependencies in time-series data have long been available (Box & Tiao, 1975; Veney, 1993). Advances in hierarchical linear modeling (HLM) in the past few decades enable researchers to model dependencies introduced when units are grouped in clusters (Bryk & Raudenbush, 1992; Newton & Llosa, 2010; Raudenbush & Bryk, 2002). In the past, ITS designs required substantial numbers of time periods if dependencies (i.e., autocorrelations) across times were to be properly modeled. But HLM methods allow ITS designs to be implemented with relatively few time points as long are there a number of recipients (or other units) at each time point. Another advantage of HLM methods is they permit covariates to be added at each level of the hierarchical groupings in the data which can increase the power and precision of the analyses. Such advances in statistical methods continue to improve the ease and accuracy with which the random effects of selection differences can be taken into account. These advances are applicable to all three types of comparisons: RA, QA, and non-QA. The same holds for non-statistical means of improving power and precision. For example, the methods Meier (2004) discusses for increasing power and precision by tailoring outcome measures to the demands of the specific research setting are equally applicable to RA, QA, and non-QA comparisons.

It is not clear that humans have sufficiently good intuition about random selection differences to reliably take account of their effects without the assistance of statistical methods, except perhaps when treatment effects are very large in relation to the heterogeneity among units. So, unless
treatment effects are large compared to the heterogeneity among units, it is not clear how qualitative researchers could credibly take account of the random effects of selection differences without the benefit of statistical procedures. Yet random effects of selection differences exist in all types of comparisons whether RA, QA, or non-QA. Perhaps, as suggested by Scriven (2009, p. 146) evaluators “are not very interested in small differences because they have a track record of never showing up on the replications at distant sites” which provides the justification for using a qualitative assessment of random selection differences “knowing that it’s a net that will catch only big fish.” From that perspective, the advantage of using statistical procedures, rather than intuition, is that statistical procedures better allow for the detection of small effects.

Coping With Selection Differences in the Presence of Attrition

As noted above, in the absence of attrition, an RA comparison is free from bias due to selection differences. In addition, no bias due to selection differences is introduced by random attrition. Nor is bias introduced in an RA comparison if attrition is the same in the two treatment groups, whether it is random or nonrandom. But if attrition is present, researchers can seldom be sure it is either random or the same in the two treatment groups. So when attrition is present in an RA comparison, researchers must entertain the possibility that the attrition introduces biases due to selection differences between the treatment groups.

The same analytic methods used to address bias due to selection differences in non-QA comparisons, as described above, can be used to address selection bias that is introduced by attrition in RA comparisons. These analytic procedures could be used either with or without imputing scores for the attrited (i.e., missing) data values (Schafer & Graham, 2002). Also as noted above, these analytic methods require additional data and assumptions that would increase uncertainty about the accuracy of estimates of treatment effects compared to estimates derived from RA comparisons in the absence of attrition. The effect of attrition is much the same in QA comparisons. That is, attrition increases uncertainty about whether the biasing effects of selection differences are well taken into account. Attrition can also make it more difficult to remove the biasing effects of selection differences in non-QA comparisons.

But although attrition tends to increases uncertainty about the results of any comparison, its effects are relatively greater in RA and QA comparisons than in non-QA comparisons. Without attrition, researchers can be more confident that bias due to selection differences is removed from RA comparisons than from QA comparisons. And without attrition, researchers can typically be more confident that bias due to selection differences is removed from QA comparisons than from non-QA comparisons. But in the presence of attrition, these differences in confidence tend to be diminished because attrition tends to make both RA and QA comparisons more like non-QA comparisons. In addition, attrition may be worse in RA and QA comparisons more like non-QA comparisons because, in comparisons between recipients, (a) the recipients in RA and QA comparisons are not allowed to choose the treatments to which they would most like to be assigned while recipients in some non-QA comparisons are allowed to select their treatments and (b) lack of freedom of choice could increase attrition.

But while attrition tends to diminish the advantage RA comparisons have over QA comparisons and diminish the advantage QA comparisons have over non-QA comparisons in removing biases due to selection differences, attrition does not tend to remove these advantages altogether. The advantages that exist without attrition tend to remain in the presence of attrition. The advantages only diminish as the amount of attrition increases. For example, consider an RA and a non-QA comparison with 10% of the outcome scores in each treatment group missing because of attrition and a statistical matching procedure used to address selection differences. First consider the RA comparison first with and then without imputation of the missing data. With imputation of the missing data,
100% of the units would be matched properly so there would be no selection bias. The only source of error would be inaccuracy in the data imputation process for the outcome scores of 20% of the units. Without imputation, 80% of the original units in each group would still be matched properly and bias due to selection differences would be present only because randomly equivalent matches for 20% of the units were missing due to attrition. In contrast, in a non-QA comparison, there is no guarantee that any of the units would be properly matched either with or without data imputation, so the statistical matching procedure would have to account for potential biases due to selection differences in all of the data.

Regardless of the type of assignment to treatment conditions (RA, QA, or non-QA), researchers should incorporate strategies to avoid attrition as much as possible (Shadish et al., 2002). Researchers should also make contingency plans in case attrition arises in spite of their efforts to avoid it. Contingency plans include collecting data needed to both diagnosis differential attrition and implement strategies to correct for the biasing effects of differential attrition which might include both data imputation strategies and matching procedures (Harris, 1998; West & Thoemmes, 2010).

**Awareness and Expectations of Treatments Received**

Both the awareness that a treatment has been implemented and expectations about the effects of a treatment can bias estimates of treatment effects. Such biases can arise because of awareness and expectations in either participants or researchers.

**Participant Expectations**

Participants are usually aware of the treatments they receive, as well as of the treatments they do not receive, and such awareness, coupled with expectations about the treatments, can affect the outcomes of the study. To the extent they are not intended as part of the active ingredients of the treatments, awareness and expectations can bias estimates of treatment effects. Placebo effects are examples of such sources of bias. Other examples of unwanted effects of awareness and expectations are resentful demoralization, compensatory rivalry, and Hawthorne effects. Resentful demoralization occurs when participants resent the treatment they receive (given their awareness that others receive a more desirable treatment) and therefore perform more poorly than they would have otherwise. Compensatory rivalry occurs when participants are aware others receive a superior treatment and, because of competitive instincts not to be outperformed, achieve better results than they would have otherwise. Hawthorne effects are caused by awareness that one’s behavior is being observed plus the (perhaps unconscious) expectation that any changes in conditions or procedures are intended to influence that behavior.

In some cases, effects due to awareness and expectations should be conceptualized as part of the active ingredients of the treatment. For example, homeopathic medications (which contain no more than miniscule amounts of active ingredients) are sometimes prescribed because doctors believe the placebo effects themselves will be therapeutic. In this case, placebo effects are part of the intended intervention. Or resentment from receiving an undesirable treatment can occur whether or not others receive a more desirable treatment, in which case the effects of resentment might appropriately be considered an inherent part of the treatment’s effect rather than an avoidable artifact. But when participants’ awareness of the treatments both received and not received introduces artifactual bias, these potential effects need to be addressed when estimating treatment effects.

One approach to coping with artifactual effects of awareness and expectations is to try to keep participants from having either awareness or expectations about the treatments. For example, awareness can be eliminated by masking (or blinding) the participants to the treatments they receive. Alternatively, artifactual biases due to resentful demoralization might be avoided by restricting the pool of participants to those who find the treatments equally appealing, although this might reduce
the generalizability of the study’s results. Or researchers can try to counteract the effects of awareness by offering compensating inducements such as the promise of receiving the more desired treatment after the study is completed.

An alternative approach to coping with artifactual effects of awareness and expectations is to measure these effects and either (a) show they are too small to matter or (b) adjust for them when estimating treatment effects. Self-report measures may be especially sensitive to detecting the presence and intensity of artifactual effects of awareness and expectations. And qualitative approaches may be better than quantitative approaches at eliciting such self-reports. But, ironically, self-reports of participants and other stakeholders may themselves be more susceptible to the effects of awareness and expectations than other measures.

Some effects of awareness, such as placebo effects, are probably just as likely to occur in RA as in QA or non-QA comparisons (and can occur even in the counterfactual comparison that defines an effect). But some effects of awareness may be less likely to occur in some comparisons than others. For example, when recipients are allowed to select the treatments they are to receive in a non-QA comparison, resentful demoralization is less likely to occur than in more “forced-choice” comparisons which restrict participants’ freedom to choose the treatments they desire, such as non-QA comparisons where administrators assign treatments and both RA and QA comparisons between recipients (Fetterman, 1982; Lam, Hartwell, & Jekel, 1994).

**Researcher Expectations**

Artifactual effects due to researchers’ expectations are a result of confirmation bias where researchers’ desires or expectations about treatments alter (perhaps unconsciously) what the researchers decide to observe and how they interpret those observations (Gilovich, 1991; Sadler, 1981). To the extent researchers have more leeway in choosing what to observe and how to interpret those observations, confirmation bias has more room to operate. If qualitative investigations tend to allow more leeway than quantitative investigations, qualitative investigations, which tend to use more non-QA comparisons than RA comparisons, tend to be susceptible to more confirmation bias than quantitative investigations. Masking (or blinding) is one approach to dealing with experimenter and observer expectancies, but masking is more likely to be feasible with quantitative methods than with qualitative methods partly because qualitative investigators tend to have more intimate familiarity with both participants and programs. Methods have been suggested for making the quantitative data analyst blind to treatment conditions (Rubin, 2005). It is not clear to me how qualitative researchers can be effectively blinded so as to avoid potential biases due to researcher expectations.

**Strength and Integrity of Treatments and Treatment Assignment**

Knowing that one program produces better outcomes than another is useful only to the extent the differences between the programs are well understood. Adequate program descriptions include specification of (a) the treatments both as planned and as implemented and (b) discrepancies between how the treatments were supposed to be assigned to units and how they were received by the units.

**Planned Interventions**

Researchers need to describe the treatments that were planned and the extent to which the treatments were or were not implemented as planned (Barnette & Wallis, 2005; Brandon, Taum, Young, Pottenger, & Speitel, 2008; Century, Rudnick, & Freeman, 2010; Mowbray, Bybee, Holter, & Lewandowski, 2006; Zvoch, 2009). Sometimes elements of treatment X which were not supposed to be part of treatment Y nonetheless become a part of treatment Y, and vice versa. Such leakage
(also called treatment diffusion or contamination) seems no more likely, in general, to occur with one type of assignment (RA, QA, or non-QA) than with another.

Heterogeneity in treatment implementation can increase variability in outcome scores which can reduce statistical power and precision. Critics of RA comparisons have sometimes claimed that heterogeneity in treatment implementation invalidates RA comparisons. But that is not true. Heterogeneity has comparable effects on power and precision across all three types assignment; RA, QA, and non-QA (Kraemer & Thiemann, 1987; Lipsey, 1990).

Noncompliance

In all types of assignment to treatment conditions (RA, QA, and non-QA), researchers need to know if the study units all received the treatment conditions to which they were assigned or if some failed to comply with the treatment assignment perhaps by crossing over to receive an unassigned treatment (West & Thoemmes, 2010). Noncompliance is a problem for all types of assignment to treatment conditions (whether RA, QA, or non-QA) in the sense that researchers need to know if noncompliance has occurred when analyzing and interpreting the results. However, noncompliance presents more of a problem for RA and QA comparisons than for non-QA comparisons because noncompliance introduces additional complexities into the analysis of the data. In addition, noncompliance may be more likely to occur in RA and QA comparisons than in some non-QA comparisons. For example, in comparisons between recipients, the participants in RA and QA comparisons are not allowed to choose the treatments to which they are assigned (unlike participants in some non-QA comparisons), and the lack of freedom of choice could increase the likelihood of noncompliance.

The problems introduced by noncompliance can be addressed using either of two analytical approaches; analyze the data according to the how the treatments were assigned or analyze the data according to how the treatments were received. For non-QA comparisons, the best option generally is to analyze the data according to how the treatments were received (rather than how they were assigned). As long as the researcher knows which treatments were received, the presence of noncompliance does not change the logic of the analysis of data from non-QA comparisons.

The option of analyzing the data according to how the treatments were received is better suited to RA than QA comparisons (so I will consider it only for RA comparisons). Grouping together everyone who receives the same treatment makes it easier to characterize the nature of the treatment difference between the groups and thereby resolves the primary problem introduced by noncompliance. The disadvantage of analyzing the data according to the treatment received is that what was originally an RA comparison is no longer an RA comparison. Because of noncompliance, the units in the treatment conditions are no longer randomly equivalent and hence could suffer from systematic effects due to selection differences. These systematic effects of selection differences can be addressed using the methods I described previously for coping with systematic selection differences in non-QA comparisons.

Because analyzing the data according to the treatment received can introduce systematic selection differences into an RA comparison, one of the relative advantages of an RA comparison is diminished—an RA comparison becomes more like a non-QA comparison. But though diminished, the relative advantage of an RA comparison compared to a non-QA comparison is not lost entirely. With non-QA comparisons, systematic effects of selection differences can be present in all of the data while in RA comparisons, systematic effects of selection differences are present only in the percentage of the units that do not comply. The remainder of the data remains free from the systematic effects of selection differences. For example, consider an RA and a non-QA comparison with 10% crossovers from each treatment group into the alternative treatment group, where a statistical matching procedure is used to take account of the systematic effects of selection differences. In the RA comparison, 80% of the units would be properly matched so the statistical matching procedure...
would have to adjust for bias in only the remaining 20%. In contrast, in a non-QA comparison there is no guarantee any of the units would be properly matched so the statistical matching procedure would have to take account of selection differences in 100% of the data.

The alternative option of analyzing the data according to how the units were assigned to treatments is well suited to either QA or RA comparisons. With this option, the data are analyzed as if the units assigned to the different treatment conditions actually received those treatments, even when they did not. This preserves the integrity of the assignment process. In the QA comparison, that mean all units are assigned to treatments using a QA rule. In RA comparisons, it means the units are assigned to treatments randomly which means systematic effects of selection differences are avoided. By preserving the integrity of the assignment process, the presence of noncompliance does not complicate the process of taking account of selection differences. However, the presence of crossovers can bias the estimate of the treatment effects in both RA and QA comparisons because not everyone receives the treatment assumed in the analysis. Adjustments have been devised to correct for this bias (Little & Yau, 1998; West & Thoemmes, 2010), though the adjustment requires assumptions which introduce additional uncertainty into the results. But while noncompliance increases the uncertainty of the analysis of data from RA and QA comparisons, that uncertainty (which is introduced by the presence of noncompliance among a portion of the units) is again usually less than the uncertainty in non-QA comparisons due to the presence of the systematic effects of selection differences among all the units.

Researchers might well consider incorporating strategies to reduce noncompliance, especially in RA and QA comparisons. Researches can also make contingency plans in RA and QA comparisons if noncompliance occurs in spite of efforts to avoid it. Contingency plans include collecting the data needed to identify noncompliance and to implement strategies to correct for its biasing effects (Little & Yau, 1998).

Incorrect Assignment to Treatments

In RA and QA comparisons, researchers need to know if the units were assigned to treatment conditions correctly and not just whether units received the treatments that were assigned. That is, researchers need to know whether the units were assigned to treatments as they were supposed to be assigned (and not just if they received treatments that might have been assigned incorrectly). In designs intended to be RA comparisons, the process of randomization can break down or be sabotaged and safeguards have been devised to ensure that assignment is truly random and is free from tampering by program personnel (Braucht & Reichardt, 1993; Ong-Dean, Hofstetter, & Strick, 2011). In designs intended to be QA comparisons, assignment based on a cutoff score on a quantitative variable can break down, for example, because participants (after learning the value of the cutoff score) fudge their scores on the QA variable so as to become eligible for the treatment they most desire. Such fudging is usually most prevalent among recipients with scores close to the cutoff value and results in what has come to be known as fuzzy assignment (Trochim, 1984). Several statistical approaches have been devised to take account of the biases introduced by fuzzy assignment (Reichardt & Henry, 2012). Researchers implementing QA comparisons between recipients are advised to avoid fuzzy assignment by keeping the value of the cutoff score confidential or using a QA variable that is difficult to manipulate.

Additional Treatments

Researchers need to know if additional treatments are implemented beyond those that are intended. Additional treatments can be unwittingly introduced by the manner in which a study is implemented. For example, if treatments X and Y consist of different curricular materials and if the materials are taught by different teachers, the differences between the teachers become part of the treatment
difference whether intended or not. Alternatively, additional treatment differences can be introduced by forces outside a research study. For example, additional treatments can be introduced by administrators who recognize that treatment X provides more services than treatment Y and attempt to compensate for that treatment difference by using discretionary funds to provide additional services to the participants assigned to treatment Y. Such external interventions to balance services have been called the compensatory equalization of treatments. No type of assignment to treatment conditions (whether RA, QA, or non-QA) seem necessarily more susceptible to the introduction of additional treatments than another.

**Ease of Implementation and Generalizability**

Non-QA comparisons constrain how units are assigned to treatments less than RA and QA comparisons. In RA comparisons, units must be assigned to treatment conditions at random. In QA comparisons, units must be assigned to treatment conditions using a cutoff score on a quantitative variable. In contrast, no constraints need be imposed on the way units are assigned to treatment conditions in non-QA comparisons. The way assignment is or is not constrained can influence how easily a comparison can be implemented and how broadly its results can be generalized.

**Ease of Implementation**

Different treatments and research settings are amenable to different types of treatment assignment (i.e., RA, QA, and non-QA). Most obviously, some treatments cannot be ethically assigned to recipients at random. For example, it would be unethical to randomly assign people to either smoke or not smoke, to take illegal drugs or not take illegal drugs, or to exercise or not exercise. In addition, administrators, such as school principals, might be unwilling to assign participants at random to certain treatments even if it were ethically permissible, for fear stakeholders, such as parents, would object. Under such circumstances, it would be easier to implement a non-QA comparison than an RA comparison. Similarly, occasions could arise where it would be easier to implement a QA comparison than an RA comparison. For example, while a principal might not allow children in need of an intervention to be assigned to a remedial reading program at random, the principal might permit assignment based on a quantitative measure of need, and hence permit a QA comparison. Of course, the opposite circumstances can also arise. For example, a principal might believe RA is fairer than any other means of distributing a limited resource, including either QA or non-QA allocations.

An important set of circumstances arises when an evaluation is performed after the fact—that is, after a program has been implemented and no more units (recipients, settings, times, or outcome variables) will receive the treatment than have already received it (Bamberger, Rugh, Church, & Fort, 2004; Mastoroudes, 1993). In this case, the evaluator is constrained to use whatever comparisons can be identified among already existing variations in treatments across units. Such circumstance arose, for example, in Patton’s (2008) evaluation of a stealth campaign to influence a Supreme Court’s decision and Scriven’s (2008) evaluation of an aid program in Africa. In such evaluations, it will often be relatively easy to uncover or create non-QA comparisons, often possible to uncover or create QA comparisons (especially QA comparisons between times in the form of ITS designs—which is what both Patton and Scriven did), but virtually impossible to uncover or create RA comparisons, although comparisons that are essentially random have sometimes been found after the fact (e.g., Langer & Rodin, 1976).

Some treatments and research circumstances are more amenable to comparisons between some types of study units than other types. For example, it might be possible, as suggested above, to assess the effects of an educational television program by assigning the treatments to different outcome variables such as letters of the alphabet. But drawing an RA comparison between outcome variables is possible only if a set of comparable outcome variables is available and only if the effects of the
treatment (e.g., an educational television program) can be directed at random to some variables (e.g., the pronunciation of different letters of the alphabet) and not to others. Unfortunately, these conditions do not often arise. Conversely, comparisons between either outcome variables or times may be possible when comparisons between recipients are not possible, because only a very few recipients are available for study. Indeed, comparisons between times or outcome variables have often been implemented with only a single recipient (called N of 1 designs).

Two conclusions can be reached. First, different settings and different treatments allow different types of assignment to treatment conditions using different types of units. No type of assignment (i.e., RA, QA, and non-QA) or unit (i.e., recipients, settings, times, and outcome variables) is equally applicable in all settings and with all treatments. Second, the ease of implementation of a comparison depends on the combination of type of assignment (i.e., RA, QA, and non-QA) and type of unit (i.e., recipients, settings, times, and outcome variables). For example, even though RA comparisons impose more constraints than non-QA comparisons, ease of implementation may depend on the type of units. So an RA comparison between times might be easier to implement in a given research setting than a non-QA comparison between recipients. That ease of implementation depends on the research circumstances provides another reason for considering all the potential comparison options in Table 1. Evaluators increase the likelihood of selecting the best option for a given set of circumstances when they consider the broadest array of options possible.

**Generalizability of Results**

Results from RA and QA comparisons tend to be less generalizable than results from non-QA comparisons, because of the different constraints imposed by the different comparisons (Shadish et al., 2002; West, Biesanz, & Pitts, 2000). For example, RA comparisons between recipients can be implemented only with recipients who are willing to accept assignment to either of the two treatment conditions that are being compared. But a non-QA comparison between recipients is not so restricted. So the results of a non-QA comparison between recipients are likely to be generalizable to a broader range of recipients than the results of an RA comparison between recipients. Similarly, the results of QA comparisons also tend to be less generalizable than results from non-QA comparisons because environments amenable to QA assignment tend to be more restricted than environments amenable to non-QA assignment. But again, differences due to units need to be taken into account. Research settings that allow comparisons between times or between outcome variables might be different than research settings that allow comparisons between recipients. As a consequence, the results from an RA comparison between times might be more generalizable than the results from a non-QA comparison between recipients. Again, there is value in explicitly considering all the options in Table 1 when designing an evaluation, especially because not all the options are equally likely to come to mind otherwise.

Of course, differences in generalizability need to be tempered by the validity of the results that are being generalized. That the results from non-QA comparisons between recipients are likely to be more generalizable than the results from RA comparisons between recipients may not be much of an advantage if estimates of treatment effects from non-QA comparisons are more biased or less precise than estimates from RA comparisons. In addition, generalizations can be more credible when based on replications across a variety of circumstances. Therefore, there can be value in using a variety of different types of comparisons for the purpose of increasing heterogeneity across replications in constraints, conditions, and biases (Mark & Reichardt, 2004).

**Other Evaluation Tasks Besides Estimating Treatment Effects**

Evaluations can serve many purposes besides estimating treatment effects. Other evaluation tasks include assessing needs, providing formative feedback to improve ongoing program operation,
assessing the features of a program or environment that impede or enhance implementation of the treatment, identifying the processes by which a treatment has its effects (i.e., program mediation—Astbury & Leeuw, 2010), identifying who benefits most and who benefits least from treatments (i.e., moderation effects), and assessing unintended consequences, among many others. For the same reasons RA comparisons can be advantageous for taking account of the systematic effects of selection differences when estimating treatment effects, RA comparisons can also be advantageous for assessing mediation, moderation, and unintended program effects, because these tasks are simply elaborations of the task of estimating treatment effects.

Critics sometimes note that RA comparisons are poorly suited for evaluation tasks unrelated to estimating effects. This may be the case, but it is not clear that QA or non-QA comparisons are any better suited than RA comparisons for such tasks. Devoting resources to a comparison for estimating treatment effects diminishes the resources that can be allocated to other tasks. If RA comparisons tend to be more expensive than other comparisons, that would disadvantage RA comparisons more than QA or non-QA comparisons when used to serve other functions. But it is not clear that RA comparisons are generally more expensive than either QA or non-QA comparisons, especially given that RA comparisons require fewer units to achieve the same level of statistical power and precision than either QA or non-QA comparisons. Other evaluation functions, besides estimating treatment effects, can often be addressed by adding research components, including various qualitative investigations, to an impact evaluation. In general, additional features, such as qualitative investigations, can be added just as easily to RA comparisons as to QA or non-QA comparisons.

Conclusions

A comparison between what happened after a treatment was implemented and what happened after an alternative treatment was implemented can be obtained using any of four different types of units of assignment to treatment conditions: recipients, settings, times, or outcome variables. And the assignment of each of these four types of units can take any of three forms: RA, QA, or non-QA. Crossing the four types of units (recipients, settings, times, and outcomes variables) with the three types of assignment to treatment conditions (random, quantitative, and non-quantitative) produces 12 types of comparisons, as presented in Table 1. I recommend that evaluators consider all 12 comparisons when selecting a design to estimate program effects.

Researchers often overlook or mischaracterize some of the 12 types of comparisons. For example, Campbell and Stanley (1966) categorized randomized comparisons between outcome measures as quasi-experiments rather than as randomized experiments. Apparently, Campbell and Stanley, like many others, assume randomized experiments entail comparisons only between people or other recipients. But the virtues of randomization are achieved whether recipients, settings, times, or outcome variables are randomly assigned to treatments. So it makes sense to classify random comparisons between settings, times, and outcome variables as randomized experiments just as random comparisons between recipients are classified as randomized experiments.

Random assignment to treatment conditions is not required to reach credible conclusions about the effects of treatments. I am highly confident, for example, that turning the key in the ignition of my car causes it to start, yet I have never performed a randomized experiment to make that assessment (though I do have very extensive data from innumerable ITS designs). Nor are randomized experiments necessarily the best comparison for estimating program effects. Which type of comparison is best under a given set of circumstances will depend on a variety of criteria including bias, precision, generalizability, ease of implementation, and cost. No one type of comparison will be best under all circumstances for all different weightings of these multiple criteria. So the potential advantages of RA must be tempered against its potential disadvantages. QA and non-QA comparisons
might be preferred to RA comparisons because comparisons based on RA may not be as feasible, results may not be as generalizable, and biases may be greater from sources other than selection differences. One or all of these disadvantages might counterbalance the advantages RA has in coping with selection differences. These are the types of considerations that must be taken into account when selecting a comparison for estimating effects.

Nonetheless, evaluators should not overlook the distinct advantages that randomized assignment to treatment conditions have compared to other types of assignment. Random assignment was designed explicitly to take account of the effects of selection differences and randomization is remarkably good at that task. Unlike other types of assignment to treatment conditions, RA eliminates bias due to selection differences in the absence of differential attrition and noncompliance to treatment assignment. And although RA does not necessarily eliminate bias due to selection differences in the presence of differential attrition and noncompliance, coping with bias due to selection differences is often performed both more credibly and more easily with RA than without RA, even in the presence of differential attrition and noncompliance.

Random assignment to treatment conditions may have no advantages, compared to other types of assignment to treatment conditions, in coping with forms of bias such as due to either experimenter or participant expectancies or to treatment diffusion or treatment equalization. In fact, RA to treatment conditions may tend to make some biases worse, compared to other forms of assignment. But other sources of bias besides selection differences are not always present. Such is not the case with selection differences. Selection differences are necessarily present in any study that estimates treatment effects by comparing what happened after a treatment was implemented to what happened after an alternative treatment, or no treatment, was implemented instead. And the biases introduced by selection differences can be, and often are, substantial. So the advantage RA has in coping with the effects of selection differences should not be dismissed lightly.

It is not clear evaluators well recognize the inevitable presence of selection differences in empirical comparisons of outcomes and the advantages of RA in coping with these differences. Because randomization has such important virtues, evaluators are well advised to consider the four different forms in which randomized experiments can be implemented. And evaluators should not reject randomized experiments without carefully considering their potential advantages. Blanket statements that randomized experiments are always to be preferred are no more defensible than blanket statements that randomized experiments are never useful. Circumstances and contingencies matter. The choice of research design requires a careful weighing of a variety of criteria under the constraints imposed by the research circumstances.

The 12 basic types of comparisons presented in Table 1 can be elaborated in at least two ways. One form of elaboration, as I have noted briefly, involves using combinations of comparisons. The other form of elaboration, as I have also noted briefly, involves adding adjunct comparisons to address threats to validity such as when bias due to selection differences is estimated and subtracted from the estimate of the treatment effect (Reichardt, 2000, 2009; Reichardt & Gollob, 1989). The depth of insight and flexibility provided by qualitative methods may be useful in identifying and implementing adjunct comparisons (e.g., Scriven, 2008, 2009). Quantitative researchers also have a long history of using additional comparisons to create complex patterns with which to disentangle treatment effects from the effects of threats to validity (e.g., Reichardt, 2000, 2009; Reynolds, 1998; Shadish & Cook, 1999; Shadish et al., 2002). Perhaps the best option is to draw from the strengths of both qualitative and quantitative approaches (Cook & Reichardt, 1979; McConney, Rudd, & Ayres, 2002; Reichardt & Cook, 1979). On the one hand, qualitative methods can be invaluable in identifying and addressing biases due to systematic selection differences, attrition, noncompliance, additional treatments, and treatment expectations. Qualitative methods can also be indispensable for detecting unintended consequences of treatments (Scriven, 2008, 2009). On the other hand, the full advantages of RA and QA comparisons cannot be obtained without quantification. And the
quantitative models of statistical methods are hard to replace for estimating the effects of random selection differences which are present in all comparisons. Each of the two traditions, qualitative and quantitative, has much to offer in coping with the many nuances and uncertainties involved in estimating effects.

In choosing a comparison, researchers should consider all the options including the three different ways of assigning treatments to units (RA, QA, and non-QA) and the four types of units (recipients, settings, times, and outcome variables) and all the potential threats to validity including selection, attrition, expectations, noncompliance, and additional treatments. Qualitative researchers do not have to use statistical methods to take account of the random effects of selection differences when treatment effects are large compared to the variability among units. Conversely, quantitative researchers do not have to use qualitative methods to understand and cope with attrition, expectations, noncompliance, and additional treatments when the effects of these threats to validity are small. But evaluators will not always be so lucky. The more evaluators appreciate both the strengths and weaknesses of the full panoply of methods, the better they will be able to craft the comparisons that are best for estimating treatment effects under different (and often difficult) research circumstances.

Decleration of Conflicting Interests
The author(s) declared no potential conflicts of interest with respect to the authorship and/or publication of this article.

Funding
The author(s) received no financial support for the research and/or authorship of this article.

Notes
1. To make the definition of an effect most relevant to evaluation, I present the definition in the form of “treatments” and use language which turns causes into treatments that can be “implemented.” But the definition of an effect need not be restricted in that way. A more general definition, which allows causes to be non-manipulable if that is desired, is the following. The effect of condition X compared to condition Y is the difference between (a) what would have happened at time 2 if condition X had been present at time 1, and (b) what would have happened at time 2 if condition Y had been present at time 1 instead of condition X, but (c) everything else at time 1 had been the same.
2. Recipients are usually persons but could be animals or groups of people or groups of animals. Recipients are the entities that receive the treatment and whose behavior or attitudes are measured as outcomes.
3. When recipients are “attached” to a specific setting (such as when students are placed in classrooms), assigning setting to treatments is equivalent to assigning recipients to treatments. But when recipients are free to move from one setting to another, such as drivers being free to drive through different intersections, assigning setting to treatments is distinguishable from assigning recipients to treatments, hence the reason for distinguishing between comparisons between recipients and comparisons between settings.
4. Note that Scriven is mistaken when he implies, at the end of the quotation, that the design lacks a matched control group because the classroom of students serves as its own matched control group.
5. However, it would not be unethical to encourage people at random to give up smoking, to stop taking illegal drugs, or to exercise. In this way, researchers could ethically use RA to estimate the effects of smoking, illegal drugs, and exercise.

References


