Two Selection Problems, One Bias Term: Experimental Sign-Up Is Treatment Choice in Disguise*

John A. List

University of Chicago and NBER

July 2026

Abstract

Consider a referee report containing the sentence: "Assignment to treatment was voluntary, though the authors acknowledge this as a limitation in the conclusion." The Editor would summarily reject. Now replace "assignment to treatment" with "entry into the experimental sample," and the sentence describes the modal published experiment. This study shows that both biases are the same mathematical object. The bias from non-random sample entry is the sorting-on-gains term from the classic selection decomposition. Interestingly, they both admit an isomorphic Roy micro-foundation and inverse-Mills characterization. A simple calibration reveals that for the typical experiment, the observed bias can be large. One solution to this bias is to conduct a natural field experiment (NFE). When an NFE is not possible, three alternative reporting practices are offered.

JEL codes: C21, C24, C90, C93

Keywords: field experiments; selection bias; internal validity; external validity; Roy model; participation

*I thank Andrew Robert Kennedy, Paul Kirch, Magne Mogstad, and Julia Seither for helpful discussions and comments. Participants at the Economic Science Association also provided useful feedback. All errors are my own.

1 Introduction

In the natural sciences, testing theory against experimental evidence is second nature: the give and take between the two is the scientific method. Economics has increasingly followed suit, with many scholars taking control of the assignment mechanism rather than modeling naturally-occurring (or historical) data alone. Presumably this is because they are unsatisfied with the identification assumptions such data demand, or because they wish to move beyond measurement to a deeper exploration of the "whys" underlying their findings.

Throughout the 1990s, the credibility revolution both accelerated and disciplined this movement. Empirical developments led to better research design (randomization foremost among its tools) and design-based identification spread from labor to development, public, health, environmental, and increasingly finance and macroeconomics (Angrist and Pischke, 2010). One central lesson of the revolution that has by now hardened into a norm that is difficult to escape is that selection into treatment is disqualifying. On identification grounds alone, a skeptic need not argue that the bias is large. Indeed, with an unknown assignment mechanism any evidence that assignment to treatment was voluntary suffices to sink even the best ideas from rising in import.

Yet, the revolution's writ has run unevenly for the past few decades. A study in which entry into the experimental sample is voluntary remains standard practice within the academy and government circles. Any resulting bias is either completely ignored, relegated to a limitations paragraph or footnote, or defined away by conditioning the estimand on the sample that showed up for the experiment. The irony of this approach in the social sciences will not be lost on the astute reader. It would be difficult to argue that experiments are not truly the credibility revolution's crown jewels, yet they are precisely the studies in which entry is voluntary by construction (in most cases). This is so because most experimentalists choose an experimental approach that generates such selection. The instrument the profession adopted to eliminate one selection problem thereby institutionalized another.

In this note, I show that the two selection problems are the same mathematical object. The bias from observing a non-random sample is, term for term, the sorting-on-gains bias from the canonical selection-into-treatment decomposition. Indeed, I show that the two bias terms are identical in every aspect: the same covariance between a voluntary choice and idiosyncratic gains, the same scaling denominator, the same Roy micro-foundation, and the same inverse-Mills correction. In fact, the only difference is that one indicator variable is swapped for another. I am unaware of any theorem that licenses the asymmetry

in our standards.

After showing that the two selection problems are not merely related concerns but the same element, I outline the experimental design consequences and practical solutions moving forward. This includes a two-pronged approach. First, searching more intensively for natural field experiment (NFE) opportunities. This solution follows from the symmetry of the selection terms: just as randomization is the design instrument that sets the selection covariance to zero, the natural field experiment (Harrison and List, 2004) is the design instrument that aligns the experiment's participation process with the target environment's, setting the sign-up covariance to zero relative to the policy population. Second, when an NFE is not viable, I advocate for three new reporting practices: i) construct a selection table, ii) explore participation shifters, and iii) leverage sensitivity reporting. While all three of these practices move our science to a higher level, ultimately, an underlying argument emerges that is true across every aspect of experimental design (List, 2026a): the first-best solution remains fixing the problem at the design stage rather than an econometric fix after conducting the experiment.

In the next section, I summarize two key experimental problems within economics. A simple model reveals the two bias decompositions and the isomorphism within those experimental problems. Section 3 provides the Roy micro-foundation of each and Section 4 calibrates bias magnitudes. The penultimate section outlines design and reporting implications. Section 6 concludes.

2 Two selection problems, one bias term

Experimental Problem 1. *Quantifying economic fundamentals, measuring treatment effects, and identifying key mediators and moderators in an ethically responsible manner.*

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

Economic experiments lend insights into two types of scientific evaluation problems, what List (2026a) denotes as Experimental Problem 1 (EP1) and Experimental Problem 2 (EP2). A key piece of EP1 is internal validity, whereas EP2 revolves around external validity and scaling. The two key selection problems discussed in this note revolve around the interplay of EP1 and EP2.

To fix ideas, I begin by considering a population of policy interest indexed by i, with

outcome equation

$$y_i = \alpha + \beta_i d_i + U_i, \tag{1}$$

where $d_i \in \{0, 1\}$ indicates treatment, β_i is the individual treatment effect, and U_i captures unobservables with $\mathrm{E}(U_i) = 0$ in the population. Let $P_i \in \{0, 1\}$ indicate whether unit i enters the experimental sample: $P_i = 1$ for those who, facing the invitation, sign up. The policy parameter is the population average effect $\bar{\beta} \equiv \mathrm{E}(\beta_i)$.

EP1: selection into treatment. Let us assume that treatment status is individually chosen rather than assigned randomly. Within any experimental sample, the naive comparison of means in this case decomposes as

$$\mathrm{E}(y \mid d = 1) - \mathrm{E}(y \mid d = 0) \;=\; \bar{\beta} \;+\; \underbrace{\frac{\mathrm{Cov}(d_i, \beta_i)}{\mathrm{E}(d_i)}}_{\text{sorting on gains}} \;+\; \underbrace{\mathrm{E}(U \mid d = 1) - \mathrm{E}(U \mid d = 0)}_{\text{baseline selection}}. \tag{2}$$

The sorting term follows from the identity $\mathrm{E}(\beta_i \mid d_i = 1) - \bar{\beta} = \mathrm{Cov}(d_i, \beta_i)/\mathrm{E}(d_i)$. Randomization is the design instrument that is the key cog in the internal validity machine of EP1: assigning d_i by any randomization technique, from Bernoulli trials to paired randomized experiments, imposes $d_i \perp (\beta_i, U_i)$. As such, both bias terms in (2) vanish by construction, as $\mathrm{E}(U \mid d = 1) = \mathrm{E}(U \mid d = 0)$. Such causal simplicity has helped to raise the experimental approach to be the crown jewel of the credibility revolution.

EP2: sign-up or selection into the sample. An aspect of EP2 that deserves mention here pertains to the generalizability of the research findings, and how we might anticipate the received insights scaling to the broader world. In this case, notice what randomization actually delivers and what it does not deliver. All empirical comparisons occur among experimental participants by definition. Thus, the experimental estimand is $\mathrm{E}(\beta_i \mid P_i = 1)$. The gap between what we estimate and what we desire to estimate is therefore

$$\mathrm{E}(\beta_i \mid P_i = 1) - \bar{\beta} \;=\; \frac{\mathrm{Cov}(P_i, \beta_i)}{\mathrm{E}(P_i)}. \tag{3}$$

Importantly, note that expression (3) is the *identical mathematical object* as the sorting-on-gains term in (2), with P substituted for d: the covariance between an indicator of a voluntary decision and idiosyncratic gains, scaled by the share who enter. There exists no formal result in econometrics of which I am aware under which covariance with a treatment-choice indicator is fatal while covariance with a sample-entry indicator is benign. Yet, in my experiences within the academy, organizations, and even in the court room discussing the merits of empirical evidence, it is universally accepted to treat (2) as

disqualifying and (3) as a footnote.[1]

As written, equation (3) takes the population average $\bar{\beta}$ as the target. This is the correct benchmark when the policy under consideration applies universally. More generally, let $P_i^* \in \{0, 1\}$ indicate participation under the rules of the target environment itself. This might be the market or policy whose effects we wish to predict. Then define the policy parameter as $\beta^* \equiv E(\beta_i \mid P_i^* = 1)$. The relevant bias becomes

$$E(\beta_i \mid P_i = 1) - E(\beta_i \mid P_i^* = 1). \tag{4}$$

Crucially, this term vanishes when the process generating entry into the experiment matches the participation process in the target environment. Equation (3) is the special case $P_i^* \equiv 1$. Throughout this study, when I write that a design "sets the sign-up covariance to zero," the statement is relative to the researcher's declared target: $\text{Cov}(P_i, \beta_i) = 0$ within the population defined by $P_i^* = 1$.

I suspect that at this point, the reader who is well versed in this area will have two initial responses. First, this is all definitional: simply relabel the estimand as "the average effect for our sample," whereupon (3) is zero by fiat. This move is superficially in the spirit of local average treatment effects, but the analogy fails at the crucial step. The LATE estimand is accepted because the complier subpopulation is carved out by variation whose assignment mechanism is known and credibly exogenous. Combining that fact with the discipline of the monotonicity condition makes the complier group well-defined. The implication is that the estimand is defined by a selection process with known provenance. This remains the case even when the compliers themselves cannot be individually named. An unrestricted volunteer sample offers no analogous instrument: participation is selected on expected gains, and relabeling the estimand does not change whose gains it averages.[2]

To be precise about where the sin lies: with internal randomization, $E(\beta_i \mid P_i = 1)$ is a perfectly well-identified parameter for a real population. The problem is not estimating it. The problem is silently substituting it for β^* without ever declaring the target, and

[1] A baseline-selection analogue of equation (3) also exists, $E(U \mid P = 1)$; or, equivalently, $(1-p)$ times the participant–nonparticipant gap $E(U \mid P = 1) - E(U \mid P = 0)$. This matters when levels rather than effects are transported (such as the case when measuring preferences, beliefs, or constraints (see List, 2026b)). For the transportability of treatment effects, the term in equation (3) is the sharp one, so I focus therein.

[2] Before moving on, I should note that even LATE's acceptance is a conditional one. The conditionality follows from a prominent critique, which holds that instrument-dependent estimands answer questions of unclear policy relevance (Deaton, 2010; Heckman and Urzua, 2010). Incidentally, I view that debate as strengthening, rather than weakening, my arguments. This follows from the fact that one of the profession's most visible debates about LATE is, at its heart, a fight about whose effects the estimand averages. In other words, it is an EP2 fight.

the bias in equation (4) reappears the moment anyone uses the estimate for the policy population, which is a key reason the estimate exists in the first place. Relabeling without a declared target is not a LATE; it is a LATE without the instrument.

Second, this is a matter of experimenter choice. We can randomize d_i but in many cases, due to the experimentalist choosing to conduct a survey, a lab, an artefactual (AFE), or a framed (FFE) field experiment, they choose not to address the P_i problem.[3] In this case, I point the reader to how we handle a key identification violation underlying internal validity, attrition. Attrition, too, is often unavoidable, yet standard practice demands measurement, bounds, and corrections (Lee, 2009) rather than silence. I return to this as a potential type of solution below.

3 A Roy model of experimental sign-up

A first-order question that arises given that the bias terms are identical mathematical objects relates to what magnitude should we expect the $\text{Cov}(P_i, \beta_i) \neq 0$ to take in practice? We can gain intuitions using the Roy machinery we already deploy for EP1 (Roy, 1951; Heckman, 1979). I begin by modeling entry into the experiment as

$$P_i = \mathbf{1}\{\, \gamma\, \text{E}(\beta_i \mid \mathcal{I}_i) + Z_i'\delta - v_i \geq 0 \,\}. \tag{5}$$

In equation (5), $\mathcal{I}_i$ represents the agent's information set when they receive the experimental invitation; Z_i includes participation shifters, some of which are controllable by the experimenter (show-up compensation) and some of which are individual parametric shifters (time costs, trust, distance); v_i is an idiosyncratic disturbance. It is convenient to work with the latent participation index directly. Define

$$s_i \equiv \gamma\, \text{E}(\beta_i \mid \mathcal{I}_i) + Z_i'\delta - v_i, \tag{6}$$

so that $P_i = \mathbf{1}\{s_i \geq 0\}$, and let $\rho \equiv \text{Corr}(\beta_i, s_i)$ denote the correlation between individual gains and the propensity to participate. The entry equation shows that if potential subjects hold even a noisy private signal about their own gains and $\gamma \neq 0$ (i.e., subjects

[3]Harrison and List (2004) define three types of field experiments: 1) an artefactual field experiment (AFE) is a conventional lab experiment, but with a non-standard subject pool; 2) A framed field experiment (FFE) builds on an AFE by adding field context in the commodity, task, and/or information that the subjects can use, and 3) a natural field experiment (NFE) builds on an FFE in that it occurs in the environment where the subjects naturally undertake these tasks and where the subjects do not know that they are taking part in an experiment. In recent work, List (2026a,b) discusses the ethics of NFEs.

sign up for the experiment because they expect the program to help them, or decline because they expect it will not), then s_i loads on β_i and $\rho \neq 0$.

If I assume joint normality of (β_i, s_i) and normalize $\text{Var}(s_i) = 1$, the standard selection formula becomes:

$$\text{E}(\beta_i \mid P_i = 1) \;=\; \bar{\beta} \;+\; \rho\,\sigma_\beta\,\lambda(c). \tag{7}$$

In equation (7), σ_β is the standard deviation of gains, $\lambda(\cdot)$ is the inverse Mills ratio, and c is the standardized participation threshold, so that the participation rate satisfies $p = \Phi(-c)$ and λ may be written as a function of p alone, $\lambda(p) = \phi(\Phi^{-1}(p))\,/p$. Equation (7) will be familiar to most readers, as it is the intuition from Heckman (1979), applied not to who takes up the treatment but to who signs up for the experiment. An implication of my model is that sign-up into an experimental trial, a survey, a lab experiment, an AFE, or an FFE is itself a Roy model, and therefore every signature on a consent form is a draw from it, or represents the $P_i = 1$ group over which inference should be made.

Defining ρ on the full index in equation (6) has a substantive payoff: it makes ρ an economic object rather than a nuisance correlation. In this form, the index inherits correlation with β_i through three channels: i) the private signal $\text{E}(\beta_i \mid \mathcal{I}_i)$, ii) any dependence between the participation shifters Z_i and gains, and iii) the disturbance v_i. This has a nice feature for our purposes because we can now pinpoint which channels a given experimental protocol activates in our design discussion below.

What naturally follows is a feature of experimental economics that is often not discussed. That is, the experimenter's leverage over P_i is itself a design choice that importantly impacts learnings around EP1 and EP2. In a lab, AFE, or FFE, the experimental subject decides P_i via participation directly, so that margin is experiment-specific by construction. The NFE, by contrast, removes the sign-up margin entirely, as under this design type subjects choose whether to enter the market. In a rational model, they do so under the rules governing the institution (Al-Ubaydli and List, 2015). In this way, the NFE returns control over the selection process to the researcher, exactly as randomization returns control over the assignment mechanism.

Equation (7) provides direct insight on when the sign-up bias, in general, has less bite. The answer is contained directly in the product of the last three terms: setting any one of these to zero makes the bias term vanish. Yet, there is a key question that the algebra conceals: sometimes the volunteers are exactly who you want to study. If the studied program will itself be offered on an opt-in basis, under a similar information environment and similar participation costs, then $\text{E}(\beta_i \mid P_i = 1)$ is not a biased estimate of the policy parameter. Indeed, it *is* the policy parameter of interest. In the notation of Section 2,

this is exactly the case $P_i = P_i^*$: the experiment's participation process replicates the policy's, the two conditioning events coincide, and the wedge in equation (4) is zero with volunteers doing the selecting.

When the two conditioning events coincide, the sorting term can be viewed as a unique empirical feature, not a bug. This is because we satisfy what is necessary for EP2 directly. Interestingly, this is the underlying logic of a unique aspect of an NFE stated in reverse: what matters is not that P_i is random, but that the experimental process generating P_i matches the natural, or targeted environment's, process. The bias in equation (7) is therefore a statement about a mismatch between two selection processes that induces an experimental subject mismatch.

One design lever that can directly impact the magnitude of sign-up bias is the information set, $\mathcal{I}_i$. A comparison of the received practices across scientific fields is instructive in this case. In a clinical trial, for example, standard practice includes informed consent, with that practice including a requirement that subjects understand the condition being studied, the potential treatment and its anticipated benefits, and any risks involved. The signal in the model, $E(\beta_i \mid \mathcal{I}_i)$, is therefore sharp, and able to operate forcefully due to regulatory construction. Accordingly, γ provides a rational reason for why patients who expect to benefit will enroll, and those who do not, will decline.[4] It is useful to contrast the standard medical protocol with the typical economics laboratory experiment. In most lab, AFEs, and FFEs, the recruitment message is deliberately uninformative. We often read "a decision-making study, earn $20, one hour" on flyers. Here $\mathcal{I}_i$ contains essentially no experiment-specific information about β_i, so $\gamma E(\beta_i \mid \mathcal{I}_i)$ collapses toward a constant and cannot generate sorting on gains. In this case, experimental entry is instead governed by the $Z_i'\delta$ shifters: time costs, the show-up fee, distance to the lab, curiosity, etc. The economics lab's much-maligned opacity is an accidental virtue, at least for the estimation of treatment effects in this genre of studies.

But, there is a catch. The protection afforded by an uninformative invitation holds only if the $Z_i'\delta$ shifters are orthogonal to the object being measured. Two cases show how this can fail. As a first case, consider a worker training program. In such instances, those individuals who sign up on the basis of low opportunity costs of time or weak labor market attachment are drawn from a region of the returns distribution that can differ systematically from the population. Importantly, in such examples, even when the

[4] I should be careful to note that this is not what Heckman (1992) described three decades ago. That discussion concerned randomization bias, which describes potential changes in the participant pool due to the act of randomization itself (some people might be averse to a coin flip). I consider randomization bias as a cousin of informative recruitment materials sharpening $\mathrm{E}(\beta_i \mid \mathcal{I}_i)$.

invitation reveals nothing about the treatments, the $Z_i'\delta$ shifters are non-orthogonal to the gain.

A second case can be found via revisiting the first part of EP1. In doing so, a tension that is subtle is revealed. Consider a laboratory experiment that measures economic primitives themselves. If the people who volunteer for experiments are more pro-social, more trusting, or differ in risk attitudes from those who do not volunteer, then the $Z_i'\delta$ shifters are not merely nuisance variation.[5] Indeed, they are the measurand. This means that if the experimentalist is measuring a primitive, such as a social preference or a risk parameter, selection on who volunteers is selection on the very object of study. In that case, no amount of vagueness in the flyer can undo it. This is, in a formal sense, the baseline-selection analogue that is noted in footnote 1: if we desire to transport primitives, the relevant term is $E(U \mid P = 1)$, and an uninformative $\mathcal{I}_i$ does nothing to set it to zero.

4 How large is the bias? A calibration

We are now in a position to estimate the size of the bias, at least as a back of the envelope exercise using expression (7). I begin by letting $p \equiv \mathrm{E}(P_i)$ denote the participation rate among those invited. Given this, under normality $\lambda(p) = \phi(\Phi^{-1}(p))/p$. The proportional bias in the experimental estimate can then be given by $\rho\,(\sigma_\beta/|\bar{\beta}|)\,\lambda(p)$. In Table 1, I report the level of bias for modest sorting ($\rho = 0.3$) and treatment-effect heterogeneity comparable in magnitude to the mean effect ($\sigma_\beta = |\bar{\beta}|$). This level of heterogeneity is consistent with the heterogeneity routinely estimated in various field settings.

The estimates in Table 1, which might be viewed uncomfortably by some, can be read as follows: suppose your intervention truly raises the outcome by 10 percent in the population. And, upon your experimental announcement, let us assume that half of those

[5]This particular concern has a long lineage across the social sciences. Levitt and List (2007) assemble the conceptual case alongside empirical evidence from the psychology literature on the "volunteer subject." That literature argues that experimental volunteers are more educated, more sociable, and higher in need for social approval than non-volunteers. Direct evidence within economics is threefold. Selection on participation shifters is well documented: in the rare designs observing the full invited population, participants had significantly lower income, more leisure time, greater interest in the tasks, and volunteered more (Slonim et al., 2013). Selection on the measurand itself is experimentally demonstrated: when subjects are given the opportunity to sort away from a sharing environment, sharing collapses among those who exit (Lazear, Malmendier, and Weber, 2012). Yet measured social and risk preferences differed little from the population in the settings where the comparison could be made (Cleave et al., 2013; Falk, Meier, and Zehnder, 2013). From a scientific discovery aspect, note that each of these facts is knowable only because these authors observed, or experimentally controlled, the participation margin. That is, because they could construct comparisons along the lines recommended in Section 5. In scanning the recent literature, it is clear that for the typical experiment, whether $\mathrm{Cov}(P_i, \beta_i) = 0$ remains unanswerable under current reporting practice.

Table 1: Proportional bias $|\mathrm{E}(\beta \mid P{=}1) - \bar{\beta}|/|\bar{\beta}|$ under $\rho = 0.3$, $\sigma_\beta = |\bar{\beta}|$

Participation rate p	0.50	0.25	0.10	0.05	0.02
$\lambda(p)$	0.80	1.27	1.75	2.06	2.42
Proportional bias	0.24	0.38	0.53	0.62	0.73

Notes: The table reports the proportional bias in an experimental estimate of the population average treatment effect computed from equation (7). I assume joint normality, modest sorting ($\rho = 0.3$), and heterogeneity equal to the mean effect ($\sigma_\beta = |\bar{\beta}|$). If I assume larger heterogeneity ($\sigma_\beta > |\bar{\beta}|$), the bias scales proportionally.

who received an invitation or read the flyer signed up. The calibration says that in this case the experiment will estimate an effect size of 12.4 percent rather than the true 10 percent. If one in ten sign up, the reported effect size will be 15.3 percent. If only one in fifty, the response rate of a typical mailed solicitation, then the measured effect is 17.3 percent.[6] In terms of optimal design considerations, nothing about the randomization, power, or execution is at fault. Rather, the distortion is an aspect of the necessary recruitment.

Table 1 reveals that the selection bias grows mechanically as participation falls (as we move from column 1 in Table 1 to column 5). This result follows from the fact that the average participant is drawn from an ever more extreme tail of the gains distribution as the participation rate shrinks. A relevant, and perverse, comparative static necessarily arises for EP2: the more difficulty the experimentalist has in recruiting (to not only fill the samples but fulfill balance requirements), the greater the wedge between their parameter estimate and what they are trying to approximate. Table 1 also speaks to aspects associated with the scaling literature. In experiments where those who expect to gain the most are the most likely to enroll ($\rho > 0$), any experiment that involves selection will overstate what the program will deliver at scale if the initial subject beliefs were directionally correct. This kind of selection represents one of the microfoundations that the scaling literature (Al-Ubaydli et al., 2017; List, 2024) has leveraged to explain causes of the voltage effect.

[6]The arithmetic behind the presented figures in Table 1 traces directly through equation (7). Note that under an assumption of normality, the inverse Mills ratio can be expressed as a function of the participation rate alone: $\lambda(p) = \phi\big(\Phi^{-1}(p)\big)/p$. So, let us consider what happens at $p = 0.50$, for example: $\Phi^{-1}(0.50) = 0$ and $\phi(0) = 0.399$, giving $\lambda = 0.80$. The proportional bias is then $\rho\,(\sigma_\beta/|\bar{\beta}|)\,\lambda(p) = 0.3 \times 1 \times 0.80 = 0.24$. As such, a true effect of 10 percent is measured as $10 \times 1.24 = 12.4$ percent. The entries at $p = 0.10$ and $p = 0.02$ follow in a similar fashion: with $\lambda = 1.75$ and 2.42, the measured effects are equal to 15.3 and 17.3 percent, respectively.

5 Design and reporting implications

What, then, is to be done? One instinct that is honed by decades of selection econometrics is to reach for a fix at the estimation stage. Perhaps a weight, a control function, a bound. But, recall a signature development of the credibility revolution is that we did not solve EP1 by perfecting the Heckman correction. The profession solved it via changing the design so that no correction was needed: randomization set $\text{Cov}(d_i, \beta_i) = 0$.

In a similar vein, the symmetry argument herein is suggestive of a remedy that is a design innovation rather than an econometric patch: conduct an NFE, which aligns P_i with P_i^* and thereby produces $\text{Cov}(P_i, \beta_i) = 0$ relative to the target population. An NFE works well in this case because when the experiment does not enter the agent's information set $\mathcal{I}_i$, the term $\gamma \text{E}(\beta_i \mid \mathcal{I}_i)$ in equation (5) contains no experiment-specific component. This has the nice feature that entry into an NFE is governed by the same participation process that operates in the policy environment itself (List, 2026c). In an NFE, the sample is still selected in the sense that there is sorting in every market, but the people are selected by the market's rules rather than the subjects' preferences, beliefs, and constraints over the experiment. Accordingly, the NFE returns control over the individual's participation decision, P_i, to the researcher, as randomization returns control over d_i. As such, my advocacy for the NFE design based solution is not merely a taste for naturalism; rather, it is the EP2 analogue of the coin flip to solve EP1 issues.

In some cases, NFEs are infeasible, inappropriate, or out of reach for the question at hand. In these instances, I offer three reporting practices to move us forward.

1. Construct a selection table. The cheapest of the three proposed practices is also the most immediately actionable. Just as no experimental paper survives review without a balance table on d_i, papers relying on volunteer samples should report the $P_i = 1$ sample against the invited (or target) population on a rich set of observables. And, if possible provide a raw participation rate p as well across the sample, including subgroups, as this can provide unique insights about heterogeneity. I should be wary not to oversell this first practice. Balance on observables is a necessary condition, not a sufficient one, and the bias in Table 1 runs entirely through unobserved gains. But the selection table disciplines the discussion in the same manner that its d_i counterpart does in modern studies. In this spirit, within current academic prose, this first reporting principle replaces a vague limitations paragraph with a set of falsifiable comparisons. And, the reported p feeds directly into the sensitivity analysis below via $\lambda(p)$.

2. Explore participation shifters. A fix that goes beyond reporting is to bring the participation equation into the design itself. One approach to do this is by randomizing

elements of Z_i directly. Design features including, but not limited to, show-up fees, recruitment intensity, and invitation framing are all useful candidates. Insightful recent work by Dutz et al. (2026) gives an indication of how far this principle can be pushed. Making use of data from a survey on labor market conditions linked to full-population administrative records, they document large participant–nonparticipant differences. Crucially, the differences survive corrections on observables. They then use randomized incentives as instruments for participation, permitting a direct test and correction of the selection bias. Their findings are sobering in the spirit of this study: selection bias was substantial, invisible from the survey data alone, and poorly handled by off-the-shelf corrections. Constructively, however, this second practice is powerful because a modest amount of design-stage randomization can convert an untestable assumption into an estimable equation. Notice also that this practice closes a loop left open in Section 2. What makes the LATE estimand acceptable is exogenous variation with known provenance acting on the choice margin. That is exactly what randomizing elements of Z_i manufactures for the sign-up margin. An unrestricted volunteer sample lacks the LATE structure; the experimenter can build it.

3. Leverage sensitivity reporting. A final reporting practice concerns outcome sensitivity. Reporting how received conclusions in the paper change across a grid of ρ values is meaningful. To operationalize this practice, I would advise using equation (7) and the observed p supplying the mapping. This is the EP2 analogue of Lee (2009) bounds for attrition. In insightful work, Andrews and Oster (2019) provide a useful complementary benchmark to my third reporting practice. They derive a simple approximation for external validity bias in which the degree of selection into the sample explained by observables informs the plausible magnitude of selection on unobserved treatment-effect heterogeneity. Whereas my proposed grid over ρ shows how conclusions vary with different assumptions, their approximation offers a data-driven way to anchor where on that grid a given study is likely to lie. The selection table that I recommended in the first reporting practice supplies precisely the observables their approach requires. That existing corrections often produce wide or misleading bounds when tested against ground truth (Dutz et al., 2026) is not an argument against this practice. Instead, it is an argument for reporting transparency concerning what the volunteer sample can and cannot support in terms of EP2 inference.

6 Conclusion

A landmark shift has occurred within the social sciences over the past several decades. Acceptable empirical work within academe, and increasingly within organizations and the court system, now calls for more credible identification. That movement has been a productive one, as we are more confident in our causal claims now than at any point in the discipline's history. Yet its reach has not extended to a mathematically isomorphic bias: individual selection into the sample itself. From the stark difference in how the two are treated, one would conclude they are species of different genera. Equations (2) and (3) show that they are the same animal wearing different subscripts.

I suspect that the asymmetry in our treatment of the two selection problems reflects both the past focus on EP1 and a history of what referees could verify empirically. This study attempts to move the treatment of EP2 selection in a similar direction by recommending starting at the experimental design stage. Where possible, conduct an NFE, since that is the design-based solution that tackles any bias directly. If an NFE is not feasible or viable to address the research question at hand, then the discussion should take the form of enhanced reporting in the areas of measurement, correction, and bounds. Such a change in reporting of empirical results arrives with low marginal cost, but importantly disciplines the scholar's leap from estimate to scaled policy. In the end, extending the credibility revolution across EP1 and EP2 should be a crucial objective for science.

References

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): 462–466.

Al-Ubaydli, Omar, John A. List, and Dana L. Suskind. 2017. "What Can We Learn from Experiments? Understanding the Threats to the Scalability of Experimental Results." *American Economic Review* 107 (5): 282–286.

Andrews, Isaiah and Emily Oster. 2019. "A Simple Approximation for Evaluating External Validity Bias." *Economics Letters* 178: 58–62.

Angrist, Joshua D. and Jörn-Steffen Pischke. 2010. "The Credibility Revolution in Empirical Economics: How Better Research Design Is Taking the Con out of Econometrics." *Journal of Economic Perspectives* 24 (2): 3–30.

Cleave, Blair L., Nikos Nikiforakis, and Robert Slonim. 2013. "Is There Selection Bias in Laboratory Experiments? The Case of Social and Risk Preferences." *Experimental Economics* 16 (3): 372–382.

Deaton, Angus. 2010. "Instruments, Randomization, and Learning about Development." *Journal of Economic Literature* 48 (2): 424–455.

Dutz, Deniz, Ingrid Huitfeldt, Santiago Lacouture, Magne Mogstad, Alexander Torgovitsky, and Winnie van Dijk. 2026. "Selection in Surveys: Using Randomized Incentives to Detect and Account for Nonresponse Bias." *Review of Economic Studies*, forthcoming.

Falk, Armin, Stephan Meier, and Christian Zehnder. 2013. "Do Lab Experiments Misrepresent Social Preferences? The Case of Self-Selected Student Samples." *Journal of the European Economic Association* 11 (4): 839–852.

Harrison, Glenn W. and John A. List. 2004. "Field Experiments." *Journal of Economic Literature* 42 (4): 1009–1055.

Heckman, James J. 1979. "Sample Selection Bias as a Specification Error." *Econometrica* 47 (1): 153–161.

Heckman, James J. 1992. "Randomization and Social Policy Evaluation." In *Evaluating Welfare and Training Programs*, edited by Charles F. Manski and Irwin Garfinkel, 201–230. Cambridge, MA: Harvard University Press.

Heckman, James J. and Sergio Urzúa. 2010. "Comparing IV with Structural Models: What Simple IV Can and Cannot Identify." *Journal of Econometrics* 156 (1): 27–37.

Lazear, Edward P., Ulrike Malmendier, and Roberto A. Weber. 2012. "Sorting in Experiments with Application to Social Preferences." *American Economic Journal: Applied Economics* 4 (1): 136–163.

Lee, David S. 2009. "Training, Wages, and Sample Selection: Estimating Sharp Bounds on Treatment Effects." *Review of Economic Studies* 76 (3): 1071–1102.

Levitt, Steven D. and John A. List. 2007. "What Do Laboratory Experiments Measuring Social Preferences Reveal About the Real World?" *Journal of Economic Perspectives* 21 (2): 153–174.

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. "Don't Give Up on Lab Experiments: Why the Field Still Needs the Lab." NBER Working Paper 35338.

List, John A. 2026c. "Make Science More Reliable: Study People as They Go About Their Lives." *Nature* 654 (8120): 863–866.

Roy, A. D. 1951. "Some Thoughts on the Distribution of Earnings." *Oxford Economic Papers* 3 (2): 135–146.

Slonim, Robert, Carmen Wang, Ellen Garbarino, and Danielle Merrett. 2013. "Opting-In: Participation Bias in Economic Experiments." *Journal of Economic Behavior & Organization* 90: 43–70.