PROJECT MUSE®

The Inheritance bequeathed to William G. Cochran that he
willed forward and left for others to will forward again:
The Limits of Observational Studies that seek to Mimic
Randomized Experiments

Thomas D. Cook

Observational Studies, Volume 1, Number 1, 2015, pp. 141-164 (Article)

Published by University of Pennsylvania Press

DOI: <https://doi.org/10.1353/obs.2015.0012>

➔ *For additional information about this article*

<https://muse.jhu.edu/article/793411/summary>

The Inheritance bequeathed to William G. Cochran that he willed forward and left for others to will forward again: The Limits of Observational Studies that seek to Mimic Randomized Experiments

Thomas D. Cook

t-cook@northwestern.edu

Northwestern University &

Mathematical Policy Research, Inc.

Introduction

Seamus Heaney had the courage to want to use the past to construct a better future from among the many potential futures always available. In “The Settle Bed” of 1991 he wrote:

And now this is “an inheritance” –
Upright, rudimentary, unshiftable planked
In the long ago, yet willable forward
Again and again and again.

To etymologists, “upright” connotes full of rectitude, being correct, while “rudimentary” connotes beginnings and first principles. Together, they signify that Heaney’s concern is with inheritances that address fundamental issues and want to be right about them. That all things are linked to the past is a truism, but that many things are “*unshiftable* planked” there is not. Firmly rooted inheritances require generation-transcending transmission mechanisms, whether objects like books, or cognitive habits like theories, or social institutions like universities, or subsequent generations like one’s own students and students’ students. Such mechanisms transcend lives, including those of individual scholars. Heaney insists that unshiftable planked inheritances have to be willed forward, not just once or twice, but “again and again and again”. So he rejects ossifying traditions that fail to accommodate new realities and instead recommends using human will to make continuous changes in an inheritance. Presumably this is by implementing possible changes, learning about their successes and failures, and incorporating their results into the original inheritance that is thereby modified. The consequence is an inheritance that is planked in the past, modified in the present, and repeatedly modified in the future by dint of human will.

Cochran was himself the beneficiary of an important inheritance that he improved and passed on. It does not diminish him to note that his writings on causation and the design of experiments and observational studies are inconceivable without Fisher. The two did not work together at Rothamsted, but Fisher often returned there after his move to Cambridge and they talked there as well as at Royal Society meetings ([Watson, 1982](#)). We even have Cochran’s own reports of conversations with Fisher, including the insight that observational studies should make the implications of a single causal hypothesis more elabo-

rated in the data (as reported, for example, in [Rosenbaum, 2005](#)). Yates mentored Cochran at Rothamsted and had earlier been a colleague of Fisher there ([Yates, 1964](#)). It seems inconceivable that Cochran and Yates did not speak in detail about Fisher’s contributions, given its general intellectual resonance and its clear relevance to their own research agendas and the mission of their research station. Cochran brought detailed knowledge of Fisher to his first stay at Iowa State; and after he emigrated he continued this dissemination process at Iowa State again and then in North Carolina, at Johns Hopkins and at Harvard. His dissemination of Fisher took place face-to-face, in teaching and in writing, perhaps most saliently in his seminal texts with Cox and Snedecor. Fisher himself did not become an intellectual giant out of nothing, but he nonetheless created most of the planks on which Cochran first stood. Here are what I think are the five major ones.

(1) *It is legitimate for statisticians to choose their intellectual problems from among the practical problems faced by those who seek to improve the physical world, be they farmers, health workers, engineers or social welfare professionals.* The main alternative to such practice-based problem choice is when research agendas emanate from puzzles in existing statistical theory or from past and emerging issues in mathematics.

(2) *Many of the practical problems practitioners face require identifying whether a manipulable action is “causally” related to possible consequences.* The inference entailed here moves Statistics beyond its historical concerns with reducing uncertainty and improving prediction ([Stigler, 1986](#)). It legitimates research on causal bias and its control, the central point of this paper. But it also legitimates working with what is perhaps a less fundamental theory of causation. Variouslly labeled by philosophers of science as the activity, manipulability or recipe theory of causation (see [Cook and Campbell, 1979](#)), it seeks the valid description of concrete If/Then connections rather than explanations of why things happen in the world, including causal connections.

(3) *Causal bias is best controlled through experimental design.* Such design requires clear null hypotheses that are then tested through a combination of how units are allocated to the treatments, how and when the study outcome is assessed, and how comparison groups are selected. The hope is that such structural elements will provide a perfect no-treatment counterfactual against which observed performance in the treatment group can be compared. Then, no other interpretation of the relationship between the independent and dependent variable is possible other than that it is causal. Many non-experimental causal methods exist, primarily “causal modeling” methods where substantive theory about a set of interdependent temporal influences is tested, essentially by estimating how well the obtained and predicted data match.

(4) *Within experimental design, random assignment is the best tool for warranting unbiased counterfactual estimates.* Random assignment balances the study groups on all observed and unobserved variables, entailing a perfect counterfactual in expectation and a probabilistically equivalent counterfactual in individual studies. Random assignment is also demonstrably implementable in real-world settings where its assumptions are often clearly met. There are other cause-probing “experimental” traditions that do not require random assignment. For instance, the experimental laboratory sciences routinely test causal hypotheses, but they rule out alternative causal interpretations by virtue (1) of closed-system test settings and apparatus that physically exclude most contending hypotheses, (2) of substantive theories whose predictions are so numerically or functionally specific that no

contending theory can explain them, and (3) of implementing procedures that have evolved, and are still evolving, to reduce the confounds from the laboratory researchers' own hopes, expectancies and interests. "Quasi-experiments" are also experimental, but by definition they are without random assignment. Instead, they seek to mimic the logic and structure of random assignment experiments, usually without the benefit of closed-system test settings. Causal ambiguities often remain with quasi-experiments, therefore, given how rare it is for the non-random process of selection into treatment to be perfectly known and measured.

(5) *When random assignment is not possible, the preference is for observational study designs that mimic random assignment as much as possible.* From Fisher's Latin Square designs on, these mimetic designs test a null hypothesis about a deliberately manipulated treatment versus a comparison group; the comparison group is deliberately selected to minimize pre-intervention group differences; the occasions of measurement are those found in many outdoor and long-lasting random assignment studies that include pretest assessments as well as posttest ones; and these pretest covariates are then "somehow" used to control for any bias remaining after the study groups have been chosen. In more modern language, the goal of the mimetic tradition is to provide the best approximations for those potential outcomes that are missing in the quasi-experiment but are available in the random assignment experiment.

Legitimate debate is possible about these five propositions, and we could add others. But they help describe and demarcate the inheritance Cochran received. Other causal inheritances were available to him at the time. One was the Galton/Pearson tradition with its emphasis on prediction, substantive modeling, multivariate data analysis and epistemological verification – an evident counterpoint to Fisher's emphasis on causation, experimentation, random assignment and what later came to be called falsification (Meehl, 1978). Cochran could also have worked on identifying the processes through which laboratory research promotes clear causal conclusions, given the strong interest at the time in how experimental physics, chemistry and biology advanced theory in their respective disciplines. But he did neither of these things, instead representing the third position just described. Each of the three has evolved as an adaptation to different parts of the world of research. For most of the 20th century, the experimental model has been preferred for open-system applications in agriculture, medicine, psychology and some engineering. More recently, it has also been imported into micro-economics, political science, education and sociology. However, it is rarely relevant in chemistry, physics or micro-biology, where closed-system experiments are used to describe and explain cause-effect relationships. Nor is the experimental model of much relevance in meteorology, macro-economics, macro-sociology, historical geology and much of population genetics. Cause is still a goal in these fields, but control through deliberate manipulation is difficult, causal modeling is the norm, and local understandings of "cause" stress explanation more than the description of If/Then relationships. Knowing about the alternatives Cochran did not choose helps us distinguish the unique boundaries of the experimental inheritance he was bequeathed, willed forward and passed on.

How he made progress is evident in several ways. One is his role in disseminating the core inheritance through his physical presence, expository writings and transmission of insights not yet committed to print (Cochran, 1965). A second is through his personal research. I am not a statistician or historian of science, but note that he sharpened thinking about how observational studies should seek to mimic the logic of randomized experiments, especially

through his work on matching (Cochran, 1983). He had begun exploring the bias-reduction role of matching much earlier (Cochran, 1950) and was particularly creative in dealing with sub-classification (Cochran, 1968) and even determining how different numbers of strata affect the amount of bias reduced (Cochran, 1965). In addition, he discovered that the mode of analyzing observational study data makes little practical difference, and that bias control is made more difficult by the size of the initial overlap between groups (Cochran, 1968). He also forwarded his inheritance through his influence on others. He trained 39 doctoral students and they (and their students) have created an even more systematic theory of causal identification and estimation that assumes the five main planks described earlier and is best embodied by the Rubin Causal Model. Expositions of it explicitly acknowledge Cochran’s influences (e.g., Rosenbaum and Rubin, 1983; Imbens and Rubin, 2015). He also influenced many other scholars of causation, including fellow statisticians like Mosteller, Moses and Tukey as well as admirers from more distant fields, like Donald T. Campbell and his students whose work often cites Cochran (see the bibliography in Shadish et al., 2002).

A full historical analysis of how Cochran directly *and indirectly* willed forward his inheritance is beyond the scope of this essay. But the point is clear. Cochran was embedded in an upright and rudimentary inheritance; he willed it forward first in one paper and then again in another and then again in another; and he bequeathed this marginally improved inheritance to others, including his own students, so that they might will it even further forward, again and again and again.

The Present Purpose

As a practitioner of mostly quasi-experimental research in complex field settings, I am denied the protections of random assignment and closed-system laboratories. So I am used to feeling vulnerable and suspect and even envious of others’ certainties about method choice. I am not formally connected to the intellectual history of causal design in Statistics. Nonetheless, I feel very comfortable with the first four propositions characterizing Cochran’s inheritance, but am ambivalent about my connections to the mimetic conception of observational study design. This is because I am aware of contrary examples I will present and of three issues central to quasi-experimental practice that the Cochran inheritance rarely discusses. I want to present the issues here and ask: (1) whether they help demarcate the inheritance’s boundaries and (2) whether one or more of them might be worth “willing forward” to incorporate into the inheritance, even if only at one of its margins. Of course, no appointed court of statisticians exists to deliberate about exclusion from the causal inheritance or inclusion into it. Scientific agenda-setting and -modification is a much more ad hoc process that almost certainly has chance components. My hope is merely to start a conversation about what should and should not be part of Cochran’s inheritance, not as he found it, but as he and his students, friends and followers have elaborated it.

The first issue from my own work as a quasi-experimental practitioner speaks to the reality that I sometimes cannot construct a single focused null hypothesis test of a causal hypothesis, even though all random assignment studies and all mimetic quasi-experiments aspire to such a test. Archetypically, this test evaluates the difference between two posttest means from two initially identical groups. Fisher himself advised Cochran that observational studies should not take this approach and should instead elaborate the same causal

hypothesis until it has multiple implications within the data that are subsequently tested. “Somehow” multiple sub-hypothesis tests have to be constructed – even from multiple data sources – and the case has to be made that they are collectively sufficient to test the hypothesis. Fisher’s advice probably surprised Cochran, for it seems to be at odds with Fisher’s own writings on null hypothesis testing. Nonetheless, Cochran (1965) provided some brief examples of causal questions that cannot be answered with a single focused test.

Other scholars have noted the same, and have worked on concepts that overlap with Fisher’s off-hand remark. Campbell’s (1966) “pattern-matching” requires postulating and testing a complex pattern of differences that might rule out all other alternative interpretations. The “critical multiplism” of Cook (1985) depends on multiple tests that are critically chosen because they collectively rule all the currently identifiable causal alternatives to the hypothesis under test. Rosenbaum’s (2005; 2009; 2011) “coherence” notion depends on the consistency of results from multiple tests across different datasets that provide a coherent link to a single causal hypothesis. Finally, the idea of Generalized Differences in Differences (GDD) involves testing statistical interaction hypotheses that are of higher order (and thus more “pattern-laden”) than the two-way interactions in standard differences in differences approaches (e.g., Imbens and Wooldridge, 2007; Chetty et al., 2009). In what follows, the national educational reform program No Child Left Behind (NCLB) illustrates a case where no single focused hypothesis test is possible but where elaborated and testable sub-hypotheses are. All the results are consistent with the hypothesis that NCLB raised academic achievement, but they flirt with verifying a predicted pattern of data and fail to falsify all possible alternative interpretations even if they do falsify all “plausible” ones. So I want to ask: Does the inheritance under discussion want to wash its hands of such inelegant and marginally less successful omnibus tests? If it does not, how might the Fisher strategy be incorporated into an inheritance where its current role is minimal.

The second issue I want to address is that causal statements require more than identifying whether two variables are “causally” related and then estimating the size and statistical significance of the obtained relationship. Also needed is a name for both the cause and effect in general language, given the impossibility of providing a comprehensive description of the cause (or effect) each time it is mentioned. Cochran (1965) briefly mentions this issue, but in the examples he cites he leaves construct validity to psychologists and sociologists, not statisticians. I do not dispute that his inheritance has made some breakthroughs in labeling causal manipulations – e.g., in discussions of comparison group types (e.g., no-treatment versus placebo control groups) and of factorial designs that decompose a complex intervention into parts that are then separately examined. Even so, randomized experiments were designed to optimize construct validity much less than internal validity, and I suspect many people would question the utility of valid cause-effect relationships in which *either the cause or effect were wrongly named*. So I present the example of an otherwise successful bail bond reform program that was discontinued because of how the manipulation was (mis-)labeled. Its benefits might have continued, however, if the treatment had been correctly labeled. What intellectual responsibility, if any, does the Cochran inheritance want to take for the construct validity of independent and dependent variables? Is this issue demarcated out, or worth eventually incorporating?

The final issue I bring up concerns the generalization of causal relationships, using the regression discontinuity design (RDD) as an example. RDD uses a deterministic treatment

assignment procedure to identify causal effects. It is deterministic because assignment to treatment or control status depends only on whether a unit scores above or below a single observed score on a continuous measure that is often of need, merit or age. Since this assignment procedure creates non-overlapping groups on each side of the cutoff, causal identification requires extrapolating the functional from the untreated regression segment into the treated segment where it estimates the missing but crucial potential outcome – what would have happened to treated units had they not been treated. Unfortunately, no independent support exists for this crucial extrapolation, and so causal inference is usually limited to the cutoff point where only a small fraction of those receiving treatment are to be found. The ensuing loss in causal generalization contrasts badly with the randomized experiment whose treatment and comparison groups totally overlap and so warrant the estimation of an average rather than a local treatment effect. The combination of RDD’s unsupported extrapolation, limited causal generalization, and dependence on modeling may explain why few statisticians have paid much attention to the design (Cook, 2008). However, simple design elements can be added to the basic RDD structure and will provide some support for the extrapolation RDD always needs. These supplementary data elements might be untreated regressions from pretest observations (Wing and Cook, 2013), from other covariates (Angrist and Rokkanen, 2015), or from non-equivalent comparison groups (Tang et al., 2015). When certain assumptions are met, we later show that RDDs with an untreated comparison function (CRD) lead to causal results that are demonstrably valid *in the whole treated area beyond the cutoff*. Do some versions of RDD, like CRD, deserve willing forward? More generally, should the inheritance pay more attention to causal generalization and raise its profile relative to causal identification and estimation?

Issue 1: When no single focused test of a causal hypothesis is possible

Random assignment experiments create a single focused test of a hypothesis about a single cause and a single effect, usually a test of mean differences. Something very similar happens in most quasi-experiments. The Rubin Causal Model seeks to create treatment and comparison groups that are equivalent conditional on covariates, thus allowing the group mean differences to be examined at posttest. I prefer such focused tests and can usually create them in observational study work. But I cannot always do so, and it is this sometime failure that motivates the issue addressed here.

Cochran was apparently fond of saying something like “Unless you can give me an example that illustrates your statistical problem, I won’t find it important enough to bother with”. In this spirit let me offer the example of NCLB. It sought to improve academic achievement in all public schools nationwide. The program began in 2002, and the law specified that by 2014 children in all schools in all states had to attain passing scores on state achievement tests. States were free to set their own time schedule for reaching the national goal, but they had to implement a system in which sanctions escalated as the number of consecutive years increased over which a school had failed to reach pre-specified annual performance levels.

Random assignment was not possible because NCLB was the product of a national law that was rolled out immediately. Nonetheless, Dee and Jacob (2011) reasoned that the key component of the national program was “consequential accountability”, the system of

Figure 1: Trends in grade 4 math in Main NAEP, by timing of accountability policy

escalating reforms a school had to undertake as a function of the number of consecutively failed years. Each state had to have such a system by 2002. But some states already had one, allowing states to be partitioned into those with and without accountability in 2002. Using the national Main NAEP math test for 4th graders over eight time points, some before and some after NCLB, [Dee and Jacob](#) constructed Figure 1.

It shows that for states that already had consequential accountability before 2002, slopes for achievement are steeper; but after 2002 the slope difference changes to become parallel, suggesting a post-2002 improvement in those states getting accountability in 2002. This is consistent with a causal impact, but only if a number of problems are addressed. First, the object being evaluated is not NCLB; it is only one mechanism within it, thus compromising the construct validity of the cause. Second, it is not clear whether other state-level, math-correlated forces might have differentially affected achievement before and after 2002 – an issue of internal validity. And finally, there are few data points during the baseline period, leading to questions about how well baseline functional form differences have been modeled – another internal validity issue.

An alternative design ([Wong et al., 2015](#)) pits national public schools that are subject to NCLB against the nation’s private schools to which NCLB hardly applied. This is not a sharply focused test, though. Prior to NCLB, public and private schools were quite different in achievement levels and maybe even slopes, and they were also subject to different historical forces that might have changed around 2002. Moreover, though the baseline time points are now more, they are still few. Nonetheless, Figures 2 and 3 plot 4th grade differences in math on Main NAEP when public schools are contrasted with Catholic schools and then with non-Catholic private schools.

For both grades, a large mean selection difference is evident at baseline, with performance higher in each type of private school. The baseline time trends are less clear, however. Simple visual tests comparing differences in differences suggest that the public schools came

Figure 2: 4th grade math for Main NAEP: Public and Catholic schools

Figure 3: 4th grade math for Main NAEP: Public and non-Catholic private schools

Figure 4: 8th grade math scores for Main NAEP: Public and Catholic schools

Figure 5: 8th grade math for Main NAEP: Public and non-Catholic private schools

to do better after 2002, narrowing the achievement gaps visible before NCLB. Statistical tests used baseline means and linear trends to examine immediate posttest mean differences, posttest linear slope differences, and final mean differences. While every estimate has a sign indicating positive NCLB effects, most are not statistically significant – perhaps due to the small number of degrees of freedom in a national level analysis.

For 8th grade math and 4th grade reading scores, the corresponding data are in Figure 4 through 7. All causal signs are again positive, but few are statistically significant. And effects seem even smaller for reading than math.

The national Main NAEP results just presented are based on items that vary over time in order to reflect national changes in teaching content. In contrast, Trend NAEP holds test items constant. Figures 8 through 10 plot the corresponding Trend NAEP differences for 4th grade math, 8th grade math and 4th grade reading. (Trend NAEP data for non-Catholic private schools are not available, and there is only one interpretable posttest time point exists due to a change in sampling design after 2004). All the results point to greater mean

Figure 6: 4th grade reading scores for Main NAEP: Public and Catholic schools

Figure 7: 4th grade reading Main NAEP: Public and non-Catholic private schools

change after 2002 in public schools and to a reduced achievement gap. Now, the two math differences in difference are statistically significant.

State level tests have advantages over national ones, particularly as regards sample size. NCLB left states free to set their own standards about how quickly to reach the ultimate 2014 national goal, and some states chose to make faster initial changes. So states varied in whether they initially had higher, moderate or lower standards for initial improvement,. (This contrast was not correlated with which states adopted consequential accountability before or after 2002). Figures 11 through 13 give the results for 4th and 8th grade math and 4th grade reading.

Immediately after 2002, all estimates of the difference of differences are positive in sign, with states taking NCLB more seriously doing better, and those in the moderate category doing next best. Adding these state-level results to the national ones in Wong et al. indicates coefficients favoring NCLB in every one of the 39 difference of difference tests across 4th or 8th grade, math or English, Main NAEP or Trend NAEP, Catholic or non-Catholic private schools, and state standard differences assessed in categorical or continuous form. Of these 39, 10 are independent tests of immediate mean and slope differences for math ($p < .001$); and eight are independent tests of differences of differences in slope and final endpoint ($p < .01$). The statistical significance levels are generally weaker for Main than Trend NAEP, but they are strong for Main NAEP when the strongest tests are conducted that combine both the initial mean and subsequent slope effects in order to examine mean differences at the last data collection point. Of the 9 such estimates, all have the same sign, and five are statistically significant at the .10 level or lower. The case is strong then, that the obtained data match the pattern predicted from the single hypothesis that NCLB raised achievement.

Yet none of the single tests warrants a strong causal conclusion; only together do they do what Fisher recommended by rendering the single causal hypothesis more complex and testing how well the predicted and obtained data patterns correspond. However, such verification does not necessarily fulfill what a well executed random assignment experiment achieves – ruling out *all* alternative interpretations. Fisher’s advice will almost always reduce the range of alternative interpretations, but it will not necessarily eliminate all of

Figure 8: 4th grade math scores for Trend NAEP: Public and Catholic schools

Figure 9: 8th grade math Trend NAEP: Public and Catholic schools

Figure 10: 4th grade reading for Trend NAEP: Public and Catholic schools

them. Indeed, [Wong et al. \(2015\)](#) discuss three possible alternatives. Given the coherent pattern of results, each is of low plausibility – but nonetheless each is still possible if a highly unlikely concatenation of alternative forces operates.

One alternative is that students suddenly began to leave Catholic schools for public schools after 2002 when scandals about the sexual behavior of priests began to emerge. This might suddenly affect mean performance in these schools and also in the other schools to which the departing students moved. Moves from Catholic schools were indeed greater just after 2002 – by about one-fifth of one percent relative to prior years. But these rates are too small to account for the public school gains, given that public schools cater to about 90% of all students nationally and Catholic schools to about 5%. Moreover, [Wong et al.](#) show that Catholic school movers had no detectable impact on the post-2002 composition of schools on three demographic measures that are usually highly correlated with achievement. In addition, for inter-school student composition changes to account for the obtained data pattern further requires that (a) the students who left Catholic schools in states with higher

Figure 11: 4th grade math for Main NAEP: High vs. med. vs. low proficiency standard states

Figure 12: 8th grade math for Main NAEP: High vs. medium vs. low proficiency standard states

Figure 13: 4th grade reading scores on Main NAEP: High vs. medium vs. low proficiency standard states

2002 standards achieved more than students who exited Catholic schools in states with lower standards, and (b) students in states with high consequential accountability were more likely to leave Catholic schools than their counterparts in states with low consequential accountability.

Another interpretation is that school officials manipulated NAEP achievement test scores after 2002, and that they were more motivated to do so in public than private schools and in states with higher standards and also in states with consequential accountability beginning in 2002. Two popular ways of manipulating test scores are to reduce how many students with disabilities (SWD) or English language learners (ELL) are tested. Table 9 of [Wong et al. \(2015\)](#) shows the percentages excused from NAEP testing did shift more in public than private schools after 2002, but that no such differential exclusion pattern is evident in the contrast of states varying in standards. Also, school officials have less motivation to manipulate the testing process for NAEP than for state achievement tests since no

consequential decisions depend on a school's NAEP results. The only manipulation-related explanation we can envisage is that high stakes testing at the state level changed the culture of all achievement test taking in and after 2002 – including for NAEP – and that this changing culture was stronger in public than Catholic schools even though both kinds of schools take part in Main NAEP testing. No data exist on this quite circumscribed version of an alternative interpretation based on manipulating test scores.

A third alternative invokes the National Council of Teachers of Mathematics (NCTM). It updated its math standards in 2000 and later claimed that this was responsible for the subsequent national improvements in NAEP math scores (National Council of Teachers Mathematics, 2008). While this alternative might explain why the math effect was larger than the reading one in 4th grade, it also requires that NCTM standards were adopted more often, or were implemented better, in high standard versus low standard states, in states adopting consequential accountability only after 2002, and in public schools more than the two private school types. Also needed are the assumptions that: (a) math standards operate with a three-year causal delay, for the standards changed in 2000 but the first evidence of a possible effect appears in 2003; and (b) issuing new standards is by itself sufficient to change performance at the national level!

The need for discussions about ruling out alternative interpretations leaves a bad taste in the mouth when compared to the simple use of a single focused hypothesis test. But consider the alternatives to this bad taste. One would be to do no evaluation. Would that leave an even worse taste? Another would be to do a randomized experiment under atypical circumstances, perhaps by issuing waivers from NCLB to exempt some school districts. However, this set-aside waiver experiment would be compromised by the kinds of districts or schools eligible for waivers and applying for them. More serious would be the nature of the schools available to parents wanting to leave a school. In the waiver context we envisage, there would be some public schools without NCLB in each school district. Yet this would not be possible in a national program and would fundamentally alter the school choices available. When study designs are evaluated by both internal *and* external validity criteria, a randomized experiment of a systematic and likely small sample of schools operating in a social context to which one does not want to generalize might not be superior to an unaesthetic bundle of multiple interrupted time-series tests conducted at both the national and state levels. Construing study quality within a *multi-attribute* decision framework, rather than today's *single-attribute* framework that prioritizes only internal validity, suggests that a sharply focused experimental test from a waiver experiment might not be superior to the messy multiple tests described above.

The epistemological bottom line with elaborating the causal hypothesis is that empirical coherence is only part of the bias control story; the rest depends on how well alternative interpretations are ruled out by the uniqueness of the predictions. In his own work, Fisher wrote about single sharp hypothesis tests, and he did not exemplify his own advice about elaborating a causal hypothesis, though I know little of his later genetics research. Nor, to my knowledge, did Cochran do anything more with Fisher's advice than pass it on, all the time advocating for observational studies based on matching procedures designed to mimic single sharp tests. Most of Cochran's students (and their students) have not heeded Fisher's advice either, with the salient exception of Rosenbaum. I grant the superiority of single focused tests. My question is whether the critical multiplist perspective described

here should be demarcated out of the inheritance, though it is consonant with Fisher's dictum? Or is it worth willing forward when single sharp hypothesis tests are not possible?

Issue 2: Simultaneously Identifying Causal Relationships and Causal Constructs

The central task of the inheritance under discussion is to identify causal relationships through a falsification process that distributes conceptually irrelevant causal forces equally across conditions. In random assignment studies, this is done unconditionally, though in quasi-experiments it is done conditionally on the covariates used. Other meanings of cause are broader. Let us leave aside the distinction between causal explanation and causal description, recognizing that experimentation always deals with causal description but only deals with causal explanation if (1) the independent and dependent variables are deliberately chosen to index explanatory concepts from a clear substantive theory, or (2) all plausible temporally mediating variables in a study are measured and the analysis identifies which of them offer a more plausible causal explanation than others. Let us instead focus on the fact that the cause and effect in each study need to be labeled in abstract language if scientific communication is to ensue. The alternative, providing a complete description of the manipulandum, is impossible because the volume of words needed is impractical and, anyway, full knowledge of causal agency is often elusive. The next example explores the conundrum that a causal relationship can be valid but useless, or even dangerous, when the causal agent is invalidly labeled.

Cook County, Illinois changed its system of bail hearings in 1999, hoping to reduce costs by substituting video hearings for face-to-face hearings with a judge. The reform applied to all charged offenses except the most serious ones – murder and serious sexual offenses, for which face-to-face hearings persisted. All the labels attached to the reform featured its video component, reflecting the universal consensus that this was the core intervention component. Some legal advocates came to dislike the reform, and one reason for this was the suspicion that it may have unintentionally increased bail amounts. By so doing, more individuals could not now afford bail and would have to remain in jail even though presumed innocent; their families and jobs would be endangered; and the county would have to pay increased jail and welfare costs.

A study was conducted using 17 years of data on daily bail amounts, but aggregated by quarter ([Diamond et al., 2010](#)). About half of the days came from before the video intervention and half after it. The charged offenses were partitioned by seriousness, with the less serious ones receiving the intervention and the most serious ones serving as no-treatment controls. The resulting comparative interrupted time series showed that, after the intervention, bail amounts changed by more for the less serious than the more serious charges. These results were first privately released to the county senior judges with information that they would soon be released to the press. Just before they were released, the video innovation of 8 years was rescinded. Face to face hearings returned for all charges.

There is little doubt that bail amounts increased after the video reform. Figure 14 gives that lowess-smoothed daily aggregate data over 17 years from [Cook et al. \(2014\)](#). The date of the reform is set at 0 and the range of days goes from plus to minus 3,000. The greater mean shift after the intervention is clear.

Figure 14: Serious and less serious offenses over time: LOWESS-smoothed daily aggregated bail amounts from before and after the bail bond reform intervention

Figure 15: LOWESS-smoothed number of judges hearing serious and less serious cases by day before and after the bail bond reform intervention

Figure 15 shows the number of judges who adjudicated bail bond hearings each day. For the less serious cases the number of officiating judges dropped precipitously from about 8 per day to three, indicating that the reform was a multi-attribute treatment package. One part was the video, and another was the reduction in the number of judges. Other evidence not produced here shows that the volume of bail bond hearings did not change over the intervention period, suggesting that the post-video judges became more specialized.

Figure 16: Percentage of all felony bail hearings held before and after the bail bond reform by the Groups of 3, 13 and 83 judges who held bail hearings both before and after the reform.

Of the 530 judges adjudicating in bail bond hearings over 17 years, 99 heard cases both before and after the video-based reform. We call these the before-after judges. All the other judges heard cases only before or only after the change. Of the 99 before-after judges, three were responsible for almost 100% of all the cases for the first year after the change and for 50% of them for the next two years. In contrast they heard 40% in the year immediately before the reform and very little before that. So they were heavily overrepresented immediately after the reform when compared to before it. This is revealed in Figure 16 which plots the number of pre- and post-intervention hearings adjudicated by the 3 most specialized, the 13 next most specialized, and the 83 who heard very few cases even after the reform.

Key issues are whether these before-after judges were disposed to set higher bail amounts than their pre-intervention colleagues even before the intervention occurred; and whether their bail amounts increased at all. Figure 17 gives the relevant data for the bail amount set by three most prolific judges, and we see no change in bail amounts from before to after the reform!

Data are more sparse in the immediate post-video period for the other before-after judges, but Figure 18 for the 83 before-after judges shows there is again no evidence that they set higher bail amounts after the intervention than before. The apparent mean increase for less serious offenses is matched by an apparent increase for the more serious offenses that were always adjudicated face-to-face.

If within-judge analyses show that bail amounts did not change from before to after the reform, and if analyses aggregated across all judges (including include the before-only and after-only judges) show that bail amounts changed from before to after the video intervention, what explains the seeming discrepancy. The most likely explanation is that the before-only judges set lower bail amounts than their before-after colleagues, and Figure 4

Figure 17: Serious and less serious offenses over time: LOWESS-smoothed individual case record data for the three most active before-after judges.

Figure 18: Serious and less serious offenses over time for the 83 before-after judges: Overlap between the LOWESS-smoothed individual case record data and quadratic order regression.

in [Cook et al. \(2014\)](#) confirms this. So the higher post-reform bail amounts found with all of the judges is probably due to the before-after change in judge composition rather than to the introduction of the new video system. Even before the video reform, the judges hearing cases both before and after the reform were prone to set higher bail amounts than their colleagues who only adjudicated them before.

What would have happened if the true causal agent had been correctly identified? Could the Cook County Court system have changed the judges who held bail hearings without sacrificing the video component and the cost savings it generated through mechanisms outlined in [Cook et al. \(2014\)](#)? Could the post-intervention judges have been randomly chosen?; Could a different set of judges with different proclivities have been selected?; Could the 3 or 99 judges have been instructed to re-consider how they set bail amounts? These are all pertinent questions, but they were never asked at the time because the universal social consensus was that the reform entailed changing only the mode in which hearings were held – from face-to-face to video. The resulting mis-identified causal agent points to a dilemma: What is the value of learning that a causal relationship is true if the cause involved in it

is wrongly named and a name has to be provided? Does the Cochran inheritance take no responsibility for mis-identified causal *agents*, only for mis-identified causal *relationships*? Many non-statisticians may find it odd to limit the idea of causation to identifying and estimating causal relationships without considering how the cause and effect are named. Is it worthwhile willing forward research on this cause-naming topic?

Issue 3: Generalizing Causal Relationships, including in the Regression Discontinuity Context

Despite research on pooling experimental results and on identifying statistical interactions between treatments and study details, the Achilles Heel of experimentation remains what is variously called external validity, heterogeneity, or causal generalization – the capacity to identify the boundary conditions on which a specific causal relationship depends, including total generalization across all boundary conditions within our space-time continuum.

Few experiments are based on samples randomly drawn from a clearly described population; most sampling particulars are products of opportunism. In research syntheses it is rare for the sample of available studies to be formally linked to a meaningful population and, anyway, formal sampling models are not relevant to all the objects of causal generalization – e.g., to different types of settings, times, and ways of operationalizing the cause and effect. Also, studies rarely have samples so large that they permit examining how the treatment interacts with all, or even many, of the possible sources of heterogeneity. Scientific decisions about causal robustness are usually more weakly supported and depend, say, on a predominantly positive causal sign or on whether effect sizes differ across the possible sources of heterogeneity that happen to be examined. When causal results are not robust, it is common to use the pattern of results to begin the process of specifying the boundary conditions that determine variation in effect sizes. The limitation here is the extensiveness of the potential moderator variables available. Longitudinal surveys typically have better sampling bases and more measures of potential moderators, but internal validity is a much bigger problem with them than with experiments. And, if researchers cannot be “reasonably” sure of a causal connection, what is the point of worrying about its generalization?

When it comes to causal generalization, RDD is in even worse shape than the usual experiment. Although the process of selection into treatment and control is perfectly known and can be modeled, the unbiased causal inference that results depends on (1) having specified the correct functional form or bandwidth size (Imbens and Lemieux, 2008); (2) the utility of a local average treatment effect at the cutoff (LATE); and (3) the willingness to tolerate less statistical power than in the experiment. To try to counteract these limitations, the comparative RDD design (CRD) adds an untreated regression to the basic RDD structure, typically to date from a pretest measure (Wing and Cook, 2013), other covariates (Angrist and Rokkanen, 2015) or a non-equivalent comparison group (Tang et al., 2015). The purposes of this untreated regression are to provide additional support for the extrapolation that RDD always needs, to increase power by adding data, and to test causal generalization not just at the cutoff but also in the entire area of the assignment variable where the treatment is available.

In Wing and Cook (2013), the key assumption of CRD is that the three untreated regression lines are parallel. These are the pretest scores on the untreated side of the

cutoff and on the treated side as well, plus the posttest observations in the untreated part of the assignment variable. When these slopes are plausibly parallel and the pretest is included in the outcome model together with the assignment variable and binary cutoff score, then the relationship of the assignment variable to the outcome will be zero. This is the ignorability condition that Angrist and Rokkanen (2015) believe is necessary for valid causal generalization beyond the cutoff score in CRD.

Tang et al. have just tested these ideas using the national Head Start impact study, a nationally representative study of 2499 children who were randomly assigned to Head Start or control status. As causal benchmarks, experimental estimates were estimated both at the RD cutoff and also in the total treated area above it. The CRD was constructed out of the experimental data by creating an assignment variable (IQ test scores), defining a cutoff value on it, and then omitting the control cases from the treated side of the assignment variable and the treated cases from the control side. One comparison regression function came from pretest scores on each of these outcomes, and the second from a cohort of non-Head Start children one year older and from the same catchment areas. There were three study outcomes measured before and after entering Head Start – performance in math, literacy and social behavior. In addition, a second assignment variable was also designated to test if the results achieved with the IQ assignment variable would be replicated. They were, and in most of what follows we present results for the single IQ assignment variable. We use CRD-Pre to designate use of a pretest no-treatment regression function and CRD-CG to designate use of a non-equivalent independent comparison group. We refer to the randomized experiment as an RCT.

Analyses not reported here showed that the three untreated regression segments were plausibly parallel for each outcome and each assignment variable. One question is how similar the experimental, basic RD and CRD estimates are *at the cutoff*. This is where theory indicates each should be unbiased and so the two causal estimates should be similar, not just to each other within the limits of sampling error, but also to the RCT when it is estimated as a LATE. More important is the second question: How similar are the RCT, CRD-Pre and CRD-CG estimates away from the cutoff, given that basic RDD is not estimated there and the internal validity of CRD designs is unknown there? A final question is how similar standard error estimates will be across the three design variants, after adjusting for the inevitably smaller RDD group due to how it was constructed from the RCT groups?

Figure 19 shows results comparing only the RCT and basic RDD for each outcome and assignment variable. They show the expected results: No bias at the cutoff, and larger confidence intervals with basic RDD.

Figure 20 shows the results for CRD-Pre at the cutoff, and Figure 21 for CRD-CG at the cutoff. In each case the estimates seem close to the experimental ones (and hence also the basic RDD). The confidence intervals are narrower than with the basic RDD and are thus closer to those from the RCT. But they are not equal to the RCT.

Figure 22 shows the causal estimates *away from the cutoff* for CRD-Pre and thus calculated for more cases than just at the cutoff. Figure 23 shows the results for CRD-CG. In each case, both the treatment estimates and confidence intervals are similar to those from the RCT. Under the conditions tested, the RCT, CRD-Pre and CRD-CG results are

Figure 19: Basic RD estimates and confidence intervals compared to the corresponding RCT benchmarks *at the cutoff* over three outcomes and two assignment variables.

Figure 20: CRD-Pre estimates and confidence intervals *at the cutoff* compared to the corresponding RCT benchmarks for one assignment variable.

virtually interchangeable with respect to both bias away from the cutoff and statistical power.

Are such findings of any relevance to the inheritance in which Cochran was embedded, that he marginally improved, and that he then passed on for others to improve again and again and again? In [Tang et al.](#), as in [Wing and Cook](#), the RCT causal estimates and standard errors are reproduced in the CRDs merely by adding supplementary data to what is otherwise a weak but theoretically unbiased design – the basic RDD. Does RDD fall

Figure 21: CRD-CG estimates and confidence intervals *at the cutoff* compared to the corresponding RCT benchmarks for one assignment variable

Figure 22: CRD-Pre estimates and confidence intervals *above the cutoff* compared to the corresponding RCT benchmarks for one assignment variable.

outside of the Cochran inheritance's bailiwick; or is it worth including in it even if only in some forms (like CRD) but not others (like basic RDD). CRD certainly needs willing forward, but will statisticians of causation take on this task?

Conclusion

The three examples presented here support causal conclusions; they may even strongly warrant them. One conclusion is about the achievement consequences of a national law; another about the consequences of a county-level reform on bail bond amounts and number of officiating judges; and the third is about the consequences of CRD for the quality of

Figure 23: CRD-CG estimates and confidence intervals *above the cutoff* compared to the corresponding RCT benchmarks for one assignment variable.

causal inferences relative to an RCT. None of these studies started from the premise of creating an observational study that mimics what would be done in an experiment. The first example involves multiple hypothesis tests, each individually imperfect, rather than a single sharply focused test. The second compares legal charges whose seriousness widely differs and whose bail amounts, unlike in an experiment, hardly overlap at all. The third is about a design (RDD) whose treatment and control groups also do not overlap and so are not matched. Yet each of these designs supported strong causal conclusions without any matching for a single strong test.

Indubitably, there are great conceptual advantages when observational studies are considered as though they were broken down experiments; and in actual research practice, there are many opportunities to match individual cases where the selection process is mostly known and where relevant and otherwise rich covariates are available. I am not against the utility of the mimetic conception of observational study design that Cochran willed forward. My concern is limited. It is whether the current dominance of this conception threatens to rule out other ways of advancing observational study design that are predicated on different principles than the pursuit of ineffable exact matching. I would like to see more principles than matching implicated in observational study design, to see more arrows in the observational study quiver. I have tried to model some of them here.

No body of statisticians is charged with making decisions about which issues to include in Cochran's inheritance or to exclude from it. So there is no mechanism for formally answering the questions this paper posed about demarcation – what should and should not be part of the inheritance?; and what might be willed forward because it seems worthy of inclusion?

The most to hope is that, over time, a consensus will emerge about (1) supplementing the traditional research agenda on internal validity by increasing the profiles of both construct and external validity, and (2) adding RDD and CRD to the Pantheon of acceptable designs, even though they are predicated on minimizing the very group overlap that experiments and

mimetic quasi-experiments seek to maximize. The issues I have described are important in the daily struggles of practitioners of observational studies who, like me, look to statisticians for help. I wish I could talk with Cochran to learn how he would make demarcation decisions. Of course, he might not be interested and simply reject the assumptions of this paper, perhaps believing I have reified the intrinsically fuzzy concept of an inheritance. But that would still be interesting to learn.

Though conversation with him is impossible, it is with some of his students and their students. Most of them are acquainted, I would guess, with the distinguished figure and record of Professor William G. Cochran of Harvard University. But it is important not to forget that in the working class Glasgow of his childhood he would have been Oor Wullie, that he was Bill to his friends on both sides of the Atlantic, and that the few students of his I know remember him with affection and with respect for his work on observational study procedures that seek to approximate the logic and practice of randomized experiments. That his students and his students' students have willed his work forward so tellingly is a blessing for his memory; but it is not something any scholar should take for granted.

Consider Galileo. He used his lens as a telescope and revealed a heliocentric universe that transformed how Man, God and Nature are related. A decade later, he realized his lens was also a microscope, but he was not interested enough to venture far into the micro worlds that eventually transformed biology and medicine. Fortunately, others were more interested in microscopes than he was, and about 50 years later “the pioneers of microscopy Antonie van Leeuwenhoek and Robert Hooke revealed that a Lilliputian universe existed all around and even inside of us. *But neither of them had students* (my italics), and their researches ended in another false dawn for microscopy. It was not until the middle of the nineteenth century... that the discovery of the very small began to alter science in fundamental ways” (Flannery, 2015, p. 30). How lucky Cochran's students and his students' students are in being so centrally networked within the extraordinarily upright and rudimentary inheritance that Cochran received, improved and passed on to them. Does this very privileged position in history entail any obligation to take seriously the demarcation tasks so amateurishly described in this paper? I guess not, but...

Acknowledgments

This work was supported by NSF grant #DRL-1228866

References

- Angrist, J. D. and Rokkanen, M. (2015). Wanna get away? Regression discontinuity estimation of exam school effects away from the cutoff. *Journal of the American Statistical Association*. (Just accepted).
- Campbell, D. T. (1966). Pattern matching as an essential in distal knowing. In Hammond, K. R., editor, *The psychology of Egon Brunswik*, pages 81–106. Holt, Rinehart & Winston, Oxford, UK.

- Chetty, R., Looney, A., and Kroft, K. (2009). Salience and taxation: Theory and evidence. *American Economic Review*, 99:1145–1177.
- Cochran, W. G. (1950). The comparison of percentages in matched samples. *Biometrika*, 37(3/4):256–266.
- Cochran, W. G. (1965). The planning of observational studies of human populations. *Journal of the Royal Statistical Society. Series A (General)*, 128(2):234–266.
- Cochran, W. G. (1968). The effectiveness of adjustment by subclassification in removing bias in observational studies. *Biometrics*, 24(2):295–313.
- Cochran, W. G. (1983). *Planning and analysis of observational studies*. Wiley, New York.
- Cook, T. D. (1985). Post-positivist critical multiplism. In Shotland, R. L. and Mark, M. M., editors, *Social Science and Social Policy*, pages 21–62. Sage Publications, Beverly Hills, CA.
- Cook, T. D. (2008). “Waiting for life to arrive”: A history of the regression-discontinuity design in psychology, statistics and economics. *Journal of Econometrics*, 142(2):636–654.
- Cook, T. D. and Campbell, D. T. (1979). *Quasi-Experimentation: Design and Analysis Issues for Field Settings*. Houghton Mifflin, Boston.
- Cook, T. D., Tang, Y., and Diamond, S. S. (2014). Causally valid relationships that invoke the wrong causal agent: Construct validity of the cause in policy research. *Journal of the Society for Social Work & Research*, 5(4):379–414.
- Dee, T. S. and Jacob, B. (2011). The impact of No Child Left Behind on student achievement. *Journal of Policy Analysis and Management*, 30:418–446.
- Diamond, S. S., Bowman, L. E., Wong, M., and Patton, M. M. (2010). Efficiency and cost: The impact of videoconferenced hearings on bail decisions. *Journal of Criminal Law and Criminology*, 100:869–902.
- Flannery, T. (2015). How you consist of thousands of tiny machines. *New York Review of Books*, LXII(12):30–32.
- Imbens, G. W. and Lemieux, T. (2008). Regression discontinuity designs: A guide to practice. *Journal of Econometrics*, 142(2):615–635.
- Imbens, G. W. and Rubin, D. B. (2015). *An Introduction to Causal Inference in Statistics, Biomedical and Social Sciences*. CAMBRIDGE UNIVERSITY PRESS, New York, NY.
- Imbens, G. W. and Wooldridge, J. (2007). What’s new in econometrics? Lecture notes, NBER Summer Institute.
- Meehl, P. (1978). Theoretical risks and tabular asterisks: Sir Karl, Sir Ronald and the slow progress of soft psychology. *Journal of Consulting and Clinical Psychology*, 46:806–834.

- Rosenbaum, P. R. (2005). Observational studies. In Everitt, B. S. and Howell, D. C., editors, *Encyclopedia of statistics in behavioral science*, Vol. 3, pages 1451–1462. Wiley & Sons, Chichester, UK.
- Rosenbaum, P. R. (2009). *Observational Studies*. Springer-Verlag, New York, NY, 2nd edition.
- Rosenbaum, P. R. (2011). Some approximate evidence factors in observational studies. *Journal of the American Statistical Association*, 106:285–295.
- Rosenbaum, P. R. and Rubin, D. B. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika*, 70(1):41–55.
- Shadish, W. R., Cook, T. D., and Campbell, D. T. (2002). *Experimental and Quasi-Experimental Designs for Generalized Causal Inference*. Houghton Mifflin, Boston.
- Stigler, S. M. (1986). *The history of statistics: The measurement of uncertainty before 1900*. Harvard University Press, Cambridge, MA, and London, England.
- Tang, Y., Cook, T. D., and Kisbu-Sakarya, Y. (2015). Reducing bias and increasing precision by adding either a pretest measure of the study outcome or a nonequivalent comparison group to the basic regression discontinuity design. Working paper, Institute for Policy Research at Northwestern University.
- Watson, G. S. (1982). William Gemmell Cochran 1909–1980. *The Annals of Statistics*, 10(1):1–10.
- Wing, C. and Cook, T. D. (2013). Strengthening the regression discontinuity design using additional design elements: A within-study comparison. *Journal of Policy Analysis and Management*, 32(4):853–877.
- Wong, M., Cook, T. D., and Steiner, P. M. (2015). Adding design elements to improve time series designs: No Child Left Behind as an example of causal pattern-matching. *Journal of Research on Educational Effectiveness*, 8(2):245–279.
- Yates, F. (1964). Sir Ronald Fisher and the design of experiments. *Biometrics*, 20(2):307–321.