Field Experiments:
Here Today Gone Tomorrow?

John A. List

May 25 2024

Abstract

Once believed to be an impossibility, field experiments in economics now occupy a central place in the empiricist's quiver. In the past few decades alone field experiments have taken on much greater import in academe, across organizations, as well as for policymakers. But is this emergence simply a fad that will soon return field experiments to obscurity? I argue in this article that there is something fundamental about the emergence of field experiments, as controlling the assignment mechanism in the field provides unparalleled power to both understand the "effects of causes" and the "causes of effects." This knowledge generation then begins to uncover the generalizability and scalability of knowledge. Quite the opposite of a withering tool that will be gone tomorrow, I urge economists to "double down" on this comparative advantage and in doing so I provide four methodological paths which I hope will cement the promise and growth of field experiments in the social sciences.

Keywords: field experiments, science, economics

List: Department of Economics, University of Chicago, Australian National University, and NBER, Chicago, IL 60637. Please send correspondence to J.A. List: jlist@uchicago.edu, (773) 702-9811. Faith Fatchen and Francesca Pagnotta provided excellent research assistance and comments on an earlier version of this manuscript.

"There is a property common to almost all the moral sciences, and by which they are distinguished from many of the physical…that it is seldom in our power to make experiments in them"
~John Stuart Mill (1844, p. 124)

Introduction

In economics, the modal approach to empirical research historically has been to write down a theoretical model and start looking for available naturally occurring data. In this sense, yesterday's economists "visited the pin factory" to help them scribe general economic theories. This was largely the result of the economic pioneers insisting that it was not possible to do economic experiments. This hit home for me as I read my first college economics textbook, Samuelson and Nordhaus (1985, p. 8):

"The economic world is extremely complicated. There are millions of people and firms, thousands of prices and industries. One possible way of figuring out economic laws in such a setting is by controlled experiments……like those done by chemists, physicists, and biologists Economists have no such luxury when testing economic laws. They cannot perform the controlled experiments of chemists or biologists because they cannot easily control other important factors. Like astronomers or meteorologists, they generally must be content largely to observe."

Yet, today, a growing number of empirical economists are searching for situations and questions in which a field experiment might offer a feasible and desirable approach. This article argues that there is methodological rhyme and reason for this surge. This is because, increasingly, the foundation to begin acquiring the knowledge necessary to test theories and inform decision makers arises in the form: does variation in X cause Y to change? If so, then in what direction and at what magnitude? Indeed, the standard problem of causal inference has been a key cog in the fundamental problem of knowledge creation.

Specifically, I argue in this study that the methodological force behind field experiments is that they permit researchers to tackle two key problems, which I denote as *Experimental Problem 1*: Measuring the causal impact of treatments, and determining relevant mediators and moderators in an ethically responsible manner; and *Experimental Problem 2*: Predicting whether the causal impacts of treatments implemented in one environment transfer to other environments, be them spatially, temporally, or scale differentiated.

My study is parsed into three sections. The next section begins with the nuts and bolts of the experimental method and provides a brief overview of how field experiments fit with the other empirical approaches. This section includes an introduction to the potential outcomes model, which helps us to understand the necessary conditions for establishing causality. I then outline four paths that social scientists can take field experiments to leverage their unique features in Section 3. While I can imagine several distinct paths forward, due to space I focus on only 4: i) design to understand heterogeneity *and* causal moderation, ii) design to understand mediation paths (or mechanisms), iii) design to understand generalizability and scaling, and iv) design in an ethically responsible manner. Each of these considerations is important, but together represent distinct advantages of field experiments.

Finally, I conclude with related paths forward should a researcher be yearning for more examples beyond the four areas of focus. Here, I focus on the uniqueness of experimentation that permits selective data generation, the longitudinal aspects of learning potential using field experiments, and how within-subject designs not only enhance experimental power but allow the researcher to estimate the full joint distribution of outcomes, allowing us to go beyond a comparison of marginal distributions (which is what is recovered when using the workhorse between-subject design).

2. Nuts and Bolts of Experiments

As a useful first step in describing field experiments, it is informative to compare other more traditional empirical approaches alongside field experiments to understand how their identification strategies relate. Let's begin by using a potential outcomes framework, and assume that individual i has covariates x_i, and the measurement approach consists of the following experimental stages:

- Z is the assignment mechanism and let z_i denote assignment. For example, $z_i = 1$ if unit i is assigned to treatment, 0 otherwise.
- D is the unit-specific treatment and let d_i be the realized treatment status, which is the treatment individual i receives (e.g., $d_i = 1$ if unit i actually takes up treatment). Note that it is possible that z_i and d_i are different due to compliance.

- Y is the observed outcome and let y_{i1} be the outcome of interest when treatment status is $d_i = 1$, and y_{i0} when treatment status is $d_i = 0$.

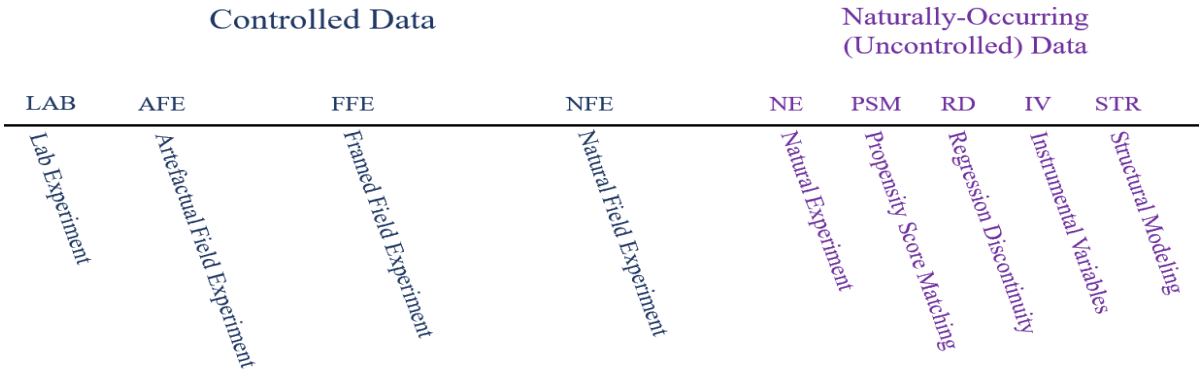
Ideally, when conducting an experiment or examining naturally-occurring data, researchers could measure individual treatment effects for each individual i, $y_{i1} - y_{i0}$, which is the difference in outcomes for individual i being in the treated versus the control group. In practice, of course, we cannot observe both outcomes simultaneously; instead, we can only observe individual outcomes in one of the treated states, and the counterfactual outcome in the other state remains unobserved. This is typically denoted as A/B testing. Instead of individual treatment effects, researchers therefore usually recover the Average Treatment Effect (ATE), given by:

$$\tau = E[y_{i1}|d_i = 1] - E[y_{i0}|d_i = 0]. \qquad (1)$$

Equation (1) is simply the difference in conditional expectations between the treatment group and the control group. This represents the fruits of the experimentalist's labor.

Exhibit 1 provides an empirical spectrum from approaches where the researcher knows and controls Z, the assignment mechanism (Lab, AFE, FFE, and NFE), to empirical approaches where the assignment mechanism is neither under the control nor known by the researcher (labeled as NE, PSM, RD, IV, and STR). I categorize these as "controlled" and "uncontrolled" because the researcher knows and controls the assignment mechanism in one case but not in the others.

Exhibit 1: Data Generation and Modeling for Causal Inference



This figure provides an empirical spectrum from approaches where the researcher controls Z, the assignment mechanism (Lab, AFE, FFE, and NFE), to empirical approaches where the assignment mechanism is neither under the control nor known by the researcher. These are categorized as "controlled" and "uncontrolled" because the researcher knows and controls the assignment mechanism in one class of approaches (Lab, AFE, FFE, and NFE), but not in the other (NE, PSM, RD, IV, and STR).

To understand the species, we must first define the species. Within economics, a conventional laboratory experiment is one that employs a standard subject pool of students, an abstract framing, and an imposed set of rules. And, following Harrison and List (2004), I define the three types of field experiments as follows:

- An artefactual field experiment (AFE) is a conventional lab experiment, but with a non-standard subject pool.
- A framed field experiment (FFE) builds on an AFE by adding field context in the commodity, task, and/or information that the subjects can use.
- A natural field experiment (NFE) builds on an FFE in that it occurs in the environment where the subjects naturally undertake these tasks and where the subjects do not know that they are taking part in an experiment.

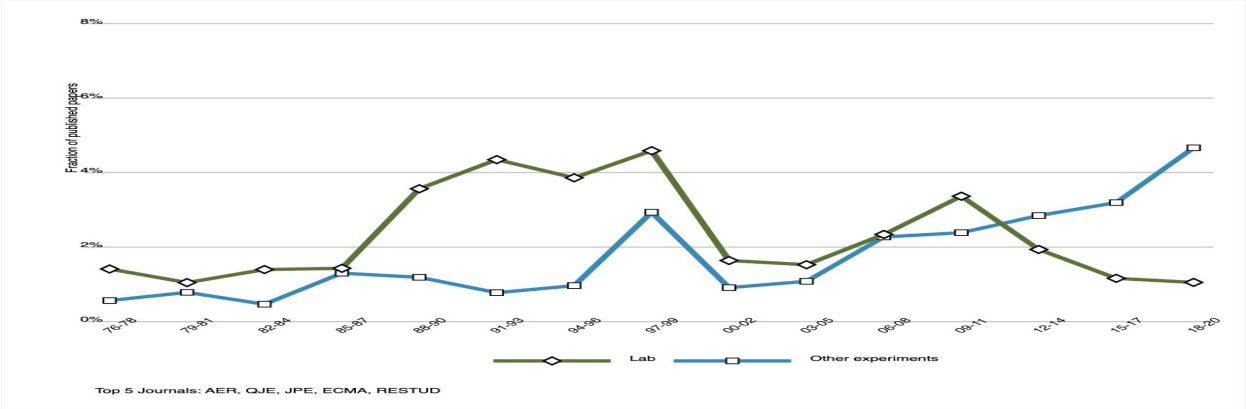
It is now well understood (see, e.g., List, 2025) that there are four exclusion restrictions necessary to ensure recovery of an internally valid parameter in randomized experiments (Lab, AFE, FFE, and NFE) as estimated in equation (1). These are the Stable Unit Treatment Value Assumption (SUTVA), Observability, Compliance, and Statistical Independence. SUTVA is an assumption that any unit's potential outcomes do not vary with the treatments assigned to or undertaken by *any* other unit and that there is no variation in the form of treatment that leads to different potential outcomes. Observability (non-attrition) requires that all participants in the study remain in the study, and for each, we observe their outcome, treatment assignment, actual treatment, and unit-level characteristics. Compliance assumes that every unit assigned to $Z_i = z$ ends up taking $D_i = z$. Statistical independence means that the assignment mechanism governing the D_i is independent of potential outcomes (i.e., the researcher controls the assignment mechanism).

Necessary assumptions to recover a causal parameter represent a key distinction between data generation in the western portion of Exhibit 1 and data analysis in the easternmost portion. The easternmost portion of Exhibit 1 is populated by several econometric approaches, of which

I include five: natural experiments (as examined via a difference-in-difference model), propensity score matching (PSM), regression discontinuity (RD), instrumental variables (IV) estimation, and structural approaches (STR). Each of these approaches considers a unit i (e.g., an individual, school, village, etc.) that has a potential outcome $Y_i(D_i)$ under treatment $D \in \{0,1\}$. The treatment effect for unit i can then be measured as $\tau_i = Y_i(1) - Y_i(0)$. The major problem, however, is one of a missing counterfactual: τ_i is unknown because for any individual, we can only observe $Y_i(0)$ or $Y_i(1)$. If we could observe the outcome for an untreated observation had it been treated, then there would be no evaluation problem. Each of these approaches has identification assumptions to establish causality. Since those assumptions are discussed broadly elsewhere, I point the reader to Harrison and List (2004), Angrist and Pischke (2009), and Imbens and Rubin (2015) for further details.

The attractiveness of the experimental method, in part due to its transparency and reasonable assumptions, has helped it to maintain a level of constant representativeness within economic journals. Exhibit 2 shows the fraction of experimental studies published in the Top 5 economics journals from 1976-2020. While the growth of studies published in this set of journals has slowed since the early years for lab experiments, there remains important representativeness of lab experiments in the first few decades of the 21st century, as can be seen from the green time series in Exhibit 2. Likewise, field experiments have shown dramatic growth in the 21st century.

Exhibit 2: Experimental Papers Published in the Top 5 Economics Journals From 1976-2020 as a Fraction of Total Papers



The graph construction follows Reuben et al. (2022) and illustrates the growth of laboratory experiments in the late 20th century in the top 5 economics journals. As can be seen, field experiments have witnessed dramatic growth over the first two decades of the 21st century in Top 5 economics journals.

Brief Field Experimental Background

As Exhibit 2 highlights, much of the early experimental work that was published in the Top 5 Economics journals took the form of laboratory experiments. Over the past few decades, however, economists have begun to depart from these roots; they are recruiting subjects in the field rather than in the classroom, using field goods rather than induced valuations, and using field context rather than abstract terminology in instructions. In this way, the systematic departure from the laboratory and the uniqueness of control that field experiments provide combine to allow field experiments to take an important place in empirical economics.

Similar to the spirit of the early laboratory experiments, field experiments typically use randomization to achieve identification. Different from laboratory experiments, however, field experiments vary the subject pool, oftentimes occur in the *natural environment* of the agent being observed, and in many cases examine behaviors that cannot be reasonably distinguished from the tasks the agent has entered the marketplace to complete. Levitt and List (2009) document three distinct waves of field experimentation within economics.

The first wave, which I denote as the dawn of field experimentation, is rarely considered to be part of the field experimental genre in economics. Considering that none of these studies were experiments with human subjects, and few were published in economics journals, this is understandable. Nevertheless, the work of Fisher and Neyman in the 1920s and 1930s is worthwhile to consider for three reasons. First, these experiments helped to answer important economic questions regarding agricultural productivity across plots of land (and thus, in the most literal sense of the word were "field" experiments). Second, these studies are generally believed to be the first to conceptualize randomization in an important field study as a key

element of the experimental method.[1] Third, Neyman's (1923) "potential yields" framework set the stage for the Rubin Causal Model, the framework advanced above.

The second wave of interest is the latter half of the 20th century, during which government agencies conducted a series of large-scale social experiments. In Europe, the early social experiments in the late 1960s included electricity pricing schemes in Great Britain. In the US, social experiments can be traced to Heather Ross, an MIT economics doctoral candidate working at the Brookings Institution. The first wave of such experiments in the US began in earnest in the late 1960s and included government agencies' attempts to evaluate programs by deliberate variations in agency policies. Such large-scale social experiments included employment programs, electricity pricing, job training programs, and housing allowances. While this early wave of social experiments tended to focus on testing new programs, since the early 1980s major social experiments have examined various reforms that test incremental changes to existing programs. These experiments have had an important influence on policy, as they were recognized as contributing to the Family Support Act of 1988, which overhauled the AFDC program. They also led to an important debate concerning the trade-off between observational and experimental data.

Conceptually, this second wave was important as the early 1970s ushered in the work of Donald Rubin (1974), who put the language and framework of potential outcomes front and center for causal inference.[2] Whereas Neyman (1923) focused on randomized experiments, Rubin (1974) made it the centerpiece of all empirical work, experimental and non-experimental. He also clarified the import of the assignment mechanism in the potential outcomes framework.

The third distinct wave of field experimentation is the surge of field experiments in economics in the past several decades. This most recent movement approaches field experiments by taking the tight controls of the lab to the field. In doing so, the analyst bridges laboratory and naturally-occurring data by systematically relaxing the controls inherent in a laboratory

[1] As Stephen Stigler (1978) notes, Peirce and Jastrow (1885) used randomization to create sequences of binary treatments in a psychological study of weights.
[2] For more on this topic see the excellent and lucid explanations within Imbens and Rubin (2020) and Cunningham (2021).

experiment. Harrison and List (2004) propose experimental features that can be used to determine the field context of an experiment. Leveraging these features, we can discuss a broad classification of three main types of field experiments that have emerged in this third wave—artefactual, framed, and natural field experiments, as defined above.

As a corpus, the literature reveals that the skepticism of the 19th and 20th centuries towards economic experimentation was ill-founded. My speculation is that these brilliant though skeptical minds, which included several Nobelists and others who did Nobel-worthy work, were simply locked into the experimental approach of the hard sciences. Recall, for example, as High Schoolers we were all taught "no clean test tubes, no clean information" in our Chemistry labs. I suspect that the skeptical economists simply generalized to economics without thinking about how randomization can solve the counterfactual problem.[3]

In this manner, it is safe to conclude from this section that rather than follow the approach of yesterday's economist, today's economist both models natural disruptions that affect the pin factory and occasionally generates data from the pin factory via experimentation to test economic theories. In a nutshell, Exhibit 1 provides an ocular glimpse of modern empirical work in economics.

3. Where to Tomorrow?

If I had my druthers, tomorrow's economist will surgically control the assignment mechanism to provide causal evidence of the inner workings of the pin factory and generate data necessary to understand if learnings from that environment will transfer to other environments, be them spatially, temporally, or scale differentiated. To operationalize that

[3] Indeed, it is not difficult to find further pessimism amongst economic greats:
"*Unfortunately, we can seldom test particular predictions in the social sciences by experiments explicitly designed to eliminate what are judged to be the most important disturbing influences. Generally, we must rely on evidence cast up by the 'experiments' that happen occur"*
~Milton Friedman (1953, p.10)
"Economists cannot make use of controlled experiments to settle their differences: they have to appeal to historical evidence"
~Joan Robinson (1977, p.1319)

objective, let us consider two types of scientific evaluation problems. First, is Experimental Problem 1 (EP1):

Experimental Problem 1: Measuring the causal impact of treatments, and determining relevant mediators and moderators in an ethically responsible manner

The first part of EP1 is understanding the "effects of causes;" or what the literature denotes as internal validity. The second part of EP1 demands learning the "causes of effects." Traditionally, the major goal of the treatment effects literature has been the first part of EP1 whereas the economics literature that builds explicit models focuses on the second part of EP1. For example, conducting a research study that shows paid parental leave affects child outcomes because it allows parents more time to read to their child (mediation), and that the estimated treatment effect is moderated by household income, is but one example of EP1 in action.

While the core of research relates to EP1, since therein lies important tests of theory and enhanced understanding of the world, in many cases our explorations should also be designed with an eye toward generalizing, leading to a second experimental task:

Experimental Problem 2: Predicting whether the causal impacts of treatments implemented in one environment transfer to other environments, be them spatially, temporally, or scale differentiated

EP2 includes external validity, but that is only part of EP2. The external validity problem relates to whether we can transport insights gained from one population (or one domain) to another, be that a behavioral response, a structural parameter, or a set of parameters. The health care program worked in Tennessee but will it work in London? The returns to a year of education are 12% per year in the US, is the return similar in Latin America, Asia, and Europe? Such questions are matters of what the literature typically denotes as external validity.

Yet, EP2 also contains elements that go beyond simply exploring treatment effects. For instance, contained within EP2 is also the notion of vertical scaling: the new curriculum worked in one school in Los Angeles, but would it be as efficient if we started the program in 1000 schools in Los Angeles? When scaling programs, policymakers need to consider both horizontal and vertical scaling, but often they consider only one (or neither). This commonly leads to a voltage effect whereby promising initial results are not delivered at scale (List 2022, 2024).

.

With these two tasks in mind, we consider several paths of advice for field experimenters. These paths are meant to "double down" on the extant comparative advantage of field experiments to help cement the promise and growth of field experiments in the social sciences.

3.1 Design to Understand Heterogeneity *and* Causal Moderation

The job training program affected low-income people more than high-income people. Older workers are less responsive to health benefits than younger workers. The middle class is more responsive to marginal income tax rate changes than the rich. The academic literature and policymaking community are littered with facts concerning policy heterogeneity of these types. Fundamentally, the ideal parameter concerning heterogeneity is the variance of the individual treatment effect, $\text{Var}[\tau_i] > 0$. Yet, more generally, standard applications of the potential outcomes framework and experimentation use "between-subject" designs (a subject is either in treatment or control, not both). This choice means that the vital problem of causal inference prevents us from obtaining information about the joint distribution of potential outcomes, leading us to make inference from information on the marginal distribution of potential outcomes.

In such cases, the analyst commonly calculates average treatment effects conditional on some pre-determined characteristics or moderators (X_i) yielding estimates of conditional average treatment effects. The conditional average treatment effect (CATE) for $X_i = x$ is defined as

$$\tau(x) \equiv \text{D}[\tau_i | X_i = x] = \text{D}[Y_i(1) - Y_i(0) | X_i = x], \qquad (2)$$

with the property that the ATE we defined in equation (1) can be re-expressed as $\text{D}[\tau(X_i)]$, taking expectations over the distribution of X_i. In the literature, individuals with a common $X_i = x$ are often referred to as a subgroup, leading to the use of subgroup average treatment effect as a synonym for CATE.[4]

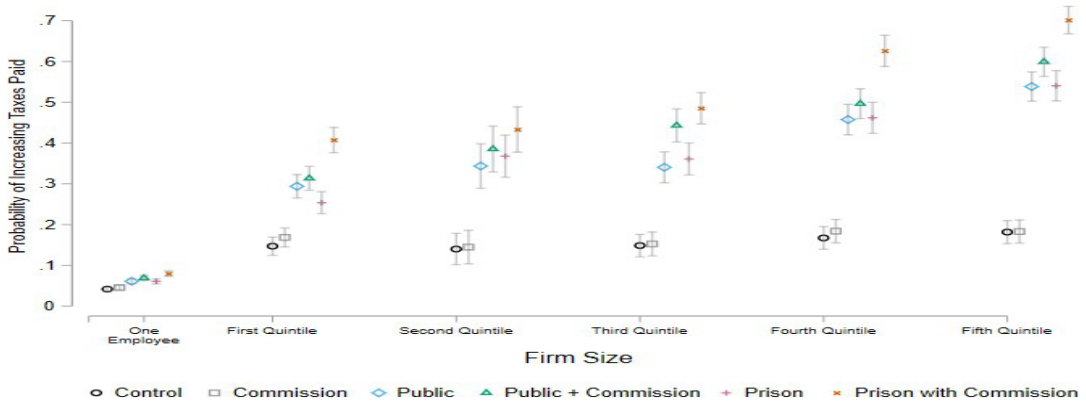
[4] Another sort of heterogeneity that is often discussed by social scientists relates to describing individuals: "women are more generous than men"; "non-whites are discriminated against in labor markets than whites"; "the elderly are more risk averse than the young". Such measures arise when we use the experimental approach to measure heterogeneity in preferences, beliefs, and constraints.

To make the points of this section most clearly, consider a recent example due to Holz et al. (2023). The authors partnered with the Dominican Republic IRS in an attempt to increase tax compliance amongst firms and the self-employed using a natural field experiment. They randomly allocated firms and self-employed workers into one of six groups. The control group is a simple reminder message sent to taxpayers before the tax deadline. To augment this baseline, they included two deterrence messages. One message increased the salience of potential prison penalties instituted under a new law. In contrast, the other increased the salience of potential social punishments by emphasizing potential public disclosure of punishments. Then, they interact with each message type by reframing tax mistakes as voluntary choices on the taxpayer's part.

Overall, the treatments increase taxes paid by $193 million (0.23% of GDP). Treatment effect heterogeneity is critical in Holz et al. (2023) because the authors can observe a unique sample of taxpayers across the entire firm size distribution. Even though the top 8% of taxpayers pay over 84% of all taxes and the top 1% of taxpayers pay 60% of all taxes, previous studies have only been able to explore behavior of small and medium-sized firms. Therefore, understanding heterogeneity in treatment effects across the full firm size distribution is quite important for policymakers. Indeed, $150 million of the $193 million extra raised funds came from the top 8% of firms. Exhibit 3 shows a summary of the result heterogeneity.

For clarity, I split firms into six categories based on their size the previous tax year. The first bin includes firms that have only a single worker. Then, I split the data into quintiles of firm size (using the population data) conditional on having more than one worker. Measured CATEs are the difference in the probability of increasing taxes paid in a treatment group relative to the control group.

Exhibit 3: Heterogeneity of Paying Taxes by Firm Size in Holz et al. (2023)



The graph reports the heterogeneity in treatment nudges across firm sizes. Firms are split into six categories based on their size the previous tax year. The figure shows that there is substantial heterogeneity in the efficacy of the deterrence nudges across firm size.

Exhibit 3 reveals unique heterogeneity but how should it be interpreted? If we return to equation (2), we find $D[\tau(X_i)]$. While this quantity has many important applications, one potential shortcoming, and the reason why this should be considered *descriptive* is that from this field experiment we do not learn whether there is something *particular* about the covariate X_i that is leading to the heterogeneity in responses, or if it is simply the case that X_i is correlated with other (potentially many) covariates that are the sources of the treatment effect heterogeneity.

Note to recover the CATE in equation (2), we require a different statistical independence condition:

$$\{Y_i(1, X_i), Y_i(0, X_i)\} \perp D_i | X_i.$$

In words, we now require that we have random assignment of D_i for each X_i. In such cases, the researcher now has the ingredients in place for a research design that is internally valid for each CATE over X_i: [5]

[5] That said, this is a property of the population, not of the sample. Researchers who calculate CATEs in a sample should check to see whether there is balance of baseline characteristics conditional on the observable characteristic they are conditioning on.

$$\begin{aligned}\tau(1) &= \mathrm{D}[\tau_i(1)|\mathrm{X}_i = 1] = \mathrm{D}[Y_i(1,1) - Y_i(0,1)|\mathrm{X}_i = 1],\\ \tau(0) &= \mathrm{D}[\tau_i(0)|\mathrm{X}_i = 0] = \mathrm{D}[Y_i(1,0) - Y_i(0,0)|\mathrm{X}_i = 0]\end{aligned} \tag{3}$$

Returning to Holz et al. (2023), the authors can obtain internally valid estimates of the CATEs for both small and large firms. However, they cannot obtain an internally valid estimate of the *causal effect of firm size* on the outcome or the treatment effect. The counterfactual question about how firm size changes the treatment effect is a statement about how the CATE changes with values X_i. To understand this quantity, we can first define the parameter of interest, $\Delta\tau(\mathrm{X}_i)$:

$$\Delta\tau(\mathrm{X}_i) \equiv \mathrm{D}[Y_i(1,1) - Y_i(0,1) - Y_i(1,0) - Y_i(0,0)] \tag{4}$$

This object is also not directly observable, both because we only observe each unit with a single potential outcome, but also because we only observe each unit with a single level of X_i. Instead, researchers can measure the difference between the CATEs, $\Delta\tilde{\tau}(\mathrm{X}_i)$.

$$\Delta\tilde{\tau}(\mathrm{X}_i) \equiv \tau(1) - \tau(0) = \mathrm{D}[\tau_i(1)|\mathrm{X}_i = 1] - \mathrm{D}[\tau_i(0)|\mathrm{X}_i = 0] \tag{5}$$

Using these equations, we can learn from our data whether the CATE is smaller or larger when firm size changes. However, this does not answer the question of whether there is a causal effect of firm size on the treatment effect. To see why, add and subtract a missing counterfactual, $\mathrm{D}[\tau_i(0)|\mathrm{X}_i = 1]$ to rewrite equation $\Delta\tilde{\tau}(\mathrm{X}_i)$ as the sum of the firm size effect on the treatment effect and component due to heterogeneity.

$$\begin{aligned}\Delta\tilde{\tau}(X_i) = &\underbrace{\mathbb{E}[\tau_i(1) - \tau_i(0)|X_i = 1]}_{Treatment\ differences\ due\ to\ firm\ size}\\ &+ \underbrace{\mathbb{E}[\tau_i(0)|X_i = 1] - \mathbb{E}[\tau_i(0)|X_i = 0]}_{Treatment\ differences\ due\ to\ heterogeneity}\end{aligned} \tag{6}$$

Equation (6) reveals that when we consider CATEs, we reintroduce a bias that looks a bit like selection bias but has a different character. The first term in equation (6) is the effect of firm size on the treatment effect. The second term is how the CATE for small firms would have differed had those units been identical to the large firms except for their size. To regain internal validity of $\Delta\tilde{\tau}(\mathrm{X}_i)$ for $\Delta\tau(\mathrm{X}_i)$, we introduce yet another distinct statistical independence assumption:

$$\{Y_i(1, \mathrm{X}_i), Y_i(0, \mathrm{X}_i)\} \perp \mathrm{X}_i. \quad (7)$$

This assumption rules out the existence of differences in potential outcomes that correlate with X_i and predict the treatment effects. For this assumption to be credible, researchers should explicitly randomize X_i. When this is impossible, researchers should attempt to account for all relevant differences between units of different types and use great caution when ascribing a causal interpretation to treatment effect heterogeneity, or to the relevant X_i.

This leads us to causal moderation. While much of our work on heterogeneity relates to descriptive analysis, building scientific knowledge cannot stop there. This is because, much like establishing causality of a specific treatment, to advance EP1 and EP2 we must understand causal moderation. To deepen our understanding of causal moderation, I urge researchers to randomize explicitly X_i in their original designs. In the Holz et al. (2023) case this might seem impossible, but developing hypothesis about why firm size matters to tax compliance might allow some unique aspects of X_i to be randomized.

For example, one might suspect that larger firms are more likely to have a tax compliance office or at least an accountant employed at the firm. This variable is ripe for randomization. Likewise, it might be the case that the firm size effect is a "deep pockets" effect whereby large firms have greater financial exposure. Again, beliefs about financial exposure are possible to randomize in an experiment.

3.2 Design to Understand Mediation Paths

In this section, I consider mediation analysis, which seeks to answer a different question than moderation. As mediators are directly changed by treatment, they may yield a causal chain, or mechanism, through which treatments yield effects. Thus, mediation analysis seeks to understand: what is the causal pathway from treatment (D) to outcome (Y)? Importantly, strong assumptions are necessary to obtain such information when mediation is not considered at the design stage. Yet, when considered in the design stage, much like the case of causal moderation, researchers can leverage weaker assumptions to understand mechanisms.

To bring things to life, consider a famous example of mediation that goes back to the 18th century, a time when diseases, such as scurvy, were often more dangerous for sailors than enemies during long sea trips. Unaware of the disease they carried, crews were plagued by symptoms of scurvy, including fungous flesh and putrid gums. However, during these times, little was known about the disease and its causes. Facing the outbreaks, an array of imaginative remedies was tried and defended. In 1747, the medical apprentice James Lind, carried out one of the first controlled clinical trials recorded in medical sciences on board HMS Salisbury.

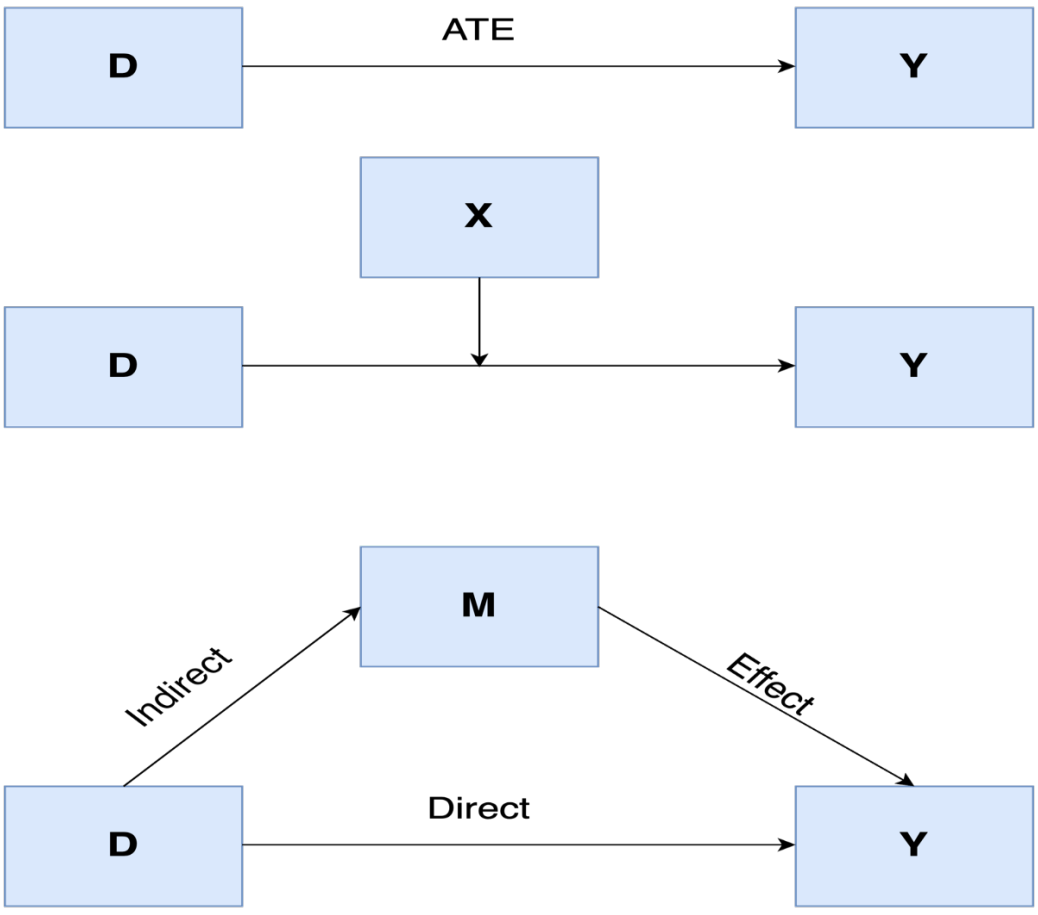
He took 12 men suffering from similar symptoms of scurvy, divided them into six pairs, and treated them with different suggested remedies: 25 drops of the elixir of vitriol three times a day, half a pint of seawater a day, two spoons of vinegar three times a day, two oranges and one lemon a day, among others. By the end of the week, those subjects on the citrus fruits were feeling much better to the point that they were able to nurse the others. Based on this framed field experiment, Dr. Lind wrote his "Treatise of the Scurvy" (Lind 1753) arguing that giving lemon juice to sailors would finally end the disease that killed more people than the French and Spanish armies combined. It was not until years later, via a 1940 study, that Vitamin C was identified as the complete mediator in the Lind experiment, and therefore represented the causal pathway to cure scurvy (Crandon et al., 1940).

At this point, it might be useful to summarize how mediation fits in with our learnings thus far. Consider the simple schematic in Exhibit 4. The top panel of Exhibit 4 shows where we started —our primary goal was to recover the ATE, $\tau = \mathrm{D}[Y_i(1) - Y_i(0)]$. Upon imposing our four identification assumptions, we had confidence in recovering the ATE. The middle panel of Exhibit 4 reveals an ocular depiction of CATEs for $\mathrm{X}_i = x$, as defined as: $\tau(x) \equiv \mathrm{D}[\tau_i(x)|\mathrm{X}_i = x] = \mathrm{D}[Y_i(1, x) - Y_i(0, x)|\mathrm{X}_i = x]$. An entirely different issue arises when the dimension of heterogeneity sits *along* the causal path, as a mediator *in between* the treatment that we experimentally vary and the outcome of interest. The bottom panel of Exhibit 4 shows this case.

Mediation refers to the transmission of the effect of treatment (*D*) through one or more other variables, termed mediators (*M*). As the bottom panel of Exhibit 4 reveals, mediation

corresponds to an indirect effect of D on Y that passes through M. Mediation can either be complete or partial. With complete mediation, the entire (or total) effect of D on Y is transmitted through one or more mediator variables. Consider the case of James Lind's work on scurvy. In this case, the entire effect of the citrus runs through enhanced Vitamin C levels. Vitamin C enhancement is the complete mediator of Lind's citrus treatments. Of course, this is a special case of mediation as in many instances within economics there might be only partial mediation, or several mediators that the scientist can explore. Consider the effect of a job training program on future wages: it is possible that job training programs can provide participants with better work attitudes, habits, skills and credentials, all of which can mediate the program's effect on future job prospects.

Exhibit 4: How Mediation Fits with ATEs and CATES



The diagram illustrates how mediation fits into the objective of recovering the ATE. The top panel shows that our objective is to obtain the ATE, and the middle panel shows that when we explore moderation we aim to calculate CATEs. In this section, we focus on the case illustrated in the bottom panel. That is, instead of the treatment (D) having *only* a direct effect on the outcome (Y), its effect also passes through other variable(s), called the mediator(s) (M), and is thus indirect.

One popular design approach that is often taken in the literature to estimate the various parameters of interest in a mediation analysis is separate randomization of the treatment D and mediator M. Concretely, separate randomization takes the general form: conduct a first experiment whereby $D$is randomized, and measure outcome M followed by a second experiment that randomizes M and measures outcome Y. In the case of Lind, this would have the experimenter randomizing both the citrus over one group and Vitamin C over another. In the end, the data would inform Lind whether Vitamin C is the mediator.

This approach is useful, but new assumptions must be invoked to recover an internally valid estimate. Concretely, one new assumption requires that the value of the mediator is as good as randomly assigned, even though the researcher did not have direct control over its level. In this case, the treatment must be assigned such that it is statistically independent of all relevant potential outcomes and mediators, and then conditional on the first-stage treatment assignment, the mediator must be statistically independent of potential outcomes. This is called the sequential ignorability assumption (see List, 2025, for a discussion).

Another type of approach used to understand mediation paths is to use surveys alongside the main experiment. Such surveys, when conducted appropriately, can help to estimate key variables that change from baseline due to the experimental treatment. Then, an appropriate analysis can decompose the direct and indirect effects of treatment (see Baron and Kenny, 1986).

A third approach to understand mechanisms is to use randomization in sub-treatments. Consider List (2004), where I explored the nature and extent of discrimination. I used a main treatment to measure whether discrimination exists in the studied market—it does—and I used sub-treatments to explore the underpinnings for the observed discrimination. Via those sub-treatments, I was able to reject Becker's (1975) theory of taste-based discrimination in favor

of statistical discrimination in the third-degree price discrimination Pigou sense (Arrow, 1972; Phelps, 1972)

3.3 Design to Understand Generalizability and Scaling

In its simplest form, questions of external validity revolve around whether the results of the experiment can be generalized to different people, situations, stimuli, and time periods. In the sciences, the issue is not a new one. Mill's assumption of the lawfulness of nature argues that the "distance" in time, space, population, and the decision environment between the study setting and the setting of ultimate interest determine behavioral connectedness. The literature has often used "external validity" and "generalizability" interchangeably because external validity pertains to the question of generalizability. In this manner, another way to characterize the generalizability of results is to describe their portability.

In general, the notion of generalizability has similar intuitions as CATEs. What is necessary is to not only allow heterogeneity by subject-specific covariates X_i, but also permit heterogeneity by domain/environmental features, or situational characteristics. For instance, if we extend equation 2 to include elements of the experimental environment, we can write the model as attempting to obtain internally valid estimates of

$$\tau = \mathrm{D}[Y_i(1) - Y_i(0) | E, A, X_i] \quad (8)$$

In equation (8), τ is the ATE for participants in a given environment, E, with a given level of experimenter scrutiny, A, and participant characteristics X_i. Taking identification of τ as given, we now want to know whether we can expect these results to hold when we vary E, A, X_i. The choice of what variables to include in equation (8) depends on theory and previous empirical evidence. In this case, if the ATE remains the parameter of interest, we are now seeking to estimate the ATE for the population of interest with characteristics, $X_i = x^\star$, in the environment of interest, $E_i = e^\star$ and where scrutiny is set at the target, $A_i = a^\star$.

One method to create an understanding of how our new factors, E and A, influence behavior is to draw a representative sample from the population E and A features. One line of thought is that with multiple potential locations, if the researcher chooses locations at random in an

initial stage of the experimental design, then this will lead to information to help generalize across these parameters. This approach, which is akin to gathering information on covariates in the experimental sample and estimating CATEs is useful when a good assortment of situational features are in the experimental sample. The researcher can then average over the distribution of A and E features in the target population, "adjusting" treatment effects accordingly. This approach is often used when the researcher constructs a random sample that is useful for interrogating heterogeneous treatment effects by using multi-site trials and then relies on probability re-weighting and effect estimation using functional form assumptions to transport insights.

Scaling

Beyond testing theories and deepening scientific knowledge, experimentalists are also attempting to speak to policymakers. One crucial question for the experimental research agenda in this area is: can this idea work at scale? In its simplest form, this question relates to the proliferation of an idea or policy from a small group—students at a certain school, for example—to a larger group in more diverse situations. While its import is undeniable, the scaling process is not simple, with pitfalls present every step of the way, running from the seed of an idea to well after policy launch. One consequence of having multiple fault lines is that ideas tend to experience a "voltage drop": the benefit-cost profile depreciates considerably when moving from the small to the large (List, 2022). One approach to combat voltage drops is to engage in "Option C Thinking" (List, 2024).

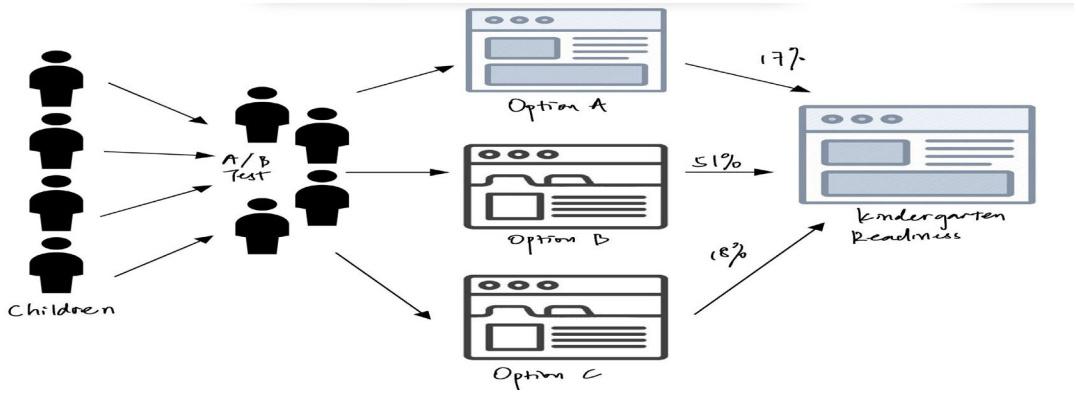
Consider Exhibit 5. In the A/B experimental test of an early childhood program summarized in the top two arms of Exhibit 5, the program is found to triple Kindergarten Readiness: from 17% to 51%! One might view this result as extraordinary, and immediately want to scale the program. To understand why that choice is not prudent, consider what exactly we have learned from this research. If it is a typical social science experiment, then it has likely been conducted as an *efficacy test*: the "best case" test of the program is arm B versus the control, arm A. To understand why more information is necessary, we must consider the incentives that the researchers faced. Those incentives are set up to create a petri dish that provides results that gives the intervention its "best shot," or likewise the largest treatment effects.

In this manner, we are answering the wrong question if we are attempting to provide policy advice. We are asking: can this idea work in the petri dish under the best-case situation rather than will this idea work at scale? This is the wrong question. We must not only do the efficacy test, but *also* relevant tests of scale within the original discovery process. The economics of many situations demand such an approach.

I refer to this new approach as Option C thinking. This is not because it represents a simple treatment arm to augment standard A/B testing. Rather, leveraging Option C thinking in our initial designs flips the traditional research model from efficacy trials to an approach that produces the type of policy-based evidence that the science of scaling demands. The Option C analogy is meant to take the initial discovery process from one of focusing purely on the details of an efficacy test to engaging in a bigger picture view, including questions such as: what constraints will the idea face at scale, what key factors can impact scaling, how can theory help me understand mediation paths and moderators in place at scale? Such evidence should be generated in the *original* design alongside the efficacy test. In this vein, theory plays an even greater role in design than usually imagined by experimentalists.

To complete our thought experiment in its simplest terms, in Exhibit 5, when following this approach, we add treatment arm C, which is a program that includes the constraints or warts that the idea will face at scale. After doing so, we find that the program that will be scaled minimally increases the Kindergarten Readiness—from 17% to 18%—a result that is not statistically significant relative to control and certainly will not pass a benefit-cost test.

Exhibit 5: Introducing Option C Thinking to the Classic A/B Testing Approach



The cornerstone of the experimental approach in the social sciences is A/B testing. For example, to test whether an education program works, the researcher takes a group of children and splits them into two groups: a control "A" group that does not receive the intervention and the "B" group that receives the early education program. This represents the top two arms of the Exhibit. If the experiment satisfies the classical identification assumptions[42], then a causal effect can be recovered. In this example, the program moves Kindergarten Readiness from 17% (Option A: control) to 51% (Option B: treatment).

3.7 Design Experiments Ethically and Responsibly

Beginning in 1932, 600 Black men from Macon County, Alabama, mostly poor and illiterate, were enrolled in a study of the progression of untreated syphilis, which at the time had no known treatment. Syphilis is a disfiguring, neurodegenerative, and deadly disease caused by the bacteria *treponema pallidum palladium* and is transmissible sexually and congenitally. By 1947, medical science established that penicillin could cure syphilis. Nonetheless, the doctors running the study withheld treatment from these men and their families for another 25 years until Jean Heller broke the story of the soon-to-be-infamous Tuskegee study.

National outrage followed, initiating landmark congressional hearings on research ethics involving human subjects. The result of those hearings was the National Research Act of 1974, which formed the National Commission for the Protection of Human Subjects of Biomedical and Behavioral Research. The Commission's main goal was to "identify the basic ethical principles that should underlie the conduct of biomedical and behavioral research involving human subjects and to develop guidelines which should be followed to assure that such research is conducted in accordance with those principles." The principles the Commission took to underlie the ethical conduct of human subjects' research were codified in the Belmont Report (1979), which formed the basis of The Common Rule (2018), which was adopted by several U.S. federal departments and agencies in 1991.[6]

Within the social sciences, there are also examples of ethically dubious experiments. One example is the Milgram experiment (1963), which studied obedience in the laboratory via application of electrical shocks to another individual. Milgram hypothesized that obedience

[6] The Belmont Report joined the other codifications of proper and responsible conduct of human experimentation. The best known of these codes are the Nuremberg Code of 1947 (re-published 1996) and the Helsinki Declaration of 1964 (re-published 1996) (whose latest iteration is the 2013 revision (2014)).

may be a deeply ingrained behavioral tendency. Upon selecting into the experiment, Milgram (1963) informed subjects that the shocks they were giving an out-of-sight person were part of a "learning experiment" to study the effects of punishment on memory. The experiment began with the subject and a confederate drawing slips of paper from a hat to determine roles. In each instance, the confederate was assigned to an "electric chair" and the subject was ordered to deliver an electric shock based on the performance of the victim.

The subject used a shock generator with clearly marked voltage levels, indicating the range from "Slight Shock" to "Danger: Severe Shock". The scene was set to convince the subject they were truly delivering shocks to the confederate. The experimenter would inform the subject that "Although the shocks can be extremely painful, they cause no permanent tissue damage." For each additional incorrect answer, the subject was instructed to deliver increasingly intense shocks. When the subject complied, this was considered obedience. If the subject protested to continuing, then the experimenter recited "prods" which were more direct the more the subject protested. If the subject still refused after the fourth "prod," a direct order, the experiment was terminated.

As one might anticipate, throughout the experiment, as noted by Milgram (1963), many subjects showed signs of "extreme tension". Milgram even noted that "uncontrollable seizures were observed for 3 subjects". Milgram (1963) found that no subject stopped before level 20 (out of 30), and out of 40 subjects a total of 14 disobeyed the experimenter at some point. After the experiment, many subjects repeatedly claimed that they were not sadistic types. Milgram (1963) transcribed some of the subject's responses with one man noting "I'd like to continue, but I can't do that to a man… You take your check…". In sum, the experiment resulted in two surprising findings. The first being the obedient tendencies, and the second being the "extraordinary tension generated by the procedures."

The astute reader might ask, where are ethical considerations in our model that recovers an ATE, as per equation (8)

$$\tau = \mathrm{D}[Y_i(1) - Y_i(0)|E, A, \mathrm{X}i]$$

Quite simply, they have not been explicitly modeled, though we likely each have our own morality barometer when it comes to human experimentation. Yet, some guideposts should be discussed, and it is not difficult to include such considerations in our framework.

Consider the case when the researcher not only values the creation of scientific knowledge, as per equation (8), but also values the welfare of experimental subjects and innocent bystanders. Within EP1, this is the "in an ethically responsible manner" consideration. We can extend the model above to create the scientist's objective function:

$$U_s = f(\pi_{kc}, \pi_s, \pi_{ib}) \qquad (9)$$

where U_s represents the utility of the scientist, which is a function of three inputs: the importance the scientist places on knowledge creation (π_{kc}), the welfare of subjects (π_s) and the welfare of innocent bystanders (π_{ib}).

For example, concerning π_s, Milgram (1963) might have reasoned that to explore obedient tendencies he needed to use outright deception so subjects would believe they were hurting the confederate, even though he saw that shocking the confederate gave the subjects disutility. In the design stage, experimentalists must make decisions within and across the three arguments in equation (9) and when they conflict with their moral code, they must decide its resolution. In such cases, it is often trading off π_{kc} against π_s and π_{ib}.

There are several ethical theories to prescribe behaviors in this instance. I follow rule consequentialism, which is a theory in ethics that centers on the idea of following certain rules whose general adoption would lead to the best possible consequences. It differs from act consequentialism, which focuses on the consequences of individual actions. Under rule consequentialism, an action is considered morally right if it complies with a set of rules that, if generally followed, would create a greater amount of good in the world than any other set of rules.

In this framework, rules are justified by the good consequences that result from their general acceptance. This approach aims to provide a more structured and predictable ethical system than consequentialism, as it emphasizes adherence to rules rather than evaluating every action

on a case-by-case basis. This also means that in certain situations, following the rule might not lead to the best possible outcome in that specific instance, but the overall adherence to the rule is believed to lead to the best outcomes in general.

I adopt rule consequentialism because it strikes a balance between strict rule-based ethics (like deontology) and the flexible, outcome-focused approach of act consequentialism. It attempts to capture the benefits of having consistent moral rules while still ensuring that those rules lead to positive outcomes overall.

My rule consequentialism approach naturally leads to a benchmarking that can be followed by experimenters. I summarize that benchmarking in Exhibit 6. Exhibit 6 summarizes the features of a Gold, Silver, Bronze, and Plutonium-239 standard economic experiment. Equation (9) indicates the interplay between π_{kc}, π_{s}, and π_{ib} in considering the various categories. A first feature of my benchmarking is that its overarching spirit is driven by rule consequentialism, but it contains features that are distinctly consequential (i.e., to be a Gold study, no subjects can be made worse off). A second feature is that there are certain non-negotiables to even be considered as a Gold, Silver, or Bronze study: truthfully, unbiasedly, and transparently reporting results and conflicts of interest and appropriate data governance and data management.

The next three set of criteria of the Gold standard revolve around subject and bystander (including proctors) protections. The burden on Gold studies might be considered above and beyond even what the Belmont Report requires, providing criteria 3-5: 3) no subject is made worse off, in either expectations or outcomes, because of your experiment, 4) no bystanders (including proctors) are made worse off, in either expectations or outcomes, because of your experiment, and 5) consider fairness concerns in both a static and dynamic sense.

Considering Milgram's (1963) study, it clearly violates the standard of no subject is made worse off because of the experiment, and therefore pushes his work to Plutonium-239 status. Clearly both within the experiment and beyond, Milgram caused anguish to his subjects, with one man in the transcripts stating "I'd like to continue, but I can't do that to a man… You take

your check…". Milgram (1963) even noted that "uncontrollable seizures were observed for 3 subjects." After the experiment, many subjects repeatedly claimed they were not sadistic types.

The next aspect of our standards in Exhibit 6 relates to fairness concerns. This feature begins with a recognition of the selection of subjects for the experiment and takes the fairness concern dynamically. Recall that the justice principle is invoked due to a concern that there is an inherent trade-off between participation in the experiment and expected welfare: those who bear the burdens of research (i.e., those who are exposed to the discomforts, inconveniences, and risks) should receive the benefits in equal measure to the burdens. Yet, given the Pareto criteria in 3) and 4) of Exhibit 6, we are effectively flipping the script with economic experiments: those who bear the fruits of taking part in an experiment should not inordinately benefit from the research. As such, the argument in many economic experiments becomes one of unfairly *withholding* treatment. In this sense, treatment should only be withheld if dynamic fairness is violated: does it cause future generations to forego valuable knowledge. The researcher should always consider both static and dynamic trade-offs.

The last three considerations, Criteria 6-8, in my standards closely relate to subject protections: informed consent, introducing risk, and outright deception. 6), That informed consent should be used unless it compromises inferential integrity (EP1 and/or EP2 are clearly disturbed), arrives directly from the Belmont Report.

At this point it is useful to consider differentiating across the various standards. Comparing Gold and Silver in Exhibit 6, the key distinctions are that subject and bystander protection moves from "absolute" to "in expectations at randomization" and in NFEs the benefit/risk trade-off is relaxed in Silver studies. For example, to operationalize how criterion 3), no subject is made worse off, differentiates studies across the various benchmarks consider a simple thought experiment.

A researcher conducts a lab experiment whereby he takes five participants and randomly gives four of them $100 but for the fifth one he takes $50 away—that is participants must literally take money from their pocket, leaving the lab with less money than when they arrived. In this particular case, it is not a Gold study, because at least one person is made worse off, but it is a Silver study, because the expected value of participating in the experiment is positive. Now,

let us reverse the game and say four participants must give $50 and one person receives $100. In this case, it is not a Gold classification because at least one person is worse off; it is not a Silver study because in expectations participants are made worse off. The determination then comes down to whether any subjects are "unduly harmed." We can imagine that in some countries taking away $50 is devastating. In that case, it is more than undue harm and the study is Plutonium-239. In other countries, it might not be deemed as indue harm, and therefore would qualify as a Bronze study for this criterion.[7]

Finally, outright deception is allowed in Silver studies if it imposes only minimal harm. Moving to Bronze studies, criteria are lessened to include only 7, and subject protections revolve around both expectations and actual outcomes, but no subject should be unduly harmed. Finally, Plutonium-239 experiments are toxic in nature, and have at least one toxic feature related to fabrication or deceptive practices in reporting or unduly harming participants (subjects, innocent bystanders, or proctors).

[7] On the sidelines of the discussion of being made better or worse off are issues of opportunity cost of time, human capital benefits from taking part in and experiment, and altruism benefits of participating. These can all be included for completeness.

Exhibit 6: A Benchmark for Research Ethics

8 Golden Rules of Experimentation	8 Silver Rules of Experimentation	7 Bronze Rules of Experimentation	Features of Plutonium-239 Experiments
1. Truthfully, unbiasedly, and transparently report results and conflicts of interest	1. Truthfully, unbiasedly, and transparently report results and conflicts of interest	1. Truthfully, unbiasedly, and transparently report results and conflicts of interest	**1. Biased reporting of results and conflicts of interest**
2. Appropriate data governance and data management	2. Appropriate data governance and data management	2. Appropriate data governance and data management	
3. No participant is made worse off because of your experiment	**3. At randomization, participants, in expectations, are not made worse off** because of your experiment	**3. In expectation, no participants are unduly harmed**	**2. Subjects, proctors, and/or innocent bystanders are unduly harmed** due to experimental procedures or mishandling of data
4. No bystanders are made worse off because of your experiment	**4. At randomization, bystanders, in expectations, are not made worse off** because of your experiment	**4. In expectation, no innocent bystanders are unduly harmed**	
5. Consider fairness concerns in both a static and dynamic sense	5. Consider fairness concerns in both a static and dynamic sense	5. Consider fairness concerns in both a static and dynamic sense	
6. Informed consent should be used unless it compromises inferential integrity (EP1 and/or EP2 are **clearly** disturbed)	6. Informed consent should be used unless it compromises inferential integrity (EP1 and/or EP2 are disturbed)	6. Informed consent should be used unless it compromises inferential integrity (EP1 and/or EP2 are disturbed)	
7. When NFEs are used, **do not introduce new risks, or enhance those risks already present in the environment** (business as usual is preferred)	7. When NFEs are used, **do not introduce more than minimal risks, or more than minimally enhance those risks already present in the environment without commensurate benefits to the exposed** (business as usual is preferred)		
8. No outright deception	8. No outright deception, **unless it imposes only minimal harm**	7. No outright deception, **unless it imposes only minimal harm**	**3. Outright deceptive**

The exhibit summarizes the different ethical standards of an economics experiment: Gold, Silver, Bronze, and Plutonium-239. The standards should be used as a rough guideline to help recognize trade-offs in design choices.

4. Epilogue

In the past few decades there is perhaps no empirical innovation that has changed economics more than field experiments. Via controlling the assignment mechanism, the experimenter sheds light on both the "effects of causes" and the "causes of effects." Yet, the scientific insights do not end there. With some imagination and theoretical guidance, the experimenter can generate data that permits an informed prediction of whether the causal impacts of treatments implemented in one environment transfer to other environments, be them spatially, temporally, or scale differentiated. When these dual goals are achieved, the power of the experimental approach is unleashed.

Perhaps the most difficult aspect of writing this study was choosing the limited number of examples to discuss to advance my argument that our field needs to "double down" on the

unique comparative advantage of field experiments. If I had another dozen pages or so I would have begun by including longitudinal aspects of learning, and how field experiments can provide unique surrogates that help us to go beyond estimating short-run substation effects and help us to understand difficult long-run issues with high causal density. Likewise, I would have expanded on how within-subject designs aid in experimental power and are able to estimate the full joint distribution of outcomes, allowing us to go beyond a comparison of marginal distributions, or two conditional expectations.

Finally, I would have developed the argument that two unique features of the experimental approach situate it well to deepen the stock of scientific knowledge: selective data generation and the ability to enhance the notion, and role, of replications. In the end, aggregating results from qualitatively different empirical methods across different settings, subject pools, and selection rules will increase our confidence in building the knowledge framework necessary for testing theories and providing empirical advice. I trust that the next generation of researchers will take the baton and continue the rapid advances of learnings. This is because unlike yesterday's economist who passively visited the pin factory, tomorrow's economist will surgically control the assignment mechanism to tinker within the pin factory.

References

Angrist, Joshua D., and Jörn-Steffen Pischke. 2009. *Mostly Harmless Econometrics: An Empiricist's Companion*. Princeton University Press.

Arrow, Kenneth J. 1972. "Some Mathematical Models of Race Discrimination in the Labor Market." *Racial Discrimination in Economic Life*, 187–204.

Baron, Reuben, and David Kenny. 1986. "The Moderator-Mediator Variable Distinction in Social Psychological Research: Conceptual, Strategic, and Statistical Considerations." *Journal of Personality and Social Psychology* 51 (January): 1173–82. https://doi.org/10.1037//0022-3514.51.6.1173.

Becker, Gary S. 1975. *The Economics of Discrimination*. Edited by 2d edition. Economic Research Studies. Chicago, IL: University of Chicago Press. https://press.uchicago.edu/ucp/books/book/chicago/E/bo22415931.html.

Crandon, John H, Charles C Lund, and David B Dill. 1940. "Experimental Human Scurvy." *New England Journal of Medicine* 223 (10): 353–69.

Cunningham, Scott. 2021. Causal Inference: The Mixtape. Yale university press.

Friedman, Milton. 1953. "The Methodology of Positive Economics." In *Essays In Positive Economics*, 3–16. Chicago: University of Chicago Press.

General Assembly of the World Medical Association. 2014. "World Medical Association Declaration of Helsinki: Ethical Principles for Medical Research Involving Human Subjects." *The Journal of the American College of Dentists* 81 (3): 14– 18.

Harrison, Glenn W., and John A. List. 2004. "Field Experiments." *Journal of Economic Literature* 42 (4): 1009–55. https://doi.org/10.1257/0022051043004577.

Holz, Justin E., John A. List, Alejandro Zentner, Marvin Cardoza, and Joaquin E. Zentner. 2023. "The $100 Million Nudge: Increasing Tax Compliance of Firms Using a Natural Field Experiment." *Journal of Public Economics* 218 (February): 104779. https://doi.org/10.1016/j.jpubeco.2022.104779.

Imbens, Guido W., and Donald B. Rubin. 2015. *Causal Inference in Statistics, Social, and Biomedical Sciences*. Cambridge University Press.

Levitt, Steven D., and John A. List. 2009. "Field Experiments in Economics: The Past, the Present, and the Future." *European Economic Review* 53 (1): 1–18. https://doi.org/10.1016/j.euroecorev.2008.12.001.

Lind, James. 1753. *A Treatise of the Scurvy in Three Parts*. Kincaid.

List, John A. 2004. "The Nature and Extent of Discrimination in the Marketplace: Evidence from the Field*." *The Quarterly Journal of Economics* 119 (1): 49–89. https://doi.org/10.1162/003355304772839524.

———. 2022. *The Voltage Effect: How to Make Good Ideas Great and Great Ideas Scale*. Crown.

———. 2024. "Optimally Generate Policy-Based Evidence before Scaling." *Nature* 626 (7999): 491–99. https://doi.org/10.1038/s41586-023-06972-y.

———.2025."Experimental Economics: Theory and Practice," University of Chicago Press.

Milgram, Stanley. 1963. "Behavioral Study of Obedience." *The Journal of Abnormal and Social Psychology* 67 (4): 371–78. https://doi.org/10.1037/h0040525.

Mill, John Stuart. 1844. “On the Definition of Political Economy; and on the Method of Investigation Proper to It.” In *Essays on Some Unsettled Questions of Political Economy*. https://www.gutenberg.org/files/12004/12004-h/12004-h.htm.

Neyman, Jerzy. 1923. “On the Application of Probability Theory to Agricultural Experiments. Essay on Principles. Section 9.” *Statistical Science* 5: 465–72.

Peirce, Charles Sanders, and Joseph Jastrow. 1885. “On Small Differences in Sensation.” *Memoirs of the National Academy of Sciences* 3 (1): 75–83.

Phelps, Edmund S. 1972. “The Statistical Theory of Racism and Sexism.” *The American Economic Review* 62 (4): 659–61.

Protections (OHRP), Office for Human Research. 2010. “The Belmont Report.” Text. January 28, 2010. https://www.hhs.gov/ohrp/regulations-and-policy/belmont-report/index.html.

Reuben, Ernesto, Sherry Xin Li, Sigrid Suetens, Andrej Svorenčík, Theodore Turocy, and Vasileios Kotsidis. 2022. “Trends in the Publication of Experimental Economics Articles.” *Journal of the Economic Science Association* 8 (1): 1–15. https://doi.org/10.1007/s40881-022-00117-z.

Robinson, Joan. 1977. “What Are the Questions?” *Journal of Economic Literature* 15 (4): 1318–39.

Rubin, Donald B. 1974. “Estimating Causal Effects of Treatments in Randomized and Nonrandomized Studies.” *Journal of Educational Psychology* 66: 688–701. https://doi.org/10.1037/h0037350.

Samuelson, Paul A., and William D. Nordhaus. 1985. *Economics*. 12th ed. New York: McGraw-Hill.

Stigler, Stephen M. 1978. “Mathematical Statistics in the Early States.” *The Annals of Statistics* 6 (2): 239–65. https://doi.org/10.1214/aos/1176344123.

The Belmont Report. 1979. “Ethical Principles and Guidelines for the Protection of Human Subjects of Research.” https://www.hhs.gov/ohrp/regulations-and-policy/belmont-report/read-the-belmont-report/index.html.

The Common Rule, Office for Human Research. 2017. “2018 Requirements (2018 Common Rule).” Text. March 7, 2017. https://www.hhs.gov/ohrp/regulations-and-policy/regulations/45-cfr-46/revised-common-rule-regulatory-text/index.html.

“The Nuremberg Code (1947).” 1996. *BMJ: British Medical Journal* 313 (7070): 1448–1448. US Dept of Health and Human Services. 1979. “The Belmont Report: Ethical Principles and Guidelines for the Protection of Human Subjects of Research.” https://www.hhs.gov/

Biography

John A. List is a Distinguished Service Professor in the Department of Economics at the University of Chicago. Among his research interests are testing and measurement using field and lab experiments within microeconomics. His work expands across several sub-areas, including behavioral economics, labor and education, personal finance, public economics, personnel economics, environment and energy, and charitable giving.