NBER WORKING PAPER SERIES

DON'T GIVE UP ON LAB EXPERIMENTS: WHY THE FIELD STILL NEEDS THE LAB

John A. List

Working Paper 35338
http://www.nber.org/papers/w35338

NATIONAL BUREAU OF ECONOMIC RESEARCH
1050 Massachusetts Avenue
Cambridge, MA 02138
June 2026

This article is drawn from various parts of my textbook: Experimental Economics, Theory and Practice. In writing that text I came to appreciate that in economics we are only in the early innings of the maturation of the experimental approach. I eventually came to the question: why would we want to give up on a major creator of knowledge? That question, and the subsequent answer to why we should not, led to this study. This study was enhanced by comments from Nour Wael Abdelbaki, Anna Dreber, Faith Fatchen, Paul Kirch, and Julia Seither. All errors remain my own. The views expressed herein are those of the author and do not necessarily reflect the views of the National Bureau of Economic Research.

Don’t Give Up on Lab Experiments: Why the Field Still Needs the Lab
John A. List
NBER Working Paper No. 35338
June 2026
JEL No. C83, C89, C9, C91, C92, C93, H0, J01

ABSTRACT

Recent enthusiasm for field experiments, and especially for natural field experiments (NFEs), in which subjects go about their daily activities unaware that any study is taking place, has sometimes been read as a verdict against the laboratory. I argue that such a verdict is wrong. Through the lens of a simple rational-choice model, I show that the four standard experimental designs (laboratory, artefactual field experiment (AFE), framed field experiment (FFE), and NFE) are comparative-static restrictions of one maximization problem, each identifying a parameter the others cannot. The model reveals that each design type has distinctive strengths and weaknesses across various dimensions of knowledge creation, including the enforcement of the conditions for causal identification, the faithfulness of the experimental environment to the theory being tested, the identification of economic primitives via theoretical structure, and the ethics of studying human subjects. On each of the dimensions, the four design types are complements rather than rivals. Nowhere is this complementarity more evident than between the two extremes. The lab enforces the conditions for causal identification that the NFE must inherit from the market; the NFE recovers the parameter that governs behaviour in the wild, free of the selection, scrutiny, and environmental distortions the lab cannot escape. A research programme using all four designs together demonstrates something no single design can produce. The framework further accommodates the recent rise of online and survey experiments as natural extensions. Our discipline’s recent drift away from laboratory evidence is leaving an important structural gap that natural field experiments, however well conceived, cannot fill.

John A. List
The University of Chicago
Department of Economics
and Australian National University
and also NBER
jlist@uchicago.edu

> *"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. 144)

> *"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 to 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)

> *"Economists . . . 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."* — Paul Samuelson and William Nordhaus (1985, p. 8)

1 Knowledge Generation in Economics: From Skepticism to Experiment

The give and take between theory and data in the natural sciences is so ingrained in modern day thought that an integral part of the scientific method, that theories must be tested against experimental evidence, is now second nature. This fact, of course, was not lost on the legends of economics, many of whom felt compelled to express their anguish by comparing empirical approaches across the social and natural sciences. Indeed, a common thread in the epigraphs is that if economists desire to engage in empirical exercises, they should start looking for available naturally-occurring (or historical) data. This is presumably because it was impossible to conduct experiments or learn anything meaningful from experimental data in economics. The belief was that economic phenomena are too entangled with each other, too dependent on context, and too difficult to isolate in controlled conditions to be amenable to the experimental method that physics, chemistry, and biology had used to such effect.

Such feelings were widely shared by economists throughout the 19th and 20th centuries, as extracting knowledge from surveys, historical data, and personal introspection represented the primary sources, and indeed in most cases the only sources, of new empirical knowledge in economics. Had this view persisted, the discipline would have been left with a single empirical strategy: the analysis of data the world happens to provide, with all of the identification challenges that such a strategy carries.

That view has since been overturned. Vernon Smith's 2002 Nobel Prize was awarded for "having established laboratory experiments as a tool in empirical economic analysis." In the years since Smith's award, field experiments have grown from a handful of pioneering studies into a substantial fraction of empirical work in top journals (Reuben et al., 2022; Levitt and List, 2009). Furthermore, various sectors of our economy, from Google, Meta, Amazon, Netflix, Lyft, Uber, to Walmart, now conduct thousands of field experiments, validating products and pricing decisions on real customers in real markets (Kohavi and Thomke, 2017). Of course, private sector firms do not have a monopoly on the method, as today the experimental approach is commonly deployed in government, non-profits, and NGOs (see, e.g., Hallsworth et al., 2017; Zhang et al., 2022). Indeed, recently field experiments have even been centrally featured in major anti-trust litigation.[1] Experimentation is no longer at the margin of how economists generate knowledge; it is increasingly at the centre.

This rising tide produces new questions that the early experimental literature did not need to address. Interestingly, such questions are not challenges in other scientific quarters either. In the academy, we call sciences such as physics, chemistry, and biology the "hard" sciences. But when considering experimenting with humans, the social sciences are in a meaningful sense, *harder*. Sheep did not select into or out of treatment in Pasteur's famous anthrax vaccination experiments at Pouilly-le-Fort, while human participants in a vaccine trial choose, on the basis of their own expectations, whether to enroll. Plants in Mendel's monastery garden grew in whatever soil he placed them in, while human participants react to the artificial conditions of a stylized experimental environment in ways that may differ sharply from how they would behave in the markets and institutions the experiment is meant to illuminate. Radium did not behave differently under Marie Curie's watch as she refined it in her shed on rue Lhomond, while human participants who know they are being observed put more weight on what an observer would think and less on the income or convenience that drives their behaviour when no observer is present.

The three foundational features that arise in these examples (selection into the study, the created environment, and the awareness of being studied) are potential sources of behavioral change that social scientists must address at some point when trying to change the world with their research. Harrison and List (2004) introduced a taxonomy that included four experimental design types (laboratory experiments, artefactual field experiments (AFEs), framed field experiments (FFEs), and natural field experiments (NFEs)) organized around precisely these three features. Each of the four designs adopts a different position on who is studied, what environment they are studied in, and

[1]The case involved the FTC attempting to break up Meta due to alleged abuse of monopoly power (Federal Trade Commission v. Meta, 2025). In his ruling, Judge Boasberg noted the following about the received field experimental evidence: "This experiment, which in the Court's judgment offers the single best evidence of what consumers consider alternatives to Meta's apps, tells a clear and consistent story: when using Facebook and Instagram becomes more costly, users turn to TikTok, YouTube, and Snapchat, with no other app notably standing out." The underlying scientific study is Goodman et al. (2026).

whether they know they are being studied. Within economics, conducting each of the four design types is commonplace today. The behavioral wedges they create are in principle observable, though most researchers treat them as latent.

Some have argued that the proliferation of experimental designs has itself revealed a hierarchy. In the extreme view of that argument, the lab is viewed as merely a relic of an earlier era, and any serious empirical question should now be taken straight to the field. The stakes of this view extend beyond methodology to the institutional ecology of experimental economics itself. On this score, a few facts reveal troubling trajectories in the publishing ecosystem. First, the share of laboratory experiments published in our flagship economics journal, the *American Economic Review*, has more than halved since 2000; yet, at the same time, field experiments, natural experiments, and online studies have grown steadily across all leading general-interest journals (Reuben et al., 2022). Second, Fréchette et al. (2022) document the same trend when showing that the center of gravity in experimental publication has shifted decisively away from the physical laboratory. Charness et al. (2023a) eloquently respond directly to this shift, arguing that laboratory experiments retain methodological advantages that online and field platforms cannot fully replicate. A key conclusion of their work is that the observed shift in publication patterns is mainly due to the supply side, reflecting features in the landscape that have recently changed, including funding pressures and the substitute set becoming more attractive (e.g., online data collection).[2]

My core argument is that this trend, and indeed any verdict against laboratory experiments as a bona fide contributor to scientific knowledge, is wrong. I make this case via a formal framework that characterizes precisely what each design identifies and why all four design types are necessary for optimal knowledge generation. The model is a simple rational-choice model, and it reveals that each of the four design types identifies a parameter the others cannot, carries enforcement guarantees that the others lack, and faces construct risks the others avoid. Importantly, the four design types fall out as comparative-static restrictions on the three key feature switches. As such, the model reframes the case for laboratory experiments around knowledge generation rather than resting on control, cost, and convenience. What has been underappreciated is that in terms of optimal knowledge creation, internal validity and external validity are complements, not substitutes.[3] By characterising precisely

[2]These facts are complemented by some of my own anecdotal evidence. Back in the early 1990s, I received an anonymous referee report noting "This author clearly does not understand experimental economics, which is best done in the laboratory. . . . . .I strongly advise rejecting this "field experimental" paper." A recent referee report was decidedly different in nature: "This author performs a field experiment, which clearly is the preferred approach to answer questions within economics...." even though the conclusion was the same "but I still advise rejecting the paper."

[3]I often hear arguments along the lines of "internal validity (IV) and external validity (EV) are substitutes". This holds some truth at the individual researcher level, where the design-space tradeoff within a single study might have tighter control coming at the cost of more naturalistic conditions, for example. Yet, my argument is at the research program level, where IV without EV gives you a reliable answer to a narrow question; EV without IV gives you an unreliable answer to a broad one. The marginal value of each rises with the other (see the (S,E,A) decomposition in List (2026a,b)).

what each design identifies, including which parameters, for which population, in which settings, a principled place emerges for each experimental type in the cumulative evidence base. In doing so, we have a rigorous basis for combining evidence across design types rather than choosing which estimate is "correct" between them.

In developing my complementarity argument, the various distinctive contributions of each experimental design type become clear. For example, the lab has a great capacity to enforce the conditions of causal identification by design and to build a stylized setting that satisfies the assumptions of the theory being tested. The NFE's distinctive contribution is to deliver the average treatment effect for the operating market the researcher is studying, free of the distortions that overt designs introduce. Neither design substitutes for the other; both are needed, and so are the AFEs and FFEs that sit between them. The case for not giving up on lab experiments, even now that we have the field, is therefore a structural one. And, so is the case for not stopping at the lab (as was made to me in the early 1990s when I started presenting field experimental work).[4]

The distinction this paper makes is not merely methodological. Indeed, I argue that it has practical consequences for at least three audiences who use experimental evidence to make decisions. For academics, the choice among the four experimental types determines what parameter the researcher has identified. Even though such parameters might seem like the same object through the lens of a parsimonious theory, they can differ substantially. For instance, a behavioural regularity established in the laboratory, or in an AFE, or in an NFE, may produce three quite different estimates. The literature typically reports them as one phenomenon. This is due to a fundamental inferential shortcut: the discipline has not yet developed the habit of reading these as estimates of three distinct objects rather than as competing measures of a single object. Until it does, the cumulative knowledge base of experimental economics will conflate parameters that, by the structure of the designs that produced them, were never the same. In this spirit, the model in this study clarifies when inferential shortcuts are inappropriate.

For policymakers, a crucial first decision that must be made is how to generate insights that match their setting of interest. A situational mismatch can turn what seems like a scalable policy into an idea that will never scale, and in fact, never had a chance to scale from the beginning. For example, a laboratory study showing that an intervention that lowers the price of generosity increases giving in a dictator game with an elasticity of 2.5 is informative about something, but it might not be informative about what would happen if a city, a state, or a federal agency rolled

[4]Generalizing any estimate from the studied setting to a different target setting is a separate problem, common to every empirical strategy in economics. This paper appreciates this as an economic problem and aims to shed new insights into that literature via a different approach. For recent discussions in economics of external validity, see Levitt and List (2007a); Guala and Mittone (2005); Deaton (2010); Bates and Glennerster (2017); Al-Ubaydli and List (2013); Bold et al. (2018); Williams (2019); Soman and Hossain (2021); Szaszi et al. (2025); List (2020b); List (2024); List (2026b). In recent work (List (forthcoming)), I argue that to tackle the generalizability crisis, we should do more natural field experiments.

out tax breaks on charitable gifts tomorrow. Field studies within operating markets, and especially ones in which the participants do not know they are being studied, measure a closer cousin of the parameter the policymaker actually needs in such instances. A policymaker who treats laboratory and field evidence as interchangeable inputs into the same decision has effectively assumed that the difference between a controlled setting and the naturally-occurring world does not matter. The evidence assembled in the broader literature and discussed in this paper suggests otherwise. The model creates intuition into why received results might differ, and how we can generate data to answer relevant questions.

For firms and other organizations, the choice of design type determines how fluidly evidence translates into performance. An AFE finding that customers prefer one product feature over another tells the organization what its subjects said when asked. This can be quite useful, but potentially not the same type of evidence as when its customers choose the product naturally, without being asked. Most large technology firms learned this lesson in the early 2000s and built infrastructure to run field experiments on operating customers at scale. A common result is that decisions based on such NFEs outperform decisions based on focus groups and surveys by margins that show up in revenue and retention. Indeed, organizations that have yet to build such infrastructure are operating inside the production frontier, overweighting stated preferences and underweighting revealed preference evidence. The framework in this paper makes the wedge between those two things precise, and explains why closing it requires a particular kind of experiment rather than just more experiments.

Turning to methodological considerations, many in the behavioural sciences have treated the replication crisis as a statistical hygiene problem. Recently, a landmark effort to replicate psychology experiments found fewer than 40 percent produced significant results, with effect sizes roughly halved (Open Science Collaboration, 2015). A parallel exercise in experimental economics found roughly two-thirds replicated, with effect sizes again attenuated (Camerer et al., 2016, 2018). From such insights, scientific reforms followed, which included the important movements of pre-registration, PAPs, larger samples, and open data. Yet, the evidence suggests the crisis had two distinct sources that were not often distinguished.

As an example, consider Many Labs 2, which tested 28 effects across 36 sites. Its data suggest that replication success varies dramatically by location: an effect that holds in 30 sites can fail to appear in 6 others. Such data patterns are consistent with contextual contingency (Klein et al., 2018). In complementary work, researchers hold the population constant and vary the experimental protocol. For example, Huber et al. (2023) examined whether competition affects moral behavior by randomizing thousands of participants across several crowd-sourced research designs. They found design heterogeneity roughly 1.6 times the average sampling error in the data, suggesting design choices alone drive substantial variation in true effect sizes. Relatedly, in important work, Holzmeister et al. (2024) generalize this logic by decomposing effect-size variation into population,

design, and analytical components across 86 meta-scientific studies. Their finding is asymmetric and quite instructive for our purposes: design and analytical heterogeneity dominate population heterogeneity. And, incorporating these sources roughly doubles standard errors and confidence intervals. The implication might be viewed uncomfortably by practitioners: standard inference treats variation across designs and analysis paths as noise to be averaged away or, more commonly, as variation that simply does not exist or is assumed away. Their results suggest it is a first-order source of uncertainty that the field systematically under values. Even within the Open Science Collaboration dataset, cross-national replications failed at substantially higher rates than domestic ones (Inbar, 2016). Yarkoni (2022) addresses the underlying problem by noting that received insights purporting to be general are typically supported by narrow samples and settings. Much of what reads as irreproducibility, therefore, is better understood as unacknowledged design-dependence which is precisely what a taxonomy distinguishing lab, AFE, FFE, and NFE designs is built to surface.

One implication of this literature is that a study obtaining a smaller effect (than the original work) with different subjects or in a different setting has not necessarily shown the original estimate was wrong for the original sample in some meaningful way. Rather, it has shown the parameter moves when the structural features change, which is predictable through the lens of the rational choice model in this study. Such findings pertain to generalizability, or external validity, not a replication failure. A recent meta-analysis (Vivalt, 2020) of development programmes documents such effect changes at scale, as does the several examples of scaling failures in List (2022). The framework herein supplies vocabulary necessary for distinguishing genuine replication failures from generalizability findings. In this sense, the framework simultaneously provides a principled basis for comparing estimates across designs without interpreting every difference in estimates as a problem to be eliminated. Taking that argument one step further, I will show that in many cases it is necessary information to understand the underlying phenomenon of interest.

At this point, I would be remiss not to mention a final clarification on this literature. None of the evidence above suggests that experiments suffer a special fragility relative to other empirical approaches. In fact, I find it quite the opposite, and the evidence is in that camp too. Examining thousands of hypothesis tests across 25 leading journals, Brodeur et al. (2016) and Brodeur et al. (2020) find that p-hacking and publication bias are considerably more severe in studies leveraging naturally-occurring data. Such issues are particularly acute in instrumental variables designs. In this light, the deeper point of this literature cuts decidedly in the experimentalist's favor: design-dependence is observable in experimental work precisely because the design is explicit and the assignment mechanism is known. In studies leveraging uncontrolled data, the analogous dependence on specification and identification choices is alive and well, but latent.

Beyond the replication crisis, the model reveals that the four experimental design types are

not rivals on a spectrum of quality but partners on five distinct dimensions of knowledge creation. Furthermore, and importantly, their strengths are mirror images of each other's weaknesses, which is perspicuous when one focuses on lab and NFEs. The first area concerns experimental control: the lab and NFEs do not differ in how much control the researcher exercises but in what she controls. The second area relates to causal identification: the lab enforces the four exclusion restrictions of causal identification by design; the NFE inherits three of them from the market and must provide credible evidence they hold. The third area concerns inference and is commonly denoted as construct validity: the lab can build the environment the theory requires but cannot guarantee subjects interpret it as intended. NFEs deliver the relevant construct the subject actually faces in the field, yet they cannot vary in a controlled way the structural parameters the theorist wishes to test or hold constant.

The fourth, and perhaps most surprising complementarity, relates to research ethics. Laboratory experiments are ethically transparent by construction, studying subjects who always know they are being studied. NFEs trade that transparency for observing behavior in the wild under conditions the Belmont Report and the Common Rule explicitly anticipate and permit. Each design is ethically defensible for reasons that depend on the existence of the other. The fifth area steps outside the model's focus on the average treatment effect to consider a different object of inference entirely: measuring economic primitives, such as risk aversion, a discount rate, or a social-preference parameter, under maintained theoretical assumptions. The lab is the design that licenses these exercises because it can impose the structural conditions the theory specifies. The NFE recovers the operating-market primitive that a structural model combining those primitives is meant to explain and extrapolate from. Taken as a whole, these five deep complementarities represent further support for my core hypothesis: our field needs the lab.

The remainder of this study proceeds as follows. Section 2 establishes the empirical spectrum and summarizes core features that distinguish them. Section 3 develops the governing model, which clarifies the distinction between the within-setting reference parameter the four designs target and the cross-setting target parameter decisionmakers seek. Section 4 summarizes the four design types in sequence, showing how inference changes with each of three design switches. Section 5 develops five key complementarities between the lab and field, showing that the designs are structural partners on every dimension that matters for knowledge creation. Section 6 outlines recent extensions, including "lab-in-field," online experiments, and surveys, showing how they fit into the theoretical framework. Section 7 concludes.

2 The Harder Science and the Identification Framework

As a useful first step in describing economic experiments, it is informative to compare other empirical approaches alongside experiments to understand how their identification strategies relate.

Throughout this paper, the analyst observes a population of agents indexed by i, each with covariates X_i, a treatment-assignment indicator $Z_i \in \{0,1\}$, a realized treatment status $D_i \in \{0,1\}$ (which may differ from Z_i if there is non-compliance), and an outcome Y_i. The fundamental object that distinguishes one empirical strategy from another is the assignment mechanism that produces Z_i.

For each agent i, let $Y_i(1)$ be the outcome that would be realized if the agent were assigned treatment and $Y_i(0)$ be the outcome that would be realized otherwise. The individual treatment effect is $\tau_i = Y_i(1) - Y_i(0)$. The fundamental problem of causal inference is that for any agent i, only one of $Y_i(1)$ and $Y_i(0)$ is observed; the counterfactual is missing.[5] The average treatment effect across a population in a particular setting is the expectation $\tau = E[Y_i(1) - Y_i(0)]$ over that population in that setting. Even given a perfect estimate of τ for the population and setting actually studied, generalizing it to a different population or a different setting requires further argument. I return to this issue below.

Within the studied population and setting, a randomized experiment delivers an unbiased estimate of τ if four exclusion restrictions hold (List, 2026a):

(i) Stable Unit Treatment Value Assumption (SUTVA).

The potential outcome of any unit depends only on its own treatment status, not on the assignments of other units, and there is a single, well-defined version of the treatment.

(ii) Statistical Independence.

The assignment mechanism is independent of potential outcomes. Because the researcher controls Z_i, this restriction is satisfied via randomization.

(iii) Compliance.

The treatment received equals the treatment assigned for every unit ($D_i = Z_i$).

(iv) Observability.

Outcomes are measured for every unit, with no selective attrition correlated with treatment.

The four restrictions are common to all four experimental design types. Indeed, they are the classic necessary conditions that all four design types rely on to identify τ for the population of subjects in the experimental setting. Together, they are what "internal validity" means in causal inference. In Section 5, I will return for a closer look at how the four design types differ in how they satisfy the four exclusion restrictions. For now, what matters is that the four designs share Statistical Independence by construction and share the goal of recovering τ within the studied sample under conditions in which the other three restrictions plausibly hold.

[5]The potential outcomes approach is commonly referred to as the "Rubin Causal Model" in the literature after (Rubin, 1974), but the framework can be found in Jerzy Neyman's Master's Thesis, which describes "potential yields" when referring to his agricultural outcomes of interest (see Neyman, 1923).

2.1 The Empirical Spectrum and the Control of the Assignment Mechanism

To set up the contrast between the four experimental types, it is useful to begin one step back, with the broader empirical spectrum on which all empirical research in economics sits. Figure 1 provides that empirical spectrum, from approaches where the researcher knows and controls Z_i (Lab, AFE, FFE, and NFE) to empirical approaches where the assignment mechanism is neither under the control nor known with certainty by the researcher. I categorize these as "controlled" and "uncontrolled" because the researcher knows and controls the assignment mechanism in one case but not in the others.

Control of the assignment mechanism is what distinguishes experiments from naturally occurring data. When using naturally occurring data, the researcher must argue (typically through an instrument, a discontinuity, or a parallel-trends assumption) for the analogue of Statistical Independence; under any of the four experimental designs, the researcher delivers it directly through the randomization protocol. This is the foundational reason why many scholars denote experiments as the "gold standard": the most fragile of the identifying restrictions when leveraging naturally-occurring data is the easiest one to enforce experimentally.

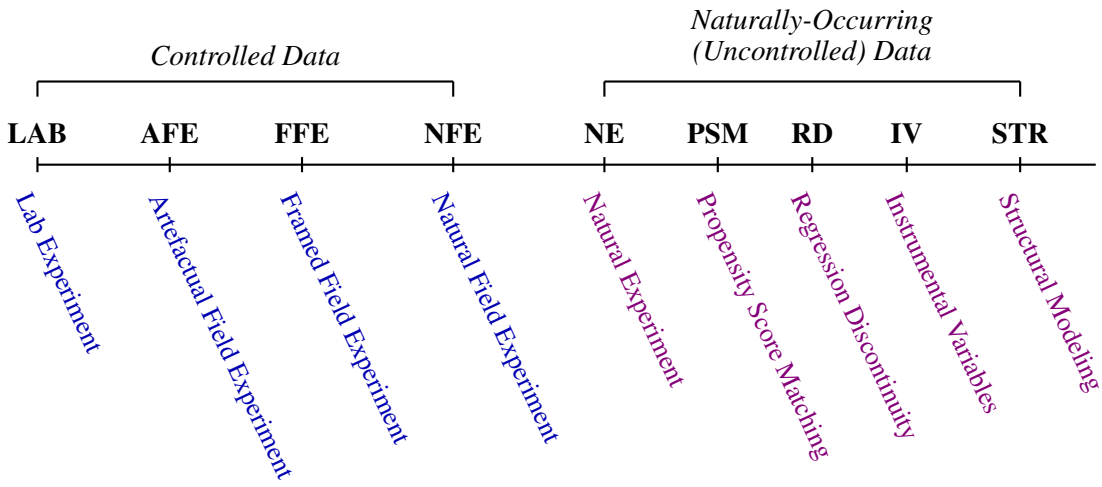

Figure 1: *Data Generation and Modeling for Causal Inference*

When the assignment mechanism is unknown to the researcher, she has access only to the variation that the world happens to provide. Substantial machinery has been developed in statistics and econometrics over the past several decades to extract causal estimates from such data. Each uncontrolled approach in Figure 1 imposes identifying assumptions to substitute for ignorance about the assignment mechanism. Each works well when those assumptions hold, yet if the assumptions fail causality cannot be inferred. Importantly, the failure modes are often hard to detect (Imbens and Angrist, 1994; List, 2026a).

When the assignment mechanism is controlled by the researcher, causal inference is structurally different. In such settings, since the researcher determines treatment assignment, selection on unobservables is impossible by construction. This is true because the assignment is independent of potential outcomes by design via our Statistical Independence assumption. That thread is what unites the four experimental types. At the heart of experimental economics is not where it is run or who participates, but who controls Z_i. In the controlled data cases of Figure 1, the researcher is in charge of data generation.

2.2 The Four Experimental Types

Over the last several decades, economists have developed many novel approaches to generate data by controlling the assignment mechanism. Harrison and List (2004) introduced four experimental types that differ along dimensions that have nothing to do with the assignment mechanism itself. They differ in who the subjects are, what environment those subjects are studied in, and whether they know they are being studied. The Harrison and List (2004) taxonomy populates the easternmost portion of Figure 1.

Laboratory experiments

recruit a convenience sample of subjects (typically university students) into an environment that the experimenter has constructed, with the subjects fully aware that a study is taking place. Induced values, restricted action sets, and stylized stakes are typical features.

Artefactual field experiments (AFEs)

retain the constructed and awareness features of the lab but draw the subject pool from the target market: farmers studying risk preferences, professional bidders studying auctions, CEOs studying negotiation.

Framed field experiments (FFEs)

move the action set from the constructed to the natural and replace induced values with naturally occurring stakes, while still recruiting target-market subjects who know they are participating in a study. Workers are hired with real wages and asked to do real work; real bidders participate in auctions.

Natural field experiments (NFEs)

go one step further: subjects are observed in their everyday activities, with treatment randomized across the natural commodities and conditions of the operating market, but without the subjects' awareness that any study is taking place.

Crucially, what the four design types share is that the researcher controls Z_i. When the other three exclusion restrictions are in place, this permits confident causal inference. Yet, as their definitions make clear, the four design types differ on three potentially key dimensions: who is studied, in what environment, and with what awareness regime. In Section 3, I formalize these differences as three feature switches in a single rational-choice model.

2.3 The Macros of Laboratory and Field Experiments

The power of an ideal laboratory experiment becomes readily apparent within this framework. Lab experiments provide the investigator with a means to directly influence the set of prices, budget sets, information set, and actions available to actors, and thus measure the impact of these factors on behavior within the context of the laboratory. Individual demand and supply curves tend to be invisible in the real world, but they come alive when the lab experimentalist uses induced values to construct a market to test, for example, if Adam Smith's invisible hand is as powerful as most surmise (see Chamberlin, 1948; Smith, 1962; List, 2004b). In this manner, randomization of induced values acts as an instrumental variable and therefore allows the analyst to make strong causal statements within the domain of study.

This attractiveness helped to make laboratory experiments a "boom industry" in the late 20th century. Indeed, as Holt (2006) documents, academic publications using the methodology were almost non-existent until the mid-1960s, surpassed 50 annually for the first time in 1982, and by 1998 exceeded 200 per year. Figure 2 shows similar growth over this period within the Top 5 Economics journals. Laboratory experiments grew from near-absence in the early 1970s through a peak in the early 2000s, then plateaued and declined, as aforementioned. Field experiments, as defined above, were essentially absent from the Top 5 before the mid-1990s, began a sustained rise in the late 1990s, crossed the laboratory series in the first decade of the new century, and have since continued to grow at an accelerating rate.

Three observations on Figure 2 motivate the rest of the paper. First, the discipline that the early skeptics (quoted in the epigraph) thought could not run experiments now conducts them at scale. Second, within experimental work the centre of gravity has shifted from the laboratory to the field, and shifted enough to stimulate this study. Third, the steepness of the field-experiments curve is itself the kind of pattern the model of Section 3 will help explain: as the discipline has internalized that field experiments operate on a different point in the empirical spectrum than laboratory experiments do, the design has spread. The argument of the rest of this paper is that the spread is appropriate but should not come at the cost of the laboratory.

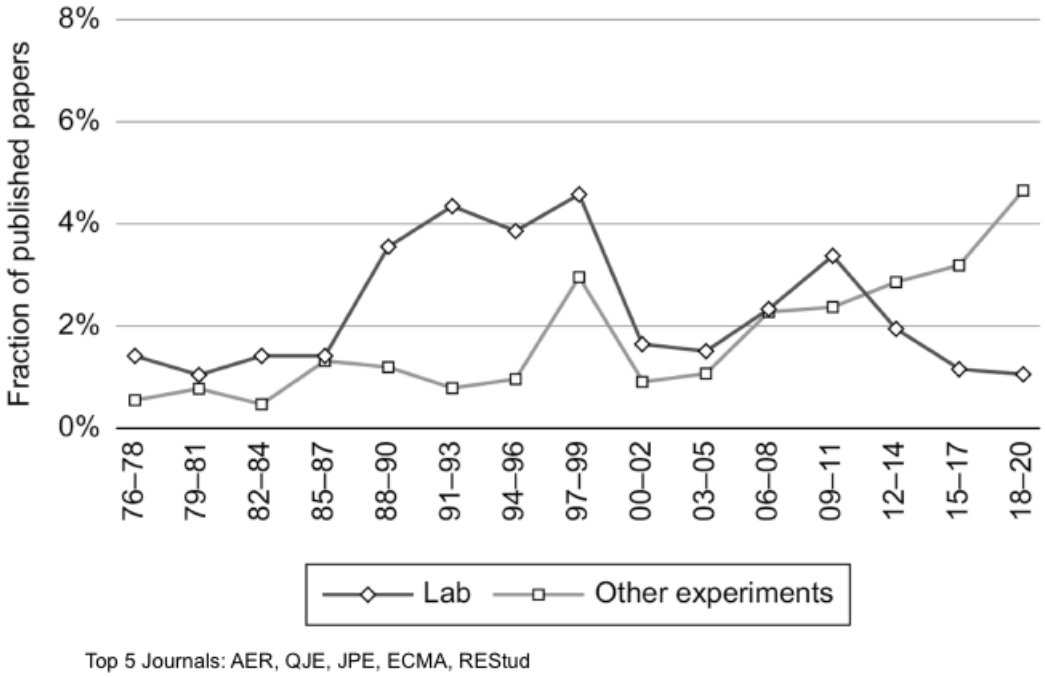

Figure 2: *Experimental publications in Top 5 economics journals, 1970–2020. Stylized adaptation of List (2026a, Chapter 2, Figure 2.2); see the source for the underlying counts.*

3 A Simple Governing Model

In this section, I take as given that the four exclusion restrictions hold to focus on a model that characterizes the distinguishing features of the four design types. I begin by considering a population of agents indexed by i. Each agent is characterized by a type θ_i drawn from a population distribution $F(\theta \mid S)$. The index $S \in \{0,1\}$ captures whether the population studied is the target population for the research question ($S = 1$) or a convenience sample drawn from a different distribution ($S = 0$). Members of S satisfy the non-zero probability condition of being in the experimental population: every member has a non-zero probability of each treatment value, and the researcher controls the functional form of assignment probabilities (List, 2026a). The type θ_i collects everything about the agent that is fixed for the duration of the experiment: preferences, beliefs, constraints, experience, and the value of the outside option.

Each agent chooses an action a from an action set $\mathscr{A}(E)$, where $E \in \{0,1\}$ indexes the environment. When $E = 1$ the agent is in the natural environment and $\mathscr{A}(1)$ is the full set of actions she would have available in the operating market. These actions are present with or without the experiment. Therefore, such actions might include buying or not buying, giving or not giving, and the like. Actions are performed with natural commodities, natural framing, natural stakes, and natural time horizon. The experimenter takes such features as given. When $E = 0$ the agent is in a

created environment in which $\mathscr{A}(0)$ has been stylized by the researcher: abstract tokens, induced values, restricted choice menus, compressed time horizons, and any other constructed features that are necessary.

Awareness is the third feature. Let $A \in \{0,1\}$ index whether the agent knows that her behaviour is being studied scientifically. The agent's objective is

$$U_i(a,D;\theta_i,S,E,A) = (1-\psi(A)) \cdot \pi(a,D;\theta_i,E) + \psi(A) \cdot M(a,D;E), \tag{1}$$

where π is the income or market component of the payoff function, M is a moral or social-image utility component (the part that responds to scrutiny, framing, and observation), and $\psi(A) \in [0,1]$ is the weight the agent places on the moral component. The crucial assumption, which is supported by an empirical literature stretching back to Orne (1962) and beyond, is

$$\psi(0) < \psi(1): \quad \text{awareness raises the weight on the moral component.} \tag{2}$$

Equation (1) is a familiar object in the rational-choice literature. Formally, it says the agent allocates effort across two components of payoff with different marginal returns, and a parameter ψ governs the relative weight she places on each when making her decision. This means that when awareness rises, the implicit shadow price of the income-maximizing action also rises. This increase causes the agent to substitute toward the action favored by the moral component of the model. This does not represent a mysterious psychological force; it is simply a common substitution effect in economics. $D \in \{0,1\}$ indexes treatment assignment.

The optimal action is therefore given by

$$a_i^*(D;\theta_i,S,E,A) = \arg\max_{a \in \mathscr{A}(E)} U_i(a,D;\theta_i,S,E,A), \tag{3}$$

and the agent's policy-relevant outcome is the realized market component of payoff, $Y_i = \pi(a_i^*;\theta_i,E,D)$. Note that this Y_i is the same outcome variable as the potential outcomes in Section 2: when $D=1$ we observe $Y_i(1) = \pi(a_i^*(1);\theta_i,E,1)$, and when $D=0$ we observe $Y_i(0) = \pi(a_i^*(0);\theta_i,E,0)$. The model just makes explicit that the potential outcomes themselves depend on θ, on E, and (through a^*) on the awareness regime. The individual treatment effect under the parameter triple (S,E,A) is therefore

$$\tau_i(S,E,A) = \pi(a_i^*(1);\theta_i,E,1) - \pi(a_i^*(0);\theta_i,E,0). \tag{4}$$

Equations (1)–(4) define the agent's problem. As I outline below, the four standard experimental designs can be defined as conditional expectations of τ_i under particular settings of the three switches, (S,E,A).

3.1 The Participation Margin

Whenever $A = 1$, the agent is informed that a study is taking place. After which, the agent makes a participation decision. This decision precedes treatment assignment to ensure the exclusion restrictions are met. Following the standard selection-on-gains formulation (List, 2026a), individual i agrees to participate ($P_i = 1$) if and only if

$$P[D{=}1] \cdot E_i[U_i(a_i^*(1)) \mid F_\tau] + P[D{=}0] \cdot U_i(a_i^*(0)) - C_i^P + W \;\geq\; U_i(\Omega_i), \tag{5}$$

where C_i^P is the participation cost (time, inconvenience, travel), W is the show-up incentive, F_τ is the agent's prior over the treatment effect distribution, and $U_i(\Omega_i)$ is the utility of the outside option. Equation (5) is the participation constraint that activates whenever $A = 1$, and it yields three intuitive comparative-statics results. First, agents with higher expected gains $E_i[\tau_i]$ from treatment are more likely to participate (selection on gains). Second, agents with lower opportunity costs $U_i(\Omega_i)$ are more likely to participate. Third, lower gain participants can be drawn into the sample with higher show-up payments W and/or higher treatment probabilities $P[D{=}1]$. Such design choices have a cost but can effectively reduce the magnitude of the selection wedge.

When $A = 0$, no announcement is made and equation (5) does not fire. The agent is in the studied environment for the same reason she would have been there absent the experiment. There is no P-conditioning to apply.

3.2 The Four Designs as Conditional Expectations

Each experimental design recovers an estimand of the form $E[\tau_i(S,E,A) \mid \text{conditioning}]$ for the population it studied in the setting it studied them in, yielding:

$$\tau^{\text{lab}} = E[\tau_i(S{=}0,E{=}0,A{=}1) \mid P{=}1, \theta \sim F(\cdot \mid S{=}0)], \tag{6}$$

$$\tau^{\text{AFE}} = E[\tau_i(S{=}1,E{=}0,A{=}1) \mid P{=}1, \theta \sim F(\cdot \mid S{=}1)], \tag{7}$$

$$\tau^{\text{FFE}} = E[\tau_i(S{=}1,E{=}1,A{=}1) \mid P{=}1, \theta \sim F(\cdot \mid S{=}1)], \tag{8}$$

$$\tau^{\text{NFE}} = E[\tau_i(S{=}1,E{=}1,A{=}0) \mid \theta \sim F(\cdot \mid S{=}1)]. \tag{9}$$

The natural reference point for the comparison, the average treatment effect in the population and operating environment that the researcher is studying, with no scrutiny and no participation filter is

$$\tau^{\circ} = E[\tau_i(S{=}1,E{=}1,A{=}0) \mid \theta \sim F(\cdot \mid S{=}1)]. \tag{10}$$

Comparing (9) and (10): $\tau^{\text{NFE}} = \tau^{\circ}$. This means that within the studied population and operating environment, an NFE recovers the full-information reference parameter. This is by construction of the NFE itself. By their very nature, the other three experimental design types differ from this reference on at least one feature switch, and each switch potentially generates an informative wedge that can be used by decisionmakers.

3.3 The Wedge Decomposition

The gap between any design's estimand and the reference parameter τ° defined in (10) can be decomposed into a sum of switch-specific wedges. Considering the notation for a lab estimate, we have

$$\tau^{\text{lab}} - \tau^{\circ} = \Delta_S + \Delta_E + \Delta_A + \Delta_P, \tag{11}$$

where $\Delta_S = E[\tau \mid S{=}0] - E[\tau \mid S{=}1]$ is the population wedge generated by drawing from a non-target distribution $F(\cdot \mid S = 0)$. Δ_E captures both the environmental wedge from the difference between $\mathscr{A}(0)$ and $\mathscr{A}(1)$ and the difference between induced and natural values in π. The term Δ_A captures the scrutiny wedge generated by $\psi(1) > \psi(0)$. Finally, Δ_P captures the additional bias from conditioning on the participation event in equation (5).

By definition, lab estimates carry all four wedges. AFE estimates carry Δ_E, Δ_A, and Δ_P. FFE estimates carry Δ_A and Δ_P. NFE estimates carry none of the four. Conveniently for our purposes, each step in the lab $\rightarrow$ AFE $\rightarrow$ FFE $\rightarrow$ NFE progression flips one switch and removes one wedge. This is the central organizing fact of the paper and illuminates why the four experimental types generate complementary data.

At this point the astute reader might wonder if the studied setting always represents the universe of target settings. Equation (10) defines τ° as the parameter for the population and operating environment that the NFE researcher studies. This could itself be a convenience sample for that researcher: perhaps these are customers of the partner firm she has obtained access to. In this case, the policymaker's parameter of interest, call it τ^{*}, may not equal τ°. A retailer running an NFE on Atlanta customers in 2026 has identified τ° for those customers in that time period. A policymaker who needs the parameter for Phoenix customers in 2030 must do further work to bridge from τ° to τ^{*}.

The further work is the science of external validity. This cross-setting generalization problem is not solved by the choice between lab, AFE, FFE, and NFE. Indeed, every empirical researcher, whether generating experimental data or using naturally-occurring data, faces it when making inference. The NFE's structural advantage is that it eliminates four within-setting wedges that potentially affect behavior (Δ_S, Δ_E, Δ_A, Δ_P). That is the key feature that separates other designs from τ° in the studied setting. The rest of the paper holds the within-setting comparison fixed and

asks what the four designs deliver against τ°. The cross-setting generalization to τ^{*} is a separate problem with its own apparatus; it is the subject of a large literature that reveals that external validity is an economic problem (see footnote 2). What this paper formalizes is the within-setting structure that makes the four experimental types distinctive.

4 The Experimental Spectrum, One Switch at a Time

This section walks through the four designs in sequence, leveraging the three switches and how they affect equations (6)–(11). While I could choose dozens of examples to illustrate, I attempt to maintain consistency with studies I am most familiar.

4.1 The Laboratory: $(S{=}0, E{=}0, A{=}1, P{=}1)$

A laboratory experiment recruits a convenience sample of university students into a created environment. The environment is created carefully to ensure that the four exclusion restrictions and the theoretical conditions necessary to test the model are met. Subjects are fully aware that a study is taking place, and consent forms are signed. All four wedges in equation (11) are therefore active. One great advantage of a laboratory experiment is what the created environment provides in terms of inference. By inducing values, restricting the action set, and standardizing the stakes, the experimenter can build an environment that satisfies the theoretical conditions the design intends to test.

The lab-specific limitation is the joint operation of Δ_E, Δ_A, and Δ_P. Behaviour elicited under $\psi(1)$ over a stylized action set $\mathscr{A}(0)$ by a student subject pool is, in general, a noisy proxy for behaviour in the operating market. The dictator-game literature provides an illustration. In a standard laboratory dictator game, around 60 to 80 percent of subjects pass a positive amount and the mean transfer is around 20 percent of the endowment (Engel, 2011). In an illuminating study, Cherry et al. (2002) show that when subjects earn the endowment first and the design has full anonymity (through the lens of the model, this lowers both $\psi(A)$ and the salience of M), over 95 percent of dictators give zero. Same agents, same task, different settings of the model's parameters. What results is a dramatically different revealed behaviour. List (2007) uses a dictator game to highlight the action-set point directly: simply enlarging $\mathscr{A}(0)$ to include a "take" option, without removing any of the original give-only choices, sharply reduces positive giving. Similarly, Liberman et al. (2004) and Ellingsen et al. (2012) make the parallel point about $\mathscr{A}(0)$: simply relabelling a prisoner's dilemma as the "Community Game" versus the "Wall Street Game"halves cooperation.

4.2 Flipping S: The Artefactual Field Experiment

An AFE retains the beauty of the created lab environment but draws the subject pool from the target market: farmers studying how risk preferences affect crop rotation or fertilizer use, CEOs studying breakdowns of high stakes negotiation, professional bidders studying auctions and the relevance of institutional features. Accordingly, in equation (11), Δ_S is removed in an AFE. This means that the non-zero probability simply applies to a different subject population. Of course, the participation decision is still made, it is just made by a different population, S. The implication is that a different set of factors might determine participation.

What does flipping the S switch provide to the researcher in terms of inference? I will focus on two crucial aspects. First, the population distribution $F(\theta \mid S = 1)$ replaces $F(\theta \mid S = 0)$. Thus, the recovered parameter is now an expectation over people whose preferences, opportunity costs, and outside options resemble those who populate the operating market of interest. Second, target-market subjects bring relevant experience to the experimental task. That experience potentially modifies the optimal action a_i^* in ways that the abstract environment may not preserve. One illustration of this idea at work is Harrison and List (2008). Their study of the winner's curse illustrates this idea cleanly. Student subjects in laboratory auctions consistently overbid and fall prey to the winner's curse. Professional bidders from the target market, studied in the same constructed environment also fall prey to the winner's curse. Yet, if the AFE construct maps into the auction format they actually face in their operating market, they avoid the winner's curse. Their accumulated experience has calibrated their behaviour on those margins in ways that students have not yet developed.

When an AFE places professionals on margins that have no counterpart in their operating experience, that protection breaks down. I return to this point in Section 5.3 in the context of construct validity. What this example highlights is that an AFE is the design type that exposes the limit of $E = 0$ most clearly. Flipping S without flipping E reveals which lab findings depend on subject inexperience and which survive when target-market subjects are inserted into the created environment. Both pieces of information are valuable for decisionmakers.

4.3 Flipping E: The Framed Field Experiment

An FFE moves the action set from the created $\mathscr{A}(0)$ to the natural $\mathscr{A}(1)$. As such, this design choice replaces aspects of the created environment, such as induced values with homegrown preferences, while still recruiting target-market subjects. Again, they sign consent forms and know they are being studied, as in the lab and in an AFE. In terms of the governing model, in equation (11), Δ_E is removed alongside Δ_S.

Flipping E buys two further things for an FFE. First, the action set itself now reflects the choices the agent would actually face. They are placed on familiar margins. Workers can be hired with real

wages and asked to do real work; donors can be approached through real solicitations; bidders can compete for real assets. Second, and in a similar spirit, the construct-validity concern is therefore attenuated. When the experimenter does not have to construct the environment from scratch, she does not have to assume that her stylized version captures the essence of the problem. Creating a task that mirrors and matches the same construct that operates in the market can be a difficult chore. What FFEs do not buy is Δ_A or Δ_P. The agent still knows she is being studied and still has to consent. The scrutiny effect $\psi(1) > \psi(0)$ operates, and equation (5)'s participation filter still selects on gains. As such, FFE estimates, like AFEs, remain conditional expectations over the participating subset of the target population, evaluated under the awareness regime.

4.4 Flipping A: The Natural Field Experiment

An NFE flips the third, and final, switch in the model. The researcher acts as an anthropologist (an "anthroconomist" if doing economic research), delivering the experiment in the marketplace of interest to agents who are unaware that a study is taking place. This means that Δ_A is removed because $\psi(0)$ replaces $\psi(1)$ and the moral-component weight returns to its market level. Also, Δ_P is removed because equation (5) is never activated: with no announcement, there is no participation margin to clear and no selection filter to apply. The recovered parameter is exactly the within-setting reference parameter τ° defined in (10): the average treatment effect for the studied population in the studied operating environment.

The empirical bite of flipping A has been well documented over the past few decades in the social sciences. For example, List (2006) tested theories of gift exchange by sending confederates to buy items at a sportscard show. The data show that the price–quality relationship varied sharply based on A. The same agents revealed different behaviour at $\psi(1)$ and at $\psi(0)$. By now, across studies in charitable giving (Karlan and List, 2007), labour markets (Flory et al., 2015), discrimination (Gneezy and List, 2006; Gneezy et al., 2012), pricing (Damgaard and Gravert, 2017), and consumer choice, various NFEs have lent insights into $\psi(0)$.

Three points about these NFEs follow from equations (6)–(11) and bear emphasis. First, the NFE's advantage is structural rather than empirical. That is, it is not that NFEs happen to produce better estimates, it is that they recover a different parameter than the other three designs recover. In many instances, that parameter is the one the decisionmaker desires. Second, the NFE does not eliminate every wedge: the population distribution $F(\theta \mid S)$ is still chosen by the researcher when they choose the market to do their work. The choice of which market to study (which retail customers, which rideshare drivers, which donors, which grade-schoolers) is itself a sampling decision. Third, the NFE design is not a panacea, as some research questions inherently require $A = 1$, some outcomes cannot be observed at $A = 0$, and the other design types retain real

comparative advantages, as discussed below.

5 The Deep Complementarity of Lab and Field Experiments

The preceding sections have shown that the four experimental designs are comparative-static restrictions of a single maximization problem. They have also revealed what each design delivers against the within-setting reference parameter τ°. While interesting, this does not necessarily show that the design types are deep complements. This section makes that case by showcasing the complementarity across five distinct dimensions. The first concerns the conventional intuition about experimental control. The second and third are formal results from the model: what the designs deliver on causal identification and on construct validity respectively. The fourth is perhaps the most surprising complementarity, research ethics. The fifth steps outside the model's focus on the average treatment effect to consider a different object of inference entirely: the structural primitives that economic theory takes as inputs. Each of these dimensions are important in their own right, and reveal a different facet of the same structural fact: the lab and the field do not compete on a single dimension. Indeed, they occupy complementary niches across every dimension.

5.1 The Control Complementarity

The conventional methodological intuition holds that laboratory experiments afford researchers "more control" than field experiments. The unified model shows that intuition is imprecise. The laboratory does retain control over the action set $\mathscr{A}$, the realization of π, and the assignment mechanism. By stylizing the environment to $E = 0$, the experimenter can guarantee that subjects face the choices the theory describes, with the values the theory assumes, in the order the theory specifies. That is real control, and it has real value for testing the internal logic of theory.

What the lab does not control, however, and indeed what no overt design can control, is the composition of the sample. Because $A = 1$, equation (5) reveals that the population distribution observed in the experiment becomes $F(\theta \mid S, P = 1)$ rather than $F(\theta \mid S)$. Selection on gains determines who shows up to the experiment and subsequently who the analyst gathers information from. Of course, as part of optimal design, the researcher can manipulate the participation incentive W. This can importantly change the treatment probability $\Pr[D = 1]$ to shift the magnitude of Δ_P, but unless the incentives induce everyone in S to sign up for the experiment, they cannot eliminate it. In this sense, the lab's control over the environment is conditional on a sample whose composition the lab does not control.

The NFE inverts this relationship. By operating at the feature of $A = 0$, the NFE bypasses equation (5) entirely. The studied sample is whomever would have been in the studied market that

day, drawn directly from $F(\theta \mid S = 1)$. Yet, the researcher's control over the environment is more limited. She takes $\mathscr{A}(1)$ as given, but her control over the sample is more complete (Al-Ubaydli and List, 2015). The supposed trade-off between control and realism is, on the dimension of sample composition, not a trade-off at all. It is a different allocation of control across two distinct objects: the environment and the sample.

The conventional language of "more control" versus "less control" across the lab and field therefore obscures what is happening in the experiment. Focusing on the ends of the experimental spectrum, this comes to life. What this section is arguing is that the lab and an NFE do not differ in how much control the researcher exercises. They differ in *what* is being controlled: the lab controls the environment and the action set at the cost of sample selection. The NFE controls sample composition at the cost of environmental stylization. Importantly, these failures and strengths are mirror images. And mirror-image failures imply mirror-image corrections: the lab corrects exactly what the NFE cannot fix; the NFE corrects exactly what the lab cannot fix. With a goal of optimal knowledge creation, what follows is that a research programme that uses both designs exercises complete control, across all the dimensions that matter, in a way that no single design can achieve.

5.2 The Identification Complementarity

Section 3 made the case that an NFE recovers the within-setting reference parameter τ° by construction. That argument holds given that the four exclusion restrictions are satisfied. Yet, it is important to understand that across the experimental spectrum the ability to ensure that each restriction holds can vary appreciably.

Statistical Independence is the one restriction that the researcher controls in every experimental design via controlling the assignment mechanism. While the researcher must choose the randomization approach, whether they opt for Bernoulli trials, completely randomized experiments, randomized block (stratified) experiments, or a paired design tends to be orthogonal to what design type they choose.

SUTVA is at risk whenever units interact. In a lab, AFE, and some FFEs, units can be physically isolated across cubicles or sessions and the experimenter can guarantee that no agent observes another's choice. In an NFE, units interact through the natural channels of the studied market: customers compare prices, drivers share information, donors talk to one another. SUTVA must be inherited from the market, not enforced in an NFE.

Compliance is at risk whenever subjects choose differently from the researcher's coded assignment. In a lab and AFE with induced values, the assigned action is essentially the only action available, and compliance is mechanically enforced. In an NFE, the worker assigned a new schedule may trade shifts, the recipient of a marketing letter may not open it, the retailer's point-of-sale system

may fail to apply a treatment-arm price. Compliance is largely inherited in an NFE. Of course, such compliance issues might be of ultimate research import, but must be recognized.

Observability is at risk whenever subjects in the field can disappear from the data in ways subjects in the lab cannot. In a lab, the experimenter is in the room and every action is recorded; rarely do subjects excuse themselves mid-session. In an NFE, actions and outcomes occur outside the researcher's watchful eye, and selective attrition becomes a live threat. For example, subjects who would have produced low outcomes may go missing at higher rates than those who would have produced high ones. If attrition correlates with treatment in this manner, then the naive estimate is biased and corrections (e.g., Lee bounds, or selection models with a valid exclusion restriction) become necessary.

This summary yields an important pattern for causal inference. It teaches us that the lab can enforce all four exclusion restrictions by construction. Furthermore, an NFE can enforce only Statistical Independence and must inherit the other three from the market. The lab purchases this confidence in internal validity at the cost of Δ_E, Δ_A, and Δ_P. The NFE inherits the three restrictions from the experimental setting and asks the researcher to provide credible evidence that the inheritance holds. This is work the lab does by construction. On the dimension of causal identification, the lab enforces what the NFE inherits, and the NFE delivers what the lab distorts.

5.3 The Construct Validity Complementarity

The four exclusion restrictions ensure that the recovered estimate is a valid estimate of τ for the studied sample. They say nothing about whether the τ that has been identified corresponds to the theoretical parameter, or the behavioral insight that the researcher believes they are estimating. That second question, whether the theoretical preconditions of the theory being tested actually apply in the studied environment, is conceptually distinct from internal validity and represents a key part of inference commonly denoted as construct validity in the literature (List, 2026a).

There are two distinct ways the theoretical preconditions of a theory can fail to hold in an experiment. First, the experimenter may not be able to construct a situation in which the structural assumptions of the theory hold. For example, the right preference structure, the right action set, the right information environment. They might be impossible to construct in a useful manner. Second, even if the experimenter believes she has constructed such a situation, the subject may not interpret it the way the experimenter intends. The lab and the field differ on these two halves of the design equation, and importantly, in opposite directions.

On the first half, the lab has a clean and distinct advantage. By inducing values, restricting the action set, and stipulating the information available to the subject, the experimenter can build a stylized environment that satisfies the theory's structural assumptions exactly, by construction. The

agent is, in effect, handed the parameters the theory takes as inputs. This is precisely what separates a laboratory test of a theory from a field test. In the lab, the experimenter tests the behavioral prediction while holding the structural assumptions fixed by design. In the field, no such separation is possible.

Consider a parochial example that played out two decades ago. Engelbrecht-Wiggans et al. (2005) desired to explore demand reduction in multi-unit auctions using a field experiment. For our purposes, they note in the context of their field experiment that it "represents a joint test of both the underlying modeling assumptions... and the behavioral assumptions." This means that if their field test failed to confirm a theoretical prediction, the failure could lie in the behavioral assumption, in one of the structural assumptions, or in both simultaneously. There is no reasonable way to know and the data cannot distinguish among them. This fact holds because in an FFE or NFE, agents' preferences are homegrown and arrive in the experiment unknown to the researcher. Further, the information they bring is whatever they happen to have, and the action set is whatever the market provides. None of this can be enforced. The lab's clean advantage on this half is therefore not merely a matter of convenience. Indeed, that control renders it as able to isolate the behavioral content of a theoretical prediction from the structural conditions under which that prediction is derived. Simply put, on these terms alone, the lab is the choice if one desires a clean theoretical test.

On the second half, the lab's advantage disappears and may even invert. That is because the experimenter cannot guarantee that the subject interprets the constructed situation as the experimenter intends, and the theory demands. One clear illustration of this fact comes from laboratory tests of social preferences. Experimenters use one-shot designs precisely to eliminate reputation-building motives: if there is no future interaction, any cooperation or generosity observed must reflect genuine social preferences (rather than strategic investment in a relationship). The logic is sound in theory. But subjects' lifetimes of experience with the kinds of interactions these games abstract (think dictator, ultimatum, trust, gift exchange, public goods) have all taken place in repeated settings. Subjects may therefore play the one-shot lab game as if it carries some repetition.[6] In such cases, they might import the behavioural norms that sustained cooperation in their natural environments into an environment the experimenter designed to exclude them. What the experimenter codes as social preferences may be reputation-building behaviour the subject has carried in from outside (see List, 2006). The experimenter has little insight into what context the subject is applying, and it is difficult if not impossible to enforce reliably the intended one.

[6]I should note that this example is well trodden. The economics version is in Levitt and List (2007b). Binmore (2005, 2007) is the canonical citation. The cleanest formalization that I have come across is due to Delton et al. (2011), who conclude that observed generosity in one-shot games has uncertainty about whether repetition will occur as a product. Yet, I should not oversell this point without a fair trial: there are researchers who occupy a strong-reciprocity camp (e.g., Gintis, Bowles, Fehr) who interpret the choices as genuine generosity, not residue of naturally-occurring forces. For example, Fehr and Henrich (2003) take on the misfired-repeated-game reading directly and reject it (see also Gintis, 2000; Bowles and Gintis, 2011).

Construct validity concerns potentially arise in every study and are not limited to preference measurement. One example outside that literature is Harrison and List (2004), who document a concern on a different dimension. There main interest relates to bidding behavior of professionals. They report that such bidders avoid the winner's curse in their familiar market setting but fall prey to it when placed in an unfamiliar laboratory format. This is not because their experience disappeared, but because the created environment cued a different decision frame than the one their experience had calibrated (see also Section 4.2). The lab enforced the experimenter's structural assumptions but could not enforce the construct the subject brought to the task.

This second half of the construct-validity problem is what FFEs and NFEs solve by design. By taking the field environment as given, the FFE and NFE deliver the construct the subject actually faces in the operating market, eliminating the risk that an experimenter-constructed situation is interpreted as something other than what the researcher intended. The market has done the translation already. There is no risk that the subject is interpreting an FFE or NFE situation as something other than what it is, because the situation is exactly what the subject would have been in regardless.

A useful takeaway from this literature and the model is therefore not that the lab dominates the field on theoretical preconditions or vice versa. The lab has the keen advantage that it can build a stylized setting that, on paper, satisfies the theory's structural assumptions. But, the subjective interpretation of this environment is difficult to enforce. The FFE and NFE inherit a setting that, by virtue of being the actual setting the theory is meant to describe, satisfies the theory's contextual assumptions by construction but whose structural assumptions the experimenter cannot vary. When a theory's structural assumptions are restrictive and contextual assumptions are robust, the lab is the right place to test it. When the contextual assumptions are restrictive and the structural assumptions are robust, the FFE or NFE is the right place. Neither design dominates. As such, the deep complementarities arise again: on the dimension of construct validity, the lab controls what the NFE cannot guarantee, and the NFE delivers what the lab cannot enforce. True complements in scientific knowledge production.

5.4 The Ethics Complementarity

In my travels, the most common negative reaction to NFEs is an ethical one: how can it be acceptable to study people without their knowledge or consent? The objection deserves serious engagement. And the ethical case for NFEs, properly understood, reveals the deepest and most surprising complementarity in this paper: the ethical case for NFEs cannot be made in isolation from the ethical case for laboratory experiments, and the two cases are best understood together.

The Asymmetry Argument. As a thought experiment, consider an asymmetry we implicitly endorse every day. Organizations, from a benevolent local government official to NGOs to Wall Street firms, routinely deploy untested policies, regulations, and pricing changes that affect millions of people daily. This is without consent and without rigorous evaluation. Indeed, we are perhaps in the beginning of the greatest social experiment of all time with recent advances in social media and AI, all propped up by organizations we implicitly trust. An NFE imposes a similar or identical experience on participants, but produces rigorous evidence about its effects. The relevant counterfactual is not an NFE versus a hypothetical world without an intervention; it is NFEs versus the real-world counterfactual of untested interventions that are scaled every day. I urge the reader to think about which is the more ethical path: testing before scaling, or scaling without testing?

Upon closer inspection of the model, a deeper version of the asymmetry argument surfaces. When awareness systematically distorts behaviour, as equations (1)–(2) and the empirical evidence in Cherry et al. (2002), Liberman et al. (2004), List (2006), Levitt and List (2007b), and scores of other work show, then requiring awareness as a precondition for research does not protect participants. It produces misleading evidence that can lead to worse policies affecting far more people at scale. The ethical case for NFEs is not that consent fails to matter. It is that valid evidence matters too, and the framework governing research ethics has long recognized this fact (as discussed below).

This set of arguments reveals an interesting complementarity that arises between the four design types: the lab, AFEs, and FFEs provide what an NFE cannot. That key provision is a fully consented, fully transparent research environment in which subjects are never studied without their knowledge. Therefore, when the research question permits, in cases where awareness and selection do not distort the parameter of interest, I argue that such designs are ethically dominant. The NFE is not a substitute for such transparency; it is the design that steps in when such transparency and selection destroy the very behaviour the research is meant to understand. The design types occupy complementary ethical niches: the lab, AFE, and FFE for parameters that survive awareness and selection, the NFE for parameters that do not.

The Letter of the Law: Belmont and the Common Rule. The foundational document of modern human-subjects research ethics in the United States is the Belmont Report (Belmont, 1979). The Report rests on three principles: beneficence, respect for persons, and justice. One might be inclined to view these principles as ruling out covert research. That is incorrect, and indeed quite the opposite, as the Report is quite explicit in this area (p. 8)

> "A special problem of consent arises where informing subjects of some pertinent aspect of the research is likely to impair the validity of the research. In many cases, it is sufficient to indicate to subjects that they are being invited to participate in research of

> which some features will not be revealed until the research is concluded. In all cases of research involving incomplete disclosure, such research is justified only if it is clear that (1) incomplete disclosure is truly necessary to accomplish the goals of the research, (2) there are no undisclosed risks to subjects that are more than minimal, and (3) there is an adequate plan for debriefing subjects, when appropriate, and for dissemination of research results to them."

This passage effectively settles the matter as a point of normative architecture. The drafters anticipated research in which informing subjects would compromise the science. In doing so, they summarized the three-part test that recognized conditions where an NFE is justified. NFEs routinely meet all three, whether the study is on charitable giving, market exchange, consumer choice, or labour supply. In each case, disclosure invalidates the research because the entire point of the design is $A = 0$. NFEs clear the minimal-risk standard by construction: because the experiment leaves the naturally occurring environment intact, subjects face no greater risk than they would have faced absent the study. My own goal goes further. When I design an NFE, I aim to leave subjects strictly better off than the status quo. On debriefing, this is commonly practiced when appropriate (and feasible).

The regulatory implementation of the Belmont Report is the Common Rule, codified at 45 C.F.R. § 46.116(f) (Common Rule, 2018). It permits an Institutional Review Board to waive informed consent under conditions that closely track the Belmont test: minimal risk, no adverse effect on the rights and welfare of subjects, infeasibility of the research without the waiver, and provision of relevant information after participation when appropriate. NFEs studying naturally occurring decisions in operating markets generally satisfy all four prongs, and IRBs across the United States have approved hundreds of NFEs under this framework over the last three decades.

Model Learnings. Combining the insights from the model with the ethics discussion yields a lesson: the science and ethics of NFEs point in the same direction. Unethical research generates biased samples, erodes participant trust, and reduces the representativeness of future research pools. A research enterprise that systematically deceives or harms participants will eventually find itself drawing only from the desperate, the unaware, and the careless. This is not only immoral, but a sampling failure that contaminates the generalizability of the work that an NFE was meant to produce. Ethical NFEs sidestep this dynamic by leaving subjects no worse off than they would have been. The same property that makes them ethically defensible is the property that makes them externally valid.

And with this arises the deepest hidden complement: the same property that makes laboratory experiments ethically transparent, the requirement that subjects know they are being studied and therefore select into the experiment, is the property that limits their external validity. The result

is that lab and NFEs are ethical complements in the most fundamental sense: each is ethically defensible for reasons that depend on the existence of the other. Without the lab as a consented benchmark, an NFE has no principled basis for claiming that the awareness distortion it avoids is real rather than simply assumed. Without the NFE as a naturalistic standard, the lab has no way to measure the cost of the transparency it provides. As a pair, they combine to form a research programme that is both scientifically valid and ethically grounded. In isolation, neither ethical or scientific claim fully holds.

5.5 The Parameter-Identification Complementarity

The four complementarities developed so far have taken the average treatment effect as the object of inference. The lab and the NFE both target τ; they differ in which wedges they must switch on and off to generate such data. There is, however, a parallel epistemic mode the framework has so far set to the side. Some experiments do not aim to identify τ at all. They aim either to test the prediction of a structural theory or to identify an economic primitive. Such primitives form the basis of economic models and include measuring a risk preference, a discount rate, a social-preference parameter, a belief, or a constraint. On this dimension as well, the lab and the field are complements rather than rivals, and the complementarity follows the same mirror-image logic as the four developed above.

Before getting started, it is useful to distinguish two approaches that experiments in this arena engage theory directly. The first is exemplified by Smith (1962) and the tradition of testing competitive equilibrium and other structural predictions in induced-value markets (Chamberlin, 1948; List, 2004b). This approach uses theory to provide a counterfactual benchmark and constructs an environment that satisfies the structural assumptions of the theory. For testing competitive equilibrium, that construction includes known supply and demand schedules, induced values, an allocation institution, and restricted action sets. The main outcome measures are whether the theoretical predictions are met. The point of the design is not τ and does not use data to estimate a counterfactual as in the examples above. Rather, the crucial aspect of the experiment is whether behaviour converges to the theoretical prediction under the conditions the theory specifies. The lab and AFEs specialize in this type of exercise. An NFE neither stipulates the primitives the theory takes as inputs nor exposes them to view. The implication is that if behaviour departs from the prediction, then the NFE cannot tell whether the structural or behavioural assumption failed. This is a crucial point discussed on the construct validity section above.

The second major approach, pioneered by Holt and Laury (2002), uses theory not to create an environment and test a prediction, but to back out an economic primitive from observed subject choices. For one strand of the literature, that approach has the experimenter presenting the subject with a menu of binary choices structured so that, under the maintained theory, the switch point

identifies a value of the parameter of interest. Expected utility theory imposed on a Holt-Laury menu recovers the CRRA coefficient r, for example. Andersen et al. (2008) extend the logic by eliciting risk and time preferences jointly, recovering the discount factor and the curvature of utility over consumption from a paired MPL design under an exponential or quasi-hyperbolic specification. Andreoni and Sprenger (2012) generalise the menu further with the convex time budget, which elicits interior allocations rather than switch points and separately identifies utility curvature from time preference. The measured object in each case is a feature of the agent's type θ_i in the language of equation (1). Importantly, again, that means it is not an average treatment effect generated by a treatment-control contrast. As with examinations of competitive equilibrium in the Vernon Smithean sense, the lab is the design that licenses this recovery. This is because identifying θ_i from observed choices requires that the choice environment satisfy the theory's structural conditions exactly. This is the half of construct validity that Section 5.3 showed only the lab can guarantee.

Of course, primitives can also be recovered using an NFE. But, importantly, in this case the analyst must layer a structural model on top of the design and assume that the model's structural conditions are satisfied by the market the agents inhabit. Karlan and List (2007) and DellaVigna et al. (2012) illustrate how far this approach can be taken, and the structural-IO tradition has been built on it (see also Cotton et al. 2026; Hedblom et al. 2019). The uniqueness of the lab is that it imposes the structural conditions by design rather than by assumption. The Holt-Laury approach does not assume that the agent faces a known menu of binary lotteries with known probabilities and known prizes. Rather, it constructs that menu. The choice between lab and field structural estimation is therefore not a choice between identifying primitives and not. It is a choice between identifying them under structural conditions enforced by design versus identifying them under structural conditions imposed by assumption. The lab pays for this enforcement in Δ_E, Δ_A, and Δ_P; the field pays via maintained assumptions the environment cannot verify.

This area of study has gained rapid advance due to its usefulness. This is because in many cases the parameter the policymaker requires is often a primitive of the agent's utility functional rather than τ. For example, in a standard economics model, the welfare gain of a charity tax break depends on the warm-glow weight, the price elasticity of altruism, and the curvature of utility over one's own consumption. Likewise, optimal pricing requires an understanding of demand curvature at counterfactual price points. And, those counterfactuals are yet to be observed by the firm. This means that risk-bearing decisions in policy design require r itself, not the observed choice rate in any single elicitation. In each case, the relevant object is a feature of θ_i, and identifying it requires precisely the structural conditions that the lab can enforce by design and the field can only assume by theory. The lab and the NFE are, on this dimension, partners in a division of scientific labour: each pursues θ_i under a different enforcement regime. Accordingly, a research programme that uses both has a more defensible primitive than one that uses either alone.

This discussion makes clear the deep complementarities in the lab and field. Once, again, they are mirror-images: to estimate primitives, the lab imposes the structural conditions identification requires by design. Alternatively, the NFE inherits them from the market, or assumes them via theory. The price that the lab pays for its enforcement arrives via Δ_E, Δ_A, and Δ_P. The price an NFE pays for its environmental authenticity is in the currency of structural assumptions. In sum, a research programme that uses both designs identifies its primitives twice. Once under conditions the experimenter has constructed; once under conditions the market has imposed. These are potentially very different estimates through the lens of the model. Therefore, when the two estimates are in accord, the discipline learns that the primitive is recoverable across the structural-and-environmental boundary that separates the lab and the field. When they disagree, we learn where the boundary is binding, which is itself informative about scaling and generalizing. Either outcome is informative; neither is recoverable from a single design.[7]

5.6 The Complementarity, Stated Precisely

The core argument of Sections 5.1-5.5 is that laboratory and field experiments are not rivals on any scientific dimension of import. In fact, it is quite the opposite theoretically and empirically. This is because they have opposing strengths across five conceptually distinct dimensions: control, identification, construct validity, research ethics, and parameter identification. One implication of this complementarity is that a research programme that generates data across the spectrum and finds similarities offers several suggestive insights: i) that the theory's causal prediction holds under the structural conditions the experimenter imposed, ii) that subjects interpreted those conditions as intended, iii) that the identified parameter survives the transition to the natural environment, iv) that the estimate is free of the awareness and participation effects that overt designs introduce, and v) that the underlying economic primitives the theory takes as inputs are themselves credibly identified, providing the structural basis on which the operating-market ATE can be extrapolated to settings not yet studied.

That is a far stronger claim than any single design can make alone, and it is the scientific case for not giving up on lab experiments even now that we have access to greater and greater numbers of natural field experiments. How this complementarity arises can come in many distinct forms. In some cases, within the same study, the researcher provides evidence of the phenomenon under consideration across the empirical spectrum (see, e.g. List, 2003, 2004a, 2006, 2009; Karlan, 2005; Benz and Meier, 2008; Burks et al., 2009; Cohn et al., 2015)

In others, the literature builds on itself. Consider the work on reference dependence. From the

[7]This intuition represents the basis of the SICC (Situation, Identification, Construct, Covariates) identification framework in List (2026a), which structures credible primitive elicitation across both lab and field. The present point is that the cumulative knowledge base of economics is best served when both apparatuses are deployed in tandem.

field arose data from surveys (contingent valuation) that suggested anomalous behavior existed and that individual endowments might matter. Kahneman and Tversky (1979) created seminal theory. There was important laboratory work in the 1980s and 1990s that showed such behaviors conformed to endowment theory. People took this to the field and found that the endowment effect is not a fundamental characteristic of preferences that is immutable by market experience. Lab and field studies replicated the experience result, and new insightful theory was introduced. Yet, the process did not stop there. Scholars then went back to the lab and dug a bit deeper into the experience results. They conducted functional MRI (fMRI) studies to investigate how experience changes neural correlates of the endowment effect. Poetry in motion (the interested reader should see List (2020a) where this is more patiently described with relevant citations).

6 Recent Extensions in Experimental Economics

The four experimental design types developed in Sections 2 through 5 do not exhaust the empirical landscape. Three additional approaches have grown rapidly within economics since Harrison and List (2004) introduced the taxonomy and warrant explicit treatment within the framework: lab-in-the-field designs, online experiments, and "representative" survey experiments. The growth has been substantial. Bohannon (2016) reports that the number of academic studies using Amazon's Mechanical Turk grew from 61 in 2011 to more than 1,200 in 2015, a roughly twentyfold growth in four years. The rate has continued to climb on Prolific and other platforms designed for academic use (Peer et al., 2022). Stated-preference and hypothetical-survey work has accumulated to a similar scale: Penn and Hu (2018) review more than 500 articles in their meta-analysis of hypothetical bias and draw 131 into the formal analysis. The "lab-in-the-field" label has become standard usage across adjacent disciplines.

Online experiments and "lab-in-the-field" studies can be located within the unified model of Section 3 by considering realized values of (S, E, A). For surveys, we must add an additional switch to the realisation of π, and consider what a "representative" subject pool means when the participation constraint is active. The aim of this section is not to relitigate the case for any of these designs, but to clarify what each identifies and where it sits relative to the within-setting reference parameter τ°. The three subsections proceed in increasing order of the framework's accommodation: lab-in-the-field designs require no expansion at all, online experiments require closer attention to two existing wedges, and survey experiments require a fourth switch.

6.1 "Lab-in-the-Field" Experiments

Within some quarters of the social sciences a "new" kind of field experiment has arisen: "lab-in-the-field" experiments. I have traced the conceptual precursor to Viceisza (2012), who coined the term 'lab-like field experiments' (LFE), with the specific phrasing 'lab-in-the-field' emerging in Viceisza (2016) and being popularized by Gneezy and Imas (2017). A formal definition, from Gneezy and Imas (2017, p. 439), is the methodology that "combines elements of both lab and field experiments in using standardized, validated paradigms from the lab in targeting relevant populations in naturalistic settings." In the Harrison and List (2004) taxonomy, this type of design corresponds to artefactual field experiments together with a subset of framed field experiments in which the experimental task is imported wholesale from the laboratory.

The framework of this paper makes the issue concrete. An AFE elicits target-market subjects' behaviour in a created environment: $(S{=}1, E{=}0, A{=}1)$. A "lab-in-the-field" design that imports laboratory paradigms (which could range from measuring generosity in dictator games to a specific estimate of τ) into a target-market population is, through the lens of the model, an AFE conducted at a site other than the campus lab. In contrast, an FFE replaces the created environment with a naturally-occurring one and moves the action set toward the natural one: $(S{=}1, E{=}1, A{=}1)$. A "lab-in-the-field" design in which farmers make real planting decisions on a target plot, workers receive real wages for real work, or donors are approached through real solicitations is an FFE, regardless of whether the data are collected on a clipboard at a market square.

As the framework in this study makes clear, the two design types identify different behavioral parameters. The AFE inherits the lab's structural-assumption advantage articulated in Section 5.3: the experimenter accepts Δ_E to permit the building of an environment that satisfies the theory's preconditions. An FFE inherits the naturalness of the field, such as the construct the subject faces, while losing the ability to vary the structural parameters in a controlled way. Reporting both as "lab-in-the-field" without further qualifier conflates two distinct estimands in the model. This effectively obstructs cumulative knowledge in just the manner this study warns against.

In the spirit of letting 1000 flowers bloom, I propose a modest amendment. This amendment to current practice preserves the label that has gained currency in the literature while restoring the design-level information the umbrella obscures: when reporting such an experiment, name the design type explicitly. In this case, either "lab-in-the-field: AFE" or "lab-in-the-field: FFE." This structure maintains the evocative shorthand but permits the reader some relief in understanding the recovered estimand in equation (11). I should note that my suggestion is not purely semantic. The behavioral parameter recovered by an AFE in a Tanzanian village differs from the parameter recovered by an FFE in the same village on Δ_E alone. Conflating them in reporting conflates them in the cumulative knowledge base. Studies in this tradition, which include Henrich et al. (2005) on cross-cultural variation in social preferences and Gneezy et al. (2009) on competitiveness across

matrilineal and patriarchal societies, are AFEs by the Harrison and List (2004) classification.

6.2 Online Experiments

In a similar spirit of lab-in-the-field:AFE and lab-in-the-field:FFE, online experiments are not a fifth design type. They reside on $(E{=}0, A{=}1)$ and typically map onto either the lab $(S{=}0)$ or the AFE $(S{=}1)$ depending on the recruited sample used by the researcher. As such, through the lens of the model, an MTurk or Prolific decision task with a stylized choice menu is a lab experiment with a different convenience pool: $F(\theta \mid S{=}0)$ has shifted relative to the campus lab population on age, geographic dispersion, screen-mediated work tolerance, prior platform tenure, and exposure to common experimental paradigms (Snowberg and Yariv, 2021; Peer et al., 2022). An online study that recruits a target population, say professional traders through LinkedIn, medical doctors via a medical society, donors through a charity mailing list, etc., is an AFE conducted online: $S{=}1$, with E and A unchanged.

A clarification is in order at the outset, since readers regularly conflate two distinct approaches in this area of study. The first is the online *field* experiments such as the A/B tests conducted at Walmart, Amazon, Lyft, Facebook or Uber. These are NFEs that happen to take place online because that is their natural ecosystem. The medium is digital but the design is $(S{=}1, E{=}1, A{=}0)$. The agent encounters the treatment in their naturally-occurring market, not in a study she enrolled in for a monetary payment. Such experiments deliver τ° directly and should be classified as NFEs rather than as a separate online category. The second approach is the online execution of a lab or AFE design. This approach is a stylised choice task hosted on MTurk, Prolific, or Qualtrics in which the subject knows she is enrolled in a study in a constructed market. It is this second case that the remainder of this subsection takes up.

With that clarification in place, the remainder of this subsection treats online execution of lab and AFE designs. Locating online experiments in the framework is not, however, the same as claiming they are identification-equivalent to their in-person counterparts. Holding (S, E, A) fixed, online execution introduces a distinct set of threats to identification: subject non-uniqueness (multiple accounts, bot and LLM-assisted submissions); unverified comprehension; absent experimenter monitoring; reduced stakes salience under delayed platform payment; and pool contamination from repeated paradigm exposure on professional respondent platforms.[8] None of these constitutes a new design type. Yet, each is an important estimand-affecting distinction that the (S, E, A) coordinates highlight. My preferred interpretation is that online execution is a feature of *implementation* of a design, not a new point in the design space. And, the methodological literature on online execution

[8]The closest predecessor to this line of argument is Charness et al. (2023a). Their defense of the comparative advantages of physical lab experiments helped to motivate the dimensions enumerated here. See also Charness et al. (2023b), who examine the identification consequences of online execution in detail.

should be read accordingly.

What online platforms do change for the lab and AFE cases, however, is the magnitude and shape of two specific wedges from equation (11). Δ_P operates differently because participation is filtered by the platform wage relative to the worker's outside option, and the marginal Prolific or MTurk participant is typically further from a representative agent on θ-relevant dimensions than the marginal undergraduate student participant (DellaVigna and Pope, 2018). Δ_S takes a particular form because the platforms select on traits, such as internet access, comfort with anonymous transactions, willingness to perform short paid tasks, prior exposure to and common experimental paradigms. That feature is of import if they correlate with the parameter of interest (Snowberg and Yariv, 2021).

This framework has a particular implication for the increasingly common practice of recruiting "nationally representative" online subject samples. They are nationally representative in that they match census quotas on observed characteristics, X. Such matching addresses the observable component of Δ_S yet does nothing about Δ_P. It is important to note that the selection that drives Δ_P is not on X but on θ. That means it is on preferences, beliefs about the study, opportunity costs, and outside options unobserved by the researcher.

A thought experiment might help to drive this point home. A 35-year-old college-educated woman who agrees to a paid online study has revealed something about her θ that her demographic twin who declined has not. Quota-matched samples eliminate the demographic component of selection (X) but leave the structural component intact. The participation filter in equation (5) of the model plays an important role whenever A=1, regardless of how the sample was advertised. A researcher who reports a quota-matched online experiment as evidence about a "representative" population has corrected one wedge; that is, the X-mappable component of Δ_S. Yet, it has left at least three other terms in the model (Δ_P, Δ_E, and the residual θ-relevant component of Δ_S) essentially untouched. In the language of the model, the estimand that is recovered is a conditional expectation. The conditionality is over the participating subset of subjects, not the population parameter the policymaker requires when crafting optimal policies.

6.3 Survey Experiments

Most of us have participated in some kind of survey, whether it was to evaluate an Amazon seller, let Yelp know about the local eatery, beauty salon, or dentist, or to express our political candidate preferences to pollsters. Such personal participation in surveys likely draws our attention to a part of the model so far left implicit: whether the income component π is realised. Whereas the two innovations in this section thus far have not required additions beyond the four canonical designs on (S,E,A), that is not the case for survey experiments.

Section 3 defines $Y_i = \pi(a^*;\theta_i,E,D)$ as the *realised* market component of payoff. Surveys

typically elicit a stated rather than realised action, and the gap between stated and revealed preference has been a stylised fact in environmental economics, marketing, transportation, and health economics for half a century (Cummings and Taylor, 1999; List and Gallet, 2001; Murphy et al., 2005; Penn and Hu, 2018). Indeed, recent work within economics has displayed this in full force. The past decade has seen a surge of survey experiments eliciting preferences and beliefs over taxation (Kuziemko et al., 2015; Stantcheva, 2021), intergenerational mobility (Alesina et al., 2018), immigration (Alesina et al., 2023), and macroeconomic expectations (Andre et al., 2022), with methodological guidance codified in Stantcheva (2023) and Haaland et al. (2023).

These designs are useful as descriptors of how respondents *reason*. However, the model reminds us that they reveal less about what respondents would *do* when the income component π is realized. Some critics view this as a matter of better wording, but that is not the challenge. In this case, the key challenge is structural. We have learned this from nearly 75 years of survey implementation within the literature of contingent valuation (CV). Such exercises involve a stated allocation in a hypothetical scenario. While it was debated for years, it is now commonly recognized that the object from that exercise is not the same object realised under π. Validation studies that compare survey responses to incentivised choices or real-world analogues find the mapping informative but partial (List, 2001; List and Gallet, 2001; Murphy et al., 2005; Hainmueller et al., 2015; Falk et al., 2018; Penn and Hu, 2018).

Given that such data generation approaches have been important within economics for decades, and have picked up steam recently in certain quarters of the profession, I accommodate survey experiments within the modeling framework herein. To do so, I define a fourth switch $V \in \{0,1\}$ indexing realization of π. The strong majority of economic experiments have $V{=}1$, as the payoffs are monetary. Under the model definition, even standard lab experiments with induced values have $V{=}1$: this means that the payoff is real even when the values are induced. Survey experiments, many psychology experiments, and most CV exercises have $V{=}0$. For inferential purposes, it is important to note that the wedge Δ_V is informative, often substantial, and decomposable from the others. And, through the lens of the model, V operates on π. The moral or social-image component M in equation (1) is realised whenever the action is performed, regardless of V. The implication is that a subject who informs a surveyor that she would donate \$50 to a hypothetical fund really does experience the self-image return on stating such in the moment. One cannot make pride hypothetical by labelling the dollars hypothetical.

To address these concerns, the model of this section delivers a structural account of the hypothetical bias the CV literature has grappled with for decades. In a stated-preference survey, $A{=}1$ (so $\psi(1) > \psi(0)$), $V{=}0$ (π drops out of the realised payoff), and the agent's effective objective collapses toward M. In such cases, behaviour should track the moral component almost exclusively; any bias will be largest precisely where M has a strong pro-social tilt. This is likely in those

sensitive question areas, such as would you give to charity, how would you vote, and healthy behaviours. And, it should be tiny or absent for goods where M barely fires. This pattern is what the meta-analytic literature documents: hypothetical bias is consistently larger for public goods and pro-social outcomes than for private goods, larger when respondents lack experience with the good, and attenuated by elicitation procedures (consequentiality scripts, cheap talk, certainty follow-ups) that dampen the social-image channel (List and Gallet, 2001; Murphy et al., 2005; Penn and Hu, 2018). Under the proposed extension, hypothetical bias is exactly Δ_V, and the meta-analytic regularities are not anomalies but predictions of equation (1) once V is added.

The same caveat about quota-matched samples that closed Section 6.2 applies *a fortiori* to survey experiments. A nationally representative survey panel matched on X has corrected the X-mappable component of Δ_S and leaves both the residual θ-component of Δ_S and Δ_P intact. The participation filter is, if anything, more aggressive in survey panels than in incentivised experiments. This is because survey panel members have selected into the panel, then into the survey, and then into responding. This represents three sequential filters operating on θ that no quota match can undo. Again, the lesson in this case is that demographic representativeness in surveys is not a substitute for behavioural representativeness. Rather, it is a partial correction whose limits should be stated when survey results are deployed for policy or theory testing. Crucially, combining insights from this section with the arguments in Section 5 reveal that combining a stated-preference question with the same incentivised choice in the same population measures Δ_V directly and is a useful empirical strategy that the literature has only begun to deploy systematically.

Early returns from exactly this strategy have yielded rich fruits. For example, across eight datasets and more than 13,000 participants, Chapman et al. (2025) examine several experimentally-validated self-assessments alongside incentivized elicitations. Their results are sobering through the lens of the model. They find that received correlations between self-assessments and their incentivized counterparts are quite modest, even after correcting for measurement error (and confounding factors account for 65 to 100 percent of the variation in the qualitative measures). More telling for the purposes of this section, correlations between self-assessments and moderating variables, such as demographics and cognitive ability, survive essentially unchanged when the incentivized elicitations are controlled for.[9] In the language of the model, what this suggests is that Δ_V is not white noise to be averaged away. Rather, it is structured variation, correlated with precisely the covariates researchers use self-assessments to study. A high validation correlation in

[9]Interestingly, two of their auxiliary findings connect directly to earlier sections of this study. First, four of their six self-assessments correlate more strongly with incentivized elicitations of constructs they were not designed to measure. This finding represents a construct-validity failure of exactly the type Section 5.3 anatomizes: the subject does not interpret the question as the question intends. Second, those respondents who have lower cognitive ability tend to rely more heavily on response heuristics (focal values of 0, 5, and 10). This result is in line with the model's prediction that when π drops out, responses tend to load on both M and on respondent-specific mappings to the scale rather than on the underlying primitive.

one sample, therefore, does not license treating a $V=0$ measure as a proxy for its $V=1$ counterpart, a lesson the CV literature has grappled with for 75 years and that the model delivers in one line.

7 Conclusion: From Skepticism to Complementarity

The economics discipline, which Mill, Friedman, Robinson, and Samuelson thought could not run experiments, now runs them at industrial scale. The approach is so ingrained in the economic science that we have now advanced to a new set of questions that the hard sciences have largely left untouched. This new set of questions relates to how received results generalize, scale, and what experimental types should be chosen to generate scientific knowledge optimally. In this study, I attempt to provide insights into the optimal experimental design question with a unified rational-choice model. The model characterizes all four standard experimental designs as comparative-static restrictions of a single maximization problem. Importantly, the simple framework reveals that each design identifies a parameter the others cannot.

The framework also produces an informative wedge via a behavioral decomposition. The gap between any design's estimand and the within-setting reference parameter can be decomposed into what I denote as switch-specific wedges. Each step in the lab $\rightarrow$ AFE $\rightarrow$ FFE $\rightarrow$ NFE progression flips one switch and removes one wedge. The NFE recovers τ° by construction precisely because all four wedges are set to zero. The wedge decomposition extends naturally to recent methodological developments. For example, "lab-in-the-field" designs locate cleanly on the existing (S,E,A) axes. Online platform experiments operate at $(E=0,A=1)$ with characteristic effects on Δ_S and Δ_P. And, survey experiments admit a fourth switch V on the realisation of π that gives hypothetical bias a structural interpretation as Δ_V. The framework is therefore not a closed taxonomy of four whimsical design choices, but an open one whose accommodation of new methods is itself a test of its scientific value.

A deeper point from the framework is an understanding about the value of the wedge decomposition. Through the lens of the model, the received wedges Δ_S, Δ_E, Δ_A, and Δ_P are not noise to be minimized or an unwanted distraction. They represent informative behavioral parameters that characterize how observed choices change as the experimental environment moves from the stylized to the natural, or when the subject pool is liberally varied. A discipline that produces estimates only at one end of the experimental spectrum, whether lab or a natural field experiment, observes only one value of each wedge. A discipline that produces estimates across the full spectrum can measure each of the wedges, understand what drives them, and use that understanding to produce insights on scaling, generalizability, mediation, moderation, and beyond. That is the science of experimentation, and it is the science that the skeptical luminaries thought economics could never perform well. The experimental revolution has made it possible.

References

Alesina, Alberto, Armando Miano, and Stefanie Stantcheva. 2023. "Immigration and Redistribution." *Review of Economic Studies* 90 (1): 1–39.

Alesina, Alberto, Stefanie Stantcheva, and Edoardo Teso. 2018. "Intergenerational Mobility and Preferences for Redistribution." *American Economic Review* 108 (2): 521–54.

Al-Ubaydli, Omar and John A. List. 2013. "On the Generalizability of Experimental Results in Economics." NBER Working Paper 19666.

Al-Ubaydli, Omar and John A. List. 2015. "Do Natural Field Experiments Afford Researchers More or Less Control than Laboratory Experiments?" *American Economic Review* 105 (5): 462–66.

Andersen, Steffen, Glenn W. Harrison, Morten I. Lau, and E. Elisabet Rutström. 2008. "Eliciting Risk and Time Preferences." *Econometrica* 76 (3): 583–618.

Andre, Peter, Carlo Pizzinelli, Christopher Roth, and Johannes Wohlfart. 2022. "Subjective Models of the Macroeconomy: Evidence from Experts and Representative Samples." *Review of Economic Studies* 89 (6): 2958–91.

Andreoni, James and Charles Sprenger. 2012. "Estimating Time Preferences from Convex Budgets." *American Economic Review* 102 (7): 3333–56.

Bates, Mary Ann and Rachel Glennerster. 2017. "The Generalizability Puzzle." *Stanford Social Innovation Review* 15 (3): 50–54.

National Commission for the Protection of Human Subjects of Biomedical and Behavioral Research. 1979. *The Belmont Report: Ethical Principles and Guidelines for the Protection of Human Subjects of Research.* Washington, DC: U.S. Department of Health, Education, and Welfare.

Benz, Matthias and Stephan Meier. 2008. "Do People Behave in Experiments as in the Field? Evidence from Donations." *Experimental Economics* 11 (3): 268–81.

Binmore, Ken. 2005. *Natural Justice.* Oxford: Oxford University Press.

Binmore, Ken. 2007. *Does Game Theory Work? The Bargaining Challenge.* Cambridge, MA: MIT Press.

Bohannon, John. 2016. "Mechanical Turk Upends Social Sciences." *Science* 352 (6291): 1263–64.

Bold, Tessa, Mwangi Kimenyi, Germano Mwabu, Alice Ng'ang'a, and Justin Sandefur. 2018. "Experimental Evidence on Scaling up Education Reforms in Kenya." *Journal of Public Economics* 168: 1–20.

Bowles, Samuel and Herbert Gintis. 2011. *A Cooperative Species: Human Reciprocity and Its Evolution.* Princeton: Princeton University Press.

Brodeur, Abel, Mathias Lé, Marc Sangnier, and Yanos Zylberberg. 2016. "Star Wars: The Empirics Strike Back." *American Economic Journal: Applied Economics* 8 (1): 1–32.

Brodeur, Abel, Nikolai Cook, and Anthony Heyes. 2020. "Methods Matter: p-Hacking and Publication Bias in Causal Analysis in Economics." *American Economic Review* 110 (11): 3634–60.

Burks, Stephen V., Jeffrey P. Carpenter, Lorenz Goette, and Aldo Rustichini. 2009. "Cognitive Skills Affect Economic Preferences, Strategic Behavior, and Job Attachment." *Proceedings of the National Academy of Sciences* 106 (19): 7745–50.

Camerer, Colin F., Anna Dreber, Eskil Forsell, et al. 2016. "Evaluating Replicability of Laboratory Experiments in Economics." *Science* 351 (6280): 1433–36.

Camerer, Colin F., Anna Dreber, Felix Holzmeister, et al. 2018. "Evaluating the Replicability of Social Science Experiments in Nature and Science between 2010 and 2015." *Nature Human Behaviour* 2 (9): 637–44.

Chamberlin, Edward H. 1948. "An Experimental Imperfect Market." *Journal of Political Economy* 56 (2): 95–108.

Chapman, Jonathan, Pietro Ortoleva, Erik Snowberg, Leeat Yariv, and Colin F. Camerer. 2025. "Reassessing Qualitative Self-Assessments and Experimental Validation." Working Paper.

Charness, Gary, James Cox, Catherine Eckel, Charles Holt, and Brian Jabarian. 2023a. "The Virtues of Lab Experiments." CESifo Working Paper Series 10796. CESifo.

Charness, Gary, Brian Jabarian, and John A. List. 2023b. "Generation Next: Experimentation with AI." NBER Working Paper 31679.

Cherry, Todd L., Peter Frykblom, and Jason F. Shogren. 2002. "Hardnose the Dictator." *American Economic Review* 92 (4): 1218–21.

Cohn, Alain, Ernst Fehr, and Lorenz Goette. 2015. "Fair Wages and Effort Provision: Combining Evidence from a Choice Experiment and a Field Experiment." *Management Science* 61 (8): 1777–94.

U.S. Department of Health and Human Services. 2018. *Federal Policy for the Protection of Human Subjects ("Common Rule"),* 45 C.F.R. Part 46.

Cotton, Christopher S., Brent R. Hickman, John A. List, Joseph Price, and Sutanuka Roy. 2026. "Why Don't Struggling Students Do Their Homework? Disentangling Motivation and Study Productivity as Drivers of Human Capital Formation." *Journal of Political Economy* 134 (1).

Cummings, Ronald G. and Laura O. Taylor. 1999. "Unbiased Value Estimates for Environmental Goods: A Cheap Talk Design for the Contingent Valuation Method." *American Economic Review* 89 (3): 649–65.

Damgaard, Mette Trier and Christina Gravert. 2017. "Now or Never! The Effect of Deadlines on Charitable Giving." *Journal of Behavioral and Experimental Economics* 66: 78–87.

Deaton, Angus. 2010. "Instruments, Randomization, and Learning about Development." *Journal of Economic Literature* 48 (2): 424–55.

DellaVigna, Stefano, John A. List, and Ulrike Malmendier. 2012. "Testing for Altruism and Social Pressure in Charitable Giving." *Quarterly Journal of Economics* 127 (1): 1–56.

DellaVigna, Stefano and Devin Pope. 2018. "What Motivates Effort? Evidence and Expert Forecasts." *Review of Economic Studies* 85 (2): 1029–69.

Delton, Andrew W., Max M. Krasnow, Leda Cosmides, and John Tooby. 2011. "Evolution of Direct Reciprocity under Uncertainty Can Explain Human Generosity in One-Shot Encounters." *Proceedings of the National Academy of Sciences* 108 (32): 13335–40.

Ellingsen, Tore, Magnus Johannesson, Johanna Möllerström, and Sara Munkhammar. 2012. "Social Framing Effects: Preferences or Beliefs?" *Games and Economic Behavior* 76 (1): 117–30.

Engel, Christoph. 2011. "Dictator Games: A Meta Study." *Experimental Economics* 14 (4): 583–610.

Engelbrecht-Wiggans, Richard, John A. List, and David H. Reiley. 2005. "Demand Reduction in Multi-Unit Auctions: Evidence from a Sportscard Field Experiment: Reply." *American Economic Review* 95 (1): 472–76.

Falk, Armin, Anke Becker, Thomas Dohmen, Benjamin Enke, David Huffman, and Uwe Sunde. 2018. "Global Evidence on Economic Preferences." *Quarterly Journal of Economics* 133 (4): 1645–92.

Federal Trade Commission v. Meta Platforms, Inc. 2025. No. 1:20-cv-03590 (D.D.C. Nov. 18, 2025) (Boasberg, C.J.).

Fehr, Ernst and Joseph Henrich. 2003. "Is Strong Reciprocity a Maladaptation? On the Evolutionary Foundations of Human Altruism." In *Genetic and Cultural Evolution of Cooperation,* edited by Peter Hammerstein, 55–82. Cambridge, MA: MIT Press.

Flory, Jeffrey A., Andreas Leibbrandt, and John A. List. 2015. "Do Competitive Workplaces Deter Female Workers? A Large-Scale Natural Field Experiment on Job Entry Decisions." *Review of Economic Studies* 82 (1): 122–55.

Fréchette, Guillaume R., Kim Sarnoff, and Leeat Yariv. 2022. "Experimental Economics: Past and Future." *Annual Review of Economics* 14: 777–794.

Friedman, Milton. 1953. "The Methodology of Positive Economics." In *Essays in Positive Economics,* 3–43. Chicago: University of Chicago Press.

Gintis, Herbert. 2000. "Strong Reciprocity and Human Sociality." *Journal of Theoretical Biology* 206 (2): 169–79.

Gneezy, Uri and Alex Imas. 2017. "Lab in the Field: Measuring Preferences in the Wild." In *Handbook of Economic Field Experiments,* vol. 1, edited by Abhijit V. Banerjee and Esther Duflo, 439–64. Amsterdam: North-Holland.

Gneezy, Uri and John A. List. 2006. "Putting Behavioral Economics to Work: Testing for Gift Exchange in Labor Markets Using Field Experiments." *Econometrica* 74 (5): 1365–84.

Gneezy, Uri, Kenneth L. Leonard, and John A. List. 2009. "Gender Differences in Competition: Evidence from a Matrilineal and a Patriarchal Society." *Econometrica* 77 (5): 1637–64.

Gneezy, Uri, John A. List, and Michael K. Price. 2012. "Toward an Understanding of Why People Discriminate: Evidence from a Series of Natural Field Experiments." NBER Working Paper 17855.

Goodman, Joseph, Lancelot Henry de Frahan, Justin E. Holz, John A. List, Niall MacMenamin, Evan C. McKay, Magne Mogstad, Sally Sadoff, and Hal Sider. 2026. "Consumer Demand and Market Competition with Time-Intensive Goods." NBER Working Paper 34743.

Guala, Francesco and Luigi Mittone. 2005. "Experiments in Economics: External Validity and the Robustness of Phenomena." *Journal of Economic Methodology* 12 (4): 495–515.

Haaland, Ingar, Christopher Roth, and Johannes Wohlfart. 2023. "Designing Information Provision Experiments." *Journal of Economic Literature* 61 (1): 3–40.

Hainmueller, Jens, Dominik Hangartner, and Teppei Yamamoto. 2015. "Validating Vignette and Conjoint Survey Experiments against Real-World Behavior." *Proceedings of the National Academy of Sciences* 112 (8): 2395–2400.

Hallsworth, Michael, John A. List, Robert D. Metcalfe, and Ivo Vlaev. 2017. "The Behavioralist as Tax Collector: Using Natural Field Experiments to Enhance Tax Compliance." *Journal of Public Economics* 148: 14–31.

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

Harrison, Glenn W. and John A. List. 2008. "Naturally Occurring Markets and Exogenous Laboratory Experiments: A Case Study of the Winner's Curse." *Economic Journal* 118 (528): 822–43.

Hedblom, Daniel, Brent R. Hickman, and John A. List. 2019. "Toward an Understanding of Corporate Social Responsibility: Theory and Field Experimental Evidence." NBER Working Paper 26222.

Henrich, Joseph, Robert Boyd, Samuel Bowles, Colin Camerer, Ernst Fehr, Herbert Gintis, Richard McElreath, Michael Alvard, Abigail Barr, Jean Ensminger, Natalie Smith Henrich, Kim Hill, Francisco Gil-White, Michael Gurven, Frank W. Marlowe, John Q. Patton, and David Tracer. 2005. "'Economic Man' in Cross-Cultural Perspective: Behavioral Experiments in 15 Small-Scale Societies." *Behavioral and Brain Sciences* 28 (6): 795–815.

Holt, Charles A. 2006. *Markets, Games, & Strategic Behavior.* Boston: Pearson Addison Wesley.

Holt, Charles A. and Susan K. Laury. 2002. "Risk Aversion and Incentive Effects." *American Economic Review* 92 (5): 1644–55.

Holzmeister, Felix, Magnus Johannesson, Robert Böhm, Anna Dreber, Jürgen Huber, and Michael Kirchler. 2024. "Heterogeneity in Effect Size Estimates." *Proceedings of the National Academy of Sciences* 121 (32): e2403490121.

Huber, Christoph, Anna Dreber, Jürgen Huber, Magnus Johannesson, Michael Kirchler, Utz Weitzel, et al. 2023. "Competition and Moral Behavior: A Meta-Analysis of Forty-Five Crowd-Sourced Experimental Designs." *Proceedings of the National Academy of Sciences* 120 (23): e2215572120.

Imbens, Guido W. and Joshua D. Angrist. 1994. "Identification and Estimation of Local Average Treatment Effects." *Econometrica* 62 (2): 467–75.

Inbar, Yoel. 2016. "Association between Contextual Dependence and Replicability in Psychology May Be Spurious." *Proceedings of the National Academy of Sciences* 113 (34): E4933–34.

Kahneman, Daniel and Amos Tversky. 1979. "Prospect Theory: An Analysis of Decision under Risk." *Econometrica* 47 (2): 263–91.

Karlan, Dean S. 2005. "Using Experimental Economics to Measure Social Capital and Predict Financial Decisions." *American Economic Review* 95 (5): 1688–99.

Karlan, Dean and John A. List. 2007. "Does Price Matter in Charitable Giving? Evidence from a Large-Scale Natural Field Experiment." *American Economic Review* 97 (5): 1774–93.

Klein, Richard A., Michelangelo Vianello, Fred Hasselman, Byron G. Adams, Reginald B. Adams, et al. 2018. "Many Labs 2: Investigating Variation in Replicability across Samples and Settings." *Advances in Methods and Practices in Psychological Science* 1 (4): 443–90.

Kohavi, Ron and Stefan Thomke. 2017. "The Surprising Power of Online Experiments." *Harvard Business Review* 95 (5): 74–82.

Kuziemko, Ilyana, Michael I. Norton, Emmanuel Saez, and Stefanie Stantcheva. 2015. "How Elastic Are Preferences for Redistribution? Evidence from Randomized Survey Experiments." *American Economic Review* 105 (4): 1478–1508.

Levitt, Steven D. and John A. List. 2007a. "Viewpoint: On the Generalizability of Lab Behaviour to the Field." *Canadian Journal of Economics* 40 (2): 347–70.

Levitt, Steven D. and John A. List. 2007b. "What Do Laboratory Experiments Measuring Social Preferences Reveal about the Real World?" *Journal of Economic Perspectives* 21 (2): 153–74.

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.

Liberman, Varda, Steven M. Samuels, and Lee Ross. 2004. "The Name of the Game: Predictive Power of Reputations versus Situational Labels in Determining Prisoner's Dilemma Game Moves." *Personality and Social Psychology Bulletin* 30 (9): 1175–85.

List, John A. 2001. "Do Explicit Warnings Eliminate the Hypothetical Bias in Elicitation Procedures? Evidence from Field Auctions for Sportscards." *American Economic Review* 91 (5): 1498–1507.

List, John A. 2003. "Does Market Experience Eliminate Market Anomalies?" *Quarterly Journal of Economics* 118 (1): 41–71.

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

List, John A. 2004b. "Testing Neoclassical Competitive Theory in Multilateral Decentralized Markets." *Journal of Political Economy* 112 (5): 1131–56.

List, John A. 2006. "The Behavioralist Meets the Market: Measuring Social Preferences and Reputation Effects in Actual Transactions." *Journal of Political Economy* 114 (1): 1–37.

List, John A. 2007. "On the Interpretation of Giving in Dictator Games." *Journal of Political Economy* 115 (3): 482–93.

List, John A. 2009. "The Economics of Open Air Markets." NBER Working Paper 15420.

List, John A. 2020a. "Experimental Tests of the Endowment Effect and the Coase Theorem." Natural Field Experiments 00687, The Field Experiments Website.

List, John A. 2020b. "Non est Disputandum de Generalizability? A Glimpse into the External Validity Trial." NBER Working Paper 27535.

List, John A. 2022. *The Voltage Effect: How to Make Good Ideas Great and Great Ideas Scale.* London: Penguin Business.

List, John A. 2024. "Optimally Generate Policy-Based Evidence before Scaling." *Nature* 626: 491–99.

List, John A. 2026a. *Experimental Economics: Theory and Practice.* Chicago: University of Chicago Press.

List, John A. 2026b. "External Validity as an Economic Problem." Working Paper, University of Chicago.

List, John A. Forthcoming. "Beyond Replication: It Is Time to Fix the Generalizability Crisis." *Nature*.

List, John A. and Craig A. Gallet. 2001. "What Experimental Protocol Influence Disparities Between Actual and Hypothetical Stated Values?" *Environmental and Resource Economics* 20 (3): 241–54.

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.* London: John W. Parker.

Murphy, James J., P. Geoffrey Allen, Thomas H. Stevens, and Darryl Weatherhead. 2005. "A Meta-Analysis of Hypothetical Bias in Stated Preference Valuation." *Environmental and Resource Economics* 30 (3): 313–25.

Neyman, Jerzy. (1923) 1990. "On the Application of Probability Theory to Agricultural Experiments: Essay on Principles; Section 9." Translated by Dorota M. Dabrowska and T. P. Speed. *Statistical Science* 5 (4): 465–72.

Open Science Collaboration. 2015. "Estimating the Reproducibility of Psychological Science." *Science* 349 (6251): aac4716.

Orne, Martin T. 1962. "On the Social Psychology of the Psychological Experiment." *American Psychologist* 17 (11): 776–83.

Peer, Eyal, David Rothschild, Andrew Gordon, Zak Evernden, and Ekaterina Damer. 2022. "Data Quality of Platforms and Panels for Online Behavioral Research." *Behavior Research Methods* 54 (4): 1643–62.

Penn, Jerrod M. and Wuyang Hu. 2018. "Understanding Hypothetical Bias: An Enhanced Meta-Analysis." *American Journal of Agricultural Economics* 100 (4): 1186–1206.

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.

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 (5): 688–701.

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

Smith, Vernon L. 1962. "An Experimental Study of Competitive Market Behavior." *Journal of Political Economy* 70 (2): 111–37.

Snowberg, Erik and Leeat Yariv. 2021. "Testing the Waters: Behavior across Participant Pools." *American Economic Review* 111 (2): 687–719.

Soman, Dilip and Tanjim Hossain. 2021. "Successfully Scaled Solutions Need Not Be Homogenous." *Behavioural Public Policy* 5 (1): 80–89.

Stantcheva, Stefanie. 2021. "Understanding Tax Policy: How Do People Reason?" *Quarterly Journal of Economics* 136 (4): 2309–69.

Stantcheva, Stefanie. 2023. "How to Run Surveys: A Guide to Creating Your Own Identifying Variation and Revealing the Invisible." *Annual Review of Economics* 15 (1): 205–34.

Szaszi, Barnabas, Daniel G. Goldstein, Dilip Soman, and Susan Michie. 2025. "Generalizability of Choice Architecture Interventions." *Nature Reviews Psychology* 4 (8): 518–29.

Viceisza, Angelino C. G. 2012. *Treating the Field as a Lab: A Basic Guide to Conducting Economics Experiments for Policymaking.* Food Security in Practice Technical Guide Series 7. Washington, DC: International Food Policy Research Institute.

Viceisza, Angelino C. G. 2016. "Creating a Lab in the Field: Economics Experiments for Policymaking." *Journal of Economic Surveys* 30 (5): 835–54.

Vivalt, Eva. 2020. "How Much Can We Generalize from Impact Evaluations?" *Journal of the European Economic Association* 18 (6): 3045–89.

Williams, Martin J. 2019. "External Validity and Policy Adaptation: From Impact Evaluation to Policy Design." *World Bank Research Observer* 35 (2): 158–91.

Yarkoni, Tal. 2022. "The Generalizability Crisis." *Behavioral and Brain Sciences* 45: e1.

Zhang, C. Yiwei, Jeffrey Hemmeter, Judd B. Kessler, Robert D. Metcalfe, and Robert Weathers. 2022. "Nudging Timely Wage Reporting: Field Experimental Evidence from the U.S. Supplemental Security Income Program." *Management Science* 69 (3): 1341–53.