Points of Intersection between Randomized Experiments and Quasi-Experiments

By
JANE GREEN

There are several contexts in which randomization in research design is not possible. Researchers may be interested in the effects of large-scale events or policy initiatives, or of complex causal processes that cannot be replicated. Threats to external validity in randomized experiments will lead to greater dividends for some research questions in a quasi-experimental design. Quasi-experiments can offer opportunities for testing causal theories, if they are carefully constructed to counter threats to internal validity. Furthermore, randomization is not a simple guarantee against threats to internal validity where few units of observation are used. An appreciation of the intersection of these approaches—between stronger quasi-experiments and weaker randomized experiments—can help guide our choice of experimental method. One can make the same argument about the benefits of pursuing good opportunities with quasi-experiments over scant opportunities with randomized experiments as is made more often in reverse.

Keywords: quasi-experiments; randomized experiments; field experiments; internal validity; external validity; causation

Observing that individuals who know they have developed a serious disease prefer treatments “proven to work” by randomized controlled trials (RCTs) over those suggested by a quasi-experimental method, Berlin and Solow (2009, 175) ask, “Why hold social programs to a lesser standard?” In pure randomized experiments, tests of causal relationships are much

Jane Green is a lecturer in politics (quantitative research methods) at the University of Manchester, UK. Her work explores the relationships between elite political strategies and policy positions, public opinion and voting, social norms and voter turnout, and the causal dynamics of elite policy competences and their relationships to the vote. She teaches various research design courses at Manchester, including Ph.D. research design, Advanced Research in Politics, and undergraduate dissertation skills.

NOTE: The author gratefully acknowledges Laura Stoker for her teaching on quasi-experimentation at the University of California, Berkeley.

DOI: 10.1177/0002716209351517

less prone to the bias caused by non-random comparison groups in quasi-experimental designs, and so, as a method of analysis for deriving reliable causal estimates, pure experiments are generally the method of choice. In a seminal work, Campbell (1973, 221) states, "True experimentation should almost always be preferred to quasi-experiments where both are available. Only occasionally are the threats to external validity so much greater for the true experiment that one would prefer a quasi-experiment."

I argue here that for some research questions and for some researchers, quasi-experiments can provide a very useful tool for research. Also, there are more cases than one might immediately think where pure randomized experiments are not feasible. True, if pure experiments provide equally generalizable insights and are also free from biases in design, their relative costs are strongly outweighed by their reliability (Gerber, Green, and Kaplan 2004). But pure experiments cannot always reproduce the types of causal relationships we are interested in exploring, and not all randomized experiments are free from systematic biases. Observing the intersections between experiments and quasi-experiments allows us to find similarities between strong quasi-experiments and weaker randomized experiments and consider the ways in which biases in quasi-experiments can be recognized and addressed. This subtle yet important change in emphasis can point to areas of research in which quasi-experiments may provide the best opportunities for the future. In sum, I argue that we may not want to bet our health and well-being on the findings of a quasi-experiment, but where randomized experiments are imperfect or not available for use, we should prefer to bet our health on the best quasi-experiment rather than on no attempts at causal estimation whatsoever.

What Are Quasi-Experiments?

A quasi-experiment is a research study in which a treatment occurs naturally in the real world, and so, in contrast to pure experiments, the treatment or the event is not distributed among people at random. In rare cases, randomization could occur at implementation, providing a natural randomized experiment, but generally quasi-experimental opportunities arise when an event or intervention is implemented among subjects in a non-random manner. A government may introduce a new welfare policy, a president may embark upon a campaign of persuasion, or an organization may conduct an advertising campaign, and none of these interventions will be targeted or observed in a randomized fashion. There may even be a major change in the institutions of government. Our task as researchers is to learn from these changes and to estimate their impacts upon important individual and social outcomes, comparing them to equivalent groups or cases in different times, cohorts, or locations for which we do not expect to find the same effects. Does implementing a new security initiative reduce the occurrence of terrorist threats? Will multilateral economic investments reduce the growth rates of unemployment in targeted counties? Do social welfare programs significantly improve the long-term life chances of recipients? These are questions that cannot be easily answered, if at all, using the randomized experiments described elsewhere

in this volume: it is difficult to instigate similar “treatments” by adequately approximating their conditions and characteristics.

In quasi-experimental research, various designs and estimation techniques are used to estimate the likely causal effect of the event or intervention, taking carefully chosen comparison groups and systematically exploring the influence of other potentially confounding factors. Although these methods do not allow for unequivocal causal inference, and are problematic to the extent that bias can only be corrected post hoc (Gerber, Green, and Kaplan 2004), the findings of quasi-experiments have been shown to be comparable to those in randomized experiments in a promising range of cases (Cook, Shadish, and Wong 2008; Cook and Steiner 2009; Shadish and Cook 2009).

An Example: Teaching Innovation and Educational Outcomes

For the purpose of illustration, let us imagine we are interested in understanding whether schoolchildren’s educational attainment improves under an innovative teaching program or a new educational investment. We could either design an RCT experiment, delivering the key components of teaching to randomly assigned schools, districts, or pupils, or we could evaluate school training programs in schools and districts in which innovative methods or investments have been introduced. Such a policy implementation would usually be introduced on the basis of need, so we can realistically assume that the schools participating in the program are those that had lower educational attainment levels at the outset. If a policymaker were ultimately interested in whether new teaching programs work among more disadvantaged children, and the new teaching program were being funded and rolled out in schools with disproportionate results of low educational attainment, we would be presented with a challenge common to the researcher: how to evaluate the causal impacts of initiatives among groups that are selected rather than randomized.

The fact that innovative school learning programs have been rolled out in some schools but not others, among some children but not others, and in some districts but not others offers the classic quasi-experimental opportunity. We could use pre- and post-intervention test scores in schools, classrooms, and districts in which the school program was delivered and compare the rate or presence/absence of educational performance change in the targeted cases with non-targeted cases. Ideally, programs would be rolled out in very similar schools with similar catchment areas at different times, allowing comparison cases to be chosen from most-similar schools and students but over different time frames (a staged innovation). In this way, we could evaluate the degree to which the educational program yields a positive result on learning outcomes. And yet, just as Berlin and Solow (2009) would not base health treatment on evidence that was not “proven to work” in a randomized experiment, would not we wish to place any child into an educational initiative that did not have a “proven” positive effect (or worse, that had not been proven not to have a negative effect). The United States’s Head Start program,

which has been running since 1964, received \$7 billion in the government's 2009 budget and another \$2 billion in February 2009 as part of the American Recovery and Reinvestment Act. It costs \$22,600 per child, compared to \$9,500 in a day care center, on average (Besharov and Call 2009). If quasi-experiments are as unreliable as some authors claim, we would want a series of the most reliable randomized experiments to justify this kind of expenditure.¹

Using an RCT, researchers would need to implement the same educational program in randomly selected schools and districts, again measuring pre- and post-treatment educational attainment indicators to compare randomly assigned treatment groups to randomly assigned controls. If, on average, we saw an increase in educational performance outcomes among all those who participated in the program compared to those who did not, we could argue that the program is effective in improving educational attainment and should be funded.

Points of Similarity in the Two Hypothetical Research Designs

In both designs we are interested in evaluating the effect of an intervention, and in both we have a control or comparison group. In the pure experiment, the control group comprises the students, schools, and districts, chosen at random, who did not participate in the scheme. In practice, the best unit of randomization would be the district, since district-level factors affecting educational attainment (e.g., other school-level teaching policies or practices and demographic factors) could then be distributed randomly across treatments and controls. In the quasi-experiment, the comparison groups would also be pupils, schools, or districts in which the teaching program was not delivered, but these would have been selected by policymakers, funders, or schools that could opt in or out of the initiative. In the pure experiment, the researcher controls the treatment timing, and if the treatment occurs alongside other school year events or initiatives, this should be common across treatment and control groups. In the quasi-experimental design, the researcher can compare pre- and post-treatment measures around the time of the initiative to determine the rate of change over the critical period, as well as the level of volatility in educational attainment pre-intervention and afterward. There are clear parallels in the experimental logic underlying the analysis of effects in field experiments and quasi-experiments.

Points of Difference: Threats to Internal Validity Using Quasi-Experiments

Despite apparent similarities in logic, these studies reveal important differences that are essential for evaluating the validity of insights gained through quasi-experimental research projects. These relate to Campbell's (1973) list of

“threats to internal validity,” which questions whether a treatment can be the cause of any change in outcome.² These are called threats to internal validity because they relate to whether we can trust the conclusions of the study as it relates to that study, rather than to its implications generalized to the wider world.

Imagine we were exploring the effects of educational programs in schools situated in more disadvantaged districts, but children in schools located in these districts were also influenced by concurrent events. For example, financial investments in similar neighborhoods could have an effect alongside the teaching program, such as nutrition programs, parental advice, housing provision, or changing employment levels. If the teaching program had little significant effect, but unemployment destabilized children’s home lives and well-being, for example, we could even draw the mistaken conclusion, comparing these districts with more affluent districts, that the program had a negative impact. This problem is an example, in Campbell’s terminology, of a “history” threat to internal validity (Campbell 1973). In this case, researchers would hope to compare school districts that participated in the program with similar districts—assuming the allocation was by chance and that the same “history” occurred in both groups alike, isolating any timely change, as far as is possible, to the intervention.

A second and related threat to internal validity arises from “maturation.” Imagine that children participating in the program were naturally on a steeper trajectory than those who were not, perhaps because children in the comparison group had lower educational attainment levels at the outset. In comparing lower- and higher-attainment groups, due to the original selection, we would see a higher increase in test scores among the intervention groups regardless of the teaching methods. Trends may already be in an upward or downward direction, or subjects or outcomes may be likely to change their behavior anyway. The absence of randomization in treatment and control groups makes the comparisons problematic, but using within-context comparisons could potentially mitigate these threats.

Measures of pupils’ educational attainment in treatment districts may also be particularly prone to fluctuation. If test scores among pupils in disadvantaged areas showed greater inherent variance, either due to a greater range of ability or truancy levels or marking volumes, a pre-post test difference in our treatment schools may result due to “instability” (Campbell 1973). Measurement error would wrongly attribute an apparent change to the teaching methods.

A “testing” threat could undermine the comparisons if pre-test procedure and scores among children in the treatment groups resulted in greater (or reduced) effort in the classroom, thereby inducing changes in post-test performance. Such testing threats are usually associated with priming respondents to take more interest in an issue following participant surveys—for example, providing more “desirable” answers to researchers in face-to-face formats, or due to the process of learning in the course of the research; but so long as the quasi-experiment and randomized experiment use the same testing procedures, both are subject to the same potential threats to internal validity. Likewise, a change in measurement of educational ability, such as a change in subject tested, test questions, or test standard, would also provide an artifact effect, a threat to quasi-experiments and randomized experiments alike. This would represent an “instrumentation” threat to internal validity.

Particularly relevant to these kinds of needs-based quasi-experimental studies are threats to internal validity caused by “regression artifacts.” When subjects or cases are chosen on a needs or merit basis, we know they are atypical in some way from the general population. But extreme scores, negative or positive, tend to moderate with time, since very high scores cannot go much higher and very low scores cannot go much lower. Schools with lower educational test scores may therefore appear to make greater improvements in learning, simply because their test scores were lower at the outset, irrespective of the intervening treatment, a problem known as a “regression artifact” because scores regress toward the mean.

Let’s imagine that children taking part in the program opt out of the scheme, due to leaving either the area or the school, or due to preference. If sufficient numbers do so, and if they leave on the basis of ability, the average post-test scores could be disproportionately higher among the treatment groups. “Experimental mortality” is a problem for all studies in which people may drop out over time, but it is particularly challenging where people drop out disproportionately by a criterion of selection. In an RCT, we would expect attrition to impact the treatment and control groups in equal measure.

The most serious of threats and the main justification for randomization is the “selection threat.” Although Campbell (1973) lists this threat separately, it should be clear from the aforementioned examples that selection is the major challenge in quasi-experiments and the main contrast to the studies using randomization, due to the potential relationships of selection with maturation (and a “selection-maturation” interaction threat) and also with measurement, instability, and history. In the hypothetical example, the selection threat cautions that schools in disadvantaged backgrounds will show responses to educational programs in atypical ways. Therefore, we cannot distinguish the treatment effect from a context effect. In a needs-based intervention, however, the primary research question would most likely be “Do innovative teaching programs raise educational attainment levels in schools with low educational attainment levels?” and so we are not interested to explore whether such programs have an effect among all children or any children. Nevertheless, it should also be clear that the primary research question cannot be answered without comparisons within relevantly similar contexts, and these similarities would need to be fairly exhaustive.

These threats to internal validity show important differences in the quasi-experimental evaluation that on first glance may cause us to reject these methods as a way of inferring causal effect. It would be possible to design a quasi-experimental study in which these threats could be *minimized*, especially if schools could be selected into both treatment and comparison groups on the basis of local context. If we were interested in examining the efficacy of teaching innovations among all or any children, the purchase gained on the question using randomization would be considerable, enabling potentially confounding factors to be present in both treatment and control comparisons and so controlling for their effects. If we were interested in observing the effects, specifically among children and schools with lower educational attainment, we would also ideally conduct randomized control trials across context, comparing unbiased treatment effects in lower educational

attaining schools and districts with moderately achieving and high-achieving schools. This ideal, however, points to one of the most practical but common constraints on randomization and arguments for quasi-experimentation: cost and resources. In the hypothetical examples, we would be hard-pressed to imagine any national or local government willing to roll out a program randomly without consideration of need or any researchers with a budget to do so themselves. For these reasons, the quasi-experiment may provide the best or only opportunity to evaluate the efficacy of the new teaching method among more disadvantaged school pupils. Furthermore, even if we were able to conduct an RCT, a series of threats to external validity might be introduced due to a design based upon randomization.

Points of Difference: Threats to External Validity Using Pure Experiments

“External validity” relates to the ability to generalize from the research context to the real world and is the main challenge to causal inference in randomized experiments instigated by researchers. Campbell (1973; Campbell and Stanley 1963) lists five possible weaknesses: incomplete treatment, artificiality, interaction effects of testing, irrelevant responsiveness of measures, and irrelevant replicability of treatments. In quasi-experiments, participants’ reactions are less likely to be confounded by these problems on all five counts. If, in the hypothetical example above, researchers could only imperfectly emulate the key ingredients of the education program, then even randomization could not mitigate against the difficulties of drawing reliable inferences about causation from educational programs conducted in schools.

Incomplete treatments occur where researchers are unable to ensure that all participants are subjected to an intervention or where some participants experience only partial treatment. Examples may be found in trials of drug rehabilitation programs or in educational programs where participants drop out midway. We could imagine that the research program trialing the innovative teaching method in a randomized design may be more prone to students and schools leaving the program, or to schools being less committed to the program: There would be fewer cases in which the treatment is implemented in a complete manner.

Artificiality presents particular challenges when the intervention is large-scale or very heavily determined by context. If, for example, researchers undertook teaching programs outside the normal school hours, or with fewer students, using different classroom settings, contexts, teachers, or materials, our ability to generalize to other cases would clearly be compromised. Such basic mistakes are not made by careful experimenters, but some research questions such as those relating to advertising campaign effects, presidential persuasion studies, and Internet campaign effects using artificial environments are distinctly different from the ways people normally interact with the various communications they encounter.

The use of pre-tests, as mentioned above, is equally problematic in randomized experiments where pre-tests sensitize participants to respond disproportionately

to treatments. These are rarely necessary using randomly selected control group participants, since pre-post test scores are seldom needed and such tests would, if used, be administered to both groups; but we could imagine that asking people about their support for a product or an organization, to get an idea of a pre-intervention baseline, may make people more attentive to that product, irrespective of any intervention.

Irrelevant responsiveness of measures and irrelevant replicability of treatments relate to a tendency in experiments to misspecify the treatment intervention, finding apparent effects or no effect when these are caused by measures and treatments, instead of real treatment effects. Measures are complex and may include components the researcher is not even aware of that may produce apparent effects. Taking again the hypothetical example, the experimenter could replicate teaching methods but fail to replicate the backing of the school or teacher and parent involvement—mediating factors or direct causal factors that may be unobservable in the quasi-experimental setting, but an essential part of the intervention nonetheless.

Naturally, if internal validity is compromised, is not acknowledged, or goes unaddressed, a study's findings have very little value, since conclusions could be based entirely on spurious relationships. A pure experiment lacking perfect external validity could still potentially tell us important things about possible relationships in the real world. However, if we can maximize internal validity using quasi-experimental research designs, the intersections between experiments with weak external validity and quasi-experiments with strong internal validity can provide less stark trade-offs than a simple and usual comparison of these two methods in which the best field experiments are contrasted with the weakest quasi-experimental designs.

Quasi-Experimental Studies with Strong Internal Validity

Using examples of quasi-experimental studies that take problems of internal validity very seriously, quasi-experimental studies can make valuable contributions to our understanding of social and political phenomena. In a quasi-experimental design in which alternative explanations can be plausibly discounted, the threats to external validity of randomized experiments are relatively more serious. There are three main tools available to researchers using quasi or natural experiments or events and interventions that occur in the real world—interrupted time series designs, control series designs, and regression discontinuity analysis.³ In combination and in careful design they can enhance the internal validity of quasi-experimental inference.

Interrupted time series designs can be used to evaluate the effects of an intervention that is rolled out across the board. Here, the only way of estimating the effect is to compare the period before with the period after the intervention. Let's imagine that the teaching program was rolled out across all need-based schools

in a given country—a major social program designed to alleviate educational inequalities but one based on selection. Taking several (ideally a realistic maximum) set of time points before the initiative and several (maximum) time points afterward, the degree to which a distinct and/or permanent increase in educational levels following the initiative would tell us something about the effect of the intervention. Any changes may also reflect an effect of history (there may be many other state or nationwide influences) or an effect of regression to the mean (educational attainment may have begun very low), and so we need many time points before the intervention and many afterward to evaluate any intervention effects in the context of pre- and post-intervention trends. Taking many educational programs that occur across a range of geographical units, such as states or even countries, and a range of time points, we could also explore the effects of interventions happening in different times, places, and among different types of school. Clear pre- and post-intervention differences across our range of cases would lend stronger support to the hypothesis and mitigate against threats to internal validity arising from history, maturation, and selection.

Control series designs use untreated groups to make comparisons between the treatment group and a control, while neither group is selected at random. The non-equivalent control group allows a researcher to examine pre- and post-test measures and examine the relative changes between the treatment and control groups pre- and post-intervention. This method (the same as suggested in the hypothetical comparison of disadvantaged and other school districts) could be prone to selection threats. However, it is also possible to use comparison groups as controls where they comprise equivalent or even the same participants. For example, implementing a program in schools from several similar districts, and comparing with schools in comparable districts in which the program is not implemented, we could suggest that any selection effects arising from socio-demographic background were reduced. This comparison requires targeting at random, but equivalence can also sometimes be determined after the fact, where covariates are analyzed testing equal mixes of key indicators.

Regression continuity analysis targets difficulties of selection, maturation, and group differences, combining control comparisons and interrupted time series designs with examination of pre- and post-test effects. If a treatment group and a comparison group have different initial educational attainment levels, the test scores will begin higher for one and lower for the other. Also, one may increase at a higher rate and the other at a lower rate, with the potential for regression artifacts, maturation effects, and selection-maturation interactions. Regression continuity designs compare the levels and the trends before and after an event or intervention to ascertain whether sharp discontinuities can be determined in these pre-existing trends. Across control groups and across time, a more persuasive treatment effect can be discerned from the threats to internal validity. However, in these studies the local average treatment effect is identified only at the cut point where the discontinuity occurs, and so generalizing from these slopes and trends could miss pre-existing fluctuations before and/or after the cut point. Nevertheless, with careful analysis over time, these tools allow researchers to study impacts of events a researcher cannot easily re-create or control by randomization.

Examples from the Field of Political Science

Researchers sometimes explore effects of phenomena in political life that are so context-specific, or choice-specific, that they preclude a randomized experimental approach. Four such studies will be considered in turn.

Mannheim and Albritton (1984) studied the effects of countries' employment of public relations (PR) companies upon positive or negative coverage of those different countries in the U.S. media. Sigelman and Rosenblatt (1996) studied the impacts of high-profile presidential speeches made by President Ronald Reagan on support for commitment of noncombat troops to Lebanon and the invasion of Grenada in 1983. Lewis-Beck and Alford (1980) studied policy outcomes arising from government safety regulation in coal mining and whether these acts resulted in fewer industrial accidents, and Stoker and Jennings (1995) studied the impacts of marriage on voter turnout and participation. A researcher wishing to conduct an RCT would find it a considerable challenge to emulate the employment of a public relations agency, a high-profile and widely discussed presidential speech, the introduction of safety legislation, or, more obviously, a marriage between a large number (or any number) of couples!

The Mannheim and Albritton (1984) study took seven country examples, six of which employed a PR company (with one control), and evaluated the coverage associated with that country in the *New York Times* prior to the employment of the PR company and afterward, after a period of three months. This research question relates to a series of events beyond a researcher's control. The study is an example of an interrupted time series design, although with a weak control design of just one case and relatively few cases in which to test an intervention. In this study, the employment of a PR firm may be indicative of other unobservable country characteristics. However, all six countries employed PR companies at different time points, and thus, threats arising from maturation, history, or instability are less persuasive rival explanations for the findings. Mannheim and Albritton show significant effects on the nature of the coverage and on the amount and the use of different time points, suggesting a systematic relationship between the intervention and time that would not be easily attributable to inter-country characteristics.

In Sigelman and Rosenblatt's (1996) study, it is not possible to examine different time points, since the authors had only one high-profile speech. On October 27, 1983, President Reagan gave a speech on nationwide television. This followed an attack on U.S. service personnel in Beirut on October 23 and a U.S. attack on October 26 on Cubans building an airfield in Grenada. A poll on October 26 showed that 54 percent of respondents approved the decision to send U.S. Marines into Lebanon, compared to 67 percent on October 27. On October 26, 65 percent of respondents viewed the government of Grenada as a threat to its neighbors, whereas on October 27, 76 percent did so; and 50 percent said sending troops was the best policy, compared with 33 percent the day before. This research question poses considerable difficulties in terms of history, maturation, instability,

and regression artifacts. Attitudes may have changed due to images of the attacks; they may have been trending upward or downward; or they may have changed due to other reasons, including even sampling methods between two survey polls. The authors seek to overcome some of these difficulties by analyzing respondents' attitudes before and after Reagan's speech for those who had and had not heard it. In this case, the control group is prone to strong selection biases—people who heard the speech could differ in important and highly relevant ways. However, analysis of relevant demographic and political differences showed that systematic differences between the "treatment" and comparison groups were negligible. By controlling for these potential biases, the authors claim an effect of presidential persuasion distinct from the possible artifact effects arising from the design.

Lewis-Beck and Alford's (1980) study is stronger still. These authors used regression discontinuity analysis to compare the introduction of different types of safety legislation on U.S. coal mining fatalities. The researchers found that two pieces of legislation, in 1941 and 1969, led to demonstrable improvements in the numbers and rates of mining fatalities, but a 1952 act showed no difference. Lewis-Beck and Alford explore the possibility that legislative change alone increased safety by considering whether increased monitoring, attention to the coal mining industry, or the number of accidents beforehand could serve as alternative explanations for the change in coal mining accidents they observed. They also control for regression artifacts, since if the first decline was due to the high number of preceding fatalities, we should not find the same decline following the 1969 legislation. There was no intermediate rise in fatalities between 1941 and 1969. It is still possible that "history" intervened—both in the two cases of decline and in the case of no decline or incline. Perhaps new technologies were introduced, or there were other innovations. Stronger still would have been a comparison of the coal industry with a similar industry or a coal industry in another country where occupational safety legislation was not introduced. Lewis-Beck and Alford do explore the same period of time series with the growth of the federal coal mine health and safety budget and find significant increases after the first and third bills but not the second. By so doing, the authors build a plausible story about the role of strong (and financially backed) legislation relative to weak (without financial increase) legislation, giving rise to an account where the role of other historical interventions or further concerns over internal validity are no longer convincing confounding effects.

Stoker and Jennings's (1995) study differs most clearly because marriage is not an externally determined event. It is a life choice that could introduce very strong selection biases into a simple comparison of married and single people's voting and participation. Stoker and Jennings's study combines regression discontinuity with interrupted time series and a control series design. They compare couples' voting and political participation in election years in which they were yet to be married with those in which they were married and find differences in propensities with the amount of time married. They use an equivalent control group. They do not compare married respondents with single respondents, but compare married respondents with the same "yet to be married" respondents in earlier time points.

By extending their analysis to multiple time points, the persuasiveness of the findings is enhanced. It is unlikely that the results are explained by fast-to-marry couples' being more likely to engage in political activities, because the same "already married/not yet married" comparisons are found among older couples, too. The study tackles threats of selection bias, maturation, history, instrumentation, and measurement and overcomes further internal validity threats by testing and retesting the theories across cohorts, time, and forms of participation. No experimenter could introduce an effect of marriage, and so here, the strengths of internal validity combined with the strengths of external validity suggest greater potential comparative design strengths to randomized experimental designs, similarly executed, with self-conscious care and attention to the limitations of the design.

Experimental Studies with Weak Internal Validity

Potential experimenters should also consider that randomized experiments are not immune to some of the problems of internal validity in quasi-experimental studies. Campbell's (1973) threats of history, maturation, instability, and experimental mortality could be shared by experiments and quasi-experiments alike (although the presence of randomized control and treatment groups should ensure that these threats are distributed evenly in each, meaning treatment effects are usually still detectable).

Randomization does not imply equal mixes of key factors in each sample, but simply means that different compositions occur by chance. The smaller the number and size of treatment groups, and the more homogeneous those groups, the greater the likelihood that the number of plausible confounding factors will be greater than the number of experimental groups. To illustrate this, we can return to the hypothetical example. Suppose, due to cost and time, we chose six schools. In two schools, no education program is delivered (the pure control), but researchers conduct an education program in another two (the treatment group), and in two others, only a "normal" program of lessons is conducted (a partial control). We could allocate the schools purely at random. The researcher would then compare the effects of these interventions on rates of test score results over time. Using a small number of large treatment groups, chosen as closely to random as possible, introduces possibilities for confounding effects. The schools in the three groups may already exhibit different upward or downward trends by group or effects due to different catchment areas, demographic compositions, and so on. We could try to control for these possibilities. Instead of selecting schools at random we may seek to match schools on income level or geographical density. Yet we would need to match on a potentially infinite and impractical number of factors to rule out all potential selection-maturation effects. Obtaining more cases than independent variables is often unavailable to the researcher wishing to use randomized field experiments. Staged innovation, where treatments are introduced at different times, is another way of implementing interventions in an experimental design,

but again, the number of units must be sufficient to exclude all possible alternative explanations for differential maturation and responsiveness effects. Using small units and large samples, randomized experiments are clearly distinct from quasi-experimental studies, but in cases where large units and small samples are essential, these types of experimental design share more potential threats to internal validity. Such randomized experiments should be conducted in as wide an array of contexts as possible. If similar results are obtained across many contexts, we can be more certain of their conclusions.

One final word relates to our understandings of causation. We have assumed that causation is a change of an outcome in the simple presence of an intervention. If pupils and schools are targeted with an innovative education program, some pupils may be more or less likely to improve test scores as a result, but for these programs to be designated as effective, their introduction in the form of a treatment should have a direct and sufficient effect when there is a change in outcome. Yet, sometimes (or always!) causation is not so straightforward. Let's assume that certain educational campaigns can be very effective, but their effectiveness depends only upon the confluence of several factors. It may be the case that children respond to some types of educational program in some subjects but only when they are backed up by educational support in other subjects, or when the school implements support packages or after school clubs, or when teachers buy into the methods of learning. These could be mediating factors, which are difficult to vary at random; but they also may have direct effects, insufficient when they are present alone, but enough to cause an outcome when they are combined—when there is “multiple conjunctural causation” (Ragin 1987). Randomized experiments enable us to estimate causal effects of one intervention that can be varied, but then those effects require extensive replication. Randomized experiments can also, potentially, co-vary two factors—evaluating the potential mediating effect of a factor, although omitted variable bias threats remain. How much more difficult would a randomized experimenter find it to design an intervention in which several factors combine to create an effect? Although quasi-experimental designs will also fail to distinguish between the separate and conjunctural effects of many factors, they provide the only context in which such factors can realistically combine. The quest for external validity in such scenarios puts much more pressure on the pure experiment and greater reward on the quasi-experimental design.

Conclusion

Returning to Campbell's (1973, 221) statement that “true experimentation should almost always be preferred to quasi-experiments where both are available” and to Berlin and Solow's (2009) comparison of the certainty we expect for high-risk decisions such as healthcare and the standards we should therefore apply to evaluating social programs, do the intersections between experiments and quasi-experiments blur these clear distinctions as posed at the outset of this essay?

The best and most repeated randomized true experiments should almost always be preferred to the least rigorous quasi-experiments, and where both are “available” in their ideal forms, randomized experiments are indeed the gold standard. However, where quasi-experiments convincingly and effectively overcome threats to internal validity, their superior external validity outweighs the pure experimental research design where external validity in pure experiments cannot be achieved and where the confluence of causal factors necessary for an outcome cannot be re-created. Furthermore, where experiments are administered with randomly assigned units that are relatively large, and where small numbers of randomized groups are used, these designs can fail to overcome the same threats to internal validity that are more prevalent in quasi-experimental designs.

One of the motivations for exploring the intersections of quasi-experiments with experiments in this volume is the expanse of policy and of research questions in which pure randomization and replication will not always be possible. To take the examples explored in this article, it is often extremely difficult to re-create or randomize large-scale policy interventions such as school innovation in teaching design. We cannot randomize an event, or its effects, when that event is a legislative initiative, a presidential address, or an institutional choice and when those events happened in the past. Furthermore, it may not always be ethical to administer a treatment at random, when such experiences or interventions lead to certain harm, such as assigning cigarette smoking to a random sample of participants. Shadish, Cook, and Campbell (2002) argue that randomization is not often useful as a research design when an answer is needed very quickly. In this article, I have argued that quasi-experiments, carefully analyzed to maximize internal validity, provide opportunities to study events and changes on a scale and in a context sometimes impossible to capture in a purely experimental form.

The ideal design would, as is so often the case, be to combine causally attributable insights from randomized field experiments with logically attributable insights from the real world in which social phenomena occur, using quasi-experiments. Perhaps by appreciating the intersections between these approaches the combination of these two approaches can be more easily achieved. It is certainly more advisable to explore good opportunities with quasi-experiments than to pursue scant opportunities with pure experiments, just as the reverse is much more commonly argued to be true.

Notes

1. Indeed, a subsequent study was conducted by the Department of Health and Human Services, finding modest improvements in learning outcomes using analyses of 383 sites among around 4,600 children (cited in Besharov and Call 2009).

2. Campbell and Stanley (1963) highlight a more comprehensive list of threats to internal validity in quasi-experimental designs, but they stem from the nine listed above, cited in Campbell (1973).

3. Statistical matching techniques and instrumental variable analyses also allow researchers, in observational settings, to statistically “control” for potential sources of bias known to the researcher.

References

- Berlin, Gordon, and Robert Solow. 2009. Get the question right, then choose the method. *Journal of Policy Analysis and Management* 28 (1): 161–181.
- Besharov, Douglas J., and Douglas M. Call. 2009. Head Start falls further behind. *New York Times*, February 8.
- Campbell, Donald T. 1973. Reforms as experiments. In *Quasi-experimental approaches*, ed. James A. Caparaso and Leslie Roos, 187–225. Evanston, IL: Northwestern University Press.
- Campbell, Donald T., and Julian C. Stanley. 1963. *Experimental and quasi-experimental designs for research*. Boston: Houghton Mifflin.
- Cook, Thomas D., William R. Shadish, and Vivian C. Wong. 2008. Three conditions under which experiments and observational studies produce comparable causal estimates: New findings from within-study comparisons. *Journal of Policy Analysis and Management* 27 (4): 724–50.
- Cook, Thomas D., and Peter M. Steiner. 2009. Some empirically viable alternatives to the randomized experiment. *Journal of Policy Analysis and Management* 28 (1): 165–66.
- Gerber, Alan S., Donald P. Green, and Edward H. Kaplan. 2004. The illusion of learning from observational research. In *Problems and methods in the study of politics*, ed. Ian Shapiro, Rogers M. Smith, and Tarek E. Masoud. Cambridge: Cambridge University Press.
- Lewis-Beck, Michael S., and John R. Alford. 1980. Can government regulate safety? The coal mine example. *American Political Science Review* 74:745–56.
- Mannheim, Jarol B., and Robert B. Albritton. 1984. Changing national images: International public relations and media agenda setting. *American Political Science Review* 78:641–57.
- Ragin, Charles. 1987. *The comparative method: Moving beyond quantitative and qualitative strategies*. Berkeley: University of California Press.
- Shadish, William R., and Thomas D. Cook. 2009. The renaissance of field experimentation in evaluating interventions. *Annual Review of Psychology* 60:607–29.
- Shadish, William R., Thomas D. Cook, and Donald T. Campbell. 2002. *Experimental and quasi-experimental designs for generalized causal inference*. New York: Houghton Mifflin.
- Sigelman, Lee, and Alan Rosenblatt. 1996. Methodological considerations in the analysis of presidential persuasion. In *Political persuasion and attitude change*, ed. Diana C. Mutz, Paul M. Sniderman, and Richard A. Brody. Ann Arbor: University of Michigan Press.
- Stoker, Laura, and Kent Jennings. 1995. Life-cycle transitions and political participation: The case of marriage. *American Political Science Review* 89:421–33.