NBER WORKING PAPER SERIES

THE EXPERIMENTALIST LOOKS WITHIN:
TOWARD AN UNDERSTANDING OF WITHIN-SUBJECT EXPERIMENTAL DESIGNS

John A. List

Working Paper 33456
http://www.nber.org/papers/w33456

NATIONAL BUREAU OF ECONOMIC RESEARCH
1050 Massachusetts Avenue
Cambridge, MA 02138
February 2025

This study was created to honor the memory of Gary Charness. It draws heavily from my forthcoming textbook (List, 2025); in particular, Chapters 3 and 10. Gary's drive for better experimental methods and enhanced understandings was an inspiration to us all. Faith Fatchen and Francesca Pagnotta ably assisted. Any errors are 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.

The Experimentalist Looks Within: Toward an Understanding of Within-Subject Experimental Designs
John A. List
NBER Working Paper No. 33456
February 2025
JEL No. C9, C90, C91, C92, C93, C99

ABSTRACT

The traditional approach in experimental economics is to use a between-subject design: the analyst places each unit in treatment or control simultaneously and recovers treatment effects via differencing conditional expectations. Within-subject designs represent a significant departure from this method, as the same unit is observed in both treatment and control conditions sequentially. While many might consider the choice straightforward (always opt for a between-subject design), given the distinct benefits of within-subject designs, I argue that researchers should meticulously weigh the advantages and disadvantages of each design type. In doing so, I propose a categorization for within-subject designs based on the plausibility of recovering an internally valid estimate. In one instance, which I denote as stealth designs, the analyst should unequivocally choose a within-subject design rather than a between-subject design.

John A. List
Department of Economics
University of Chicago
1126 East 59th
Chicago, IL 60637
and Australian National University
and also NBER
jlist@uchicago.edu

In a landmark experiment conducted in 1747, James Lind, a Scottish naval surgeon, sought to identify a cure for scurvy, a disease that plagued sailors on long voyages. Lind divided twelve sailors suffering from scurvy into six pairs and administered different treatments to each pair, including cider, vinegar, seawater, oranges, lemons, and a medicinal paste. Remarkably, the sailors who received citrus fruits showed significant improvement, while the others did not. This pioneering clinical trial demonstrated the effectiveness of citrus fruits in preventing and treating scurvy, leading to their widespread adoption in the British Navy and the eventual eradication of the disease among sailors. Perhaps taking the lead from Lind, more than a century later, in 1882, Louis Pasteur designated half of a group of 50 sheep as controls and vaccinated the other half. After all the animals received a lethal dose of anthrax, the results were clear and compelling. Two days post-inoculation, every one of the 25 control sheep had succumbed to the disease, whereas the 25 vaccinated sheep were alive and well, vividly demonstrating the power and importance of vaccination (Levitt and List, 2009).

One common thread that connects these classic experiments to the economic experiments of today is the use of a between-subject experimental design: some subjects are placed in treatment while others are simultaneously placed in a control group. After which, treatment effects are measured post treatment by estimating conditional expectations. That is, the analyst measures the average treatment effect (ATE), $\bar{Y}_i(1) - \bar{Y}_i(0)$, by taking the difference between the average outcome in the treatment group and the average outcome in the control group.[2] This means that while we cannot observe the *individual* treatment effects, we can observe the difference between the means in the two groups. Or, likewise, we can observe the difference between two marginal distributions.

As such, if the researcher desires to estimate other informative outcomes, such as median effects or various percentile effects, a different design must be used, or further assumptions must be invoked. One such approach that provides richer information is to conduct a within-subject (WS) experimental design (Charness et al., 2012). Under this approach, each unit receives both the treatment and control but in a sequential fashion, and the treatment effect is identified by comparing the same unit's outcomes across the two conditions. While a straightforward design

[2] The theorem that delivers this result is remarkable because it requires so few assumptions. One key aspect of the theorem, however, is that the key driver is that the mean is a linear operator, thus the difference-in-means is the mean of differences (List, 2025).

approach to implement, to recover an internally valid estimate, the WS design relies on three additional assumptions to those necessary to recover the ATE in a between-subject design.

In this study, to provide a deeper understanding of WS designs, I present an overview of the marginal benefits and costs of WS designs, focusing on the three new assumptions. These new assumptions are balanced panel, temporal stability, and causal transience. A central consideration in recovering an internally valid estimate is whether these three key identification assumptions are reasonably met. The first assumption, balanced panel, requires that all participants remain in the study for all $\mathcal{T}$, and for each we observe their outcome, treatment assignment, actual treatment, and unit-level characteristics. The second assumption, temporal stability, requires that the potential outcome not be a function of time. The third assumption, causal transience, is typically the most challenging because it requires that the effect of the contemporary treatment not depend on the regime. Under causal transience, the effect of treatments *do not* persist over time. When these three assumptions are met, the WS design not only reaps greater power than a between-subject design but also provides insights beyond estimating marginal treatment effects, as the analyst learns about the full joint distribution of treatment effects.[3]

After discussing the three new assumptions, this study explores the advantages and disadvantages of WS designs and provides a playbook for the analyst interested in generating data optimally using a WS design. To make matters concrete, I use three running examples. These examples are useful to help draw out the assumptions underlying WS experimental approaches. In doing so, this study showcases the two key benefits of a WS design but cautions about the potential baggage that comes with certain WS designs. Specifically, there are cases where WS designs stretch the bounds of credulity concerning our identification assumptions on causal transience (denoted as overt WS designs) or temporal stability (denoted as epochal WS designs), yet there are examples where the assumptions are more tenable (denoted as stealth WS designs). I denote such WS designs as "stealth" because they satisfy the new identification assumptions with a "stealth-like" design approach.

[3] Even though WS designs can achieve greater precision, in practice, researchers do not employ them regularly. Surveying two recent volumes of the journal *Experimental Economics*, Bellemare et al. (2014) report that most studies in their survey (41 out of 58) use a between-subject design.

The remainder of this study proceeds as follows. The next Section outlines the basic notation of the potential outcomes model and summarizes the exclusion restrictions. Section III presents the three running examples. In general, but not always, the literature reveals that compared to behavior generated in between-subject designs, subjects in a WS design behave more rationally, behave more in line with neoclassical theory, and tend to conform to social norms more closely when they have a comparative context, or likewise an evaluability baseline. Section IV discusses the various threats to WS designs. Section V summarizes key advantages of WS designs. Section VI concludes.

II. Potential Outcomes Basics

In this section, I introduce the potential outcomes framework to define a causal effect.[4] In doing so, I present a set of exclusion restrictions that are sufficient, in a between-subject design, to interpret the difference in conditional expectations of an outcome between two groups with distinct treatment assignments as a causal effect. In this framework, causality arises from a treatment being assigned to a unit. The definition of a unit is, typically, uncontroversial. A unit can be a plot of land, a firm, a household, a collection of individuals, or a market. For concreteness, I use $i \in \mathcal{I} = \{1,2,\ldots,N\}$ to denote an individual unit. I denote the observed outcome for unit i as Y_i**.** There is no time subscript because, importantly, the same unit observed in different time periods is assumed to be different, or separate, units. I relax this assumption below when I introduce WS experimental designs.

The unit-specific treatment is denoted by D_i, where a realization is $d \in \mathcal{D}$. However, the treatment assignment of all other units (i.e., who gets what) will also be relevant. Let the vector

$$\boldsymbol{D}_{-i} = (D_1, D_2, D_{i-1}, D_{i+1}, \ldots, D_N) \in \mathcal{D}^{N-1}$$

record the treatments of units other than i, but such that order matters, and each element is tied to the treatment of a particular unit. This vector takes realized values $\boldsymbol{d}_{-i} \in \mathcal{D}^{N-1}$. Thus, the complete description of the treatment program for unit i will be $\boldsymbol{D_i} = (D_i, \boldsymbol{D}_{-i}) \in \mathcal{D}^N$, with realizations $\boldsymbol{d} = (d, \boldsymbol{d}_{-i})$. This treatment program can take the form of an intervention, an

[4] The potential outcomes approach is commonly referred to as the "Rubin Causal Model" in the literature, but the framework can be found in Jerzy Neyman's master's thesis ([1923] 1990), which describes "potential yields" when referring to his agricultural outcomes of interest. The interested reader should see Rubin (1974; 1975; 1978).

inducement, a manipulation, actions taken, or decisions made – the key is that, thus far, we allow the precise distribution of treatment across units to matter to each individual unit.

A key piece of intuition in the potential outcomes framework is that for each $\boldsymbol{d}$, there is a random variable $Y_i(\boldsymbol{d})$, denoted as the potential outcome. The potential outcome corresponds to what the outcome *would have* been had the realized state been $\boldsymbol{d}$. Each potential outcome is *hypothetically* observable. However, after the occurrence of treatment $\boldsymbol{d}$, researchers can observe at most only one potential outcome for each unit. The other potential outcomes are unobservable because the treatment that would have led to that potential outcome's realization did not occur. This is the fundamental problem of causal inference—a missing data problem.

An astute reader has already noticed that the above definition of treatment and potential outcomes leaves the door open to an enormous variety of possible unique treatment programs arising through minute changes to the treatment itself. To make progress, I now focus on a more restrictive definition of treatment which takes a strong stance on what a treatment is and whose treatment matters for whom. I do so by introducing the first key assumption, which constrains the definition of treatment and links potential outcomes to observed outcomes.

Assumption 1 (Stable Unit Treatment Value Assumption, or SUTVA): For all $\boldsymbol{d} \in \mathcal{D}^N$, $Y_i(\boldsymbol{d}) = Y_i(d)$.

The Stable Unit Treatment Value Assumption (SUTVA), sometimes called the no-interference assumption, incorporates two distinct assumptions. The first is that any unit's potential outcomes do not vary with the treatments assigned to or undertaken by *any* other unit. This means that for any set of units with a treatment program $\boldsymbol{d}$ we can simply write $Y_i(\boldsymbol{d}) = Y_i(d)$ for each unit. Put another way, under this assumption, we no longer need to speak of treatment programs and can instead refer unambiguously to the effect of a specific unit's treatment.

The second part of Assumption 1 is that there is no variation in the form or version of the treatment that leads to different potential outcomes. When this sense of SUTVA does not hold, $\mathcal{D}$ no longer represents mutually exhaustive states of treatment. For example, if a lab experiment has multiple experimental proctors who vary in quality in a way that affects potential outcomes, then that experiment would run the risk of two potential treatment dimensions: the recipient and the proctor. SUTVA rules out variations of the proctor that affect potential outcomes. Put together, the two parts of SUTVA imply that we can simply speak of $d \in \mathcal{D}$ as a particular treatment of a particular unit.

At this point, it is useful to make one normalization. Specifically, I will refer to the state of the world where $d = 0$ as the untreated (or *control*) condition and the states of the world where $d \neq 0$ as the *treated* conditions. In this spirit, our objective is to learn the causal effect of assignment to state $d \neq 0$ on some observed outcome of interest relative to a fixed alternative where $d = 0$. With SUTVA, we now have a concise way to completely characterize the building blocks of this causal effect. In particular, if we let D_i denote an individual unit's possible treatment, then we have the following equation linking observed outcomes to potential outcomes:

$$Y_i = \sum_{d \in \mathcal{D}} Y_i(d) \mathbb{I}[D_i = d], \tag{1}$$

where $\mathbb{I}[D_i = d]$ is an indicator function for the receipt of treatment level d by unit i. Given equation (1), we can define the individual-level treatment effect as the difference in outcomes for unit i when the individual receives treatment versus control, $\tau_i \equiv Y_i(1) - Y_i(0)$.

Equation (1) also builds in three important features of this model. First, in writing $Y_i(d)$ as a function of the treatment d, we require that the treatment precedes measurement of the outcome. Second, the researcher must be able to clearly define the action that would have made the alternative potential outcome the realized potential outcome. Third, we immediately face the "fundamental problem of causal inference." That is, a causal effect is the comparison of potential outcomes for the same unit in the same moment, but with different treatments. Since we observe only $Y_i = Y_i(d)$ for the realized $D_i = d$, we can learn about causal effects only by observing multiple units and comparing $Y_i(d)$ to $Y_j(0)$ for some $i \neq j$ and $d \neq 0$.

Ideally, $Y_j(0)$ is equivalent to the outcome of someone treated, $Y_j(1)$, had we withheld their treatment. We cannot observe this counterfactual as units exposed to $D_i = d$ are not the identical units exposed to $D_i = 0$. Thus, the central task of empirical research is to find reasonable approximations to the relevant counterfactual potential outcome. Doing so requires a series of *exclusion restrictions*. SUTVA is the first of four exclusion restrictions necessary to identify a causal effect within the potential outcomes framework of a between-subject design.

To explore the other three, I begin by introducing the assignment mechanism. Let $Z_i \in \{0,1\}$ be the assignment mechanism that allocates treatment to a particular unit. When this treatment assignment corresponds to the undertaken treatment D_i we say that $D_i = Z_i$. Absent this assumption, we can interpret the received results as the causal effect of the assignment mechanism, Z, rather than how the actual treatment, D, affected outcomes. This is important because in most cases we want to know the effect of the actual treatment not the effect of treatment assignment.

Let $\boldsymbol{X}_i$ represent a vector of characteristics that the researcher measures prior to treatment assignment (that is, predetermined characteristics, or covariates).[5] Finally, let R_i denote the post-treatment decision to remain in the study, that is not to attrit the study, where $R_i = 1$ indicates units that remain while $R_i = 0$ indicates units that attrit the study. Importantly, Y_i will be observed only for those with $R_i = 1$. With this notation in hand, we arrive at the next assumption needed to recover the ATE in a between-subject design.

Assumption 2 (Observability): For all units $i \in \mathcal{I}$, we have $\mathbb{P}[R_i = 1] = 1$, and for each unit $i \in \mathcal{I}$ the researcher observes $(Y_i, D_i, Z_i, \boldsymbol{X}_i)$.

In other words, Assumption 2 requires that all participants originally in the study remain in the study, and for each, we observe their outcome, treatment assignment, actual treatment, and unit-level characteristics.[6] In what follows, we will maintain $\mathbb{P}[R_i = 1] = 1$ (unless we explicitly state otherwise), and so omit R_i.

If we observed the individual treatment effect, $Y_i(1) - Y_i(0)$, directly then we could simply calculate the individual treatment effect for those with observable data and through that comparison understand the causal effect of the treatment on those units. However, the necessity of observing multiple units requires that we observe information on all units in our sample. We require this information to calculate the average values of $Y_i(1)$ and $Y_i(0)$. If the treatment assignment or outcome is missing due to a decision by the units, then there is an additional concern that the decision to have a measurable outcome is a potential outcome itself. In this case, equation

[5] When units are people, typical examples of such predetermined characteristics include race, sex, gender, or age at the start of the study.

[6] It is not essential to include $\boldsymbol{X}_i$ in this assumption, but because many studies explore treatment effect heterogeneity across different values of $\boldsymbol{X}_i$, I include it here for completeness.

(1) does not represent mutually exhaustive states of the world and there are unmodeled factors that impact the subject's potential outcomes.

In addition to considering observability, we also must make an assumption on treatment compliance. This leads to our next assumption needed to recover the ATE.

Assumption 3 (Complete Compliance): For all units $i \in \mathcal{I}$, we have $\mathbb{P}[D_i = Z_i] = 1$.

This assumption states that every unit assigned to $Z_i = z$ ends up taking $D_i = z$. When this assumption is satisfied, we can condition on D_i alone, rather than on both D_i and Z_i. Thus, when invoking this assumption, we can omit Z_i, focusing solely on D_i. In general, there are two types of violations to this assumption.

To highlight these violations, we maintain that $\mathcal{D} = \mathcal{Z} = \{0,1\}$. The first violation is that some units assigned to the control group ($Z_i = 0$) acquire the treatment by other means ($D_i = 1$). The second violation is that some units assigned to the treatment group ($Z_i = 1$) end up not taking the treatment ($D_i = 0$). These types of non-compliance are referred to as one-sided non-compliance. When both violations occur together (i.e., both directions of one-sided non-compliance), the literature denotes this as two-sided non-compliance. We can use the complete compliance assumption to write the conditional expectations of the outcome variable in either treatment group as purely a function of the assignment rather than a function of both assignment and take-up. As mentioned above, when this assumption is violated, researchers commonly move to study the assignment mechanism's causal effect rather than the treatment's causal effect (i.e., the effect of Z on Y rather than D on Y). While in that case we can make valid causal inference about Z, such causal effects are typically not of primary interest.

When Assumptions 1, 2, and 3 hold, we can again write the observed outcome as a function of the treatments. Because we have only two levels of D_i, equation (1) can be expressed as

$$Y_i = D_i Y_i(1) + (1 - D_i) Y_i(0). \tag{2}$$

Equation (2) allows us to express the conditional expectations of the observed outcome as a function purely of potential outcomes for those who participate in the experiment. That is, we can express the average value of the outcome in our treatment group as $\mathbb{E}[Y_i(1)|D_i = 1] = \mathbb{E}[Y_i|D_i = 1]$ and the average value of our control group as $\mathbb{E}[Y_i(0)|D_i = 0] = \mathbb{E}[Y_i|D_i = 0]$. With this in mind, we can express the differences in conditional expectations between those in the treatment group and those in the control group as

$$\tilde{\tau} \equiv \mathbb{E}[Y_i(1)|D_i = 1] - \mathbb{E}[Y_i(0)|D_i = 0] = \mathbb{E}[Y_i|D_i = 1] - \mathbb{E}[Y_i|D_i = 0]. \tag{3}$$

We define a study as having internal validity if this observed difference, $\tilde{\tau}$, is equal to the average causal effect of D_i on Y_i, defined as:

$$\tau \equiv \mathbb{E}[Y_i(1) - Y_i(0)] \tag{4}$$

Through the lens of our framework, $\tilde{\tau}$ is internally valid for τ if the observed difference in conditional expectations between treatment and control groups is equal to the average causal effect of D_i on Y_i. To better understand the relationship between $\tilde{\tau}$ and τ, there is a particularly illuminating representation we can arrive at with a bit of manipulation. To see this, it will be useful to introduce one more piece of notation: $\mathbb{P}[D_i = 1]$, the fraction of the population that is treated.

Just as a property of expectations we know that we can split the ATE (τ) as follows:

$$\begin{aligned}
&\tau \equiv \mathbb{E}[Y_i(1) - Y_i(0)] \\
&= \mathbb{P}[D_i = 1]\, \underbrace{\mathbb{E}[Y_i(1) - Y_i(0)|D_i = 1]}_{ATE\ on\ the\ treated\ (ATT)} + (1 - \mathbb{P}[D_i = 1])\, \underbrace{\mathbb{E}[Y_i(1) - Y_i(0)|D_i = 0]}_{ATE\ on\ the\ untreated\ (ATU)} \\
&= \mathbb{P}[D_i = 1]\{\mathbb{E}[Y_i(1)|D_i = 1] - \mathbb{E}[Y_i(0)|D_i = 1]\} \\
&\qquad + (1 - \mathbb{P}[D_i = 1])\{\mathbb{E}[Y_i(1)|D_i = 0] - \mathbb{E}[Y_i(0)|D_i = 0]\}
\end{aligned}$$

Now, we can consider the components driving a wedge between τ and $\tilde{\tau}$. With a bit of algebraic manipulation, we find that

$$\begin{aligned}
\tilde{\tau} - \tau = (1 - \mathbb{P}[D_i = 1]) &\underbrace{\{ATT - ATU\}}_{differential\ effects\ by\ treatment\ choice} \\
&+ \underbrace{(\mathbb{E}[Y_i(0)|D_i = 1] - \mathbb{E}[Y_i(0)|D_i = 0])}_{selection\ bias}
\end{aligned} \tag{5}$$

Equation (5) reveals that generally, there is a wedge between τ and $\tilde{\tau}$, meaning that the difference in outcomes between the treated and control groups does not recover the average causal effect of treatment. The wedge comprises the sum of two terms. The first term reflects the potentially differential average effect of the treatment for those who choose treatment (ATT) and those who do not (ATU). The second term is the selection bias term, $\mathbb{E}[Y_i(0)|D_i = 1] - \mathbb{E}[Y_i(0)|D_i = 0]$. This term reflects differences in *untreated* outcomes between those who undertook treatment ($D_i = 1$) and those who undertook the control condition ($D_i = 0$) that would have existed absent treatment. That is, selection bias of this kind is entirely the result of outcome differences in the untreated state.

Since the choice to undertake treatment may result from an underlying optimization or decision problem, there is usually no compelling reason to assume that such units would have the same outcomes absent treatment. In naturally occurring data, the assumption that the selection bias term is zero might stretch the bounds of credulity. Understanding the difference between τ and $\tilde{\tau}$ requires making assumptions about the assignment mechanism, Z, that is, making assumptions about how each unit came to receive its realized treatment. With the maintained Assumptions 1-3 in place, a sufficient condition for $\tilde{\tau}$ to be an internally valid estimate of τ is statistical independence.

Assumption 4 (Statistical Independence): $\{Y_i(1), Y_i(0)\} \perp D_i$.

Statistical independence means that the assignment mechanism governing the D_i is independent of potential outcomes. Under this assumption, $\mathbb{E}[Y_i(0)|D_i = 1] = \mathbb{E}[Y_i(0)|D_i = 0]$, so that the selection bias is zero due to randomization. Randomization acts as an instrumental variable in this case. Moreover, the ATE on the treated among participants (ATT) is equal to the ATE among participants:

$$\begin{aligned} \mathbb{E}[Y_i(1) - Y_i(0)|D_i = 1] &= \mathbb{E}[Y_i(1) - Y_i(0)|D_i = 0] \\ &= \mathbb{E}[Y_i(1) - Y_i(0)] \end{aligned} \tag{6}$$

And similarly, the ATU is also equal to the ATE. Thus, ATE = ATT = ATU. Accordingly, the wedge between τ and $\tilde{\tau}$ is zero, and $\tilde{\tau}$ is internally valid and recovers τ. Crucially, random assignment to treatment automatically implies statistical independence as required by Assumption 4, and hence solves the issue of selection bias.

Within-Subject Design Assumptions

Assumptions 1-4 represent the exclusion restrictions necessary to recover the ATE from a between-subject experimental design. Moving to a WS design demands us to consider data over multiple periods. Now, we consider subjects $i \in \{1, \dots, N\}$ with the outcome, Y_{it}, measured in periods $t \in \mathcal{T} = \{0, \dots, T\}$. Every individual receives a treatment regime $D_i = (D_{i1}, \dots, D_{iT})$ where D_{it} represents unit $i's$ treatment assignment in period t. The experimenter observes Y_{it} in each period after the administration of D_{it}. The potential outcome for unit i in period t now depends on the contemporary treatment D_{it}, the treatment regiment $\boldsymbol{D}_i$, along with all time-specific effects, t, and can be expressed as $Y_{it}(D_{it}, \boldsymbol{D}_{i,-t}, T)$ where $\boldsymbol{D}_{i,-t}$ denotes the vector of treatment statuses in all periods other than t.

Several potential treatment effect parameters can be identified using a WS design depending on a willingness to make certain assumptions. With only assumptions 1-4, the researcher can identify the treatment regime's causal effect on the outcome in each of the experimental periods because this corresponds to a between-subject comparison of $\boldsymbol{D}_i$ and $\boldsymbol{D'}_i$. Yet, a keen advantage that arises from the fact that the experimenter can observe both Y_{i1} and Y_{i0} for a single unit is that it permits them to identify the individual treatment effect. The subsequent richness in the outcome space provides information beyond estimating marginal distributions; in WS designs, the full joint distribution of outcomes is recoverable. Note that we can write $\mathrm{Var}[\tau_i]$ as a function of potential outcomes:

$$\begin{aligned}\mathrm{Var}[\tau_i] &= \mathrm{Var}[Y_i(1) - Y_i(0)] \\ &= \mathrm{Var}[Y_i(1)] + \mathrm{Var}[Y_i(0)] - 2\mathrm{Cov}[Y_i(1), Y_i(0)]\end{aligned} \tag{7}$$

A between-subject experiment provides information on the marginal distribution of potential outcomes, providing the necessary ingredients to recover $\mathrm{Var}[Y_i(D_i)]$. However, the fundamental problem of causal inference prevents us from obtaining information about the joint distribution of potential outcomes in a between-subject design: $\mathrm{Cov}[Y_i(1), Y_i(0)]$ is unobserved in such experiments. Yet, a WS design generates such information naturally. Within the experimental economics community, for example, Isaac and Walker (1988a; 1988b) explore how to measure behavioral response to a change in stakes in a WS design.

Yet, to remain internally valid, the WS design demands three further assumptions. A first new assumption is the balanced panel assumption.

Assumption 5 (Balanced Panel): For all $(i, t) \in \mathcal{I} \times \mathcal{T}$, we have $\mathbb{P}[R_{it} = 1] = 1$, and for all $(i, t) \in \mathcal{I} \times \mathcal{T}$ with $\mathbb{P}[R_{it} = 1] = 1$ the researcher observes $(Y_{it}, D_{it}, Z_{it}, \boldsymbol{X}_{it})$.

Assumption 5 requires that all participants remain in the study for all $\mathcal{T}$ periods and for each we observe their outcome, treatment assignment, actual treatment, and unit-level characteristics. This is stronger than Assumption 2 (observability) because we require it for each period rather than the single period used in a between-subject design.

The second assumption we must make is the temporal stability assumption.

Assumption 6 (Temporal Stability): For all $t \in \mathcal{T}, Y_{it}(D_{it}, \boldsymbol{D}_{i,-t}, t) = Y_{it}\left(D_{it}, \boldsymbol{D}_{i,-t}\right)$.

Temporal stability rules out any time-varying effects that affect potential outcomes.[7] The main threats to temporal stability are discussed below. Yet, I will note that this assumption is more likely to hold when the time between measurements is short (on the order of minutes). Our final new assumption we must add to arrive at an internally valid WS parameter is causal transience.

Assumption 7 (Causal Transience): For all $\boldsymbol{D}_i,\ Y_{it}(D_{it}, \boldsymbol{D}_i, T) = Y_{it}(D_{it}, T)$.

The causal transience assumption states that the effect of the contemporary treatment does not depend on the regime.[8] When causal transience is satisfied, the effect of each treatment does not persist over time. This assumption implies that the treatment effects, or outcomes, do not depend on the order in which they are implemented. This assumption assumes that future treatments do not influence current behavior and no "carryover" effects from previous treatments exist. That is, future outcomes are orthogonal to all past treatments and outcomes. Experimenter demand effects represent a key factor that can cause carryover effects that lead to a violation of Assumption 7 (Rosenthal, 1976; Levitt and List, 2007).

III. Three Running Examples

To make the above matters concrete, I use three running examples. These examples also help to draw out certain marginal benefits and marginal costs that the analyst should consider when deciding on conducting a between-subject or WS design. The first experimental example comes from the laboratory, a clever study on deception due to Gneezy (2005). This example is

[7] When there are time-varying factors that influence potential outcomes, one can still estimate an ATE when these factors are constant across individuals and do not moderate the treatment effects by controlling for time-fixed effects. However, the presence or absence of time-fixed effects are not relevant for the between-subjects design.

[8] In the nonexperimental literature, this assumption is often denoted as "impersistent outcomes." The interested reader should see Hull (2018). In parts of the experimental literature, this assumption is stated as the no-anticipation assumption: $Y_{i1}(D_{i1}, D_{i,-t}, t) = Y_{i1}(D_{i1}, t)$. The no-anticipation assumption states that in period 1, treatments implemented in periods $t > 1$ do not affect potential outcomes in period $t = 1$. Under this assumption, only the contemporary treatment matters in period 1, and we can identify the treatment effect in the standard way,

$$\tau_1 \equiv \mathbb{E}\left[Y_{i1}\left(1, \boldsymbol{D}_{i,-t}, 1\right) | D_{i1} = 1, \boldsymbol{D}_{i,-t}\right] - \mathbb{E}\left[Y_{i1}\left(0, \boldsymbol{D'}_{i,-t}, 1\ \right) | D_{i1} = 0, \boldsymbol{D'}_{i,-t}\right] = \mathbb{E}[Y_{i1}(1,1) - Y_{i1}(0,1)].$$

τ_1 is the same treatment effect one estimates with a between-subject design. Since Assumption 7 subsumes the no-anticipation assumption, I do not include it as a formal assumption.

pedagogically viable because for our purposes Gneezy (2005), by design, provides a test of the most problematic WS assumption: causal transience. Gneezy leveraged a WS design wherein experimental participants judged scenarios in which a car salesman lied about the car's condition. The cost of the lie was randomized across treatments. In one condition, the buyer's repairs cost was $250, and in the other, it was $1000.

Fifty subjects were assigned to the WS design, in which they were presented with both scenarios whereas 50 different subjects were placed in a between-subject design and were presented only one of the conditions. Gneezy found that while participants generally consider the cost of the lie when rating its fairness, he observed dramatic differences in the results across the WS and between-subject designs. In the between-subject design, 36% of the subjects called the lie very unfair in the low-cost scenario, in contrast to only 18% of subjects in the WS design. This pattern of data is consonant with a violation of causal transience. For this reason, I refer to the Gneezy (2005) study as an overt within-subject design.

The second experimental example is a study that used a natural field experiment to explore the nature and extent of discrimination in the sports card market (List, 2004). By having confederates approach and purchase various pieces of memorabilia from unsuspecting dealers, List leveraged the key features of a WS design by having the same dealer bargain with women and men of different ages and races. I will focus on observed gender differences in this study.

List's WS design allowed exploration of aspects of the entire treatment distribution, such as what fraction of dealers discriminated, how much the 10th percentile dealer discriminated, and how much discrimination was observed amongst the most discriminating dealers. The experimental design, which included data gathered from more than 1100 market participants, provided findings that suggest there was a strong tendency for women to receive initial and final offers that were inferior to those received by men. The observed discrimination was not due to animus but represented statistical discrimination. I refer to List's (2004) design as a stealth within-subject design because the three new identification assumptions in this study were reasonably met. This is in part due to the stealth-like features of the NFE to send buyers to dealers over a few-hour period in a manner that is not different from what would happen in the normal course of business.

Lastly, in the third example, Rodemeier (2023) used a novel NFE to estimate willingness to pay (WTP) for carbon mitigation. His NFE leveraged a grocery and beverage delivery service website.

When a subject visited the website, they were randomized into one of several groups in a between-subject design. The treatment groups were offered a chance to compensate carbon emissions by buying a carbon offset either at market price, with a reduced price (50% or 75%), or with a quantity increase (100% or 300%). Empirical results showed that subjects increased demand for mitigation when prices fell, but not when the quantity increased. Importantly, the NFE revealed that subjects were willing to pay roughly 16 EUR/tCO2 for carbon offsets.

Rodemeier (2023) then conducted an interesting follow-up survey that was sent to the same subjects 10-11 months after the NFE was completed, providing a key WS design feature. He found a considerable difference between results in the NFE and those in the hypothetical valuation survey months later: The average WTP in the survey was 200 EUR/tCO2, roughly 1,150% above the revealed preference estimate from the NFE. I view Rodemeier's (2023) WS design as a useful mixture of a stealth/overt approach (which I denote as an epochal within-subject design) in that his subjects were most likely unaware of the connection between his NFE and the survey because of the washout period, but the balanced panel and temporal stability assumptions are called into question because of this design choice. This element highlights a key trade-off faced by many experimenters conducting a WS design.

The Three Running Examples and the Within-Subject Design Assumptions

Even though for each of the three running examples I highlight the WS aspect of their designs, a key feature is that they all contain a between-subject nature that allows them to identify the ATE. For example, consider a subset of data from List's (2004) stealth WS design, as in Panel A of Exhibit 1. Panel A reveals the data generated from the first six dealers approached, three by men (denoted $Y_i(0)$) and three by women (denoted $Y_i(1)$). Dealer #1 was approached by a female buyer first, and he offered her a price of $125 for the good. Dealer #2 was approached by a male buyer first, and he offered him a price of $95 for the good. In this case, the marginal distributions are recovered, and an ATE of $1.67 can be computed if we generated data via a between-subject design, as shown in the last row of Panel A.

Yet, a keen advantage that arises from the fact that the experimenter can observe both Y_{i1} and Y_{i0} for a single unit is that it permits them to identify the individual treatment effect. To explore this point further, let us consider Panel B in Exhibit 1. This panel completes the missing information in Panel A by providing the data generated from List's WS stealth design. Whereas dealer #1 offered the female buyer $125, when he was later approached by a male buyer, he offered the good to him for $105. Dealer #1, therefore, had an individual treatment effect ($Y_i(1) - Y_i(0)$) equal to $20. Panel B shows that in these WS data, we can quickly compute what fraction (50%), and which dealers (dealers #1, #2, and #5), gave women a higher price offer than men, as well as the relevant price differences. Alternatively, Panel A in Exhibit 1 reveals the marginal distributions that are recovered from a between-subject design, and an ATE of $1.67 is computed. Had this been List's chosen design, his result would remain directionally in favor of discrimination against women. Still, the informational content is considerably lower and very different from the generated data by the WS design contained in Panel B of Exhibit 1.

Exhibit 1: ATEs Using Between- and WS Designs

Panel A: Between-Dealer Data

	$\boldsymbol{Y_i(0)}$	$\boldsymbol{Y_i(1)}$	$\boldsymbol{\tau_i}$
Dealer #1	?	$125	?
Dealer #2	$95	?	?
Dealer #3	?	$115	?
Dealer #4	?	$100	?
Dealer #5	$110	?	?
Dealer #6	$130	?	?
Average	$111.67	$113.13	$1.67

Panel B: Within-Dealer Data

	$\boldsymbol{Y_i(0)}$	$\boldsymbol{Y_i(1)}$	$\boldsymbol{\tau_i}$
Dealer #1	$105	$125	$20
Dealer #2	$95	$120	$25
Dealer #3	$120	$115	-$5
Dealer #4	$100	$100	0
Dealer #5	$110	$130	$20
Dealer #6	$130	$130	0
Average	$110	$120	$10

The table highlights the different treatment effects that can be recovered from a between- versus a within-subject (WS) design using data from List (2004). In this setting, $Y_i(0)$ denotes a male while, $Y_i(1)$ a female customer. Panel A focuses on the between-subject design where we observed only the first-period data – for each dealer we have only the price they offered to the first customer that approached them (e.g., dealer #1 offered a female customer $125). Thus, we can recover only the marginal distributions and, as shown in the last row, the ATE ($1.67). Conversely, as shown in panel B, with a WS design we observed data over multiple periods – for each dealer we have the price they offered to both buyers that approached them (e.g., dealer #1 offered the female [male] buyer $125 [$105]). Thus, on top of the ATE, as show in the last column, we can recover the individual (per dealer) treatment effect as the difference $Y_i(1) - Y_i(0)$.

While gaining insights on the entire distribution of potential outcomes is attractive, there are inherent trade-offs when employing a WS design. Consider the overt design of Gneezy (2005) and the designs of List (2004) and Rodemeier (2023). When exposing subjects to both treatment conditions, researchers introduce several complications that are not present in between-subject designs. The first complication arises because the treatment conditions occur at different times. This separation across time means that (1) there is an extra burden of data collection whereby attrition might become a concern and (2) treatment status will not be independent of potential outcomes in the presence of any time-varying factors that influence outcomes.

For example, suppose that individuals in Rodemeier (2023) were difficult to contact 10-11 months after his original NFE. And, even if they were successfully contacted, perhaps they were not interested in filling out a survey. Or perhaps there was an informational shock pertaining to the threats of climate change among his sampled population during the 10–11-month period. Then, the comparison between the two treatments would reflect the effect of the changing treatment and the changing nature of the climate threat due to the informational shock. This can be accounted for via a WS design that changes order to measure, and control for, such temporal effects but in the case of Rodemeier (2023) the NFE was done first, and the survey was completed 10-11 months later. Such a design choice is a traditional one in the literature.

A second issue is a more difficult problem in many practical settings: Treatments may have more than a transient impact. Otherwise stated, the effect of each treatment may not be temporary or occur upon administration. This issue is most severe for overt WS designs, as the subject is aware of treatment application in such instances. In contrast, in a stealth WS design, by definition, the treatments themselves are not salient (in List's case, dealers were approached by hundreds, if not thousands, of patrons per day, so bargaining with one extra 30-year-old man or woman who might have approached them in the market naturally was not noticeable). Charness et al. (2012) provide an excellent discussion of such issues drawing from the evaluability literature (see also Grice, 1966; Poulton, 1973; Greenwald, 1976; Hsee, 1996; Frederick and Fischhoff, 1998; and List, 2002)

Let us dig into Assumptions 5-7 more patiently considering the running examples. Assumption 5 requires that all participants remain in the study for all $\mathcal{T}$ periods and for each we observe their outcome, treatment assignment, actual treatment, and unit-level characteristics. This is a tenable assumption for laboratory experiments, such as Gneezy (2005), yet as discussed more fully below, this assumption might present greater difficulties for the two NFEs that we consider—measuring discrimination amongst dealers over the day and willingness to pay for carbon offsets 10-11 months after the first treatment was imposed.

The second assumption is temporal stability. For Gneezy (2005), this assumption rules out factors such as fatigue affecting outcomes in late periods. This assumption is more likely to hold when the time between measurements is short (on the order of minutes). Again, this assumption might present greater difficulties for the two NFEs we consider. The final assumption is causal transience. For the running examples, suppose subjects in Gneezy's lab experiment were unaware that their treatment would change in the second period. In that case, there is no reason to expect that they would change their behavior in the first period. Likewise, in List (2004), given that he was conducting an NFE and randomly allocated confederates to dealers, he could test whether this assumption was met (it was). Finally, since Rodemeier (2023) deployed his survey 10-11 months after his NFE there was little concern about violation of this assumption. Researchers can increase the plausibility of the no-anticipation assumption by concealing the information about treatments until they are implemented or by conducting a stealth WS design (whether it is an NFE or other experiment type). This is generally achieved by not revealing what remains in the experiment until the experiment reaches that point. The main threats to causal transience running in the other direction, no carryover effects, are discussed in the next section.

IV. Threats to the Internal Validity of Within-Subject Designs

Understanding threats to internal validity is crucial for causal inference, and for WS designs we must be keenly aware of threats to our three new identification assumptions. In this section, I discuss in greater detail a few of the most important threats using the three running examples.

Threats to Balanced Panel

An oft-cited advantage of WS designs is that the researcher gathers twice the amount of data per subject compared to a between-subjects design. In this manner, economies of scale can provide a tangible cost advantage. Yet, a potential trade-off is that Assumption 2 (observability) is violated. In a WS design, the added temporal component places an additional strain on identification. In terms of the three running examples, Gneezy's lab experiment had little threat of violating the balanced panel assumption and provided List's (2004) dealers did not leave the market early or sell all the goods over a few-hour period, his design satisfied this assumption as well. Yet, since Rodemeier (2023) deployed his survey 10-11 months after his NFE there was a potential attrition concern. Since for privacy reasons he was unable to match survey respondents with choices in his NFE, this potential threat is an open issue. One approach to minimize attrition in such designs is to backload incentives.

Threats to Temporal Stability

Under temporal stability, a unit's potential outcome does not depend on the period in which the experimenter measures the outcome. The literature has enumerated various threats to temporal stability and here I focus on three key threats: (1) time-specific shocks to the outcome, (2) time trends such as learning or regression to the mean, and (3) the possibility that measurement outcome might change over time.[9] While the first two threats are well documented in the economics literature, the third threat is new and deserves some explanation. By way of example, consider educational interventions that use WS design to explore children's human capital. An assessment at three years of age is not appropriate for children who are four years old. If a WS design measures students' outcomes over these periods, then differences in the test might lead to differences in implications about treatment effects. Thus, how the measurement outcome changes over time is an important consideration.

For Gneezy (2005), these three key threats likely posed few issues. In general, this is true because the assumption is more likely to hold when the time between measurements is short (on the order of minutes). This seems plausible in most lab experiments. For example, whether Gneezy ran his treatments on Tuesday or Wednesday should conceivably not have mattered. However, if one of the treatments was run on one day, and the second treatment on another, any time-varying factors, such as the day of the week, would then be correlated with treatment and bias estimates of the treatment effect. Another exception is time trends, such as learning or natural regression to the mean. Experiments that measure outcomes that are subject to such features, such as a productivity experiment, for example, might find differences depending on whether it is measured before or

[9] See Campbell and Stanley (1966) and Biesanz and Kwok (2003).

after a lunch break or after learning-by-doing on the job occurred. In that case, we would expect workers' productivity in the second period to be higher than in the first, absent any changes in incentives. Alternatively, workers may become fatigued, making their productivity lower in the second period absent any changes in incentives.

Such threats potentially present themselves more forcefully in List (2004) and Rodemeier (2023). An intermediate case is List (2004); what is required in his NFE is that, during the few hours between when the dealers were approached by a man or a woman, there was not a shock in demand or supply that affected the nature and extent of measured discrimination. He carefully chose items to ensure that demand and supply considerations would not exert undue influence for this very reason.

Finally, perhaps the most demanding example is Rodemeier (2023), who needed the valuation of carbon offsets not to undergo a substantial change over the 10–11-month period for Assumption 6 to be satisfied. Importantly, there can be no time-specific shocks to the outcome or time trends, such as learning, that influence potential outcomes. Whereas Gneezy (2005) and List (2004) could explore whether such assumptions held in their data because they generated different random orders of the treatment conditions, Rodemeier (2023) has revealed preference data only in period 1 and only hypothetical data in period 2 (10-11 months later). This highlights a useful demarcation of WS designs and points toward an optimal design approach.

One approach to control for time effects is to generate different random orders of the treatment conditions. For example, for Gneezy (2005), one order might assign low-cost scenario in period 1 and then the high-cost scenario in period 2. Another might be the opposite. Participants are then randomly assigned to one of the orders. Of course, the experimenter can do this with more than two treatments. With N_D treatment conditions, there are $N_D!$ different unique orders. When multiple orders appear in the same period, one can use time-fixed effects to net out time from the treatment effect. Ideally, all treatment conditions will appear in each of the periods. However, this is not always feasible with many treatments.

The optimal assignment of orders should follow the Latin square design when possible. In a Latin square, each treatment appears once on each trial order and in each order across trials. For example, when $D_{it} = \{0,1,2,3\}$, one potential Latin square design is shown in Exhibit 2. Gneezy (2005) and

List (2004) used an approach like a Latin square design. Such an approach is taken to ensure that there is balance of treatments in the lab or throughout the day.

Exhibit 2: An Example of a Latin Squares Design

	Period 1	Period 2	Period 3	Period 4
Order 1	$D_{it} = 3$	$D_{it} = 1$	$D_{it} = 0$	$D_{it} = 2$
Order 2	$D_{it} = 2$	$D_{it} = 0$	$D_{it} = 1$	$D_{it} = 3$
Order 3	$D_{it} = 1$	$D_{it} = 2$	$D_{it} = 3$	$D_{it} = 0$
Order 4	$D_{it} = 0$	$D_{it} = 3$	$D_{it} = 2$	$D_{it} = 1$

The table captures one potential Latin square design when we have four treatments $D_{it} = \{0,1,2,3\}$. The key take away is that in a Latin square design each treatment appears once on each trial order and in each order across trials.

In Exhibit 2, the balance of treatment conditions implies that many types of violation of the temporal stability and causal transience assumptions will have small effects on the experiment's internal validity. However, it will not always ensure the internal validity of the results. For example, when learning-by-doing occurs over time rather than over output, the learning confound will no longer bias the treatment effect. Conversely, causal transience might be violated when learning-by-doing occurs over output. The causal effect of treatment will still be biased.

Latin square designs also permit researchers to test certain implications of our assumptions. Under causal transience and temporal stability or no-anticipation, the treatment effect should not depend on when the researcher observes the outcome. Thus, researchers can test whether treatment effects vary with time or with the order. The assumptions are more credible when these effects do not vary with these factors. In List's (2004) NFE, where he leveraged a stealth WS design, for example, the data suggested few temporal aspects of discrimination.

Threats to Causal Transience

The main threats to causal transience are anticipation and carryover effects. Anticipation effects occur when subjects change their behavior because of treatments that will happen in the future. This identification requirement usually represents the most troubling assumption in WS designs. In some cases, using an NFE can make it more likely that causal transience holds. Consider, for example, the NFE of List (2004). This design is stealth because the dealers making the offer decision did not know they were part of an experiment. Therefore, the gender of one confederate should not have impacted the decision the dealer made about the other confederates. If instead, the dealer had known that they would be confronted with another confederate later in the day of a different gender, they might have changed their behavior in a way that they believe would maximize profits.

Causal transience is potentially vexing across many settings. For example, Uber drivers might respond to anticipated wage increases in the future. The no-anticipation assumption rules out this possibility and allows for an unbiased estimate of the treatment effect in the first period. This "have your cake and it too" design is attractive because the first period from a between-subject design provides information even if the second period in a WS design is compromised. Of course, if the analyst knows the WS design's assumptions will not be met, there remains an opportunity cost and resource cost of executing the second period treatment in such cases.

Carryover effects occur when the impact of treatment persists through the measurement. Consider Gneezy (2005). If his sole purpose was to estimate the treatment effect of the size of lie—the cost of the repairs for the buyer is $250 versus $1000—over his 50 subjects in his overt WS design, then he must assume that carryover effects are zero. His results suggested otherwise, which was the basis of his contribution (to show that the assumption was not met). Unlike Gneezy (2005), List's (2004) inference relied on the zero-carryover effect assumption to hold, and his data suggested it did (no observed effect of time or sequence in dealer offers, which makes sense given his design). Likewise, proper inference about how WTP for carbon offsets changes across the NFE and survey relies on this assumption to be met in Rodemeier (2023), and that is why he spaced these choices in time (by 10-11 months).

Another form of carryover effect is sensitization. Sensitization occurs when subjects discriminate between differences in treatments and, therefore, may be more responsive than when they are exposed to only one treatment. For example, when subjects are inattentive to some factor that varies across treatments, researchers increase the salience of that factor by making it the only feature that changes across periods. Subjects may be more attuned to the differences caused by that variable than they would be naturally, or in a between-subject design. Gneezy (2005) found evidence consistent with subjects being sensitive to the size of the lie. Sensitization may also increase the threat of experimenter demand effects. Economic experiments that estimate price elasticities are particularly susceptible to such effects, for example, as within-person elasticity estimates nearly always are larger than between-person elasticity estimates.

Researchers may detect carryover effects if they counterbalance treatment regimes and compare the effects of a given treatment by position. If effects are asymmetric, carryover effects may be confounding the results. One way to reduce the size of carryover effects is to use "washout periods," as in List (2004) and Rodemeier (2023). This approach increases the amount of time between treatments. However, the benefits of washout periods are context specific and subject to trade-offs. For example, washout periods are likely to remove carryover and fatigue effects but come at the expense of potentially introducing violations of temporal stability.

The logic behind washout periods is that if the treatment effects of interventions fade over time, then the experimenter can recover an estimate of the treatment effect absent carryover effects. Washout periods are common in pharmaceutical trials. In these settings, the length of washout periods is usually determined using some function of the half-life of a pharmaceutical product as measured by the drug's concentration in the blood. Yet, it is difficult to find measurements that are analogous to the half-life that one can use to determine the proper length of time in an economic experiment. Instead, the proper length of washout periods is context specific. For example, in List (2004) several hours were used. In Rodemeier (2023), 10-11 months were used. In Gneezy (2005), several minutes separated treatments, clearly not long enough to be a useful washout period, which was by design. A researcher may determine the proper washout period in pilots by varying the length of time between treatments and testing for symmetry across different treatment regimens.

V. Key Advantages of Within-Subject Designs

When the 7 key assumptions are met, we can observe $Y_{it}(1)$ when we expose unit i to $D_i = 1$ and $Y_{it}(0)$ when we expose unit i to $D_i = 0.$ Because we can observe both potential outcomes for the same unit, we can estimate the original parameter of interest, $\tau_i \equiv Y_{it}(1) - Y_{it}(0).$ Of course, from τ_i, we can estimate the ATE, which was the objective of the between-subject design, but we can also recover the full distribution of treatment effects. This distribution can help to identify whether the average masks important heterogeneities. This is one key scientific advantage of WS designs. A complementary advantage is the gain in power from leveraging a WS design. I discuss both in turn in this section.

Heterogeneity and the Full Distribution of Treatment Effects

Recall from equation (7) that treatment effect heterogeneity refers to positive variance in the individual treatment effect. This means that researchers can easily calculate conditional average treatment effects (CATEs), the ATE conditional on some predetermined covariates. Under a WS design satisfying all the necessary assumptions, $\text{Cov}[Y_i(1), Y_i(0)]$ is observable. This means that the heterogeneity in the treatment effects can be studied directly. That is, the heterogeneity within each CATE is now recovered.

There are numerous benefits gained from understanding such heterogeneity. First, going beyond ATEs and understanding who benefits and who is potentially harmed by treatment can allow for better distribution of the treatment or program. For example, suppose the government knew the type of unemployed people whose outcomes changed the most in response to job retraining programs. In that case, they could pinpoint those people for the program and develop alternative programs for others. Outside of economics, this can also be important. In medical trials, one might want to understand whether some individuals respond positively to a drug while others do not.

Second, understanding heterogeneity across outcomes helps to uncover mediators. For example, in Exhibit 1, panel B, if dealers #1, #2, and #5 have different beliefs about gender valuation or bargaining distributions than dealers #3, #4, and #6, this might help us understand the underpinnings of the observed discrimination. In that case, they did, and the findings supported the notion that the discriminating dealers believed that women and men had different valuation distributions. Finally, it can help deepen our understanding of the generalizability of treatment effects, predicting when a policy is scalable or it has external validity (List, 2024). In this sense, when treatment effects vary based on observables, understanding which characteristics are correlated with treatment can suggest other settings where similar programs should be tested.

Experimental Power

Beyond providing richer information, WS designs have the additional benefit in that data collection is potentially cheaper and for the same budget the researcher can identify smaller effect sizes. This is because WS designs allow for estimations that control for individual-specific effects, reducing the variance of the treatment effect estimator to the extent that WS correlations explain the outcome.

In this manner, WS designs are advocated for their potential to yield more powerful tests for the same cost than between-subject designs. The literature suggests an approach based on Monte Carlo simulations to compare the power achieved in between-subject versus WS designs. In certain cases, a between-subject design requires four to eight times as many subjects as a WS design to reach an acceptable level of statistical power.[10]

Also, note that we have largely ignored cost considerations in our discussion thus far, assuming that collecting data from N subjects twice (WS design) is as costly as collecting data from $2N$ subjects (between-subject design). In practice, however, this is often not the case; in laboratory experiments, adding additional periods often comes at a small additional monetary cost for the researchers. Likewise, many field experiments also have large fixed costs (e.g., hiring and training surveyors) that make additional rounds of data collection less expensive on the margin. I consider these aspects of the design below.

Minimum Detectable Effects for Within-Subject Designs

Statistical power computation for between-subject designs is covered extensively in the literature. In this section, I show how to use simulation methods to approximate the statistical power of WS designs under relatively general assumptions about the data-generating process. I begin by assuming that the analyst is interested in estimating the ATE using the following regression model:

$$Y_{it} = \pi_0 + \tau D_{it} + \mu_i + \epsilon_{it}, \tag{8}$$

Treatment, D_{it}, is a time-varying treatment variable where $D_{it} = 1$ when unit i receives treatment at period t and is 0 otherwise. Time-invariant subject-level heterogeneity is captured by $\mu_i \sim F_\mu$ and is assumed to be independent of treatment under randomization. Idiosyncratic errors, $\epsilon_{it} \sim F_{\epsilon|D}$, are assumed to be homoscedastic.[11]

Computing statistical power requires following three steps:

1. Choose the number of subjects, N, the number of periods, T, the values of (π_0, τ), and the distributions (F_μ, F_ϵ). Given this information, generate the sample $\{\{(Y_{it}, D_{it}): t = 1, .., T\}: i = 1, \ldots, N\}$.

[10] See, for example, Bellemare et al. (2016).

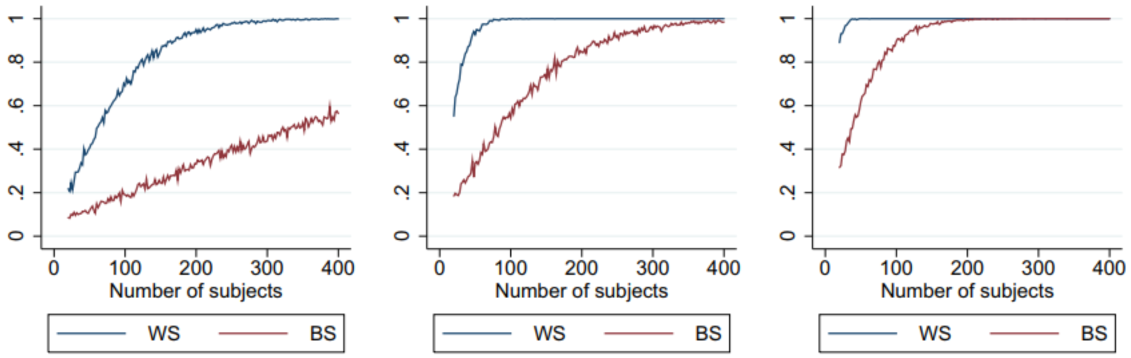
[11] Following similar logic as with between-subject designs, it is reasonable to use balanced designs when errors are homoscedastic. When errors are heteroscedastic, one can increase power by allocating subjects to the noisier conditions for a higher number of periods.

2. Estimate equation (8) and compute $z = \widehat{\tau}_t / se(\widehat{\tau}_t)$ and the corresponding p-value of the null hypothesis $H_0: \widehat{\tau}_t = 0$ against a one-sided or two-sided alternative.[12]
3. Repeat steps 1 and 2 for a large number of samples. In each iteration, compute the fraction of the p-values less than the significance level of the test. This fraction represents the power of the test.

By repeating these steps for various N and T, the analyst can determine the statistical power for different sample sizes and calculate the minimal sample size needed to reach a certain statistical power. This can be achieved separately for each design or to examine the effect of the number of periods and how to balance the number of subjects across the treatments.

For illustrative purposes, I assume that the parameters in the model are given by $\pi_0 = 0.37$, $(\hat{\sigma}_\mu)^2 = 0.045,\ (\hat{\sigma}_\epsilon)^2 = 0.02$. Exhibit 3 shows the statistical power attained from a WS design or a between-subject design with two periods for various treatment effects. In the between-subject case, I assume that the design allocates the same number of subjects to treatment and control conditions for all periods.

Exhibit 3: Statistical Power for Within- Subject and Between-Subject Designs



$\tau = 0.05$ $\tau = 0.1$ $\tau = 0.15$

The graphs compare the statistical power obtained from a WS (blue line) and between-subject (red line) design for three treatment effects – 0.05, 0.1, and 0.15. Both designs have two periods, the same sample size, and model parameters $\pi_0 = 0.37,\ (\hat{\sigma}_\mu)^2 = 0.045,\ (\hat{\sigma}_\epsilon)^2 = 0.02$. A comparison of the two lines in each panel demonstrates that the within-subject (WS) design surpasses the between-subject design. The latter requires a greater number of subjects to achieve the same statistical power. Additionally, when examining results across panels, we observe that the effectiveness of the WS design becomes more pronounced as the effect size diminishes.

Empirical results from power calculations appear in Exhibit 3. The blue line represents the statistical power for given sample sizes using a WS design. In contrast, the red line represents the

[12] When there are multiple outcomes or multiple treatments, these p-values should be adjusted for multiple hypothesis testing.

statistical power for the same sample sizes using a between-subject design. The left-most panel shows the power assuming that the treatment effect size is 0.05, the middle panel assumes the treatment effect size is 0.1, and the right-most panel assumes the treatment effect size is 0.15.

Across all three panels of Exhibit 3, we learn that the WS design outperforms the between-subject design substantially in each scenario. For example, when the treatment effect size is 0.05, researchers must collect data from up to four times as many subjects to obtain the same statistical power as a WS design. Importantly, as we move from large detectable effect sizes (0.15) to smaller ones (0.1 and 0.05), or as we move from the right-most panel leftward, the efficacy of WS designs improve comparatively.

Exhibit 3 shows that for our example, the largest gains appear when the budget necessitates a small sample size, or when the effect sizes are smaller. This means that as the feasible between-subject sample size grows, the gains in power from using a WS design diminish. Likewise, the gains from using a WS design are smaller when the postulated effect size is larger. However, the exact gains will depend on the parameters chosen by the researcher. For this reason, I recommend forming priors of the parameters the researcher needs through a pilot before deciding on the optimal sample size and which design choice to make.

VI Discussion

Once considered unattainable, experiments in economics have now become a cornerstone of empirical research. Over the past few decades, their significance has grown immensely, resonating through academia, organizations, and policy-making circles. A dominant theme in this research agenda is the method employed to recover and estimate treatment effects. Subjects are simultaneously assigned to treatment and control groups, followed by the administration of the experimental treatment. Treatment effects are then measured by estimating conditional expectations, allowing for a precise analysis of the treatment's impact. This study explores an alternative to this classic model: a within-subject experimental design.

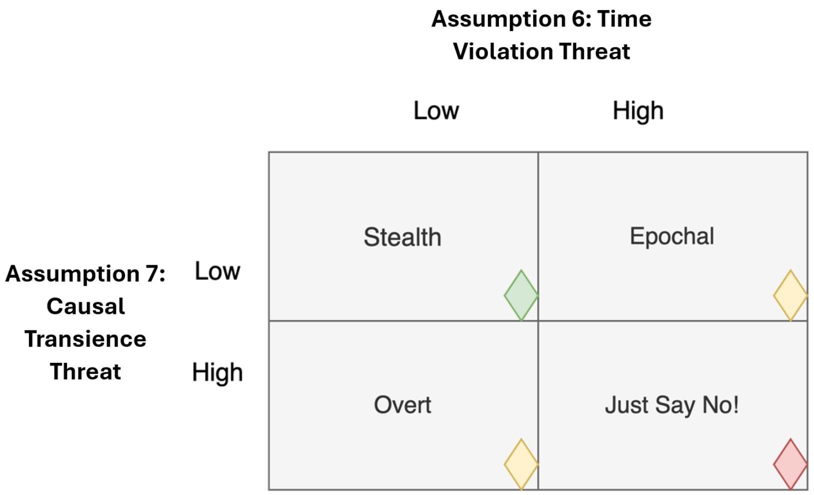
The relative unpopularity of WS designs is likely due to the strong assumption they require for inference. When the same subject is exposed to different treatment conditions, WS comparisons only provide a causal estimate if the new exclusion restrictions I discuss in this study are in place.

Many analysts find these new assumptions untenable and therefore opt for between-subject designs.

Yet, the theme of this study is that the decision should be made case-by-case. There are some instances wherein there can be a strong argument for a WS design. In such cases, I strongly urge the analyst to take full advantage of the WS enhanced features, both through its additional experimental power and the rich information provided. I find that in many instances scholars who use WS designs do not fully realize the benefits of their data and simply present ATEs. That is a fine beginning, but the analysis should not stop there since the full joint distribution of outcomes can be recovered. Another aspect of the literature is that, in general, but not always, compared to behavior generated in between-subject designs, subjects in a WS design behave more rationally, behave more in line with neoclassical theory, and tend to conform to social norms more closely when they have a comparative context, or likewise an evaluability baseline.

In the end, I encourage researchers to carefully weigh each design's advantages and disadvantages along the dimensions discussed in this study and select the design that is best suited to answer the research question at hand. To aid in that choice, I provide Exhibit 4, which includes a summary of the various WS designs and my guidance on design choice. When considering Exhibit 4, there is usually not a universally preferred approach unless the stealth WS design is in reach. In that case, my guidance is to choose that approach whenever possible. Intermediate cases are overt and epochal WS designs. Before choosing one of these approaches, the analyst should keep in mind that our North Star as experimentalists is achieving internal validity. The analyst should never put potential information acquisition or power above succeeding to accomplish that key goal. In most situations, the choice to vary treatment conditions between or within subjects should depend on the experimental characteristics, the nature of what information is sought, and the trade-offs the experimenter is willing to make. If a stealth WS design is not in reach, however, great care should be taken before choosing a WS design.

Exhibit 4: Four Types of Within-Subject Experimental Designs



		Assumption 6: Time Violation Threat	
		Low	High
Assumption 7: Causal Transience Threat	Low	Stealth 🟩	Epochal 🟨
	High	Overt 🟨	Just Say No! 🟥

The diagram offers guidance on selecting the appropriate within-subject (WS) design, considering threats to two key identification assumptions: temporal stability and casual transience. Colors in the bottom-right corner of each quadrant indicate a preference ranking among the possible designs. If conditions permit, the preferred design is the stealth approach, followed by the overt and epochal approaches.

References

Bellemare, Charles, Luc Bissonnette, and Sabine Kröger. 2014. "Statistical Power of Within and Between-Subjects Designs in Economic Experiments." SSRN Electronic Journal, January. https://doi.org/10.2139/ssrn.2529895.

———2016. "Simulating Power of Economic Experiments: The powerBBK Package." Journal of the Economic Science Association 2 (2): 157–68. https://doi.org/10.1007/s40881-016-0028-4.

Biesanz, Jeremy C., Stephen G. West, and Oi-Man Kwok. 2003. "Personality Over Time: Methodological Approaches to the Study of Short-Term and Long-Term Development and Change." Journal of Personality 71 (6): 905–42. https://doi.org/10.1111/1467-6494.7106002.

Campbell, Donald Thomas, and Julian C. Stanley. 1966. Experimental and Quasi-Experimental Designs for Research. Edited by (Nathaniel Lees) Gage. Chicago: R. McNally.

Charness, Gary, Uri Gneezy, and Michael A. Kuhn. 2012. "Experimental Methods: Between-Subject and within-Subject Design." *Journal of Economic Behavior & Organization* 81 (1): 1–8. https://doi.org/10.1016/j.jebo.2011.08.009.

Frederick, S.F. and Fischhoff, B., 1998. Scope (in) sensitivity in elicited valuations. *Risk Decision and Policy*, *3*(2), pp.109-123.

Gneezy, Uri. 2005. "Deception: The Role of Consequences." American Economic Review 95 (1): 384–94. https://doi.org/10.1257/0002828053828662.

Greenwald, A.G., 1976. Within-subjects designs: To use or not to use?. *Psychological Bulletin*, *83*(2), p.314.

Grice, G. Robert. 1966. "Dependence of empirical laws upon the source of experimental variation." Psychological Bulletin, 66 (1966), pp. 488-498

Hsee, C.K., 1996. The evaluability hypothesis: An explanation for preference reversals between joint and separate evaluations of alternatives. *Organizational behavior and human decision processes*, *67*(3), pp.247-257.

Hull, Peter. 2018. "Estimating Treatment Effects in Mover Designs."

Isaac, R.M. and Walker, J.M., 1988a. Group size effects in public goods provision: The voluntary contributions mechanism. *The Quarterly Journal of Economics*, *103*(1), pp.179-199.

———1988b. Communication and free-riding behavior: The voluntary contribution mechanism. *Economic inquiry*, *26*(4), pp.585-608.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): 153–74. https://doi.org/10.1257/jep.21.2.153.

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): 153–74. https://doi.org/10.1257/jep.21.2.153.

———2009. “Field Experiments in Economics: The Past, the Present, and the Future.” *European Economic Review* 53 (1): 1–18. https://doi.org/10.1016/j.euroecorev.2008.12.001.

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

———2002. Preference reversals of a different kind: The “more is less” phenomenon. *American Economic Review*, *92*(5), pp.1636-1643.

———2024. Optimally generate policy-based evidence before scaling. *Nature*, *626*(7999), pp.491-499.

———2025. *Experimental Economics: Theory and Practice*. The University of Chicago Press.

Neyman, J. 1923 trans. 1990. "On the application of probability theory to agricultural experiments. Essay on principles. Section 9." Statistical Science 5 (4): 465-472. https://www.jstor.org/stable/2245382.

Poulton, E.C., 1973. Unwanted range effects from using within-subject experimental designs. *Psychological Bulletin*, *80*(2), p.113.

Rodemeier, Matthias. 2023. “Willingness to Pay for Carbon Mitigation: Field Evidence from the Market for Carbon Offsets.” SSRN Scholarly Paper. Rochester, NY. Available at SSRN: https://papers.ssrn.com/abstract=4360822.

Rosenthal, R., 1976. Experimenter effects in behavioral research. (2nd ed.), Wiley, New York

Rubin, Donald B. 1974. “Estimating Causal Effects of Treatments in Randomized and Nonrandomized Studies.” *Journal of Educational Psychology* 66 (5): 688–701. https://doi.org/10.1037/h0037350.

———. 1975. “Bayesian Inference for Causality: The Importance of Randomization.” *The Proceedings of the Social Statistics Section of the American Statistical Association*, 233–239.

———. 1978. “Bayesian Inference for Causal Effects: The Role of Randomization.” *The Annals of Statistics* 6 (1): 34–58. https://doi.org/10.1214/aos/1176344064.