natural experiments or **quasi-natural experiments**: a *serendipitous* situation in which persons are assigned randomly to a treatment (or multiple treatments) and a control group, and outcomes are analyzed for the purposes of putting a hypothesis to a severe test; *also* situations that "approximate" this situation.

Contents

1 Introduction

The term "natural experiment" has been used in many, often, contradictory, ways. It is not unfair to say that the term is frequently employed to describe situations that are neither "natural" nor "experiments" or situations which are "natural, but not experiments" or vice versa.

It will serve the interests of clarity to initially direct most of our attention to the second term – experiment. A useful, albeit philosophically charged definition of an experiment "is a set of actions and observations, performed in the context of solving a particular problem or question, to support or falsify a hypothesis or research concerning phenomena."(Wikipedia 2005).

With such a broad definition in hand, it may not be surprising to observe a wide range of views among economists about whether they perform experiments or not. Smith's (1987) entry on "experimental methods" in the New Palgrave, for example, begins with the premise that "historically, the method and subject matter of economics have *presupposed* that it was a *non–experimental science like astronomy or meteorology* [emphasis added]." As he makes clear, his observation implies that *today*, economics is a experimental science. Bastable's (1925) entry on the same subject in the "Old Palgrave" overlaps only superficially with Smith's and divides experiments along the lines suggested by Bacon: *experimenta lucifera* in which "theoretical" concerns dominate and *experimenta fructifera* which concern themselves with "practical" matters. In sharp contrast to Smith, Bastable concludes that *experimenta lucifera* are "a very slight resource" in economics.

These two views of experiment, however, do not seem to helpful in understanding the controversy regarding natural experiments. "Experiment" is our context is merely the notion of putting one's view to the most "severe" test possible. A good summary of the the "spirit of experiment" (natural or otherwise) comes from the American philosopher Charles Sanders Peirce:

> [After posing a question or theory], the next business in order is to commence deducing from it whatever experimental predictions are extremest and most unlikely . . . in order to subject them to the *test of experiment.*
>
> The process of testing it will consist, not in examining the facts, in order to see how well they accord with the hypothesis, but on the contrary in examining such of the probable consequences of the hypothesis as would be capable of direct verification, especially those consequences which would be very unlikely or surprising in case the hypothesis were not true.
>
> When the hypothesis has sustained a testing as severe as the present state of our knowledge . . . renders imperative, it will be admitted provisionally . . . subject of course to reconsideration.[1]

[1] Peirce (1958) 7.182 (emphasis added) and 7.231 as cited in Mayo (1996). See Mayo (1996) for a nice exposition of this and related points.

1.1 The Philosophy of Experimentation in Natural Science

In the emergence of modern natural science during the 16th century, experiments represented an important break with a long historical tradition in which observation of phenomenon was used *in* theories as a way to justify or support *a priori* reasoning. In Drake's (1981) view: "The Aristotelian principle of appealing to experience had degenerated among philosophers into dependence on reasoning supported by casual examples among philosophers and the refutation of opponents by pointing to apparent exceptions not carefully examined."[2] Shadish et al. (2002) suggest that this "break" was twofold: First, experiments were frequently employed to "correct" or "refute" theories. This naturally led to conflict with political and religious authorities: Galileo Galilei's conflict with the Church and his fate at the hands of the Inquisition is among the best known examples of this conflict. Second, experiments increasingly involved "manipulation" to learn about "causes". Passive observation was not sufficient: As Hacking (1983) says of early experimenter Sir Francis Bacon: "He taught that not only must we observe nature in the raw, but that we must also 'twist the lion's tale', that is, manipulate our world in order to learn its secrets."

Indeed, at some level, in the natural sciences there has been comparatively little debate about the centrality of experiment.[3] Until the 19th century, the term experiment was typically reserved for studies in the natural sciences.

In the low sciences such as economics and medicine, the role of experiment is been the subject of extensive debate, much of tied up with debate on whether all the types of experiments possible in real science are possible in economics as well as well as debates about the many meanings of the word "cause."

A key distinction between much real science and economics involves the centrality of "randomization." No randomization is required, for example, to study whether certain actions will produce nuclear fission since "control" is possible: If a set of procedures applied to a piece of plutonium – under certain pre–specified experimental conditions – regularly produces nuclear fission, as long as agreement exists on the pre–specified conditions and on what constitutes plutonium, etc. it is possible to put the implied propositions to the type of severe test that would gain widespread assent – all without randomization. Put in a different way, randomization is only required when it is difficult to put a proposition to a severe test without it.

A related issue is whether a study of "causes" requires some notion of "manipulation." Most definitions of "cause" in social science involve some notion of "manipulation" (Heckman 2005) – Bacon's "twisting of the tail", so to speak. In physics, by way of contrast, some important "causes" do not involve manipulation *per se.* One might argue that Newton's law of gravitation was an example of a "mere" empirical regularity that became a "cause".[4] In

[2] For a useful introductory discussion with direct relevance to the human sciences and an extensive bibliography, see Shadish, Cook and Campbell (2002). Much of the historical account presented here follows their discussion.

[3] Ironically, it has typically been only philosophers of science who have downplayed the importance of experiment. Hacking (1983) makes a strong case that philosophers typically have exhibited a remarkably high degree of bias in minimizing their importance in favor of "theory."

[4] Leibniz objected vehemently to Newton's law: in an intellectual environment where the world could be understood as "mechanical pushes and pulls" Newton's law required the invocation of "occult pow-

this essay, we take the view that even if manipulation weren't necessary to *define* causality, manipulation is central to whether it is possible to discuss the idea intelligibly in social sciences and whether some kind of "severe test" is possible.(DiNardo 2005)[5]

2 Randomization: an attempt to evade the problems of imperfect"control"

If one accepts the centrality of manipulation (or something like it), it will not be surprising that the application of principles of experimentation to humans who have free will, make choices, etc. entails a host of issues that, *inter alia*, sharply constrains what might be reasonable to expect of experiments, natural or otherwise.

If it isn't possible, desirable, or ethical to "control" humans or their "environment" as it sometimes is in the natural sciences, is it possible to learn anything at all from experiment broadly construed? *Randomization* in experiments developed in part to try to evade the usual problems of isolating the role of the single phenomenon in situations. In the 19th century it was discovered that by the use of "artificial randomizers" (such as a coin toss) it was possible, in principle, to to create two groups of individuals which were the same "on average" apart from a single "treatment" (cause) which was under (at least partial) control of the experimenter. Hacking (1988) has observed that their use began primarily in contexts "marked by complete ignorance": the economist F.Y. Edgeworth was early to apply the mathematical logic of both Bayesian and "classical" statistics to a randomized trial of the existence of "telepathy."[6]

ers."(Hacking 1983). It is also interesting to note that although it a subject of debate, many have associated Voltaire's Dr. Pangloss with Leibniz. Dr. Pangloss "proved admirably that there is no effect without a cause . . . in this the best of all possible worlds" – a very different notion of cause! (Voltaire 1796), Chapter 1.

[5] Some philosophers have sought to *define* science around issues related to "control", arguing that the phenomenon that economists are trying to investigate are impossible to study scientifically at all. Philosophers have articulated numerous reasons for the difference between social and natural science. A few examples may be helpful: (Nelson 1990) argues, for example, that the objects of inquiry by the economist do not constitute "a natural kind". Put very crudely, the issue is the extent to which all the phenomena that we lump into the category "commodity", for example, can be refined to some essence that is sufficiently "similar" so that a scientific theory about commodities is possible in the same way as a "body" is in Newtonian mechanics. This is often discussed as the issue of whether the relevant taxonomy results in "carving nature at the joints." Hacking (2000) introduces the notions of "indifferent kinds" – the objects in the physical science – atoms, quarks, etc. with "interactive" kinds – the objects of study in medicine or the social sciences. We might interact with plutonium or bacteria, but neither the plutonium nor the bacteria are aware of how we are classifying them or what we are doing to them. This can be contrasted with "interactive kinds" which are aware and for which "looping" is possible. For example, mental retardation might lead to segregation of those so designated. This segregation might lead to new behaviors which then might not fall under the old label, etc. Consequently, investigation of such phenomena might be likened to "trying to hit a moving target." Searle (1995) on the other hand, notes that the objects of interest in social science while epistemologically objective, are ontologically subjective. While the loss of 100 dollars may be very "real" to someone, the notion of money requires groups of individual to accept money as a medium of exchange. Again the existence of atoms does not require us to recognize their existence.

[6] Although economists played an important role in the development of randomization, economists as a whole were quite slow to embrace the new tools. In an echo of debates that faced natural sciences in the 1600s, this was due in part "because the theory [of economics] was not in doubt, applied workers sought

Over time, the term "experiment" evolved to include both experiments of the "hard sciences" where a measure of control was possible as well as situations in which artificial randomizers were used to assign individuals (or plots of land, etc.) to different "treatments." A key role was played by R. A. Fisher and his seminal (1935) *Design of Experiments* as well subsequent publications which discussed the theory and practice of using artificial randomizers to learn about causes.

There are at least two key limitations of randomized experiments relative to experiments where "scientific" control is possible:

- Absent real control, one only has a weak understanding of the "cause" in question. For instance, one can do a randomized controlled trial of the effect of aspirin on heart failure understanding nothing of the mechanism by which aspirin affects the outcome. Moreover, it is clear that the experiment is "context specific." One's generalization about atoms in a laboratory often extends to atoms in other contexts in a way not possible in social science.

- Any single experiment – even under the ideal situation – does not always reveal the true answer. In the logic of randomized design, the usual inference procedure is merely one that *would* give the right answer on average *if* the experiment were repeated. At best, the true answer is just a "long–run tendency" in repeated identical experiments.

2.1 Social Experiments: Why not do a "real" randomized trial?

Even absent these limitations, there are a long list of reasons why economists frequently have little interest in randomized trials. The most important reason is that many of the real randomized experiments (often called "social experiments") of which one one could conceive (or have been implemented), are immoral or unethical. At a most basic level, who decides to "perform an experiment" and who "decides" or is recruited to be experimented upon often reflects deep–seated social injustice. Even Brandeisian (see below) experiments can take on a sinister cast – state governments surely do not consider the interests of all its citizens equally.

Indeed, historically the conduct of experiments on persons has told us as much or more about the structure of society than anything else: one well–known example is the "experiments" conducted by the U.S. Public Health Service from 1932 to 1972 on about 400 poor black men who had advanced syphilis. Among the aims of the experiment was to determine the the effect of untreated syphilis. To this end, the medical doctors misrepresented themselves to the subjects (the sons and grandsons of slaves) as providing free medical care. For example, when penicillin became the standard of care, the subjects were deliberately not provided the medication: rather, the doctors were content to observe the horrific progress of the disease as some went blind or insane.

Another set of reasons are practical – experiments are costly to administer. Another reason is attrition: often people drop out of such experiments (often in non–random ways), greatly complicating the problem of inference. A distinct, although sometimes related, reason is that results of social experiments involving randomization are sometimes difficult to

neither to verify nor to disprove."(Morgan 1987).

interpret. One often cited reason is that those recruited to participate in such experiments may be different than those for whom the policy is ultimately intended. In even the simplest experiments, "compliance" is imperfect. Not everyone assigned to a treatment takes it up – indeed, it is often the case that analysis is done on an "intent to treat" basis. That is, those "assigned" to treatment are compared to those assigned to the control whether or not those assigned to treatment actually "took" the treatment. Another often cited reason is that what is likely when a social experiment is conducted with a small number of persons might be very different when applied to much larger numbers of persons. Persons, unlike atomic particles, have free will. In the world of persons, the "experiment" doesn't necessarily stop after the experimenters have made their observations.[7]

3 Types of Natural Experiments

Thus far we have seen that the word "experiment" can be used in two very different senses: one to denote situations where real "control" is possible and second involving artificial randomizers. As a consequence, the term "natural experiment" has been used in very different senses. I take a look at the origins of the term and the different ways the term has been used, although we focus on natural experiments most frequently arising in economics.

3.1 Natural Experiments in Natural Science

An early use of the term "Natural Experiment" in English describes an investigation into the functioning of "nature". The term comes from a translation "Saggi di naturali esperienze fatte nell'Accademia del Cimento" published in Italian in 1667 which appeared in an English translation by Richard Waller in 1684 as "Essayes of natural experiments made in the Academie del Cimento"(Waller 1964). The short-lived Accademia del Cimento was founded in Florence in 1657 by the Medici brothers, Prince Leopold and Grand Duke Ferdinand II and the *Saggi* record a small subset of the large number of experiments by the Cimento that involved such issues as "smells do not traverse Glass", and "the failure to confirm Existence of Atoms of cold." Although the experiments of the Academy included trials involving humans, they did not involve randomization. Indeed, the legacy of these investigations into humans is more relevant to the study of 16th century culture and authority relations than 16th century science.[8]

Over time, in the hard sciences, the term natural experiment has also come to describe both cases where "nature" provides an experiment that resembles the controlled situation that scientists would like observe but are unable to create themselves. An unsuccessful experiment may help make the point clear: in a famous quote by Albert Einstein to Erwin

[7] For example, even in the context of a true randomized experiment, those denied treatment often have the opportunity to find it elsewhere. (See Heckman and Smith (1995) for one discussion of the merits of randomized trials in the social science with references.)

[8] Tribby (1994), for example, discusses an investigation into a "gentler" laxative that could "satisfy" the needs of Grand Duke Ferdinand II as well as those of the many "delicate persons" who visited or had dealings with the court that involved experimentation on individuals described variously as "a mercenary", "a vagrant", "the Little Moor", etc.

Findlay Freundlich (who was attempting to assess the whether path of ray of light was affected by gravity), Einstein wrote: “If only we had a considerably larger planet than Jupiter! But nature has not made it a priority to make it easy for us to discover its laws.”

3.2 Natural Experiments as Serendipitous Randomized Trials

In contrast to the natural experiment of the hard sciences, the term natural experiment is often used by economists to denote a situation where real randomization was employed, without the intent of providing a randomized experiment. For example, in 1970 –1972 men from specific birth cohorts were conscripted into the U.S. military by way of a draft lottery. Each day of the year was randomly assigned a number which (in part) determined whether or not you were at risk of being inducted into the military service to fight in the U.S. war on Indochina. As a consequence, men of specific birth cohorts born only a day apart, for example, had very different risks of serving in the military. In Hearst, Newman and Hulley (1986), the authors asked whether the war continued to kill after the warrior returned home. The authors compared, among other things, the suicide rates individuals who on average were *ex ante* similar, but who had very different probabilities of having completed military service.

The example is sufficiently simple to make a number of points about the limitations of natural experiments. *If* one can assume that the mere fact of having such a birth date that put one at high risk of military duty, and that having a birth date raised (or did not lower) any person’s risk of serving in the military, then it is possible to use something akin to two stage least squares (2SLS) to estimate an “average” effect of military service for those who were induced to serve in the military by the draft lottery. However, Hearst et al. (1986) are quick to observe that *whether or not* one actually served in the military, the mere fact of having been put at risk of the lottery might have had an effect on delayed mortality.[9]

Returning to how one might go from an estimate generated this way to more general inference, one has a number of other obstacles. For example, the delayed mortality effects of military service on those *induced* to serve by an unlucky birth date might be different than the effect on those who *volunteered* to fight in the war. If the effects are very different, it would obviously be incorrect to use estimates generated by those induced to serve to extrapolate to the broader population of interest.

More generally, our ability to generalize the valid results of an experiment is much more limited when we can only manipulate the cause indirectly (such as in the example above) than in the we can manipulate the cause directly: there is often the possibility that there are important differences between persons who take up the treatment as a result of having been encourage to do so and those who were similarly encouraged but did not take up the treatment.

[9] In econometric terms, this would be a violation of the “exclusion restriction” of 2SLS. If such is the case, it is apparent that a comparison of men with high risk birthdays to those with low risk birthdays will be an admixture of the effect of the military service on later mortality *and* any direct effect of the lottery itself. An additional problem is the possibility of non-random selection induced by men dying while at war. This was judged to be small due since the fraction of U.S. soldiers who died while serving in action was a small fraction of the total.

3.3 The Regression Discontinuity Design as a Natural Experiment

One research design which involves the "serendipitous" randomization of individuals into a treatment is called the regression discontinuity design. Since it is a relatively "clean" example of something that approaches a truly randomized experiment without involving explicit randomization, it provides a good illustration of the strengths and weaknesses of natural experiments.[10] For illustration let's consider trying to analyze the causal effect of "unionization" on firms in the U.S. The naive approach would be to compare unionized firms to non-unionized firms.

The basis of the regression discontinuity design is the existence of a "score" or a "vote" which assigns persons to one treatment or another. In the U.S. context, workers at a firm can win the right to form a labor union by means of a secret ballot election. If fifty percent plus one of the workers votes in favor of the union, the workers win the right to be represented by a union. Less than that, and they are denied such rights.

To understand how this works consider elections at two different sets of work sites that employ large numbers of workers. In one set, $0.5 + \Delta$ of the workers vote in favor of the union and win the right to bargain collectively where Δ is some small number. In another set, slightly less than fifty percent vote in favor of the union, and are denied the right to bargain collectively. The vote share in these sites is $0.5 - \Delta$. Suppose we have large amounts of data on such elections and can accurately estimate the average outcome (say the fraction of firms that continue to exist 15 years after the vote).

Using almost exactly the same set up as before, we compare those places where the union wins to those where the union loses:

$$E[\overline{y}_{\text{Union}} - \overline{y}_{\text{No Union}}] \;=\; E[y|\text{vote} = .5 + \Delta] - E[y|\text{vote} = .5 - \Delta]$$

If firm survival is described by the same "model" as in $\star$ above, where now $T = 1$ denotes winning the right to bargain collectively, we get:

$$\begin{aligned} E[\overline{y}_{\text{Union}} - \overline{y}_{\text{No Union}}] \;=\; & \beta + \left(\underbrace{E[f(X)|\text{vote} = .5 + \Delta] - E[f(X)|\text{vote} = .5 - \Delta]}_{\text{Observable Differences}} \right) + \\ & \left(\underbrace{E[\epsilon|\text{vote} = .5 + \Delta] - E[\epsilon|\text{vote} = .5 - \Delta]}_{\text{Unobservable Differences}} \right) \end{aligned}$$

The "trick" is that if we choose Δ to be small enough (i.e. close to zero), then

$$E[f(X)|\text{vote} = .5 + \Delta] \;\approx\; E[f(X)|\text{vote} = .5 - \Delta]$$

and

$$E[\epsilon|\text{vote} = .5 + \Delta] \;\approx\; E[\epsilon|\text{vote} = .5 - \Delta]$$

[10] For an analysis of the relationship between the regression discontinuity design and randomized controlled trials see Lee (2006).

and we get a “good” estimate of the “effect of unions” in the same sense that we get a good estimate of the effect of a treatment in a randomized controlled trial. That is, if we focus our attention on the difference in outcomes between “near winners” and “near losers” such a contrast is formally equivalent to a randomized controlled trial if there is at least some “random” component to the vote share. For example, sometimes people take ill on the day of the vote – if that happens randomly in some sites, two sites which otherwise would have had the same final vote tally had everyone shown up are now different. When such differences are the difference between recognition or not, one has the practical equivalent of a randomized controlled trial.[11]

A few moments’ reflection will make clear both the appeal of such experiments and their limits. Advocates of a natural experiment approach point to the fact that the implicit randomization involved in this design means that we can be more confident with such a comparison than a naive comparison that merely compares unionized to non–unionized firms. This would almost certainly confound the true “effect” with pre–existing differences in unionized and non–unionized firms with “unionization.” Advocates will also point to the fact that the experiment is relevant to a potential policy – say lowering the threshold required to win representation rights by a small amount.

Detractors will observe many limitations. Is the effect of a union that is set into a place by a 51 percent vote the same as the effect of a union where the workers vote unanimously? Possibly not. Stipulating the validity of the estimate, is it reasonable to suggest that the effect of unionization would be the same if all workplaces were allowed to vote on a union? Probably not. Is it possible that a union at one worksite affects other worksites? What about the effect on the firm’s competitors. Indeed, it is even possible to question the premise that a union is a “treatment” at all. Does it make sense to talk of a single effect of a labor union when there is such heterogeneity in what the notion “labor union” represents. Is the I.W.W. of Joe Hill (a famous radical labor union) the same thing as AFL-CIO of George Meany (a “conservative” union?)

More generally, “causes”, “treatments”, etc. are much more fragile objects for the types of things usually interesting to economists than the types of things interesting in natural science. The concepts of natural science are often capable of quite substantial refinement in a way that concepts in the human sciences rarely are.

3.4 “Natural Natural Experiments”?

As I have already mentioned, the term “natural experiment” has been used in several different ways inconsistent with our definition. It seems silly, however, to claim that our definition is the “true” or correct one. It will therefore behoove us to consider some cases which the term is used that do not obviously involve randomization of a treatment or something that approximates such randomization.

[11] The mere existence of a “score” that discontinuously exposes one to a treatment is not enough. This design would not be appropriate, for example, to analyze the causal effects of U.S. Congressional votes on various issues. Substantial “manipulation” – i.e. through negotiation, etc – of the final vote tally is common and suggests that individuals near but on opposite sides the threshold are not otherwise similar. See the entry on “regression discontinuity.”

Rosenzweig and Wolpin (2000) for instance, have coined the expression “natural natural experiments” to denote a wide range of studies involving the use of twins. The emphasis on the word natural is intended to highlight the role of “nature” in providing the variation. Twins have been of inordinate interest to the social scientists since they seem to offer the possibility of “controlling” for “genetics” Consider one case of interest to economists: “returns to schooling.” Does acquiring an additional year of school result in higher wages in the labor market? How much higher? To fix ideas consider a simple model of the sort:

$$y_{ij} = \beta S_{ij} + a_j + \epsilon_{ij}$$

We are interested in some outcome, say hourly wages, and the causal effect of years of schooling S. It will greatly simplify the discussion if we assume that all persons “treated” with “schooling” experience the same increase in their wages – i.e. the treatment effect is a constant across individuals. We have gathered a random sample of $j = 1 \ldots J$ “identical” (monozygotic) twins ($i = 1, 2$). The term a_j is not directly observable but includes everything that the twins have in common – genetics, environment, etc. The error term ϵ_{ij} includes everything that the twins do not have in common and can’t be observed as well as the effects of misspecification, etc. Though this simple setup can be greatly elaborated (see Ashenfelter and Krueger (1994) for a clear exposition) the essential idea is that the *difference* between the twins purges the outcome of the a_j term so that an ordinary least squares regression of the difference in wages Δy_{ij} on ΔS_{ij} yields a good estimate of

$$\widehat{\beta} \quad \text{is a good estimate of} \quad \beta + \frac{\text{Var}(\Delta\epsilon, \Delta S)}{\text{Var}(\Delta S)}$$

The first term is the goal of such studies. The second term points to the possibility that there are other influences which might be correlated both with schooling and that affect the outcome. [12]

When will $\widehat{\beta}$ to be a good estimate of the returns to schooling β? The conditions are essentially the same as for the randomized controlled trial: if we can treat the assignment of schooling to the two twins as if it were determined by a random coin toss then differences in the level of schooling between the two twins – ΔS_{ij} – will be independent of differences between the two twins in unobserved influences on wages – $\Delta\epsilon_{ij}$. Detractors of this approach doubt that such an assumption is plausible. In simple language, if the twins are so “identical” why do they have different levels of schooling? Perhaps the parents noticed that one twin was more interested or had more “aptitude” for school work than another. If that were the case, estimates of the returns to schooling would be confounded with differences in the aptitude for schooling despite the fact that we had “controlled” for a large number of other factors. The key difference between this case and what I have identified as a natural experiment is the lack of an obvious approximation to randomization.[13]

[12] The second term can be viewed as the slope coefficient from this *gedanken* OLS regression: $\epsilon = \text{constant} + S\delta + \text{error}$

[13] Bound and Solon (1999) discuss *inter alia* a host of difficulties in treating twin differences as experimental variation. I do not discuss twins studies which utilize twins as a “surprise” to family size which have some element of randomization.

3.5 Other Research Designs: Quasi-experiments

Finally, I should make note of the fact that some authors use the term natural experiment more broadly than I have construed it here. Meyer (1995) for instance, considers natural experiments the broad class of research designs "patterned after randomized experiments" but not (generally) involving actual randomization. One term often used for such situations is "quasi–experiment." The relationship between these quasi-experiments and the natural experiments I have been describing is quite varied and ranges from those whose difference from the standard of randomized assignment is merely a matter of "degree" to those which in which assignment to treatment differs so much from the standard of randomization that it is really a difference in in "kind."

Most of these quasi-experiments are variants of a "before and after" where an observation is made before and after a treatment. Often a before–after comparison for one set of observations (the treatment – T) is compared to another set (the control –C). A typical setup might compute a treatment effect by taking the difference in two differences:

$$\text{Treatment Effect} = \left\{\overline{y}_{T,\text{after}} - \overline{y}_{T,\text{before}}\right\} - \left\{\overline{y}_{C,\text{after}} - \overline{y}_{C,\text{before}}\right\}$$

For this reason, such quasi-experiments are described as using "difference–in differences" approach to identifying a causal relationship.

In the U.S. the fact that the state (or city) governments of have some liberty to enact laws independently of the federal government, for example, has led to a great deal of research using "Brandeisian" experiments. The term comes by way of U.S. Supreme Court Justice Louis Brandeis, in the case *New State Ice vs Liebmann*

> There must be power in the States and the Nation to remould, through experimentation, our economic practices and institutions to meet changing social and economic needs. . . . It is one of the happy incidents of the federal system that a single courageous State may, if its citizens choose, serve as a laboratory; and try novel social and economic experiments without risk to the rest of the country.

To give one such example, consider evaluating the effect of changing the age at which it is legal to purchase alcohol on the consumption of marijuana. At the beginning of the 1980s states generally two types of legal regimes. In one set, alcohol could not be legally sold to those under the age of 21. In another, the legal minimum drinking age (LMDA) was 18. In the mid 1980s, the Federal Government put a great deal of pressure on those states with legal minimum drinking ages of 18 to raise them to 21 and by the end of the 1980s, all states had 21-year old drinking ages.

The assignment of drinking age statutes to the states at the beginning of the 1980s could not be considered "approximately" random[14]. However, due to a Federal Policy of implemented in the mid 1980s of eventually denying Federal Highway funds to states with legal minimums less than 21-years old, something perhaps approximating an "experiment"

[14] Utah, for example, which is home to a large number of adherents to the Mormon religion – which proscribes alcohol use – had a 21-year drinking age at the beginning of the 1980s

can be arrived at by comparing *changes* in alcohol or marijuana consumption during the 1980s in those states who were forced to change (and changed early) to those who were forced to but raised their drinking age later.

Let Δy_t denote the change in the fraction of 18-21 year olds who reported smoking marijuana in the last month from 1980 to 1990 in states that had 18 year old drinking ages that were increased, and Δy_c denote the similar change in states whose drinking age was always 21-years old. Then an estimate of the effect of the drinking age might be:

$$\Delta y_t - \Delta y_c = \text{Effect of LMDA}$$

Although randomization is not employed *per se*, the credibility of these exercises can be at least partially evaluated. For instance, if the outcome of interest has been approximately constant in both the treatment and control groups for a long time preceding the change in legal regime the estimate is generally more credible. Less credible is the case in which the outcomes in the control group and the treatment group are quite variable over time, the control group and the treatment group do not follow similar patterns *before* the proposed experiment, or when both are true.

4 Controversies: Concluding Remarks

Natural experiments and their kin have been at the heart of much work in economics. Nonetheless, they are the subject of considerable debate. One of the most cited limitations of natural experiments – by both supporters and detractors – is that such experiments are context specific. Indeed, one frequently encountered "strength" of natural experiments is that often concern the evaluation of an actual policy. Nonetheless, there are limitations. Assuming that the experiment is "internally valid" we still have to ask: How do we generalize from one experiment to the broader questions of policy? In the foregoing has suggested that it is difficult. There are at least three broad class of reasons:

1. While a natural experiment might provide a credible estimate of some particular serendipitous "intervention", this may have only a weak relation to the type of interventions being contemplated as policies. Many of potential reasons for a weak relationship are similar to those encountered in social experiments.[15]

2. Some interesting questions are unanswerable with such an approach because serendipitous randomized experiments are few and far between. The extent to which this criticism is warranted, of course, depends on the availability of alternative ways of putting our views to a severe test.

3. More generally, without a "theory", estimates from natural experiments are uninterpretable.

[15] Among other things, for example, the effect of a treatment in a demonstration program might be quite different from the outcome that would obtain if the treatment were applied more broadly or to different persons.

I am sympathetic with all three criticisms although (3) deserves some qualification. While it has been argued that even in the natural sciences it is impossible to have “pre–theoretical” observations or experiments, Hacking (1983) makes a strong case that experimentation has a life of its own: sometimes suggesting ideas in advance of theory, other times the consequence of theory, and sometimes testing theories. Much of this debate in the natural sciences revolves around the notion of what constitutes a “theory.” Whatever the validity of the view that one can not experiment in advance of “theory” in the natural sciences, in the social sciences, it is clear that no theory has the same standing as, say, general relativity in physics. This is the sense which Noam Chomsky observes that “as soon as questions of will or decision or reason or choice of action arise, human science is at a loss.” Indeed, the standing of randomized experiments – in some fields of inquiry regarded as “the gold standard” of evidence – is a great deal lower than the best experiments of natural science; they are most often useful in situations otherwise marked by “complete ignorance.”(Hacking 1988) In short, while the human sciences might have the same ambition as natural science, the status of what we know will almost surely be quite limited.

Nonetheless, one does not need a “correct” theory in hand, nor an understanding as rich as that found in some of the natural sciences to find an experiment useful. At the risk of over–using such metaphors, the fact that the Michelson–Morley experiments were in part about testing for the existence of “ether” did not make them uninteresting. Experiments are just ways to use things we (think we) understand to learn about something we do not. And while the sorts of “natural” experiments “serendipitously” provided by society may be very limited and are often the product of unhappy social realities, they can sometimes perhaps serve a small role in enhancing our understanding.

Any assessment of the usefulness of natural experiments depends on how one judges the power of other methods of inquiry. Such a discussion is well beyond the scope of this essay. Nonetheless, not discounting their many limitations, one benefit of natural experiments I have tried to highlight is that for some they might open up the possibility of revising their beliefs in light of evidence or suggest new ways to think about old problems, however limited. A key aspect of experiments (natural or otherwise) is the willingness to put one’s ideas “to the test.” Often careful study of a natural experiment, however limited, may also make one aware of how complicated and difficult are the problems we call “economics.” Even if the success we might have in generalizing natural experiments more broadly may be quite limited, if they bring nothing but humility to the claims social scientists make about much we actually understand, that alone would justify an interest in natural experiments.

5 Notes

The example of the effect of the legal minimum drinking age comes from DiNardo and Lemieux (2001). The example on the use of the regression discontinuity design comes from DiNardo and Lee (2004) and DiNardo and Lee (2002).

References

Ashenfelter, Orley and Alan B. Krueger, "Estimates of the Economic Returns to Schooling from a New Sample of Identical Twins," *American Economic Review*, December 1994, *84* (5), 1157–1173.

Bastable, Charles F., "Experimental Methods in Economics," in Henry Higgs, ed., *Dictionary of Political Economy*, Palgrave's Dictionary of Political Economy, 1925, chapter Experimental Methods in Economics.

Bound, John and Gary Solon, "Double Trouble: On the Value of Twins–Based Estimation of the Return to Schooling," *Economics of Education Review*, April 1999, *18* (2), 169–182.

DiNardo, John, "A Review of Freakonomics," Unpublished Manuscript, University of Michigan January 15 2005.

___ **and David S. Lee**, "The Impact of Unionization on Establishment Closure: A Regression Discontinuity Analysis of Representation Elections," Working Paper 8993, National Bureau of Economic Research June 2002.

___ **and** ___ , "Economic Impacts of New Unionization on Private Sector Employers: 1984-2001," *Quarterly Journal of Economics*, November 2004, *119* (4), 1383 – 1441.

___ **and Thomas Lemieux**, "Alcohol, marijuana, and American youth: the unintended consequences of government regulation," *Journal of Health Economics*, November 2001, *20* (6), 991–1010.

Drake, Stillman, *Cause, Experiment, and Science: A Galilean Dialogue, Incorporating a New English Translation of Galileo's Bodies That Stay Atop Water, or Move in It*, Chicago: University of Chicago Press, November 1981.

Fisher, Sir Ronald Aylmer, *Design of Experiments*, Edinburgh, London: Oliver and Boyd, 1935.

Hacking, Ian, *Representing and intervening: Introductory topics in the philosophy of natural science.*, Cambridge, England: Cambridge University Press, 1983.

___ , "Telepathy: Origins of Randomization in Experimental Design," *Isis*, September 1988, *79* (3), 427–451.

___ , *The Social Construction of What?*, Cambridge, MA: Harvard University Press, 2000.

Hearst, Norman, Tom B. Newman, and Stephen B. Hulley, "Delayed Effects of the Military Draft on Mortality:A Randomized Natural Experiment," *New England Journal of Medicine*, March 6 1986, *314*, 620–624.

Heckman, James J., "The Scientific Model of Causality," Unpublished Paper, University of Chicago, University College London, and the American Bar Foundation April 28 2005.

____ **and Jeffrey A. Smith**, "Assessing the Case for Social Experiments," *Journal of Economic Perspectives*, 1995, *9* (2), 85–110.

Lee, David S., "Randomized Experiments from Non-random Selection in U.S. House Elections," *Journal of Econometrics*, Forthcoming 2006.

Mayo, Deborah G., *Error and the Growth of Experimental Knowledge* Science and Its Conceptual Foundations, Chicago: University of Chicago Press, 1996.

Meyer, Bruce, "Natural and quasi-experiments in economics," *Journal of Business and Economic Statistics*, April 1995, *13* (2), 151–161.

Morgan, Mary S., "Statistics without Probability and Haavelmo's Revolution in Econometrics," in Lorenz Krüger, Gerd Gigerenzer, and Mary S. Morgan, eds., *The Probabilistic Revolution: Ideas in the Sciences*, Vol. 2, Cambridge, MA: The MIT Press, 1987, chapter 8, pp. 171–200.

Nelson, Alan, "Are Economic Kinds Natural?," in C. Wade Savage, ed., *Scientific Theories*, Vol. 14 of *Minnesota Studies in the Philosophy of Science*, Minneapolis: University of Minnesota Press, 1990, pp. 102–135.

Peirce, Charles S., in A. Burks, ed., *Collected Papers*, Vol. 7–8, Cambridge, MA: Harvard University Press, 1958.

Rosenzweig, Mark R. and Kenneth I. Wolpin, "Natural 'Natural Experiments' In Economics," *Journal of Economic Literature*, December 2000, *38* (4), 827–874.

Searle, John, *The Construction of Social Reality*, New York: The Free Press, 1995.

Shadish, William R., Thomas D. Cook, and Donald T. Campbell, *Experimental and Quasi–Experimental Designs for Generalized Causal Inference*, Boston: Houghton Mifflin Company, 2002.

Smith, Vernon L., "Experimental Methods in Economics," in John Eatwell, Peter Newman, and Murray Milgate, eds., *New Palgrave : A Dictionary of Economics*, Palgrave Macmilan, 1987, pp. 241–249.

Tribby, Jay, "Club Medici: Natural Experiment and the Imagineering of "Tuscany"," *Configurations*, 1994, *2* (2), 215–235.

Voltaire, *The History of Candid; or All for the Best*, Cooke's ed., London: C. Cooke, 1796. Translated from the French of M. Voltaire. Embellished with superb engravings.

Waller, Richard, "Essayes of Natural Experiments Made in the Academie Del Cimento, Under the Protection of the Most Serene Prince Leopold of Tuscany," A facsimile of the 1684 Edition with a new Introduction by A. Rupert Hall, New York and London 1964. Translated by Richard Waller. Written in Italian by the Secretary of the Academy.

Wikipedia, "Experiment — Wikipedia, the free encyclopedia," [Online: Accessed September 28, 2006] 2005.