Chapter 9: Longitudinal Designs

A Course in Experimental Economics

John A. List

All content is derived from the Experimental Economics Textbook by John A. List (2026).

Key Ideas

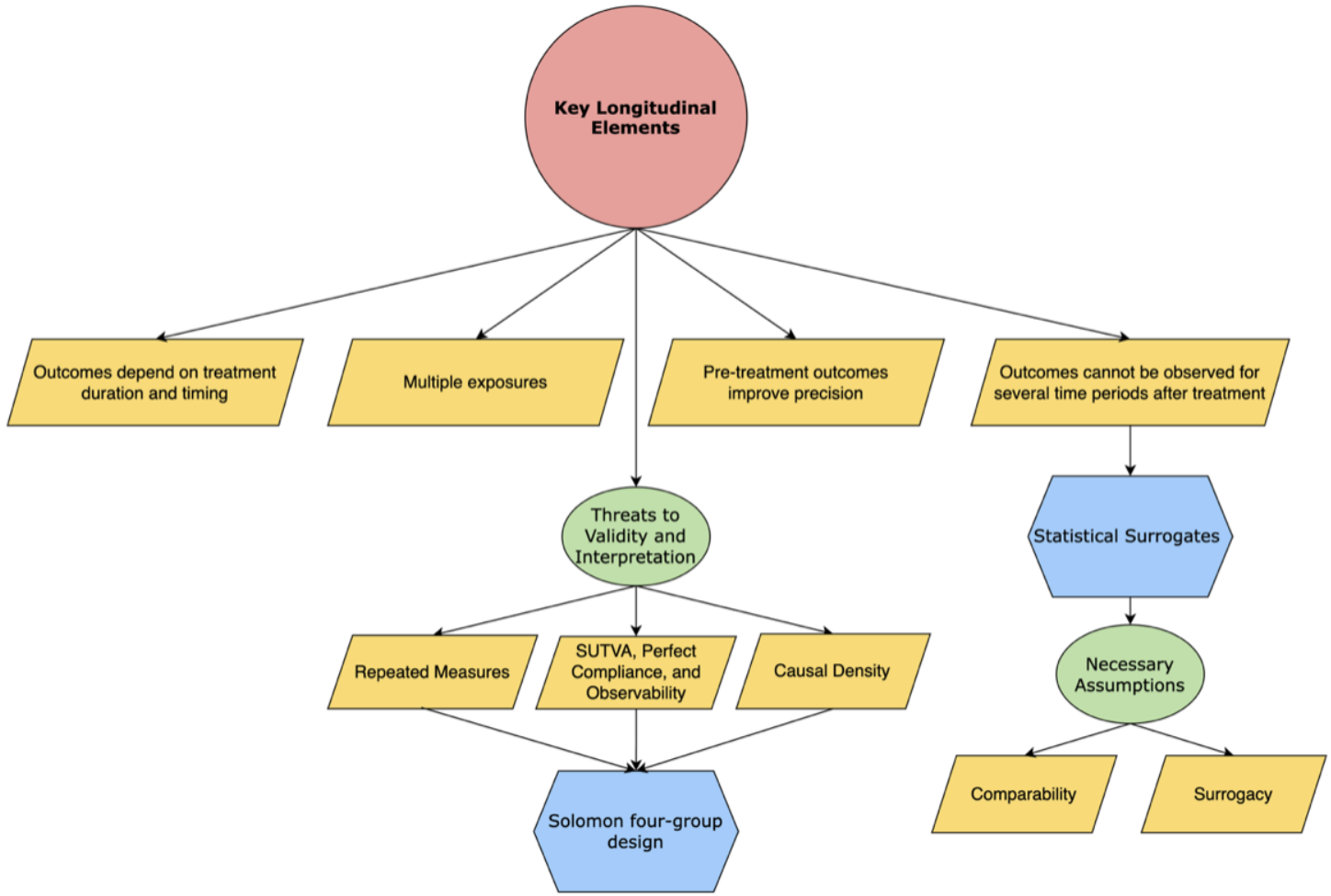
1. Certain questions call upon a time component to explore relevant hypotheses.

2. Examples of key longitudinal features include multiple exposures, repeated choices, staggered designs, pre-treatment outcomes, and measuring outcomes long after treatment.

3. Recovering treatment effects from designs with longitudinal features often requires additional identification assumptions.

4. In certain cases, the analyst can enhance statistical power by taking account of data and experimental opportunities in preliminary design stages.

Mind's Eye


Key Longitudinal Elements
Outcomes depend on treatment duration and timing
Multiple exposures
Pre-treatment outcomes improve precision
Outcomes cannot be observed for several time periods after treatment
Threats to Validity and Interpretation
Repeated Measures
SUTVA, Perfect Compliance, and Observability
Causal Density
Solomon four-group design
Statistical Surrogates
Necessary Assumptions
Comparability
Surrogacy

Incorporating Time into Experimental Designs

- Some economic questions call for the experimental design to go beyond a **static** treatment

- Two major classes of experiments that have a time element
  1. Adding time to the **treatment** domain
     - Examples: multiple exposures, repeated choices, staggered designs
  2. Adding time to the **outcome** domain
     - Examples: pre-treatment outcomes, long-run outcomes

Adding Time to the Treatment Domain

- When adding time to the design, it is useful to construct a **treatment regime**, where D_{it} represents unit i's treatment assignment in period t:

$$\mathbf{D}_i = (D_{i1}, ..., D_{iT})$$

- Potential outcomes can thus be expressed as $Y_{it}(\mathbf{D_{i,t}}, T)$

- In the classic **one-shot** experimental approach, the treatment and control regimes are given by (1) and (0)

Adding Time to the Treatment Domain

- A first departure from the one-shot approach is to have subjects exposed to **multiple periods** and make decisions at the end of each

- Such settings are commonly denoted as **repeated choice**, **repeated games**, or **n-shot games**
  - In the case of a 5-period experiment, we now have treatment and control regimes given by $(1, 1, 1, 1, 1)$ and $(0, 0, 0, 0, 0)$
- A second departure from the classic one-shot design is **staggered designs**, which are conducted on a set of units over multiple time periods where the starting time of the treatment varies by unit
  - A treatment regime of (1,1,1,1,1) versus (0,0,1,1,1) is one such example

Adding Time to the Outcome Domain

- Crucial longitudinal elements can arise even in cases wherein treatment does not have a time element
- Our first consideration in this area is leveraging pre-treatment outcomes, either through controlling for them in estimation or estimating a difference-in-differences model
- A second consideration involves cases wherein the key outcome(s) is measured several periods after treatment ends
- In such cases, proxy variables known as **statistical surrogates** are useful

Structure of the Chapter

- We use three examples to draw out assumptions underlying experiments that involve the important element of time

- We discuss:
  - Important treatment effect parameters one can recover in such experiments
  - Threats to validity and interpretation
  - Key advantages and disadvantages of experiments that introduce a time component

Outline

Three Running Examples

- Our first example is a study published by James Andreoni (1988), who designs a public goods game experiment to explore free-riding behavior amongst undergraduate students
- Previous studies provided evidence of free-riding behavior being present in repeated games but not in single-shot games
- Andreoni (1988) designed a lab experiment to examine the import of strategies and learning as underlying mechanisms for the observed data patterns
- **Example of adding time to the treatment domain (multiple periods)**

Textbook Section 9.1

Andreoni (1988)

- **Learning** refers to the idea that a one-shot game is not sufficient for subjects to learn the incentives of the game
- **Strategies** refers to the idea that repeated games allow for schemes such as signals of future moves or cooperative behaviors
- To separate the two hypotheses, Andreoni conducted an experiment in which subjects were split into two types of games:

- A 10-round, "repeated single-shot" game ("Strangers" design) in which composition of groups changed after each repetition
- A standard 10-round repeated public goods game ("Partners" design) in which groups were maintained throughout all repetitions

Andreoni (1988)

- Empirical results of the experiment contradict both hypotheses, and Andreoni (1988) concludes that neither can be supported as an explanation of decay in giving in public goods games

Miguel and Kremer (2004)

- Our second example is Miguel and Kremer's (2004) evaluation of the Primary School Deworming Project (PSDP)
- PSDP was designed and conducted by a Dutch nonprofit organization (ICS) in a rural farming region of western Kenya – an area with the highest helminth infection rates in its district
- In total, over 30,000 students ages 6-18 were enrolled in the project across 75 primary schools in the region
- Randomization took place at the school-level and was staggered, allowing Miguel and Kremer to estimate externality benefits and effects on education outcomes such as school participation and academic test scores

Miguel and Kremer (2004)

- The schools in PSDP were randomly divided into three groups using an FFE, and the deworming treatment was rolled out to each group in phases across several years due to ICS's financial and administrative constraints

- As a result, Miguel and Kremer (2004) were able to compare schools treated earlier with schools that had not yet received the treatment acting as controls

- **Example of adding time to the treatment domain (staggered design) and outcome domain**

Miguel and Kremer (2004)

- They found that the deworming treatment increased primary school participation by an average of 7.5 percentage points, reducing absenteeism by one-quarter

- Further, the treatment created positive externalities. By reducing the burden of worms, school participation rates increased for students in schools neighboring the treated schools

Miguel and Kremer (2004)

- They found that the deworming treatment increased primary school participation by an average of 7.5 percentage points, reducing absenteeism by one-quarter
- Further, the treatment created positive externalities. By reducing the burden of worms, school participation rates increased for students in schools neighboring the treated schools
- Deworming had no effect on test scores

Schweinhart, Barnes and Weikart (1993)

- Our third experimental example is Schweinhart et al.'s (1993) analysis of the famous Perry Preschool Project, a preschool program launched by psychologist David Weikart in Ypsilanti, Michigan in 1962

- The project began as a relatively small-scale study to test a high-quality education approach on 123 low-income African American children aged 3-4 years old, that were at risk of failing school

- The FFE was designed as a combination preschool from Monday-Friday for 2.5 hours per day, as well as a 1.5 hour home visit each week, for eight months out of the year across a span of 2 years

Schweinhart, Barnes and Weikart (1993)

- While the Perry Preschool Project began as an early childhood FFE to understand whether high-quality education would benefit at-risk children and their communities, it soon became a longitudinal study documenting the lasting impact of the intervention on adult outcomes
- The researchers collected data on the children each year until they reached age 11, then again at ages 14, 15, 19, and 27
- They found that the project produced dramatic benefits for academic achievement, graduation rates, and earnings
- It also lowered rates of crime, delinquency, teenage pregnancy, and welfare dependency when compared to the control
- **Example of adding time to the outcome domain (long-run outcomes)**

Three Running Examples

Study	Experiment Type	Units	Control Condition	Treatment Condition	Outcome(s)	Longitudinal Elements
Andreoni (1988)	Lab	College Undergraduate Students	Fixed partners	Random partners	Public Good Contributions	Yes; treatment domain
Primary School Deworming Project (PSDP; Miguel and Kremer, 2004)	Framed field experiment	Rural Kenyan primary schools	No deworming treatment	Mass school-based deworming treatment	Infection rates, school participation rates, and test scores	Yes; treatment domain and outcome domain
The 1962 Perry Preschool Project (Schweinhart, Barnes, and Weikart, 1993)	Framed field experiment	Low-income African American Families with 3-4 year-old children	No preschool or home visiting program	High-quality preschool and home visiting program, 8 months a year, for 2 years	Several short run outcomes such as grades, test scores, classroom behaviors and eventually longer-term outcomes such as graduation rates	Yes; outcome domain

[Table 9.1 in the Textbook]

Outline

Potential Outcomes in Repeated Exposure Designs

- Using only the assumptions from Chapter 3, the analyst can identify the average treatment effect for participants in the first period of the experiment,

$$\tau_1 = \mathbb{E}[Y_{i1}(1) - Y_{i1}(0)|P_i = 1]$$

- For the rest of this chapter, we will add an additional assumption that the units' observations are observed in every period

- We refer to this as the **balanced panel** assumption

Textbook Section 9.2

Treatment Effects in the Presence of Repeated Exposure

Assumption 9.1 (Balanced Panel):

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

- Assumption 9.1 requires that all participants in the study remain in the study for all t and for each, we observe their outcome, treatment assignment, actual treatment, and unit-level characteristics

- This is a stronger assumption than Assumption 2 (Observability) in Chapter 3 because we require it for each period rather than the single period that was used in that chapter

Treatment Effects in the Presence of Repeated Exposures

- Under Assumption 9.1 one can estimate the average treatment effect in any period, $t \in \mathcal{T}$

$$\tau_t = \mathbb{E}[Y_{it}(1) - Y_{it}(0)]$$

- Further, Assumption 9.1 allows us to write

$$\tau_t^{DID} = \mathbb{E}[(Y_{it}(1) - Y_{it-1}(1)) - (Y_{it}(0) - Y_{it-1}(0))]$$

which measures the differential change in outcomes over time between the treatment and control groups

Outline

Staggered Experimental Designs

- The analysis of n-shot games is especially useful in lab experiments where it is possible to ensure a constant treatment assignment (1,1,1,1,1) and repeat the same group for multiple rounds.
- In some field experiments, a unit's treatment assignment may change over time (0,0,1,1,1).
- This class of experiment is commonly referred to as a **staggered design**
- Staggered designs are used for instrumental reasons – a staggered design is desired *per se* – or for practical reasons

Textbook Section 9.3

Staggered Experimental Designs

- One benefit is that staggered designs allow you to measure **instantaneous** and **lagged** effects

- For example, in Miguel and Kremer (2004):
  - 1998: Group 1 is treated, Groups 2 and 3 are control
  - 1999: Groups 1 and 2 are treated, Group 3 is control
  - 2001: Groups 1, 2, and 3 are treated
- For the remainder of the chapter, we only consider treatment regimes in which units are assigned to the control group or receive treatment after a certain number of periods (e.g., $(0, 0, 1, 1, ..., 1)$).

Staggered Experimental Designs

More formally we define this assumption as follows:

Assumption 9.2 (Irreversible Treatment Adoption):

For all $(i, t) \in \mathcal{J} \times \mathcal{T}$, $D_{it} \leq D_{it+1}$

- Assumption 9.2 means that once a unit has been exposed to the treatment, we consider that unit to be treated forever
- In many settings, this assumption may hold due to the restrictions imposed by the experimental design
- For example, if the treatment occurred at an aggregate level, it is not possible for individual units to switch back to the control group

Staggered Experimental Designs

- Given Assumption 9.2, potential outcomes can thus be expressed as $Y_{it}(\mathbf{D_{i,t}}, T)$, which implicitly ensures that a unit's treatment assignment satisfies SUTVA

- Moreover, this guarantees that units do not anticipate their treatment assignments and thus, their current behavior cannot be influenced by their future treatment assignment **(preview: causal transience)**

- In the appendix, we provide rules of thumb to maximize power for staggered designs with purely instantaneous treatment effects and explore designs that attempt to estimate long-term effects (Xiong et al., 2023)

Outline

Leveraging Pre- and Post-Treatment Outcomes to Increase Power

- While the experimental family in economics shares the interest in the temporal effects of treatment, there is great disparity in how information in period 0 is used
- In many cases, before period 1 starts, laboratory and AFE experiments use a period 0
  - These are commonly referred to as 'practice rounds' that are financially inconsequential to allow subjects a 'warm-up', and to allow questions before, during, and after to ensure a full understanding of the rules, institutions, or mechanisms before the choices for payment begin
- Usually **baseline data** are not reported because period 0 is purely meant to aid in subject comprehension
- In certain cases, data gathered in period 0 where all subjects are treated identically, are useful

---

Textbook Section 9.4

Leveraging Pre and Post Treatment Outcomes to Increase Power

- There are many ways in which baseline data can be useful
  - Increase power (Chapter 5)
  - Stratify on the baseline outcome in the randomization (Chapter 6)
  - Improve the credibility of small sample designs or heterogeneity analysis
- In this section, we show two key approaches to using pre-treatment outcome information

1. The first is to extend our discussion in Chapter 5 using pre-treatment covariates to increase power

2. The second is to collect multiple pre-treatment outcome measures and estimate a panel data model

Including Covariates and Pre-Treatment Outcomes in the Estimation Model

- Recall (from Chapter 5):

$$var(\hat{\tau}) = \frac{\sigma^2}{N} = \frac{var(\epsilon)}{N * var(D)}$$

- In Chapter 5.7.1 and Appendix 5.2, we explored how inclusion of pre-treatment covariates using efficient regression adjustment improves precision in the Oregon Health Insurance Experiment data examined by Finkelstein et al. (2016)

- The gains from adding pre-treatment outcomes as controls nearly doubles the gains we previously observed: researchers can now reduce sample sizes by about 5-6% by simply adding information on pre-treatment outcomes

Leveraging Pre-Treatment Outcomes in a Panel Data Estimation Model

- We consider a difference-in-differences specification for a field experiment

$$Y_{it} = \pi_0 \mathbb{I}[\sum_t D_{it} = 1] + \tau^{DID} D_{it} + \gamma_t + \epsilon_{it}. \tag{9.1}$$

- $\mathbb{I}[\Sigma_t D_{it} = 1]$ is an indicator representing whether unit i has ever been assigned to the treatment group, γ_t are time period fixed effects, and D_{it} is unit i's treatment status in period t

- Assume that ϵ_{it} is independent across units with cross-sectional variance σ^2

Leveraging Pre-Treatment Outcomes in a Panel Data Estimation Model

- Errors may be autocorrelated over time for the same unit and we assume that τ^{DID} is constant over time

- The treatment effect parameter estimated in the difference-in-differences specification can be written as

$$\begin{aligned}\tau^{DID} &= \tau^{post} - \tau^{pre} \\ &= \sum_{t>0}(\mathbb{E}[Y_{it} \mid D_{it} = 1] - \mathbb{E}[Y_{it} \mid D_{it} = 0]) \\ &- \sum_{t\leq 0}(\mathbb{E}[Y_{it} \mid D_{it} = 1] - \mathbb{E}[Y_{it} \mid D_{it} = 0])\end{aligned} \tag{9.2}$$

Leveraging Pre-Treatment Outcomes in a Panel Data Estimation Model

- Alternatively, one can ignore the pre-treatment measures of the outcome because randomization ensures that

$$\mathbb{E}[Y_{it} \mid D_{it} = 1] = \mathbb{E}[Y_{it} \mid D_{it} = 0]$$

for all $t \leq 0$

- In this case, one can estimate an alternative parameter,

$$\tau^{Post} = \sum_{t>0} (\mathbb{E}[Y_{it} \mid D_{it} = 1] - \mathbb{E}[Y_{it} \mid D_{it} = 0]) \qquad (9.3)$$

- Either of these treatment parameters can be identified in a randomized experiment with any amount of pre- or post-periods

- However, the **number** of pre- and post-periods and the **choice of parameter** will have important implications for statistical power

The Gain from Pre-Period Measures

- We begin by making several assumptions on the error term from the difference-in-differences equation
  - Assume that the variance is constant across all time periods and that there are equal variances between all pairs of time points
  - Assume that the autocorrelation in outcomes does not depend on treatment status
  - Assume that the treatment and control groups are each of identical cross-sectional size, $n_d = n_0 = n_1$

The Gain from Pre-Period Measures

- Under these conditions,

$$Var[\tau^{Post}] = \frac{2\sigma^2}{n_d}\left(\frac{1+(N_{Post}-1)\rho}{N_{Post}}\right) \tag{9.4}$$

where N_{post} is the number of periods observed after treatment is assigned and ρ is the degree of autocorrelation

- Similarly

$$Var[\tau^{DID}] = \frac{2\sigma^2}{n_d}\left(\frac{1+(N_{Post}-1)\rho}{N_{Post}} - \frac{(N_{Pre}+1)\,\rho-1}{N_{Pre}}\right) \tag{9.5}$$

where N_{Pre} is the number of periods observed before treatment

The Gain from Pre-Period Measures

- The gain from incorporating pre-periods can be seen by comparing the two variances:

$$Var[\tau^{Post}] - Var[\tau^{DID}] > 0 \text{ when } \rho > \frac{1}{N_{Pre} + 1} \tag{9.6}$$

- This relation reveals that the difference-in-differences estimator only provides more power than the post estimator when the autocorrelation is sufficiently high

- If there is no autocorrelation in the outcomes, only post-treatment measurements will be beneficial from a power perspective [1]

[1] See also Appendix A9.2 in the Textbook for more information behind the derivations of optimal sample rules to increase statistical power.

Intuition for Choosing Difference-in-Means versus Difference-in-Differences

ρ	N_{pre}	**Variance Ordering**	**Preferred Choice?**
0	1	$Var[\tau^{DiD}] > Var[\tau^{post}]$	Difference-in-means
	5	$Var[\tau^{DiD}] > Var[\tau^{post}]$	Difference-in-means
	10	$Var[\tau^{DiD}] > Var[\tau^{post}]$	Difference-in-means
0.25	1	$Var[\tau^{DiD}] > Var[\tau^{post}]$	Difference-in-means
	5	$Var[\tau^{post}] > Var[\tau^{DiD}]$	Difference-in-differences
	10	$Var[\tau^{post}] > Var[\tau^{DiD}]$	Difference-in-differences
0.5	1	$Var[\tau^{DiD}] > Var[\tau^{post}]$	Difference-in-means
	5	$Var[\tau^{post}] > Var[\tau^{DiD}]$	Difference-in-differences
	10	$Var[\tau^{post}] > Var[\tau^{DiD}]$	Difference-in-differences

Note: The table reports simple rules of thumbs for whether to choose difference-in-means or difference-in-differences based on ρ, N_{pre}, and the variance ordering. In cases where $Var[\tau^{DiD}] > Var[\tau^{post}]$ and $\rho = 0$, the preferred approach is difference-in-means regardless of N_{pre}. Likewise, whenever $N_{pre} = 1$, the preferred approach is difference-in-means. Alternatively, the preferred choice begins to shift to DiD for larger values of N_{pre} and when $Var[\tau^{post}] > Var[\tau^{DiD}]$.

[Table 9.2 in the Textbook]

Threats to Validity and Interpretation

- A criticism of collecting repeated measures (e.g., pre-test and post-test) is that the pre-test may sensitize subjects to the subsequent treatment, making it more effective
- Repeated measures of the same outcome may change behavior for other reasons as well
- Respondents may report unfamiliar or unclear survey outcomes more accurately after additional rounds of measurement
- Alternatively, units may grow fatigued with repeated measures
- In situations where this is a severe threat to identification, researchers can use a Solomon four-group design (Solomon, 1949)

Solomon Four-Group Design

- In the Solomon four-group design, the experimenter splits each of the treatment groups into two subgroups
- Half the subjects in each treatment take both the pre-test and the post-test, while the other half takes only the post-test
- Then, the researcher compares the post-test treatment effects within an experiment to test whether the pre-test altered the results
- The researcher can pool the results if they are similar across groups or impute the baseline outcome for unmeasured subjects under a missing at random assumption that the design ensures
- In cases where they differ, the researcher can report both results and discuss each's generalizability

Outline

Experimental Designs with Outcomes Measured Long After Treatment

- Economic experiments typically focus on estimating "short-term" outcomes.
- this is for two reasons:
  1. Conducting an experiment that follows subjects for several months or years is substantially more costly
  2. in many cases τ_1 or τ_1^{DiD} is the ultimate parameter of interest in a theoretical and practical sense
- In this section we are trying to answer the question: **What are the benefits and challenges of measuring long-run outcomes?**[2]

[2] See also chapter 13 and 19 for a discussion on attrition and how to incentivize subjects to come back for the follow-ups.

CHECC: A Modern-Day Perry Preschool Project

Benefits of Long-Run Outcomes

- From a policy and scientific perspective, understanding the long-term effects of many interventions is critical
- In such cases, it is fundamental to consider the spillover and other scaling effects of the implemented policies and programs
- However, such effects often need time to manifest, so measuring the long-term effect of interventions is even more important
- Moreover, Hawthorne, John Henry, resentful demoralization, and experimenter demand effects can in principle be identified by collecting long-run data
- Finally, the return on investment will be much larger if the effects persist longer-term, affecting relevant **benefit-cost** calculations

Challenges with Long-Run Outcomes

- The actual measurement of a long run treatment effect and the discovered underlying mechanisms at work depends on the outcome's **causal density**
- The causal density of an outcome variable refers to how well the determinants of the outcome are understood
- As the time between experiment and observation of the outcome increases, so do the number of potential mechanisms driving the effect and the causal density
- For example, consider the question of whether financial incentives increase test scores

Identification Assumptions When Outcomes are Far Removed From Treatment

- From Chapter 3, identifying the ATE in a cross-section of data depends on four key assumptions: SUTVA, observability, perfect compliance, and statistical independence
- Our assumptions are also required to identify an average treatment effect in a longitudinal experiment
- However, SUTVA, observability, and perfect compliance may be less likely to hold when the researcher measures the outcome long after administering treatment

Identification Assumptions When Outcomes are Far Removed From Treatment

- Recall that SUTVA has two parts: no heterogeneity in the treatment version and no spillovers across units
  - An experiment that takes a while to implement has more opportunity to differ than one administered instantaneously
  - Moreover, participants' possibility of communicating about the experiment or generating peer effects is more likely when subjects have more opportunities to communicate
- Additionally, non-compliance may also be more likely when the outcome occurs long after randomization

- Finally, collecting the outcome data multiple times increases the burden on respondents, leading to potential attrition problems (discussed in Chapter 12)

Statistical Surrogates

- An issue arises when one can only observe the outcome of interest long after the administration of the treatment but evidence in the interim is invaluable

- For example, one primary outcome of interest to policy makers in the Chicago Heights Early Childhood Center (CHECC) is the child's lifetime earnings which we will denote by Y_i

- The average treatment effect for participants in CHECC is given by

$$\tau_{P=1} = \mathbb{E}[Y_i(1) - Y_i(0) \mid P_i = 1] \tag{9.9}$$

- The central challenge in estimating $\tau_{P=1}$ is that the outcome emerges decades after the experimental treatment ends

Statistical Surrogates

- A common approach to this problem is to measure the treatment effect on a more proximate and easily obtainable variable referred to as a **statistical surrogate**, denoted by Y_i^S

- **Can you think of any surrogates for the child's lifetime earnings?**

Statistical Surrogates

- A common approach to this problem is to measure the treatment effect on a more proximate and easily obtainable variable referred to as a **statistical surrogate**, denoted by Y_i^S

- In the CHECC example, 3rd grade math test scores might serve as a useful statistical surrogate for lifetime earnings

- The treatment effect measured in the experiment is denoted by

$$\tau^S_{P=1} = \mathbb{E}[Y_i^S(1) - Y_i^S(0) \mid P_i = 1]. \tag{9.8}$$

Surrogate Outcomes

- Say we observe (Y_i^S, Y_i) for a sample of non-participants (observational data)

- Researchers can use the relationship between Y_i^S and Y_i in this naturally-occurring data set to extrapolate from the treatment effect relationship in the former equation to estimate the treatment effect on the outcome of ultimate interest

- While convenient, this extrapolation requires two strong assumptions about the relationship between participation, treatment, the surrogate, and the outcome of ultimate interest

Internal Validity of Statistical Surrogates

- The next step is to determine conditions under which $\tau^S_{P=1}$ is informative of τ

- First, we must assume that the sample of subjects we observe (Y_i^S, Y_i) is comparable to the sample of subjects for whom we observe (Y_i^S, D_i)

- This can be formally expressed as

Assumption 9.3 (Comparability):

$P_i \perp\!\!\!\perp Y_i \mid Y_i^S$

- Assumption 9.3 requires that the conditional distribution of the primary outcome given the surrogates is the same in the observational and experimental samples

Internal Validity of Statistical Surrogates

- Next, we assume that there is a linear relationship between the outcome and the surrogate

- Further, without loss of generality, we assume that

$$Var[Y_i^S \mid D_i] = Var[Y_i \mid D_i] = 1$$

so that the effect sizes of both the outcome and the surrogate are expressed as standardized differences from the means of those variables, respectively

Internal Validity of Statistical Surrogates

- Assuming linearity, we can write the conditional expectation of the outcome given the surrogate and the treatment as

$$\begin{aligned}\mathbb{E}[Y_i \mid Y_i^S, D_i] &= \mathbb{E}[Y_i \mid D_i] + \frac{Cov[Y_i, Y_i^S \mid D_i]}{Var[Y_i^S \mid D_i]}(Y_i^S - \mathbb{E}[Y_i^s \mid D_i]) \\ &= \mathbb{E}[Y_i \mid D_i] + \rho_{Y_i, Y_i^s \mid D_i}(Y_i^S - \mathbb{E}[Y_i^S \mid D_i])\end{aligned} \tag{9.11}$$

- Now we can use this to write

$$\mathbb{E}[Y_i \mid Y_i^S, D_i = 1] - \mathbb{E}[Y_i \mid Y_i^S, D_i = 0] = \tau - \rho_{Y_i, Y_i^S} \cdot \tau^s$$

- Which we can rearrange to be

$$\tau = \rho_{Y_i, Y_i^S} \cdot \tau^s + (\mathbb{E}[Y_i \mid Y_i^S, D_i = 1] - \mathbb{E}[Y_i \mid Y_i^S, D_i = 0]) \tag{9.12}$$

Internal Validity of Statistical Surrogates

- The above equation for τ highlights that the relationship between the average treatment effect on the outcome and the average treatment effect on the surrogate is confounded by

$$\mathbb{E}[Y_i \mid Y_i^S, D_i = 1] - \mathbb{E}[Y_i \mid Y_i^S, D_i = 0]$$

- This expression reflects variation in the outcome arising from treatment that is not accounted for by the surrogate

Internal Validity of Statistical Surrogates

- This brings us to our final assumption, which is referred to as the Surrogacy, or Prentice, condition

- If there is a violation of the Surrogacy condition, the treatment may have a positive effect on the surrogate, but a negative effect on the true outcome

- This is particularly a concern when ρ_{Y_i, Y_i^S} is small

Assumption 9.4 (Surrogacy Condition):

$D_i \perp\!\!\!\perp Y_i \mid Y_i^S$

- The surrogacy condition requires that the surrogate(s) fully capture the causal link between the treatment and the primary outcome

Internal Validity of Statistical Surrogates

- Under the surrogacy condition, there remains no effect of the treatment on the outcome after conditioning on the surrogate(s)
- As such, Assumption 9.4 rules out any direct effects of the treatment on the outcome that do not pass through the surrogate
- In the CHECC example, a violation of this assumption would arise if the pre-school or parent meetings affected skills important to labor market outcomes that were not captured in test scores (e.g. non-cognitive skills)

Internal Validity of Statistical Surrogates

- Under this condition, $\mathbb{E}[Y_i \mid Y_i^S, D_i = 1] - \mathbb{E}[Y_i \mid Y_i^S, D_i = 0] = 0$, and the relationship between the outcome and surrogate becomes

$$\tau = \tau^s \rho_{Y_i, Y_i^S} \tag{9.13}$$

- The treatment effect on the outcome of interest is equal to the treatment effect on the surrogate times the correlation between the surrogate and the outcome of interest in the observational sample

Interpreting Surrogate Experiments

- However, interpret with caution!
  - The surrogacy assumption seems unlikely to hold in most settings
  - Without a restriction on the relationship between the treatment and the outcome, the treatment effect on the surrogate may have the opposite sign as the treatment's effect on the outcome
  - Even when the comparability and surrogacy conditions hold, the degree to which the surrogate informs about the outcome of interest depends directly on the correlation between the two outcomes

Interpreting Surrogate Experiments

- We recommend that researchers only use surrogate outcomes when
  1. Direct outcomes are infeasible
  2. A strong correlation between the outcome and the surrogate has been established
  3. The comparability and surrogacy conditions hold

Multiple Surrogates

- Thus far, we have considered a single surrogate and a single outcome
- However, it is possible to measure multiple surrogates for the same outcome
- When multiple short-term surrogates are observed, the researcher can potentially combine them into a surrogate index (Athey et al., 2019)
- The surrogate index is the conditional expectation of the outcome given the surrogates and baseline outcomes in the observational data set
- One can construct the index as the predicted value from a regression of the long-term outcome on the surrogates

$$Y_i^{index} = \beta_0 + \sum_k \beta_k Y_i^{S,k} + \epsilon_i$$

Multiple Surrogates

- Then, one can estimate the treatment effect on the effects of each of the individual surrogates
- However, the researcher should note that in order to construct the surrogate index, she must measure all of the variables in both the observational and experimental data set
- That means when designing an experiment, the researcher should first determine what variables are highly correlated with the outcome of interest in the observational data set
- Then, based on that information, decide which outcomes to measure in the experiment
- When the number of potential surrogates become large, LASSO or other machine learning methods might be useful to choose the specification or the index

Choice and Consequence: Surrogates are Not Created Equally

- A key assumption underlying a deeper understanding of how treatment affects a long-term outcome is the surrogacy condition
  - Surrogates fully capture the causal link between the treatment and the primary outcome
- Consider our running example, the CHECC intervention, which randomized students to a pre-school/parent program or a control group

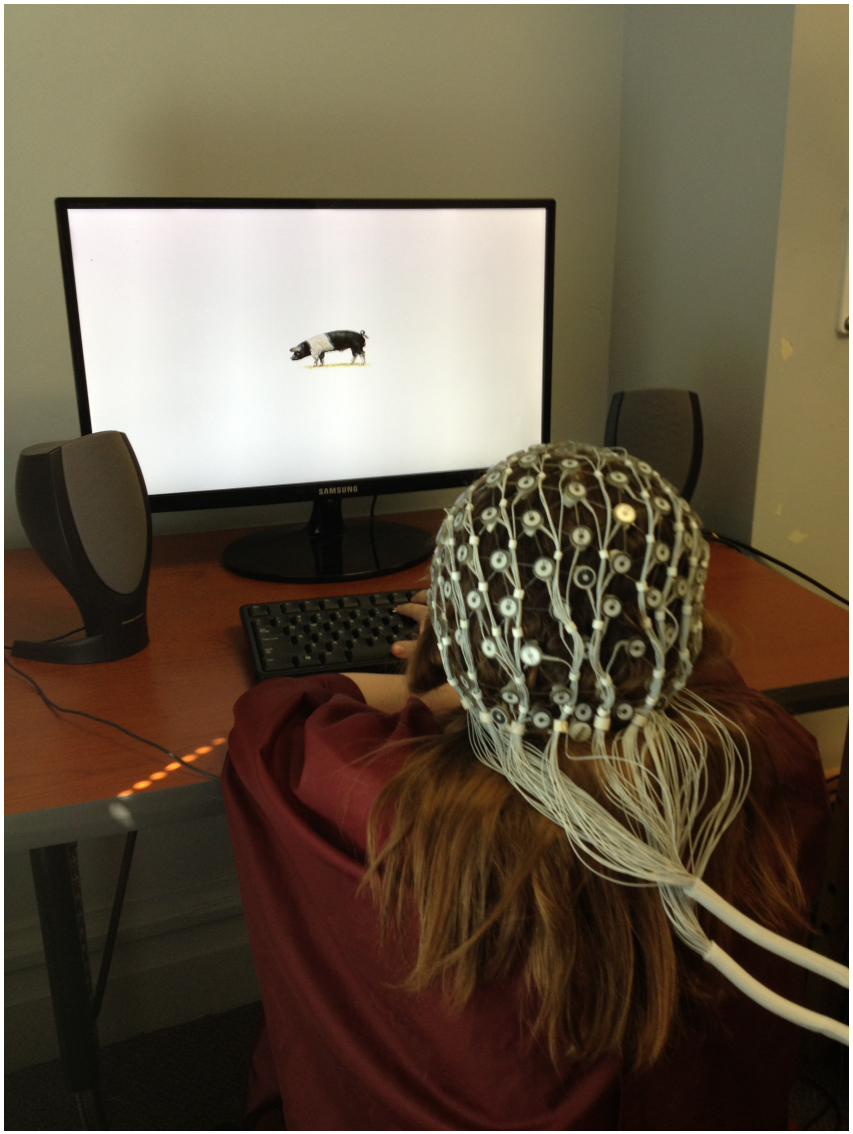
- Are test scores alone valid statistical surrogates for lifetime earnings?

Choice and Consequence: Surrogates are Not Created Equally

- Another way to pose the surrogacy condition question is: can one imagine another link between treatment and the primary outcome?

- Ye et al. (2022) explored this question by measuring brain activity using EEG from a subset of CHECC participants at the end of the one-year program period (when they were 5) and comparing these measures to a control group
  - Brain activity associated with cognitive and noncognitive skills

Measuring children's brain activity



Choice and Consequence: Surrogates are Not Created Equally

- The authors found that the preschool program administered by CHECC improved children's brain activity associated with noncognitive skills
  - Higher ERP amplitude, implying more attention and higher effort
- Additionally, four years after the EEG experiment was conducted, Ye et al. (2022) collected follow-up behavioral assessments of their cognitive and noncognitive skills

- The authors found that EEG measures were highly predictive of these traits

Choice and Consequence: Surrogates are Not Created Equally

- Altogether, these results provide evidence that the preschool program caused measurable effects in brain activity, which are predictive of non-cognitive skills and long-term outcomes

- This finding suggests that pairing test scores and brain imaging measures into a surrogacy index might yield a measure that is more likely to satisfy the surrogacy condition

Outline

What have we learned?

- While the four identifying assumptions discussed in Chapter 3 still hold, the longitudinal aspects provide new considerations

- In practice, the surrogacy assumptions might be difficult to satisfy, but they may be more tenable with a surrogate index and/or future discoveries in this active area of research

What have we learned?

- A natural question arises concerning longitudinal exercises: have we left any key areas of research on the sidelines?
- The answer is yes, within the traditional experimental design we capture causal parameters using a between-subject design: each person is placed into either treatment or control and we estimate the marginal distributions
- We can add a longitudinal dimension to this aspect of design to yield even greater insights, but this comes at the cost of layering on further identification assumptions
- In the next chapter we augment the traditional between-subject design by examining within-subject experimental designs

References I

Andreoni, J. (1988). Why free ride? strategies and learning in public goods experiments. *Journal of Public Economics 37*(3), 291–304.

Athey, S., R. Chetty, G. W. Imbens, and H. Kang (2019). The surrogate index: Combining short-term proxies to estimate long-term treatment effects more rapidly and precisely. Working Paper 26463, National Bureau of Economic Research.

Barter, P. J., M. Caulfield, M. Eriksson, S. M. Grundy, J. J. P. Kastelein, M. Komajda, J. Lopez-Sendon, et al. (2007). Effects of torcetrapib in patients at high risk for coronary events. *New England Journal of Medicine 357*(21), 2109–2122.

Begg, C. B. and D. H. Y. Leung (2000). On the use of surrogate end points in randomized trials. *Journal of the Royal Statistical Society: Series A (Statistics in Society) 163*(1), 15–28.

Burlig, F., L. Preonas, and M. Woerman (2020, May). Panel data and experimental design. *Journal of Development Economics 144*, 102458.

Chandar, B. K., A. Hortaçsu, J. A. List, I. Muir, and J. M. Wooldridge (2019). Design and analysis of cluster-randomized field experiments in panel data settings. Working Paper 26389, National Bureau of Economic Research.

Currie, J. (2001). Early childhood education programs. *Journal of Economic Perspectives 15*(2), 213–238.

Czibor, E., D. Jimenez-Gomez, and J. A. List (2019). The dozen things experimental economists should do (more of). *Southern Economic Journal 86*(2), 371–432.

Day, N. E. and S. W. Duffy (1996). Trial design based on surrogate end points—application to comparison of different breast screening frequencies. *Journal of the Royal Statistical Society: Series A (Statistics in Society) 159*(1), 49–60.

Fafchamps, M., D. McKenzie, S. Quinn, and C. Woodruff (2012). Using pda consistency checks to increase the precision of profits and sales measurement in panels. *Symposium on Measurement and Survey Design 98*(1), 51–57.

Finkelstein, A. N., S. L. Taubman, H. L. Allen, B. J. Wright, and K. Baicker (2016). Effect of medicaid coverage on ed use—further evidence from oregon's experiment. *New England Journal of Medicine 375*(16), 1505–1507.

Frison, L. and S. J. Pocock (1992). Repeated measures in clinical trials: Analysis using mean summary statistics and its implications for design. *Statistics in Medicine 11*(13), 1685–1704.

Heckman, J. J., S. H. Moon, R. Pinto, P. A. Savelyev, and A. Yavitz (2010). The rate of return to the HighScope Perry Preschool program. *Journal of Public Economics 94*(1), 114–128.

Heckman, J. J., R. Pinto, and P. Savelyev (2013). Understanding the mechanisms through which an influential early childhood program boosted adult outcomes. *American Economic Review 103*(6), 2052–2086.

Houser, D. and R. Kurzban (2002). Revisiting kindness and confusion in public goods experiments. *American Economic Review 92*(4), 1062–1069.

Houser, D., E. Xiao, K. McCabe, and V. Smith (2008). When punishment fails: Research on sanctions, intentions and non-cooperation. *Games and Economic Behavior 62*(2), 509–532.

References II

Huang, J., D. Reiley, and N. Riabov (2018). Measuring consumer sensitivity to audio advertising: A field experiment on Pandora internet radio.

Levitt, S. D., J. A. List, S. Neckermann, and S. Sadoff (2016). The behavioralist goes to school: Leveraging behavioral economics to improve educational performance. *American Economic Journal: Economic Policy 8*(4), 183–219.

Manzi, J. (2012). *Uncontrolled: The Surprising Payoff of Trial-and-Error for Business, Politics, and Society* (1 ed.). New York: Basic Books.

McKenzie, D. (2012). Beyond baseline and follow-up: The case for more t in experiments. *Journal of Development Economics 99*(2), 210–221.

Miguel, E. and M. Kremer (2004). Worms: Identifying impacts on education and health in the presence of treatment externalities. *Econometrica 72*(1), 159–217.

Parks, G. (2000). *The High/Scope Perry Preschool Project*. Washington, DC: US Department of Justice, Office of Juvenile Justice and Delinquency Prevention.

Prentice, R. L. (1989). Surrogate endpoints in clinical trials: Definition and operational criteria. *Statistics in Medicine 8*(4), 431–440.

Samphantharak, K. and R. M. Townsend (2009). *Households as Corporate Firms: An Analysis of Household Finance Using Integrated Household Surveys and Corporate Financial Accounting*. Cambridge: Cambridge University Press.

Schweinhart, L. J., H. V. Barnes, and D. P. Weikart (1993). *Significant Benefits: The High/Scope Perry Preschool Study through Age 27*. Monographs of the High/Scope Educational Research Foundation. Ypsilanti, MI: High/Scope Press.

Solomon, R. L. (1949). An extension of control group design. *Psychological Bulletin 46*(2), 137–150.

Woolcock, M. (2013). Using case studies to explore the external validity of 'complex' development interventions. *Evaluation 19*(3), 229–248.

Xiong, R., S. Athey, M. Bayati, and G. Imbens (2023). Optimal experimental design for staggered rollouts. *Management Science 70*(8), 5317–5336.

Ye, K. J., A. Samek, K. J. Yoder, J. Decety, and A. Hortaçsu (2021). Early childhood programs change test scores but do they change brain activity? Unpublished manuscript.

Zwane, A. P., J. Zinman, E. Van Dusen, W. Pariente, C. Null, E. Miguel, M. Kremer, et al. (2011). Being surveyed can change later behavior and related parameter estimates. *Proceedings of the National Academy of Sciences 108*(5), 1821–1826.