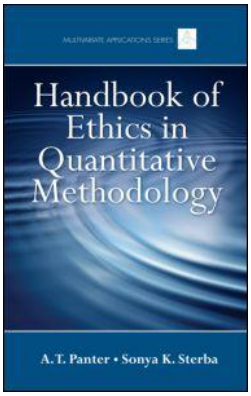

Handbook of Ethics in Quantitative Methodology

A.T. Panter, Sonya K. Sterba

Ethics and the Conduct of Randomized Experiments and Quasi-Experiments in Field Settings

Publication details
https://www.routledgehandbooks.com/doi/10.4324/9780203840023.ch7
Melvin M. Mark, Aurora L. Lenz-Watson
Published online on: 20 Jan 2011

7

Ethics and the Conduct of Randomized Experiments and Quasi-Experiments in Field Settings

Melvin M. Mark
The Pennsylvania State University

Aurora L. Lenz-Watson
The Pennsylvania State University

In 1995, the Administration on Children, Youth and Families (ACYF) implemented the Early Head Start program at sites across the United States. Essentially a younger sibling of the long-standing Head Start program, Early Head Start was created with the primary goal of enhancing the health and development of younger children from low-income families through the provision of services to low-income families with pregnant women, infants, and toddlers, and through the training of service deliverers. One aspect of the rollout of Early Head Start was an experimental evaluation of its implementation and effectiveness. Eligible families and children from 17 communities were randomly assigned either to participate or not participate in the local Early Head Start offerings. This research design allowed the researchers to estimate the effect of participating in the program. Findings from this Early Head Start evaluation were generally positive. In particular, children who were assigned to participate in Early Head Start had higher levels of cognitive and social–emotional development and displayed a larger vocabulary than their comparison group peers (Mathematica Policy Research, 2002).

From one perspective, the Early Head Start evaluation can be viewed as a clear social good, in that it provides information that has the potential to enlighten important democratic deliberations. A study of this sort provides strong evidence of program impact (or lack thereof). When programs are shown to have positive outcomes, this evidence can be used in support of efforts to continue and expand the program. For example, evidence from a study like the Early Head Start evaluation can be cited in

legislative debates about program funding. In contrast, if the program is found to be ineffective or harmful, decision-making processes are again informed, presumably pointing to the need to revise the intervention or find other solutions. Whatever the results, useful information is injected into deliberative processes. However, from another vantage point, serious ethical questions can be raised. To test the effectiveness of Early Head Start, by design some children were randomly assigned to receive no Early Head Start services, and the findings indicate that these children in the comparison group were disadvantaged relative to the Early Head Start participants in terms of cognitive and social development and vocabulary. Is it ethical that such differences in children's performance—which could have consequences for longer-term developmental trajectories—arise as a direct result of a research study? Do the potential benefits of the study offset the withholding of potentially beneficial services at random?

In this chapter, we examine such ethical considerations as they arise with *randomized experiments* and *quasi-experiments* in field settings. Because randomized experiments often receive more criticism on ethical grounds, we address these studies more than their quasi-experiments cousins. In the next section, we define randomized experiments and quasi-experiments and examine the rationale for their use. This rationale is important because it is related to the argument that advocates of experiments provide in response to the most common ethical criticism. In a subsequent section, we discuss randomized experiments and quasi-experiments in field settings in terms of ethical considerations, reviewing both an ethical argument for and ethical arguments against such studies. We address ethical challenges in part by considering how contemporary methodological developments and practices can ameliorate ethical concerns that have been raised. Finally, we explore two issues related to ethics that we believe deserve future attention. By way of preview, a theme that emerges is that methodological quality is not simply a technical consideration, but rather can have important implications for ethics. Throughout the chapter, we return to the Early Head Start example and occasionally refer to other evaluations of social programs. However, the discussion applies to other applications of experiments and quasi-experiments in field settings as well.

Randomized Experiments and Quasi-Experiments: A Primer

Randomized experiments and quasi-experiments are tools that can help answer a particular type of causal question (see Pearl, Chapter 15, this volume, for further discussion on establishing causality). In particular, they are relevant when one wants to know whether and to what extent

a potential causal variable makes a difference in one or more outcomes of interest. For example, in the Early Head Start example, policy makers and others were interested in whether participation in Early Head Start (a potential causal variable) makes a difference in children's school readiness and other specific measures (the outcomes of interest).

In randomized experiments, more than one "treatment" is administered. Put differently, individuals are in different "conditions" of the experiment. Sometimes the experiment compares one named treatment (e.g., Early Head Start) with a control or comparison group that receives no explicit treatment, or perhaps a placebo, or "treatment as usual" (i.e., whatever happens naturally in the absence of individuals being assigned to the named treatment). In other studies, multiple named treatments are compared (e.g., participants could be assigned to either Early Head Start or to a package of 15 home visits by a social worker). Historically, individuals are assigned to conditions, but assignment can instead take place with other units, such as classrooms, workgroups, or even communities. Which condition a given unit is in is determined by random process, such as the flip of a fair coin, the use of a random number table, or a computer program's random generator. In the case of Early Head Start, prekindergarten children were randomly assigned either to a condition in which they were enrolled in a Early Head Start program or to a treatment-as-usual comparison condition in which they were not enrolled in a Early Head Start but instead received whatever care their family provided or arranged.

Quasi-experiments are similar to randomized experiments in the sense that they compare how different treatment conditions perform on the outcome(s) of interest. Quasi-experiments differ from randomized experiments, however, in that they do not involve the random assignment of experimental units to treatment conditions. Instead, they incorporate various types of comparisons across conditions, across time, and perhaps across different kinds of outcomes and participants. Quasi-experiments often also incorporate statistical adjustments intended to correct for biases that can result from the absence of random assignment. Quasi-experiments are an option when either ethical concerns or pragmatic reasons prevent randomized assignment, but the question of a treatment's effect on outcomes is of interest (Cook & Campbell, 1979).

As noted previously, the Early Head Start evaluation involved a randomized experiment. The key benefit of a randomized experiment, using the language of Campbell and his colleagues (Campbell & Stanley, 1966; Cook & Campbell, 1979; Shadish, Cook, & Campbell, 2002), is that it minimizes *internal validity threats*. Internal validity threats are factors other than the treatment variable of interest that could influence the outcome(s). If the Early Head Start study had not used random assignment, for example, perhaps different kinds of families would have enrolled their children

in the program, relative to those families that did not. For example, the families that entered their children into Early Head Start might have been more interested in education, or they might have been more committed to their children's development in general. Or perhaps these families were more likely to have a working parent (and thereby also better off financially), or they were better connected socially and so more likely to be aware of opportunities such as Early Head Start, or less likely to be facing life challenges that hinder their ability to get the child to the Early Head Start Center. Without random assignment these or other factors might affect which children go into which condition. Moreover, the same factors might influence the outcomes measured in the study. For example, families with a greater interest in education might tend to enroll their children in Early Head Start, and the family's interest in education might also lead to greater school readiness (apart from any effect of the program). This would be an example of the internal validity threat of *selection*. Selection occurs when a preexisting difference between the treatment groups affects the outcome variable such that the true treatment effect is obscured. Random assignment renders selection statistically implausible. If children are randomly assigned to conditions, the statistical expectation is that no preexisting factors will be systematically related to condition. Only random differences between the groups are expected, and this possibility is addressed through traditional hypothesis testing statistics (Boruch, 1997).

The strengths of the randomized experiment can be contrasted with the potential strengths and weaknesses of quasi-experiments. In fact, there are numerous quasi-experimental designs, ranging from a few that are close to the randomized experiment in terms of internal validity, to ones far more susceptible to internal validity threats. We will describe a strong quasi-experimental design later, as a potential alternative to randomized experiments that may satisfy certain ethical objections. Here we review one quasi-experimental design that is typically relatively weak in terms of internal validity, the "one-group, pretest–posttest design." In this quasi-experiment design, participants are measured on the outcome of interest both before and after receipt of the treatment. Pretest and posttest scores are compared in an effort to determine the effectiveness of the treatment. For example, if children scored better after Early Head Start than before, one might be tempted to conclude that the program was effective. However, several internal validity threats other than selection would commonly apply (Campbell & Stanley, 1966; Cook & Campbell, 1979; Shadish et al., 2002). In this hypothetical Early Head Start quasi-experiment, the threat of *maturation* would almost certainly apply. Maturation operates when the true treatment effect is obscured because of naturally occurring changes that occur in participants over time. Because children normally improve in terms of social and academic development between, say, the age of 1 and 3 years, seeing improvement from pretest to posttest would not necessarily

suggest that Early Head Start is effective. Because maturation and other internal validity threats (including history, testing, instrumentation, and statistical regression) are frequently plausible when a one-group, pretest–posttest design is used, it is generally not a good choice for research field settings (although exceptions exist, as noted later).

Randomized Experiments and Quasi-Experiments in Field Settings: Key Ethical Issues

Many of the ethical considerations that apply to randomized experiments and quasi-experiments in field settings are common to other forms of social research. Not surprisingly then, thoughtful discussions of ethical guidelines for experimental methods (e.g., Boruch, 1997; Shadish et al., 2002) typically draw in part on general statements about research ethics, such as the *Belmont Report* (Department of Health, Education, and Welfare, 1978). The *Belmont Report* emphasizes three principles for the ethical conduct of research: beneficence, respect for participants, and justice. In practice, these three principles translate into relatively familiar practices in research ethics. In general, prospective participants should voluntarily provide informed consent before participation, where consent includes clear information about the nature of the study and its risks and benefits; fair remuneration for participation can be given but care is needed to avoid having incentives become coercive; potential risks to participants should be minimized; efforts should be taken to maximize the study's benefits for the participant; and more generally, participants' privacy should be respected, typically including confidentiality for any information gathered, especially sensitive information. Because these topics are discussed in some detail in other chapters of this volume (e.g., Gardenier, Chapter 2; Leviton, Chapter 9), we focus here on topics that apply primarily to randomized experiments and their quasi-experimental cousins.

An Ethical Argument for Randomized Experiments

Notably, a general argument has been put forward that ethical considerations support the conduct of randomized experiments in applied research. In sum, the argument is that (a) there is a compelling need to know about the effectiveness of various treatments, and (b) the randomized experiment is especially beneficial in addressing this need. The

presumed benefit of randomized experiments is typically framed in terms of their providing the most trustworthy information about the effects of treatments, but sometimes comes in the form of a belief that findings from randomized experiments may be more influential on subsequent actions. We focus initially on the more common form of the argument, that the need for information about treatment effectiveness is best met by randomized experiments.

The basic argument was articulated decades ago by early advocates of the use of randomized trials in medicine (e.g., Gilbert, McPeak, & Mosteller, 1977). Regarding the first part of this argument, about the need for treatment effect information, it seems clear that there is a need to know whether a new treatment for lung cancer is effective relative to current best practices, or whether stents or bypass surgery are more effective for a particular type of cardiovascular blockage. Without good evidence, uncertainty prevails about the best course of action. Or, even worse, historical happenstance or persuasive advocacy by an ostensible expert, combined perhaps with anecdotal evidence, can result in a particular treatment being widely used—even though it may be ineffective or even harmful.

This need to know about effective interventions is not limited to medicine (Henry, 2009). For example, Gersten and Hitchcock (2009, p. 82) summarize an argument for randomized experiments in education "so that professional educators would have an evidence base for making decisions about selecting curricula, intervention programs for students with learning problems, selecting professional development approaches and determining effective structures for establishing tutoring programs." In the criminal justice domain, Farrington and Welsh (2005) indicate "there is a moral imperative for randomized experiments in crime and justice because of our professional obligation to provide valid answers to questions about the effectiveness of interventions" (p. 31). Similar assertions can and have been made about many other areas of practice that are studied by evaluators and applied social researchers.

A second general claim underlying the ethical argument for randomized experiments is that this method has value relative to other ways of addressing the causal question. The most common form of this argument, made by Gilbert et al. (1977) and others, involves the assertion that randomized experiments are the preferred method for obtaining an unbiased estimate of the effect of a treatment of interest. The argument that randomized experiments are needed to get the right answer typically draws on the notion that internal validity threats are more likely to apply to findings from other methods, as discussed previously (e.g., Gilbert et al., 1977; Cook, 2002). Sometimes this argument for randomized experiments includes an empirical component, showing that randomized experiments in certain areas provide different results than those obtained from quasi-experimental or other types of studies (e.g., Boruch, 1997; Cook, 2002; Mark

& Reichardt, 2009). In addition, advocates of randomized experiments may highlight the cost of incorrect conclusions, which are presumably more likely with other methods. For example, inaccurate findings can result in an ineffective (or even harmful) intervention being adopted widely, as well as in opportunity costs in terms of other potentially effective interventions not being considered; alternatively, inaccurate findings can lead to an effective program being dropped or needlessly redesigned (Bickman & Reich, 2009).

As noted previously, there is a variant on the second portion of the ethical argument presented by Gilbert et al. (1977) and others for randomized experiments. Rather than (or in addition to) arguing for the greater validity of randomized experiments, some advocates of experiments claim that findings from this method will have greater capacity to affect subsequent action. In the case of Early Head Start, the use of a randomized experiment was mandated by Congress, at the very least suggesting that legislative attention to the study findings would be lessened if an alternative method were used. More generally, it can be argued that randomized experiments add value in the sense that they are more likely to be taken seriously in policy deliberations (Henry, 2009) or in motivating practitioners to change their behavior (Gersten & Hitchcock, 2009).

In short, arguing from ethical considerations, a case can be made that it is good to carry out experimental studies of the effectiveness of various treatments. For example, consider Bickman and Reich's (2009) claim that weaker methods are more likely to provide the wrong conclusion about whether a social program is effective and that such an error can have serious negative consequences. Might there not then be an ethical mandate for researchers to try to minimize such risks? Or consider the ethical principles from the *Belmont Report*. Both beneficence (the maximization of good outcomes and the minimization of risk) and justice (the fair distribution of risks and benefits) would seem to be reduced if researchers use biased methods that lead to the wrong conclusion about treatment effects. The greater likelihood of harm from weaker methods, as well as the accompanying reduction of beneficence and justice, is greater if one included implications, not simply for study participants, but also future potential clients. For example, imagine that a weaker quasi-experiment found that Early Head Start was ineffective, but this finding resulted from selection bias. Under this scenario, children who could have benefited from the program may be relatively disadvantaged for years to come, reducing the good outcomes and attenuating the benefits that otherwise could have arisen from the study. As this example suggests, a case can be made that ethics supports the use of research methods that will give the best answer about program effectiveness because having such an answer can increase the likelihood of good outcomes, especially for those initially disadvantaged.

Randomized experiments have received more attention than quasi-experiments in terms of ethicality. In part, this is because random assignment is more intrusive. It is because randomized experiments, by definition, determine at random which treatment participants receive—and thus potentially affect important outcomes such as school readiness—that such studies are more of a target for criticism. In contrast, in quasi-experiments the force(s) that determine each participant's condition, such as self-selection or the happenstance of which program is offered in one's community, often seem more natural.

Ethical Criticisms of Randomized Experiments, and Responses to Them

Numerous criticisms have been made regarding the use of randomized experiments and (to a lesser extent) quasi-experiments in field settings. Some of these criticisms are explicitly framed as ethical, whereas others have an implicit ethical component. In this section, we review several ethically laden critiques of randomized experiments, as well as the responses to these critiques. The ethical challenges are organized here in relation to five criteria promulgated by the *Federal Judicial Center* (1981) *Advisory Committee on Experimentation in the Law*. In short, the five criteria are that (a) the proposed study needs to address an important problem; (b) real uncertainty must exist about what the best course of action is; (c) alternatives to an experiment should be expected to be less beneficial than the experiment; (d) it should be plausible that study results will have influence, such as by informing possible changes in policy or practice; and (e) the experiment should respect participants' rights, for example, by not being coercive and maximizing benefits for participants in the experiment. From one vantage, when met in practice these five criteria can be seen as a more detailed elaboration of the ethical argument in support of conducting a randomized experiment in a field setting (Boruch, 2005; Shadish et al., 2002). Here, they help organize the discussion of ethical critiques and responses to them.

The first criterion of the Federal Judicial Center (1981) is that the proposed study addresses an important problem, that is, that the study addresses something in society that needs to be improved. A corresponding form of criticism is that, either in general or in the particular case, the question of treatment effectiveness is not of sufficient interest. Sometimes such criticisms are intertwined with concerns about whose interests are being represented in research, with this concern typically framed in terms of the interests of those who already have power versus those who are more

disadvantaged. Greene (2009, p. 157), for example, claims that "questions about the causal effects of social interventions are characteristically those of policy and decision makers, while other stakeholders have other legitimate and important questions.… This privileging of the interests of the elite in evaluation and research is radically undemocratic" and, one might surmise from Greene's comments, ethically problematic (see also Leviton, Chapter 9, this volume). The alternative position, as suggested previously, is that good answers to the question of program effectiveness may be important for potential program beneficiaries.

The second criterion from the Federal Judicial Center is that there is real uncertainty about the best course of action. For example, if extensive evidence indicates that Early Head Start is beneficial relative to treatment as usual, then it would be unethical to randomly assign children to these two conditions. Of course, after the fact, if positive effects occur it is easy with hindsight to claim that it was known all along that the treatment was effective. This tendency may be even stronger because a treatment probably would not be tested without at least some sense (even if not with sufficient evidence) that it will be beneficial. Without a strong evidence base, however, the uncertainty that exists about a given intervention's effectiveness is likely to be considerably greater than some observers might presume, especially program advocates. Indeed, some reviewers of social interventions suggest that ineffectiveness is the norm (e.g., Rossi, 1987). Reviewing earlier literature on medical interventions, Chalmers (1968, p. 910) is almost poetic in suggesting that uncertainty is commonly warranted in advance of rigorous experimentation: "One only has to review the graveyard of discarded therapies to discover how many patients might have benefited from being randomly assigned to a control group." Thus, when confronted with the criticism that an experiment involved withholding an effective treatment from members of the control condition, it is important to assess whether it was clear in advance that the other condition's treatment was effective (Burtless, 2002).

Another more specific ethical criticism also falls under the umbrella of the Federal Judicial Center's second criterion. That is, in some experiments, a treatment of interest is compared not with best practice but with a placebo or some other treatment thought to be relatively ineffective. For example, a new pain reliever might be compared with a placebo rather than an effective pain reliever already on the market. This choice of a less potent comparison will increase statistical power and increase the likelihood that a significant difference will be observed. However, the results can be misleading as to the relative effectiveness of the new treatment, and there is generally less certainty about the performance of a new treatment relative to best practice than relative to a placebo. Even worse from an ethical perspective, members of the comparison group are denied access to a more effective treatment simply for the purpose of the experiment. Thus,

good ethics often argues for the use of a "best practice" comparison group. In some cases, however, no "best practice" treatment may be known, or the reality may be that any treatment thought to be beneficial would be rarely used, so a "practice as usual" condition can be justified. In the Early Head Start example, perhaps a best practice comparison could have been identified, such as assigning children to a well-funded preschool with a good teacher:child ratio. However, absent the public funding that would occur under Early Head Start, the reality is such that the ostensible best practice option would be available to only a small minority of the disadvantaged families in the study population, if any. Thus, the practice-as-usual condition provides a policy-relevant counterfactual while not denying anyone a potentially effective treatment that they may have selected in the absence of the experiment.

The third criterion from the Federal Judicial Center is that a randomized experiment is expected to be better able to answer the causal question than are alternatives. Much criticism of randomized experiments (and some of more rigorous quasi-experiment) falls under this criterion. This criticism includes claims that alternative methods suffice for assessing the effectiveness of a treatment. Such claims are not always framed in ethical terms, but they imply an ethical criticism, for example, by suggesting that a cost–benefit assessment of a proposed experiment would tilt toward an alternative method. Critics of randomized experiments sometimes point out, quite accurately, that in everyday experience experiments are not required to determine causal impact (Scriven, 2009). For example, no controlled study is needed to learn the effect of touching a red hot electric burner. (On the other hand, one could argue that such examples implicitly involve strong quasi-experimental designs, with a long time series of data with no burning of the hand before touching the red burner. Control observations from past touching of other items perhaps even include the burner when it is not red, with a special comparison observation. That is, it was the hand that touched the burner but not the other hand that was burned, etc.).

Indeed, a case can be made that quasi-experiments, even relatively weak quasi-experiments, provide acceptable evidence about treatment effects *in certain cases*. For example, Eckert (2000) argued that a simple one-group, pretest–posttest design sufficed for evaluating the effectiveness of training programs being carried out by a unit of the World Bank. Eckert considered each of the internal validity threats that can apply to the design and argued that the threats would not plausibly apply to the studies of the training programs in question. For example, for the threat of maturation, Eckert argued that the nature of the outcomes measured was such that it was implausible for naturally occurring shifts in knowledge to occur in the short time between pretest and posttest. In contrast, for many, if not most, of the issues that might be addressed by field experiments, it will not

be so easy to rule out internal validity threats in advance. To the contrary, in many contexts, such threats are plausible, which is why, unlike with the electric burner, experimental procedures are often needed. Similarly, observation or self-report can sometimes provide accurate information about a treatment's effect. In general, however, the plausibility is substantial that internal validity threats will affect such methods, at least for the kind of treatment effects that social researchers and evaluators are likely to be called on to assess.

The standard way of stating this need for randomized experiments or strong quasi-experiments is that these methods are most needed when internal validity threats are plausible. Perhaps it is useful also to try to specify when such validity threats are likely to be plausible. Put differently, under what conditions are randomized experiments and their closest quasi-experimental approximations most likely to be needed (i.e., where will alternative methods least suffice)? Beyond the obvious requirement that one is interested in the effect of a treatment on identified outcomes, there are conditions under which experiments, rather than alternatives, are likely to be most useful (Mark, 2009). First, experimental methods will be relatively more useful when people are interested in being able to detect small or modest effects. If the only effects of interest are huge, other methods may suffice. For example, if people would support Early Head Start only if it resulted in children at age 3 performing at a fifth grade level in reading and math, simpler methods would probably suffice. When the effect of interest is so big, it could probably be detected with simpler methods than a randomized experiment—and potential internal validity threats would not be plausible for such a huge expected increase in achievement (even though they might create some bias in the estimate of the precise size of the treatment effect). In contrast, when people are interested in small effects, techniques such as random assignment are needed. Given a potentially small treatment effect, plausible validity threats would not only create bias in the estimate of the size of the treatment effect, but also could lead to completely misleading findings about whether a positive treatment effect exists at all. Second, experimental methods will be more useful when the causal field is complex, that is, when (a) multiple factors affect the outcome of interest, (b) the outcome may change over time as a result of the effects of factors other than the treatment, and (c) people naturally vary on the outcome—all of which are close to standard circumstances for the kinds of phenomena examined in randomized field trials. For example, the multiplicity of factors that can affect outcomes such as vocabulary size and the other outcomes measured in the Early Head Start evaluation, along with the routine nature of changeover time and of individual differences, especially when combined with interest in effects that are not extremely large relative to existing variation, argues against claims (e.g., by Scriven, 2009) that treatment

effects can be observed directly and without methods such as the randomized experiment.

A potentially more compelling critique of the relative value of randomized experiments is based not on internal validity considerations but rather on *external validity*. Generally speaking, external validity refers to the accuracy of attempts to apply the findings of a study to persons, settings, or times other than those examined in the study. One form of this general criticism is that the conditions which allow for random assignment may be atypical, making attempts at generalization dubious (Cook & Campbell, 1979). For example, perhaps the communities that are willing to participate in a randomized experiment of Early Head Start differ systematically from most other communities and in ways that would lead to a different treatment effect than elsewhere.

A related criticism is that the experiment enables a focus on the average effect size (i.e., the treatment effect averaged across all participants), even though the relevant processes may be contingent on the specific characteristics of the individual person, the context, and the vagaries of treatment implementation (e.g., Greene, 2009). That is, randomized experiments at least need not open the "black box" to examine the process by which the treatment has its effects. For example, an article in *The Economist* (2008) notes:

> A randomized trial can prove that a remedy works, without necessarily showing why. It may not do much to illuminate the mechanism between the lever the experimenters pull and the results they measure. This makes it harder to predict how other people would respond to the remedy or how the same people would respond to an alternative. And even if the trial works on average, that does not mean it will work for any particular individual. (*The Economist*, 2008, p. 2)

Again, even when such criticisms are not framed explicitly in terms of ethics, they have an ethical dimension; for if the findings of a randomized experiment are not valuable in terms of guiding future action, then the rationale for their conduct is diminished.

There are several ways to respond to these criticisms of randomized experiments. One is to recall, as the Federal Judicial Center's criteria made explicit, that the focus should be on the *relative* ability of randomized experiments and of alternative methods to provide useful information. Thus, in assessing the appropriateness of a randomized experiment and an alternative method, it would be necessary to argue that the combined internal and external validity of the alternative equals or surpasses that of the experiment. Notably, many alternatives, such as case studies, may have merits, but these merits are not such that case studies better facilitate generalization to other sites. A second response is to review the

representativeness or the diversity of the cases within an experiment as a way of arguing that the study's findings should inform action elsewhere. For example, if faced with external validity criticisms in the case of Early Head Start, the researchers could point to the geographical and other forms of diversity across the participating sites, perhaps examining statistically the extent to which the participating children and their families are similar to potentially eligible participants nationwide. A third response would involve going beyond the bare-bones randomized experiment by (a) testing for possible moderated effects (i.e., interactions of key characteristics with the treatment variable), (b) conducting meditational tests of possible mechanisms by which the treatment effect would occur, and/or (c) more generally, using multiple and mixed methods to complement the strengths and weaknesses of the randomized experiment.

The fourth criterion specified by the Federal Judicial Center (1981) for use of a randomized experiment is that study results should, or at least plausibly could, have influence, such as by informing possible changes in policy or practice. It appears that criticisms related to this criterion are for the most part based on another of the five criteria. For example, concern about generalizability of findings, just discussed, can contribute to an argument that the finding of, say, an Early Head Start evaluation could not fruitfully inform decisions by a prospective program site or a specific family about whether to enroll their child (*The Economist*, 2008). Beyond such concerns, it is notable that the literature on the use of research findings, although demonstrating that use occurs, does not give great confidence about a prospective prediction of the use of any particular study of treatment effects (e.g., Nutley, Walter, & Davies, 2007). Thus, we think the right threshold involves there being a reasonable possibility that the study will be influential.

The fifth requirement from the Federal Judicial Center (1981) is that a prospective experiment should respect participants' rights, for example, by not being coercive and by maximizing benefits for participants in the experiment. This requirement involves many considerations, such as informed consent, that are far from unique to field experiment and so are given little attention here. One notable consideration is highlighted by this criterion's emphasis on study *participants*: At least in some cases the risks of a study are borne by study participants, whereas the benefits may accrue largely to others after the study is over. For example, assume that the Early Head Start study shows the benefits of that treatment for eligible children and also leads to increases in funding for the program. Future generations would benefit from the study, as would have children in the Early Head Start condition, but all this would do little for the children assigned to the treatment-as-usual comparison group. Gilbert et al. (1977) suggested that the benefit to future individuals matters greatly in assessing the ethical (and pragmatic) argument for randomized trials,

alluding to the debt that a current generation inevitably owes to past ones. Nevertheless, the distribution of risks and benefits to study participants, including in a condition that proves less desirable in terms of outcomes, is an important issue.

There are several ways of increasing benefits to study participants, even if they are assigned to what proves to be the less effective condition (e.g., Shadish et al., 2002). These include comparing a new treatment option with the best available treatment rather than a no-treatment comparison; offering the more effective treatment to those in the other group after the study is over, when this is practical; providing participants with benefits that leave them better off than they would have been without participating in the study (e.g., payment for participation; health services in a study of job training), even if these are unrelated to the primary outcome variable. In addition, in assessing the benefit:risk ratio for study participants, it seems appropriate to consider what treatment opportunities would have been present if the study had not been conducted. For example, the Early Head Start study created preschool opportunities for those in the treatment group that would not have existed in the absence of the experiment, and the study did not deny access to any opportunities that comparison group members could avail themselves of. In the next section's discussion of methodological quality as an ethical consideration, we return to possible techniques for minimizing risk and increasing benefits for study participants.

Research Quality as an Ethical Issue

Methodological quality is usually seen as a technical consideration, the subject of methods courses and technical critiques in journals and conferences, but not as an ethical matter. Rosenthal concluded that there are ethical implications of methodological quality, referring to psychological research generally, "Bad science makes for bad ethics" (1994, p. 128). When research is designed to address the impact of social and educational programs, such as Early Head Start, which have the opportunity to change children's lifelong behavioral, emotional, and occupational trajectories, an even stronger argument can be made that methodological quality has an ethical import. This was implicit in the earlier discussion of the ethics of randomized experiments. Bickman and Reich (2009) point out that getting the wrong answer about a treatment's outcomes can change a study's risks and benefits dramatically. In the context of program evaluation, Mark, Eyssell, and Campbell (1999) argued that if wrong answers could result in harm to subsequent program participants and if methodological limits can increase the risk of getting the wrong answer, then methodological

shortcomings appear to be an ethical concern. Indeed, one can make a plausible argument that the ethical implications of methodological quality are greater for research with applied implications (see Cizak & Rosenberg, Chapter 8, this volume, and Leviton, Chapter 9, this volume, on the topic of high-stakes applied settings). For example, if a methodologically flawed evaluation is used, serious costs may arise for real people; in contrast, in more traditional basic arenas such as mainstream cognitive psychology, the self-corrective mechanisms within scholarly communities are likely to correct erroneous conclusions over time.

Implications of Methodological Advances for Meeting Ethical Challenges

A corollary of the position that methodological quality is an ethical matter is worth considering: Methodological advancements may be able to attenuate ethical challenges to the conduct of randomized experiments or quasi-experiments. Indeed, quality practices—some now familiar and some still in development—may suffice for addressing ethical critiques of randomized experiments noted in the previous section.

One illustrative methodological advance concerns the now-widespread use of power analysis to identify the minimum number of persons needed to test study hypotheses, without unnecessarily exposing extra persons to a potentially harmful treatment. Consider the risk to participants in, and the potential benefits of, the Early Head Start intervention. What if 1,000 children were assigned to the comparison group, even though the study would have been able to detect a treatment effect with only 200 per condition. This would expose far more children than needed to whatever risk exists. Conversely, imagine that only 200 children were assigned per condition, even though the study would not have reasonable statistical power to detect a meaningful treatment effect without 1,000 per condition? In this case, the study would not have a reasonable chance of providing the benefit of detecting program effects and contributing to better decision making about the program. The widespread application of power analysis attenuates these problems. Power analyses can estimate the number of participants needed to observe a treatment effect of a given size. Consequently, the number of participants used can be selected in a way that allows for the benefit of meaningful findings while minimizing the likelihood that too many participants are needlessly exposed to any risk the study might have (Maxwell & Kelley, Chapter 6, this volume).

Another methodological advance that can help minimize risk in an experiment is the "stop rule" (called *adaptive sample size planning* in

Maxwell & Kelley, Chapter 6, this volume). With a stop rule, analyses are conducted at various planned points (typically with an adjustment for the error rate from conducting multiple tests). If, say, a significant treatment effect is observed, the experiment is halted. Otherwise, it continues. Stop rules are especially likely to be used in an experiment in which participants enter over time, such as an evaluation of a surgical procedure, rather than in a study in which all participants take place at the same time, such as with a new curriculum that is implemented in randomly chosen classrooms in a given school year. Even with studies in which all participants enter at the same time, a *stop rule* can be implemented if the outcome is measured repeatedly over time. That is, the stop rule could end the study as soon as a significant effect (or effect of a prespecified size) is observed, even if additional measurement waves had been planned. In a study with stop rules, it may be possible to add a delayed treatment component, whereby participants in the less effective condition receive the more effective treatment after the original study is halted. In the case of Early Head Start, these design ancillaries would have involved (a) conducting analyses at multiple points in time and, if the significant positive effects of Early Head Start were observed midway through the study, then (b) enrolling the practice-as-usual comparison group children in Early Head Start. This approach is not always feasible (e.g., a key outcome variable might not be reliable until the children are at least 3 years old, or there may not be Early Head Start spaces available for the comparison group children). When feasible, however, stop rules minimize risk for participants and, when used in conjunction with a delayed treatment feature, allow participants who had been in the less beneficial conditions group to receive the treatment method that their participation helped demonstrate is effective.

Adaptive randomization is another methodological advance that holds promise for reducing the ethical concerns about assigning people at random to the less effective condition in an experiment (Hu & Rosenberger, 2006). Adaptive randomization begins by assigning equal numbers of participants to each condition. However, in an adaptive randomization scheme, interim analyses are used to adjust the probability of assignment to each condition based on the apparent effectiveness to that point. For example, if the treatment group has interim outcomes that are 1.5 times as good as the outcomes in the treatment-as-usual comparison group, then the assignment probabilities would be adjusted so that 1.5 times as many participants would be assigned to the treatment group as the study continues. At this explanation suggests, adaptive randomization applies when participants enter over time. To date, it appears that adaptive randomization has been used primarily in early-phase medical trials (Hu & Rosenberger, 2006), but the technique may find its way into the toolkit of

applied social researchers for certain kinds of field experiments, such as in legal settings where cases tend to trickle in over time.

Faced with real or potential criticism about the use of a randomized experiment, another general approach is to consider methodological advances that do not involve random assignment to condition but hold promise for giving an unbiased estimate to the treatment effect. One strong alternative of this sort is the *regression–discontinuity design* (Imbens & Lemieux, 2008; Shadish et al, 2002). In this quasi-experimental design, a set of study participants is measured on a "quantitative assignment variable" (QAV), and all individuals on one side of a cutoff score on the QAV are assigned to one condition, whereas those on the other side are assigned to the other condition. In this way, a treatment can be assigned based on need, circumstances, or merit, rather than at random. For example, in the case of Early Head Start, researchers might start by identifying children and families willing to participate in the study. Then researchers might measure the QAV, such as a measure of the children's initial cognitive development (alternatively, the QAV could be a measure of families' adjusted income, or a composite based on several indicators). A cutoff score would be established, and children with scores below the cutoff would be assigned to Early Head Start, with children scoring above the cutoff assigned to the treatment-as-usual comparison condition. In essence, a treatment effect is observed when the outcome scores of the children in the treatment group are higher than would be expected based on the trend of scores in the comparison group. Put differently, if there is a discontinuity in the regression line (with the QAV predicting the outcome) at the cutoff, the only plausible explanation in most cases is that the program made a difference. The regression–discontinuity design escapes much of the ethical criticism of the randomized experiment because it assigns the treatment to those with greater need (or in some cases, greater merit). In the past, the design has been used rarely. However, the regression–discontinuity design has received increased attention, including by economists who are advancing statistical design and validity checks (e.g., Hann, Todd, & Van der Klaauw, 2001; Imbens & Lemieux, 2008). Thus, the design may become a more common alternative to randomized experiments.

Another design-based methodological advance relies on random assignment but does not involve direct assignment into treatment conditions. Rather, in the random encouragement design, participants are assigned at random either to receive or not receive extensive recruitment efforts encouraging them to participate in the treatment of interest. Although not yet implemented in enough field studies to be confident about its practicality, the design appears to hold promise in avoiding or minimizing ethical criticisms about withholding potentially beneficial treatments in a randomized experiment. The *random encouragement design* was implemented by Wells et al. (2000), who used multiple forms of encouragement,

including education and reduced fees, to solicit patients at a randomly assigned set of clinics to participate in a quality improvement program for the treatment of depression. Statistical methods such as instrumental variables analysis are then used in an effort to provide good estimates of the treatment effect (see Schoenbaum et al., 2002, for an illustration of instrumental variables analysis with the Wells et al. data). In essence, these analyses assume that any effect of encouragement arises only by way of the increased program participation that the encouragement creates; this assumption facilitates statistical estimates of the effect of the program itself. The random encouragement design alleviates ethical concerns that can arise from procedures that restrict participants' access to multiple treatment options. It can reduce any concerns about coercion. These ethical benefits *may* occur with little loss of validity, although further experience with the design is needed.

In short, a set of methodological practices, including recent developments, offers promise for ameliorating some ethical concerns about randomized experiments and their quasi-experimental cousins. Power analysis, adaptive randomization, and use of the "stop rule" reduce unnecessary exposure to a potentially harmful treatment. When the "stop rule" is used with a delayed treatment component, participants in less effective conditions receive treatment that was shown to be effective. Regression–discontinuity design assigns participants who are most in need (or most deserving) of treatment to be assigned to what is hypothesized to be the more effective treatment condition. The random encouragement design reduces ethical concern about the study constraining participants' ability to make choices among treatment options. Thus, the general theme that methodological quality has ethical implications can be expanded. Methodological advances can serve to resolve ethical criticisms, if they satisfactorily address the underlying ethical problem.

Three Topics That Warrant Future Attention in Applied Research Studies and Program Evaluations

In this section, we highlight three issues that appear to deserve future attention. For the first of these, attention is needed from methodologists and statisticians. For the second, consideration is required from those involved in the design of applied research and evaluation studies, especially those involved in the selection of measures, as well as measurement specialists. For the third issue noted in this section, a variety of parties could contribute, including scholars conducting empirical research on applied research and evaluation itself, group process researchers, and the

broader community of those interested in how decisions are to be made about the ethicality of proposed research.

Moving Further Beyond Average Effect Sizes

As noted previously, randomized experiments and quasi-experiments can be criticized for their focus on average treatment effects, which may not provide adequate guidance for action if the effects of the treatment are moderated by undetected interactions (Greene, 2009; *The Economist*, 2008). In the face of such interactions, a program might benefit some participants but have no effect or even be harmful for others. The ethical concern seems obvious: If there is a harmful effect for a subset of participants and if the experiment fails to detect this, the harmful effects will not be ameliorated; even worse, the study results could lead to the program being administered universally in the future despite its harmful effects to some participants. Even without a harmful effect, if the treatment is ineffective for some participants, the failure to detect the differential effects could have serious opportunity costs by keeping the relevant subgroup from obtaining an alternative treatment that might be beneficial for them.

One proposed but not fully developed response is to conduct "*principled discovery*" (Mark, 2003; Mark, Henry, & Julnes, 2000). Principled discovery holds potential for addressing one form of ethical criticism of randomized experiment and quasi-experiments. More generally, it could help increase the ability of studies to guide future action, in the face of moderated relationships that limit the guidance that can be taken from average treatment effects. The basic idea of subsequent principled discovery is to engage in two phases (possibly with further iteration between the two). One would begin analyses, before principled discovery, by conducting the planned analyses to test an a priori hypothesis (e.g., that Early Head Start will lead to better outcomes than treatment as usual). In the first phase of principled discovery, the researcher would then carry out exploratory analyses. For example, the Early Head Start evaluator might examine whether the program has differential effects by looking for interaction effects using one after another of the variables on which participants have been measured (e.g., gender, race, age, family composition, etc.). A wide variety of statistical techniques can be used for the exploratory analyses of this first phase of principled discovery (Julnes & Mark, 1998; Mark, 2003). The exploration of phase 1 is not without risks, however, especially the possibility of being misled by chance. Statistical significance of course simply means that a given finding is unlikely to have arisen by chance if there really were no difference. But the conduct of many exploratory tests creates a

risk that some finding will be significant because of chance. Stigler's (1987, p. 148) admonition is apt: "Beware of testing too many hypotheses; the more you torture the data, the more likely they are to confess, but confession obtained under duress may not be admissible in the court of scientific opinion" (see also Maxwell & Kelley, Chapter 6, this volume).

If the exploratory analyses of phase 1 result in an interesting discovery, the classic admonition is to try to replicate the discovery in another study. However, this will often be infeasible in the case of field studies such as program evaluations, where any use of the study findings is likely to occur before replication is possible. Thus, the second phase of principled discovery would be called for, in which the researcher seeks one or another form of independent (or quasi-independent) confirmation of the discovery. In many instances, this will involve other tests that can be carried out within the same data set (although data might be drawn from other data sets, or new data might be collected after phase 1). For example, if an interaction were observed such that Early Head Start has a smaller effect for children in families with relatively less parental education, this could lead to another prediction that a similar and probably stronger interaction will be obtained with a composite variable (drawn from home visits) based on the amount of children's books and educational material in the children's homes. As this example illustrates, phase 2 of principled discovery will generally require an interpretation of the finding from the phase 1 exploration. This interpretation in turn gives rise to the phase 2 hypothesis. The value of the phase 2 test is that, if the original discovery is not real but instead is only the result of chance, then there is generally no reason to expect the phase 2 test to be confirmed. Future application of the approach, including further investigation of techniques for controlling for error rate in the two phases of principled discovery, seems warranted.

Changes Over Time in Value-Based Outcomes

The outcomes that should be examined in a study are not magically revealed. Moreover, the concerns and values that drive the selection of outcomes are not historically invariant. What people care about in relation to a kind of intervention can change over time, and use of outcomes that do not reflect current values can lead to a waste of participant time and research resources and to a limited potential for the study to make a difference. Of course, in applied research the key outcomes for a study often derive rather directly from the study purpose. For example, for an evaluation of Early Head Start, measures of cognitive development seem to derive naturally from the program and its goals. However, the outcomes

that people care about and their expectations about what programs will achieve are not static. As an example, measures of social development are more common today than in the early days of preschool evaluation.

The selection of outcome variables has ethical implications, even if these are indirect. For example, if sound decision making about a preschool program would require measures of both cognitive and social development, but only cognitive development is assessed, problems can occur. The benefits of the study may be curtailed. At the extreme, a study may lead to the selection of the wrong treatment, for example, if one program has a slight benefit with respect to cognitive outcomes but performs far worse on the (unmeasured) social development outcomes. In this light, the ethical import of the value-based selection of outcomes seems evident.

A contemporary example merits attention. With growing concern about global climate change, it seems possible that environmental impact may gain in importance even for programs and policies that do not have primarily an environmental focus. For example, one can imagine a future in which a program such as Early Head Start would include procedures for assessing the environmental impact of the intervention. This would include attention to such things as power use at program sites and energy use in transportation to program sites, relative to the estimated energy impact of whatever disparate arrangements individual children in the comparison group have. Future work on tracking environmental impact for social and educational programs may be fruitful, as may more general work on social methods for identifying the outcomes that people value for a given kind of program.

Procedures for Making Judgments About the Ethicality of a Proposed Study

Throughout this chapter we have encountered a set of questions (e.g., whether the treatment effect question is important, whether a randomized experiment will provide a better or more influential answer), the answers to which help determine the ethicality of a potential randomized experiment or of its various quasi-experimental cousins. But how are these questions to be answered? Absent a compelling protocol for researchers to judge the ethicality of a proposed study, a standard part of the answer is to rely on *institutional review boards* (IRBs) to provide answers. However, perhaps the most challenging aspect of assessing the ethicality of a randomized experiment involves the question of how to go about trying to answer questions, such as the importance of the treatment effect question and the relative benefits of a randomized experiment, in a particular case.

The inclusion of community members on IRBs in part recognizes that the matters to be judged are not only technical ones. (Regulations require both a member who is not affiliated with the institution and a member who is not a scientist, although in practice it appears these are usually represented in a single community member). However, both the political or values questions (e.g., Is the question of the treatment's effects on possible outcomes sufficiently important to justify the conduct of the study?) and the more technical ones (e.g., Is a random assignment experiment needed, relative to alternative methods?) are addressed by the same group, at least some of whom may not have the requisite skills and/or information for making thoughtful judgments about both kind of questions. Moreover, the IRB typically enters into the process after a study has been fully planned, making adjustments costly and painful.

Future research may be able to inform specific judgments related to the ethicality of future research. For example, studies could assess the extent to which potential program beneficiaries or their proxies (e.g., parents of preschool children), across a range of program types, are interested in obtaining valid answers to the question of program's treatment effect, in contrast to claims of critics such as Greene (2009). Research could also assess the practicality and worth of procedures that might expand on traditional IRB procedures, such as variants on the deliberative polling methods used in recent years by political scientists. More generally, group process researchers could fruitfully apply their expertise in an effort to improve IRB (or complementary) procedures.

Conclusion

This chapter has addressed ethical issues related to the conduct of randomized experiments and quasi-experiments in applied field settings, including program evaluation. The ethical argument for randomized experiments and their strongest quasi-experimental cousins has been reviewed. Ethical criticisms of randomized experiments, and responses to them, have been presented. We have reviewed the potential for existing and emerging methodological advances to ameliorate certain of the ethical challenges to experiments. We have also briefly considered three topics we believe deserve further attention in the future. Although the discussion has been general, in practice ethical judgments are made about specific studies, the details of which matter. Nevertheless, we hope that the presentation of the ethical arguments for and against experiments and the other topics addressed in the chapter will help in framing more thoughtful consideration of the ethics of proposed or actual randomized experiments and quasi-experiments.

References

Bickman, L., & Reich, S. (2009). Randomized control trials: A gold standard with feet of clay. In S. Donaldson, T. C. Christie, & M. M. Mark (Eds.), *What counts as credible evidence in applied research and evaluation practice?* (pp. 51–77). Thousand Oaks, CA: Sage.

Boruch, R. F. (1997). *Randomized experiments for planning and evaluation*. Thousand Oaks, CA: Sage.

Boruch, R. F. (2005). Comments on 'Use of randomization in the evaluation of development effectiveness.' In G. K. Pitman, O. N. Feinstein, & G. K. Ingram (Eds.), *World Bank series on evaluation and development, Vol. 7: Evaluating development effectiveness* (pp. 205–231). New Brunswick, NJ: Transaction.

Burtless, G. (2002). Randomized field trials for policy evaluation: Why not in education? In F. Mosteller & R. F. Boruch (Eds.), *Evidence matters: Randomized trials in education research*. Washington, DC: Brookings Institution Press.

Campbell, D. T., & Stanley, J. C. (1966). *Experimental and quasi-experimental designs for research*. Skokie, IL: Rand McNally.

Ceci, S. J., & Papierno, P. B. (2005). The rhetoric and reality of gap closing: When "have nots" gain but the "haves" gain even more. *American Psychologist, 60,* 149–160.

Chalmers, T. C. (1968). Prophylactic treatment of Wilson's disease. *New England Journal of Medicine, 278,* 910–911.

Cook, T. D. (2002). Randomized experiments in educational policy research: A critical examination of the reasons the educational evaluation community has offered for not doing them. *Educational Evaluation and Policy Analysis, 24,* 175–199.

Cook, T. D., & Campbell, D. T. (1979). *Quasi-experimentation: Design and analysis issues for field settings.* Skokie, IL: Rand McNally.

Department of Health, Education, and Welfare (1978). *The Belmont report: Ethical principles and guidelines for the protection of human subjects of research.* Washington, DC: U.S. Government Printing Office.

Donaldson, S., Christie, T. C., & Mark, M. M. (Eds.). (2009). *What counts as credible evidence in applied research and evaluation practice?* Thousand Oaks, CA: Sage.

Eckert, W. A. (2000). Situational enhancement of design validity: The case of training evaluation at the World Bank Institute. *American Journal of Evaluation, 21,* 185–193.

Farrington, D. P., & Welsh, B. C. (2005). Randomized experiments in criminology: What have we learned in the last two decades? *Journal of Experimental Criminology, 1,* 9–38.

Federal Judicial Center. (1981). *Experimentation in the law: Report of the Federal Judicial Center Advisory Committee on Experimentation in the Law*. Washington, DC: U.S. Government Printing Office.

Gersten, R., & Hitchcock, J. (2009). What is credible evidence in education? The role of the What Works Clearinghouse in informing the process. In S. Donaldson, T. C. Christie, & M. M. Mark (Eds.), *What counts as credible evidence in applied research and evaluation practice?* (pp. 78–95). Thousand Oaks, CA: Sage.

Gilbert, J. P., McPeak, B., & Mosteller, F. (1977). Statistics and ethics in surgery and anesthesia. *Science, 198,* 684–689.

Greene, J. C. (2009). Evidence as "proof" and evidence as "inkling." In S. Donaldson, T. C. Christie, & M. M. Mark (Eds.), *What counts as credible evidence in applied research and evaluation practice?* (pp. 153–167). Thousand Oaks, CA: Sage.

Hann, J., Todd, P., & Van der Klaauw, W. (2001). Identification and estimation of treatment effects with a regression-discontinuity design. *Econometrica, 69,* 200–209.

Henry, G. T. (2009). When getting it right matters: The case for high quality policy and program impact evaluations. In S. Donaldson, T. C. Christie, & M. M. Mark (Eds.), *What counts as credible evidence in applied research and evaluation practice?* (pp. 32–50). Thousand Oaks, CA: Sage.

Hu, F., & Rosenberger, W. F. (2006). The theory of response-adaptive randomization in clinical trials. Hoboken, NJ: Wiley Interscience.

Imbens, G. W., & Lemieux, T. (2008). Regression-discontinuity designs: A guide to practice. *Journal of Econometrics, 142,* 615–635.

Julnes, G. J., & Mark, M. M. (1998). Evaluation as sensemaking: Knowledge construction in a realist world. In G. Henry, G. W. Julnes, & M. M. Mark (Eds.), *Realist evaluation: An emerging theory in support of practice* (pp. 33–52). San Francisco: Jossey Bass.

Mark, M. M. (2003). Program evaluation. In S. A. Schinka & W. Velicer (Eds.), *Comprehensive Handbook of Psychology* (Vol. 2, pp. 323–347). New York: Wiley.

Mark, M. M. (2009). Credible evidence: Changing the terms of the debate. In S. Donaldson, T. C. Christie, & M. M. Mark (Eds.), *What counts as credible evidence in applied research and evaluation practice?* (pp. 214–238). Thousand Oaks, CA: Sage.

Mark, M. M., Eyssell, K. M., & Campbell, B. J. (1999). The ethics of data collection and analysis. In J. L. Fitzpatrick & M. Morris (Eds.), *Ethical issues in program evaluation* (pp. 47–56). San Francisco: Jossey Bass.

Mark, M. M., Henry, G. T., & Julnes, G. (2000). *Evaluation: An integrated framework for understanding, guiding, and improving policies and programs*. San Francisco: Jossey Bass.

Mark, M. M., & Reichardt, C. S. (2009). Quasi-experimentation. In L. Bickman & D. Rog (Eds.), *The Sage handbook of applied social research methods* (2nd ed., pp. 182–213). Thousand Oaks, CA: Sage.

Mathematica Policy Research. (2002). *Early Head Start research: Making a difference in the lives of infants and toddlers and their families: The impacts of Early Head Start*. Available at http://www.mathematica-mpr.com/publications/pdfs/ehsfinalsumm.pdf

Nutley, S. M., Walter, I., & Davies, H. T. O. (2007). *Using evidence: How research can inform public services.* Bristol, UK: Policy Press.

Rosenthal, R. (1994). Science and ethics in conducting, analyzing, and reporting psychological research. *Psychological Science, 5,* 127–134.

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

Schoenbaum, M., Unutzer, J., McCaffrey, D., Duan, N., Sherbourne, C., & Wells, K. B. (2002). The effects of primary care depression treatment on patients' clinical status and employment. *Human Service Research, 37,* 1145–1158.

Scriven, R. (2009). Demythologizing causation and evidence. In S. Donaldson, T. C. Christie, & M. M. Mark (Eds.), *What counts as credible evidence in applied research and evaluation practice?* (pp. 134–152). Thousand Oaks, CA: Sage.

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

Stigler, S. M. (1987). Testing hypotheses or fitting models: Another look at mass extinction. In M. H. Nitecki & A. Hoffman (Eds.), *Neutral models in biology* (pp. 145–149). Oxford, UK: Oxford University Press.

The Economist. (2008, December 30). The bright young thing of economics. Retrieved from http://www.economist.com/finance/displayStory.cfm?story_id=12851150

Wells, K. B., Sherbourne, C., Schoenbaum, M., Duan, N., Merideth, L., Unutzer, J.,… Rubenstein, L. V. (2000). Impact of disseminating quality improvement programs for depression in managed care: A randomized controlled trial. *Journal of the American Medical Association, 238,* 212–220.