Randomized Experiments as the Bronze Standard

Richard A. Berk

CCPR-030-05

August 2005

Randomized Experiments as the Bronze Standard[*]

Richard A. Berk

Department of Statistics
UCLA

August 2, 2005

Abstract

In this paper, the strengths and weakness of randomized field experiments are discussed. Although it seems to be common knowledge that random assignment balances experimental and control groups on all confounders, other features of randomized field experiments are somewhat less appreciated. These include the role of random assignment in statistical inference and representations of the mechanisms by which the treatment has its impact. Randomized experiments also have important limitations and are subject to the fidelity with which they are implemented. In the end, randomized field experiments are still the best way to estimate causal effects, but are a considerable distance from perfection.

1 Introduction

The benefits of randomized field experiments have been widely advertised for generations. Fisher's famous book, *The Design of Experiments*, was published in 1935. Hundreds of textbook treatments followed (e.g., Cox, 1958;

[*] Thanks go to David Weisburd and two reviewers for very helpful comments on an earlier draft of this paper. Thanks also go to the National Science Foundation for their support (SES - 0437169: "Ensemble Methods for Data Analysis in the Behavioral, Social and Economic Sciences").

Box, Hunter and Hunter, 1978). Even textbooks and monographs emphasizing design alternatives to randomized experiments often begin with a tribute to random assignment (Campbell and Stanley, 1963; Rosenbaum, 2002.)

But the message has been selectively received. It is generally recognized that random assignment balances treatment and control groups on potential confounders. Unbiased estimates of causal effects can follow directly. No other research design confers these benefits so reliably. At the same time, other features of randomized experiments are overlooked or misunderstood. As a result, the expectations imposed on randomized experiments are sometimes unrealistic: sometimes unreasonably positive and sometimes unreasonably negative. Complicating matters is that whatever happens to be the methodological tool du jour is too often oversold. The result is that the rhetoric surrounding randomized experiments and its critics tends to inflate as well.

The goal of this paper is to ratchet down the claims and criticisms. A careful look at randomized experiments will make clear that they are not the gold standard. But then, nothing is. And the alternatives are usually worse.

2 Randomized Experiments and Causal Inference

It will prove useful in the following discussion first to review briefly the counter-factual framework in which randomized experiments can be placed and to consider how random assignment solves the ”fundamental problem of causal inference.” The result will be a useful *definition* of a “causal effect,” coupled with a rationale for how causal effects can be estimated from data. The formulation is sometimes called the Neyman-Rubin Model (Neyman, 1923; Rubin, 1986; 1990).

For ease of exposition, consider a binary intervention. There is a *single* individual and a treatment. In this instance, the individual is a juvenile offender, and the treatment is boot camp. The control condition is the usual confinement in a state youth facility.[1]

[1] The control condition is far more than the absence of the treatment. There is real content. More generally, it might be better if the concepts “treatment” and “control” were replaced by the notion of different treatments. The absence of the treatment is not nothing.

Let y be the response, which is whether the individual is arrested for a new offense after release. Rearrest is coded as a 1, and no rearrest is coded as a 0.[2] The observed intervention t equals 1 if the individual is sentenced to boot camp and 0 if not. Finally, and a source of potential confusion, let t^* equal 1 if the *hypothetical* intervention is boot camp and 0 if not. Thus, t^* represents what the intervention could be, not what it is.

Asking what would happen as a result of boot camp (compared with conventional incarceration) is to ask about the hypothetical intervention. It is the causal effect of this hypothetical intervention that we wish to estimate. We are seeking an answer to a "what if" question. What would happen if the juvenile were sent to a boot camp? What would happen if the juvenile were sent to a conventional juvenile facility? And how would those two outcome compare? It is in the comparison that a causal effect will shortly be defined.

Consider now the actual (not hypothetical) intervention, which is what the juvenile actually experiences as either a sentence to boot camp or a sentence a state youth facility. There are four possible pairings between the intervention that was received and the hypothetical intervention. The outcome, conditional on these four pairs, can be represented as follows:

1. $y|(t^* = 1, t = 1)$: the outcome if hypothetically boot camp was the sentence, and it actually was the sentence.

2. $y|(t^* = 1, t = 0)$: the outcome if hypothetically boot camp was the sentence, but it actually was not the sentence.

3. $y|(t^* = 0, t = 1)$: the outcome if hypothetically boot camp was not the sentence, but it actually was the sentence.

4. $y|(t^* = 0, t = 0)$: the outcome if hypothetically boot camp was not the sentence, and it actually was not the sentence.

The point of the pairings is to examine how the hypothetical treatment effects, representing what we want to learn about, map on to what can actually be observed. Only the first and fourth conditional relationships are observable even in principle. The second and third reflect "counterfactuals." One cannot observe the outcome for an individual sentenced to boot camp, for instance, if that individual was not sentenced to boot camp.

[2] Working with a binary outcome helps to keep the exposition simple, and there is no important loss in generality.

Moreover, even for the first and fourth conditional relationships, only one can be observed for a given individual. Either the juvenile was sentenced to boot camp or the juvenile was not. So, for a single individual, one cannot compare the outcome under the boot camp intervention to the outcome under the incarceration intervention. This leads to the "fundamental problem of causal inference." Causal effects are *defined* at the level of a given unit (here, a juvenile appearing in juvenile court). But defined in this manner, casual effects cannot be observed directly.

At the level of a single unit, it is difficult to imagine a compelling solution. For example, what sensible use could be made of a before-after comparison? The condition that came first would likely have an impact on the condition that came second. And other things would be changing as well.

Suppose our juvenile served time at a boot camp, was released, rearrested, and then sentenced to a term in a state youth facility. For the comparison between the two interventions to be valid, the "pre-existing" state of the juvenile before the boot camp experience would need to be same as the "pre-exiting" state of the juvenile before the incarceration experience. At the very least, the juvenile would be older the second time around, and age is a well known correlate of criminal behavior. Moreover, it is likely that the first sentence would affect the impact of the second sentence. If nothing else, the juvenile would be carrying the label of a repeat offender.

The practical solution is move to the level of groups. If two groups of juveniles could be made comparable before any intervention, an average disparity in outcome between the two groups could be attributed to the particular intervention received. Then, a useful estimate of the causal effect of boot camp compared to conventional incarceration might be the difference in the proportions rearrested for the two groups. In other words, for a particular juvenile i, a causal effect is defined as $(y_i|t_i^* = 1) - (y_i|t_i^* = 0)$. What we get to know is $(\bar{y}|t = 1) - (\bar{y}|t = 0)$.[3] Average effects are substituted for individual effects, and real interventions are substituted for hypothetical interventions.

Thus, we do not learn about the response of the given juvenile to the two conditions. But if we select our groups carefully, we can learn about the *average* response to the two conditions. And this is not bad. In general, it is unlikely that every unit being studied would have precisely the same response to an intervention. Therefore, computing an average treatment effect makes considerable scientific sense. Moreover, scientific interest is typically about

[3] If a binary variable is coded as a 1 or a 0, the mean is the proportion.

types of units, not individual units. Therefore, as a scientific matter, knowing the average causal effect for a meaningful representation of juvenile offenders is likely to be far more useful than knowing about the causal effect for a given juvenile, even if that could be determined. So, let's play this through.

Suppose we select a random sample of juveniles arrested and taken to juvenile court. The average response to "what if" sentence boot camp is then

$$\begin{aligned}(\bar{y}|t^* = 1) = (\bar{y}|t^* = 1, t = 1) \times p(t = 1) +& \\ (\bar{y}|t^* = 1, t = 0) \times p(t = 0).&\end{aligned} \tag{1}$$

In equation 2, the proportion rearrested *if* the sentence is boot camp depends on the outcome for those actually sentenced to boot camp and the outcome for those actually sentenced to confinement in a juvenile facility, both weighted by particular proportions. For those sentenced to boot camp, the proportion getting the treatment is used. For those sentenced to conventional incarceration, the proportion incarcerated is used. Note that the second part of the right-hand side is unobservable and as such, is counterfactual.

For the "what if" sentence of incarceration in a conventional youth facility, same kind of expression can be written:

$$\begin{aligned}(\bar{y}|t^* = 0) = (\bar{y}|t^* = 0, t = 1) \times p(t = 1) +& \\ (\bar{y}|t^* = 0, t = 0) \times p(t = 0).&\end{aligned} \tag{2}$$

In equation 3, the proportion rearrested if the sentence is conventional incarceration depends on the outcome for those actually sentenced to incarceration and on the outcome for those actually sentenced to boot camp, both weighted, once again, by appropriate proportions. Now, it is the first piece of the right-hand side that is counterfactual.

If the two counterfactual conditions in equations 2 and 3 could be observed, one could compute the two problematic terms, and $(\bar{y}|t^* = 1) - (\bar{y}|t^* = 0)$ could be estimated. But counterfactuals are by definition unobservable. As a fallback position, one might use the juveniles not sentenced to boot camp to provide an estimate of what would have happened had those sentenced to boot camp instead been incarcerated. Likewise, one could use the juveniles sentenced to boot camp to provide an estimate of what would have happened had the incarcerated juveniles instead been sentenced to boot camp.

The difficulty with that strategy is that the juveniles sent to boot camp would probably differ on the average from those sent to a conventional youth facility. For example, a judge might try to match the needs of each juvenile with the setting in which a constructive experience would be most likely. So, the "less hardened" juveniles might be sent to boot camp. Then, if a smaller proportion of juveniles sent to boot camp were rearrested, one could not determine whether this resulted from the treatment or the proclivities brought to the treatment. Perhaps the boot camp intervention was just getting a mix of better risks.

Random assignment to boot camp or incarceration can solve this problem by making the two groups of juveniles on the average comparable before one of the two interventions is imposed. In other words, random assignment will on the average balance the two groups on all possible confounders. As a result, the estimate $(\bar{y}|t^* = 1) - (\bar{y}|t^* = 0)$ can be unbiased. Indeed, most of the justifications for random assignment rely this reasoning.

But there is a catch. One must meet the assumption of "no interference," or what in fancier language is called the "stable unit treatment value assumption" (SUTVA). This assumption requires that the treatment or control condition to which a unit is assigned has no impact on the response of another unit. For example, whether a given juvenile is assigned to boot camp or incarceration has no impact on the chances of rearrest for another juvenile, whatever the second juvenile's assigned intervention happens to be.

How might SUTVA be violated? Suppose the random assignment happens to place a substantial number of rival gang members in the same boot camp. That could dramatically alter the nature of the treatment and the subsequent response, and perhaps bias the study against finding recidivism reductions for boot camp.

Suppose within a conventional state youth facility, some juveniles are assigned at random to an innovative drug treatment program and some are not. Because the juveniles mingle, the potential coping skills learned by those assigned to the drug treatment program could easily diffuse to those not assigned to the drug treatment program. Estimates of any beneficial treatment effects might then be biased downward; some of the controls got an important piece of the treatment.

A few violations of SUTVA in a study with many subjects is not likely to matter in practice. But the violations can create substantial bias when they are widespread or if given the opportunity to snowball. Thus, group randomized trials are especially vulnerable. In group randomized trials, units are assigned to interventions in clusters defined by convenience. For example, all the individuals within the same youth facility are assigned at random to the same intervention. Because all of the individuals in a given facility are organizationally linked, it is easier to treat them all in the same fashion. But it is precisely because of their connections to one another that that SUTVA is placed at risk.

Building on the last illustration, suppose a drug treatment program was randomly provided to some youth facilities and not others. The potential benefits of the drug treatment program for individuals could be enhanced or reduced because of peer pressure among juveniles, all of whom were participating in the same program. For example, if some chose to only go through the motions, it could affect how seriously others committed in the program.

There is currently no statistical fix when SUTVA is violated. There may be solutions in the future, but a major obstacle would seem to be the need to characterize exactly how the interference occurs. That is, one would need a plausible model of the interference, data with which to estimate its key features, and a statistical procedure that would at least produce consistent estimates.

For now, it is better to design and implement experiments in which the chances of such interference are small.[4] However, for group randomized experiments, there is the option of treating the group, not the individual, as the unit of interest. Thus, if the units are youth facilities, the recidivism rate for each facility characterizes that facility as a whole, not the individuals within it. SUTVA now addresses how different *facilities* interact. Does the treatment assigned to one facility affect how another facility responds? It might if the head of one facility knew the treatment assigned to another facility. The manner in which the juveniles within a given facility interact, however, is no longer relevant.

A concern about shifting to the group as the analysis unit is that the substantive questions are changed. The experiment is now about the performance of youth facilities not about the performance of the individual juveniles within them. Another concern is the reduction in sample size. If the unit is the youth facility, the sample size is the number of such facilities. For these reasons at least, researchers who design group randomized trials do not usually make the group the unit of analysis. But then, SUTVA remains at

[4] SUTVA is no less of a potential problem in quasi-experiments and observational studies.

risk.

3 Randomized Experiments and Statistical Inference

A reading of R.A. Fisher's *The Design of Experiments* makes clear that sound statistical inference is a very important part of the rationale for random assignment. The statistical inference problems is this: if the study were undertaken again, and the same subjects were reassigned to treatment and control conditions, the difference in means (or proportions) between the experimental and control groups would almost certainly be a bit different. The random assignment would likely change the composition of the two groups. For example, in the second study the group sent to boot camp may include a few more juveniles who were bad risks.[5]

Stated more broadly, any observed comparisons between the outcome for the experimental group and the outcome for the control group confounds possible treatment effects with random variation in group composition. One of Fisher's key contributions was to provide a rationale and a set of tools for taking the random composition effects into account. The rationale is based on the following thought experiment.

Imagine that there was no treatment effect. So, the proportions that are rearrested are really the same under the two interventions. Now imagine that all possible random assignments of the juveniles to the two conditions were constructed and for each, the difference in the two proportions calculated. A histogram of these differences would be necessarily centered on zero, but there would a number of times in which the proportion rearrested would be higher for the boot camp condition and a number of times the proportion rearrested would be higher for the incarcerated condition. Sometimes the absolute value of the difference in the proportions could be quite large.

The distribution of these differences is an example of a permutation distribution. The standard deviation of this distribution, usually called the standard error, conveys approximately the average disparity between the true treatment effect of zero and the composition effects resulting from the very large number of random assignments. So, one is able to gauge how much variation one can expect in the observed differences in proportions caused by

[5] A more formal representation of this will be presented shortly.

random assignment alone. Equally important, the middle 95% of this permutation distribution defines the 95% confidence interval. Finally, if it were to turn out that an observed difference in the two proportions from a real experiment fell outside of this interval, there would be grounds to reject the premise of no treatment effect. One would have the makings of a test of a null hypothesis that the response to the treatment and control conditions was the same.

The main problem with Fisher's rationale is that the permutation distribution cannot be constructed with data. It is the product of a thought experiment. However, with some combinatorial calculations, the key properties of the permutation distribution can be computed.[6] Conventional confidence intervals and statistical tests follow. The latter are often called "exact tests" because as a matter of convenience, approximations to the permutation distribution are often used, such as the t-distibuition or the χ^2-distribution.

Note what has been accomplished. We have a way to directly address the uncertainty produced by random assignment. If exact tests based on the permutation distribution are used, there are no distributional assumptions. In the case of a t-test for the difference between means, for example, one can avoid of the common concern about whether the response variable is at least approximately normal.

How can this go wrong? Perhaps the major problem is associated with group randomization. Even if the SUTVA assumption is met, it is still likely that the responses to either the treatment or the control condition will be more alike within the groups used for random assignment (e.g., different youth facilities) than between groups. The effective sample size within each group is then smaller than the nominal sample size within each group, and there is less statistical power. Conventional statistical inference will tend to produce overly optimistic conclusions. Fortunately, if each subject's group is known, statistical adjustments can be made. A recent discussion of such adjustments can be found in a pair of articles by Blitstein and his colleagues (2005a; 2005b).

4 Models of Treatment Effects

Up to this point, a causal effect for individual i has been based on the idea that each individual can have one response under the treatment condition

[6] For an excellent technical discussion see Rosenbaum, 2002, section 2.4.

and another under the control condition, and that the two responses can vary across individuals. This formulation is represented by

$$Y_{i,T}, \quad Y_{i,C}, \tag{3}$$

where "Y" is the response, "T" stands for the treatment condition, and "C" stands for the control condition. An estimate of the average treatment effect presumes this formulation. But, the *mechanism* by which effects differ across individuals is not represented either in the conceptualization or the estimate.

Sometimes we want to be more specific. We want to characterize how the responses are produced. Consider a simple additive model,

$$Y_{i,T} = Y_{i,C} + \tau. \tag{4}$$

As a result of the treatment, a constant, τ, is added to what would have been the response under the control condition. For this formulation of the treatment effect, τ is not subscripted. Thus, the impact of the treatment is the *same* for all subjects in the experiment. The response observed under the treatment can differ across subjects because the subjects do not have identical responses under the control condition. That is, they start off being different. Then, all are affected by the intervention is the exact same way. The value of τ is usually what is being estimated when the difference between the mean (or proportion) for the experimentals and controls is calculated.

Equation 4 might seem too restrictive. Perhaps one should allow for different subjects to respond to the treatment in different ways. One way to do this is with

$$Y_{i,T} = Y_{i,C} + \tau_i | x_i. \tag{5}$$

The response to the treatment for subject i depends on the value of some covariate x_i. This is an example of an interaction effect. For instance, how a juvenile responds to boot camp may depend on the age at which the juvenile is committed; the age of subject i is x_i.

The variable x_i has a very different status from the intervention assigned at random. It is not assigned at random and consequently, is likely to be correlated with many other variables that are, in turn, correlated with the response. The number of prior arrests is one illustration; it is likely to be correlated with age and the chances of rearrest after release. In short, the problem of confounders resurfaces despite an intervention assigned at random.

Equally important, if x_i cannot be manipulated, it by definition cannot be a cause, and the manipulation must be plausible for the setting in which the experiment is conducted. Thinking again about a given individual i, could the value of x_i be set to some other value? One could easily imagine altering the length of the sentence either to a boot camp or to a youth facility. One can easily imagine this because sentence length can, within bounds, be determined by a judge. In contrast, a judge cannot alter a given juvenile's age, race, or sex. These attributes cannot, therefore, be causes.

Although fixed attributes cannot be causes, they can still help define the setting and circumstances in which and intervention operates. And as such, they can alter an intervention's impact. For instance, boot camp could have one effect on boys and another effect on girls.

Consider the following example. Suppose, a staff person at the boot camp felt that the program was not working well for a particular female "recruit." The in-your-face intimidation that might be productive for boys is perhaps counter-productive for girls. It would certainly be possible reduce the amount intimidation to a level that might be appropriate for girls. However, turning a girl into a boy is not an option. The level of intimidation is manipulable, sex is not.

If x is suspected of altering how the treatment affects the response, and if x is measured as part of the experiment, x's role can be examined. In perhaps the most straightforward manner, block randomization can be employed and an average treatment effect can be estimated separately within each block. That is, subjects are arranged into groups by different values of x. Within each group, subjects are assigned at random to the experimental and control condition. Each group can then be analyzed separately. If desirable, an overall weighted average effect can be computed as well. For the boot camp example, one block might be boys and another block might be girls.

Far more risky are a post hoc analyses searching for differential treatment effects across individuals who vary on one or more covariates. Because individuals will almost certainly differ in their observed responses, whether under the treatment or control conditions, it will always be possible to construct groups that differ as well. And if one looks long enough and is sufficiently creative, predictors will be found to justify after-the-fact the groups that are identified. Data snooping of this kind is a useful form of exploratory data analysis. But the findings are likely to capitalize on idiosyncratic patterns in the data so that they will not replicate; an apparent finding of different treatment effects for different groups can be artifact of the data snooping.

Moreover, any p-values for formal tests likely will be too optimistic and any confidence intervals likely too narrow. The best solution is to mount a new experiment in which the results inductively discovered in the first experiment become hypothesis to be tested in the second.

Another kind of treatment effect that can vary across subjects rests on a multiplicative formulation. Thus, one can have

$$Y_{i,T} = \beta Y_{i,C}, \tag{6}$$

where $\beta > 1$ is usually assumed.

The response under the control condition is multiplied by β to produce the response to the treatment. Therefore, the difference between the responses under the two conditions depends on the response under the control condition. A larger response under the control condition leads to larger gap between the two responses. Moreover, because of the multiplication, the distribution of the responses under the treatment is more dispersed. Nevertheless, the value of β can be estimated in much the same manner as an additive effect.[7]

There many other ways to formulate models of treatment effects. Options get especially rich if there are a number of different treatments that can be represented quantitatively. For example, the treatment might be boot camp sentences of different lengths. One is then within a dose-response framework where, for instance, logistic response curves are popular.

In summary, there is often much more to extract from an experiment than the average causal effect. It can be helpful to consider how the treatment effect is generated and then try to estimate one or more parameters of interest. In criminal justice applications, this is rarely done.

5 Some Operational Benefits and Costs from Randomized Experiments

To this point, the message has been that random assignment confers a number of significant technical benefits, although some depend on important

[7] Sometimes researchers will assume a non-additive effect, but proceed using functions order statistics, such as the median rather than the mean, to summarize responses. One motivation can be to avoid problems caused by a few outliers. Rosenbaum (2003: section 5.3) can be consulted for details.

assumptions. There are also a number of operational benefits that have been discussed in earlier writings (e.g., Berk et al., 1985; Boruch, 1997; Shadish et al., 2002). Some examples deserve brief mention here. They are not the primary focus of the paper, but can be especially important for mounting randomized field experiments and then presenting the results.

1. The idea of random assignment is relatively easy to explain to stakeholders. Of late, state lotteries have become common across the United States, and there is widespread understanding that in a fair lottery, each ticket has the same chance of winning. Random assignment described as a fair lottery is easily understood, or at least more easily than most of alternatives research designs (e.g., regression-discontinuity designs or nearest neighbor matching).

2. Randomized experiments have "face validity." If subjects are assigned at random, there can be no cherry picking either by subjects who self-select into programs they favor or by administrators who have a professional investment in the outcome.

3. Random assignment plays no favorites. All subjects are treated alike by the assignment mechanisms. In that sense, randomized experiments can be seen as fair.

4. Randomized experiments are often relatively easy to analyze, and one important consequence is results that are likely to be accessible to stakeholders. Simple comparisons between means or medians, coupled with significance tests, will commonly suffice. When more demanding procedures are needed, the cause is usually a difficult response variable (e.g., time to failure) or the need to capture a dose response relationship. Advanced multivariate methods can sometimes by helpful, but they are rarely mandatory. And this is good.

This is not to argue that there is no operational downside. Perhaps the largest operational problem is inflexibility. The integrity of a randomized experiment depends on implementing a consistent protocol. The treatment randomly assigned must be the treatment delivered.

Sometimes, a bit of flexibility can be built in, but one must anticipate potential problems and design workable solutions. For example, an experiment testing certain criminal justice responses to domestic violence might be viewed as inappropriate for felony domestic violence. Dropping from the study *all* offenders accused of felony domestic violence would not bias the results, even after random assignment, as long as a clear definition of felony domestic violence were provided. The main cost would be a smaller sample size, which in some cases could be remedied. But if a clear definition were not provided, significant bias could creep in. For instance, police officers might void the experiment for individuals thought to be particularly dangerous, but only when those individuals were assigned one of the less restrictive treatments. Then, the experiment would be biased in favor of the less restrictive treatments. In short, randomized experiments require that their designers anticipate a range of potential problems and develop contingency plans. This is difficult to do well.[8]

Perhaps the major public relations problem is applying random assignment when stakeholders think they already know which intervention works best. Under these circumstances, assigning subjects at random means denying benefits to individuals who need them. If the stakeholders are correct, there is no rationale for experimenting. When the evidence is equivocal (or essentially absent), stakeholders may still argue for assignment based on need, or more broadly, on matching interventions with the different needs or characteristics of particular subjects. Thus, individuals arrested for felony domestic violence may "need" more restrictive interventions than individuals arrested for misdemeanor domestic violence.

If stakeholders are firm in their convictions, it may be impossible to implement conventional random assignment. At the very least some creative options would need to be introduced. For example, one can employ a block randomized design with blocks defined by need. In the high-need block, 80% of the individuals might be assigned to the treatment thought to be more beneficial. In the low-need block, 20% of the individuals might be assigned to the treatment thought to be more beneficial. It is also sometimes possible to compensate individuals after the experiment, who in retrospect were denied an appropriate treatment. A popular example is to make arrangements for the control group to receive the intervention thought by some to be beneficial after the experiment is over.

[8] Some might argue that another major operational problem is cost. However, randomized experiments are not necessarily more costly than the next best alternative. And in any case, there appears to be no definitive accounting one way or the other.

6 Generalizing from Experiments

About a generation ago, Lee Cronbach (1982) argued that the emphasis on causal inference, a key technical justification for random assignment, was misplaced. Unless it was possible to generalize from an experiment, the experiment was probably not worth doing. These concerns have been echoed in recent discussion by some econometricians (Heckman and Smith, 1995).

There are several kinds of generalization that might be desirable:

1. to different subjects;
2. to different settings;
3. to different times;
4. to related interventions; and
5. to related outcomes.

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

There is nothing inherent in randomized experiments that precludes generalization, and generalizations are routinely made. But there can be important complications. In particular, subjects who know they are part of an experiment may behave differently than had the intervention delivered been "real." A common illustration is interventions with police; they may react differently to an intervention that is temporary than to one that is permanent. For example, police officers may just go through the motions if required on an experimental basis to refer certain domestic violence victims to a restraining order clinic. If such referrals were an ongoing component of the range of interventions police were required to consider, they might provide the referral information in an effective manner.

However, the problem is *not* random assignment. *Any* intervention that is experimental presents the very same risks. In principle, one might "blind" study subjects to the reality that they are part of an experimental program. However, ethical concerns will usually eliminate this option. The only apparent solution is to work with observational data gathered on programs that are ongoing and stable. But this is a very high price to pay. Innovative programs or programs undergoing substantial changes are ruled out. Moreover, there can be additional complications if the individuals who are already participating do not exhaust the population of individuals the program is meant to cover. The "early adopters" may be atypical.

Likewise the difficulties associated with generalizing over subjects, settings, times, interventions, and outcomes are not unique to randomized experiments. They can be problematic for any evaluation, even those based on observational data. And the potential solutions are often shared as well. These include the following approaches.

1. Using probability samples of subjects so that generalizations can be made from the sample to the population from which the sample was drawn.

2. Using suites of studies carefully designed so that variants in the interventions could be tested with different mixes of subjects, in different settings, and with related outcomes, all selected to document useful generalization targets. A proper analysis would then be based on a pooled analysis using the raw data from all of the studies as one large data set (Berk et al., 1992; Bloom et al., 2002).

3. Using suites of studies that are a mix of true experiments, quasi-experiments, and observational studies so that the comparative advantages of each can be exploited.[9]

4. Using credible theory to justify what kinds of generalizations are plausible.

There is not space here to consider these options in any detail. A good and accessible discussion can be found in a recent book by William Shadish and his colleagues (Shadish et al., 2002). In addition, Cronbach's original writings on generalization (1982) are still a good read.

Shadish and his collaborators also offer an endorsement of sorts for meta-analysis as a generalization tool. Even with the caveats they provide, their assessment is rather too positive. A range of serious concerns about meta-analysis have been lodged over the past 15 years (e.g., Wachter, 1988; Petitti, 1993; Briggs, 2005) that go to the heart of the technique, and for which there

[9] This was suggested by one of the anonymous reviewers.

seems to be no good rebuttal. Thus, when Shadish and his colleagues advise readers "to be critical of meta-analysis as you would any scientific method, but no more so" (2002: 446), they are being much too gentle. There is a gapping whole between what the formal mathematics of meta-analysis require and what is possible even under the best of circumstances. A discussion of the issues would take us far afield, but David Freedman and I offer the following bottom line (Berk and Friedman, 2003: 247):

> The interesting question is why the technique is so widely used. One possible answer it this. Meta-analysis would be a wonderful method *if* the assumptions held. However, the assumptions are so esoteric as to be unfathomable and hence, immune from rational consideration: the rest is history [emphasis in the original].

7 Implementation Problems

The most serious vulnerability of randomized experiments is their implementation. It is difficult to do randomized experiments well. Some especially nettlesome problems include the following.

1. The research design is usually premised on a certain number of subjects. But far fewer sometimes show up. As a result, there can be a substantial loss in statistical power and serious difficulties generalizing the results to those who should have participated, but did not. It is good practice, therefore, to work hard as the experiment is designed to determine what the sample size is likely to be, and if necessary, make arrangements to provide special incentives for participation.

2. A related problem is attrition from the study; cases are lost as the study progresses. The price can be the reduced statistical power and if there is differential attribution with respect to the interventions, bias as well. The biases can sometimes be reduced with a proper analysis of the attrition process (Foster and Fang., 2004), but the requisite assumptions are often heroic.

3. Random assignment is not actually implemented. This can happen in many different ways. For example, the overall protocol may be misunderstood and used incorrectly. As a result, one cannot know

for each subject what the randomly assigned intervention should have been. At best, one then has a quasi-experiment, but more likely, little more than an observational study.

4. The treatment assigned may not be the treatment implemented. For example, an inmate assigned to a job training program in prison may fail to show up. This can turn a randomized field experiment into an "intention-to-treat" experiment. The intention to deliver a particular treatment is randomly assigned, but not the treatment itself. If there is at least some correspondence between the treatment assigned at random and the treatment implemented, instrumental variable procedures can sometimes provide an effective solution (Angrist et al., 1996). The key obstacle is reduced statistical power.

5. The response variable (s) may be poorly measured. For example, self-reports of crime victimization are subject to all of the usual problems with self-report data and if these problems are associated with intervention assigned, can produce very misleading results. The best advice is to measure well. A fallback position is to conduct auxiliary studies documenting the performance characteristics of the measures used. For example, self-report data may be compared to crime reports made to the police. Both have flaws, but through careful comparisons, much can be learned about the kinds of events captured and missed by each.

In summary, there are sometimes constructive ways respond to each of these problems and in a few cases, the fix can be very effective. But more typically, implementation difficulties weaken an experiment substantially. At the same time, the implementation problems associated with randomized experiments have close parallels with the implementation problems associated with a wide variety of other designs, and sometimes the problems are even much the same. For example, a random sampling design developed to obtain a representative set of subjects in an observational study may stumble for many of the same reasons that random assignment can fail.[10] Quasi-experiments can become under-powered for many of the same reasons as randomized experiments. Attrition is at least as problematic in panel studies as in randomized experiments. And serious measurement error is widespread regardless of the research design.

[10] And response rates have been dropping like a stone for well over a decade.

8 Conclusions

Perhaps the key retort to individuals who claim randomized experiments are undesirable is to ask "compared to what?" Quasi-experiments, such as the generalized regression-discontinuity design (Berk and de Leeuw, 1999), have many of the same implementation problems as randomized experiments and more. For example, there is still an assignment protocol that must be followed, and the design requires a credible model of how the response is related to the covariate used for assignment.

If the alternative to a randomized experiment is an observational study, the difficulties are likely to be even worse. There are a host of potential implementation problems and the modeling required can be highly suspect (Rosenbaum 2002; Berk 2003). For example, propensity score adjustments assume the variables related to selection into the various intervention groups are known, included in the data set, measured without error (random or systematic), and are appropriately introduced into the selection model. These can be daunting requirements even if some allowance is made for modest departures from the ideal.

Randomized experiments rest on more complicated, subtle, and fragile foundations than some researchers appreciate. Proper implementation of randomized experiments is demanding. Textbook requirements are rarely met. Thus, randomized experiments are not the gold standard. But if the truth be told, there is no gold standard.

There can be settings in which the strengths and weaknesses of potential research designs favor an alternative to randomized experiments. Randomized experiments should not be the automatic choice whenever they are feasible. But the alternatives to randomized experiments are likely to be worse and should be employed only after thorough comparisons showing that they clearly preferable.

9 References

Angrist, J., Imbens, G.W., and D.B. Rubin (1996) "Identification of Causal Effects Using Instrumental Variables" (with discussion). *Journal of the American Statistical Association* 91: 441-472.

Berk, R.A., Boruch, R., Chambers, D., Rossi, P.H. and A. Witte (1985) "Social Policy Experimentation: A Position Paper." *Evaluation Review* 9: 387-430.

Berk, R.A., Campbell, A., Klapp, R. and B. Western (1992) "The Differential Deterrent Effects of An Arrest in Incidents of Domestic Violence: A Bayesian Analysis of Four Randomized Field Experiments," *American Sociological Review*:5: 689-708, 1992

Berk, R.A. and J. de Leeuw (1999) "An Evaluation of California's Inmate Classification System Using a Generalized Regression Discontinuity Design." *Journal of the American Statistical Association*, 94: 1045-1052.

Berk, R.A., and D.A. Freedman (2003) "Statistical Assumptions as Empirical Commitments" In T.G. Blomberg and S. Cohen (eds.), *Punishment and Social Control: Essays in Honor of Sheldon Messinger.*, second edition: 235-254. New York: Aldine de Gruyter.

Berk, R.A., (2003) *Regression Analysis: A Constructive Critique.* Newbury Park: Sage Publications.

Blitstein, J.L., Hannab, P.J., Murry, D.M., and W.R. Shadish (2005a) "Increasing the Degrees of Freedom in Existing Group Randomized Trials: the df Approach." *Evaluation Review* 29: 241-267.

Blitstein, J.L., Hannab, P.J., Murry, D.M., and W.R. Shadish (2005b) "Increasing the Degrees of Freedom in Future Group Randomized Trials: the df Approach." *Evaluation Review* 29: 268-286.

Bloom, H.S., Hill, C.J., and J.A. Riccio. (2002) "Linking Program Implementation and Effectiveness: Lessons from a Pooled Sample of Welfare-to-Work Experiments." *Journal of Policy Analysis and Management* 22: 551-575.

Boruch, R.F., (1997) *Randomized Field Experiments for Planning and Evaluation: A practical Guide*. Newbury Park, CA: Sage Publications.

Box, G.E.P, Hunter, W.G., and J. S. Hunter (1978) *Statistics for Experimenters*. New York: John Wiley.

Briggs, D.C., (2005) "Meta-Analysis: A Case Study." *Evaluation Review* 29: 87-127.

Campbell, D.T., and J.C. Stanley (1963) *Experimental and Quasi-Experimental Designs for Research*. Boston: Houghton Mifflin.

Cronbach, L.J. (1982) *Designing Evaluation of Educational and Social Programs*. San Francisco: Jossey-Bass.

Cox, D.R. (1958) *Planning of Experiments*. New York: John Wiley.

Fisher, R.A. (1935) *The Design of Experiments*. New York: Hafner Press.

Foster, M.E., G.Y. Fang (2004) "Alternatives to Handling Attrition: An Illustration Using Data from the Fast Track Evaluation." *Evaluation Review* 28: 434-464.

Heckman, J.J., and J.A. Smith (1995) "Assessing the Case for Randomized Social Experiments. "*Journal of Economic Perspectives* 9: 85-110.

Neyman, J. [1923] (1990) "On the Application of Probability Theory to Agricultural Experiments. Essay on Principles. Section 9." Translated and edited by D.M. Dabrowska and T.P. Speed, *Statistical Science*, 5: 465-471.

Petitti, D.E. (1993) *Meta-Analysis, Decision Analysis, Cost-Effectiveness*, second edition. New York: Oxford University Press.

Rosenbaum, P.R. (2002) *Observational Studies*, second edition. New York: Springer-Verlag.

Rubin, D. B. (1986) "Which Ifs Have Causal Answers." *Journal of the American Statistical Association* 81: 961-962.

Rubin, D.B. (1990) "Comment: Neyman (1923) and Causal Inference in Experiments and Observational Studies." *Statistical Science* 5: 472-480.

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

Wachter, K.W. (1988) “Disturbed about Meta-Analysis?” *Science* 241: 1407-1408.