Chapter 2: A Primer on Economic Experiment

A Course in Experimental Economics

John A. List

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

Key Ideas

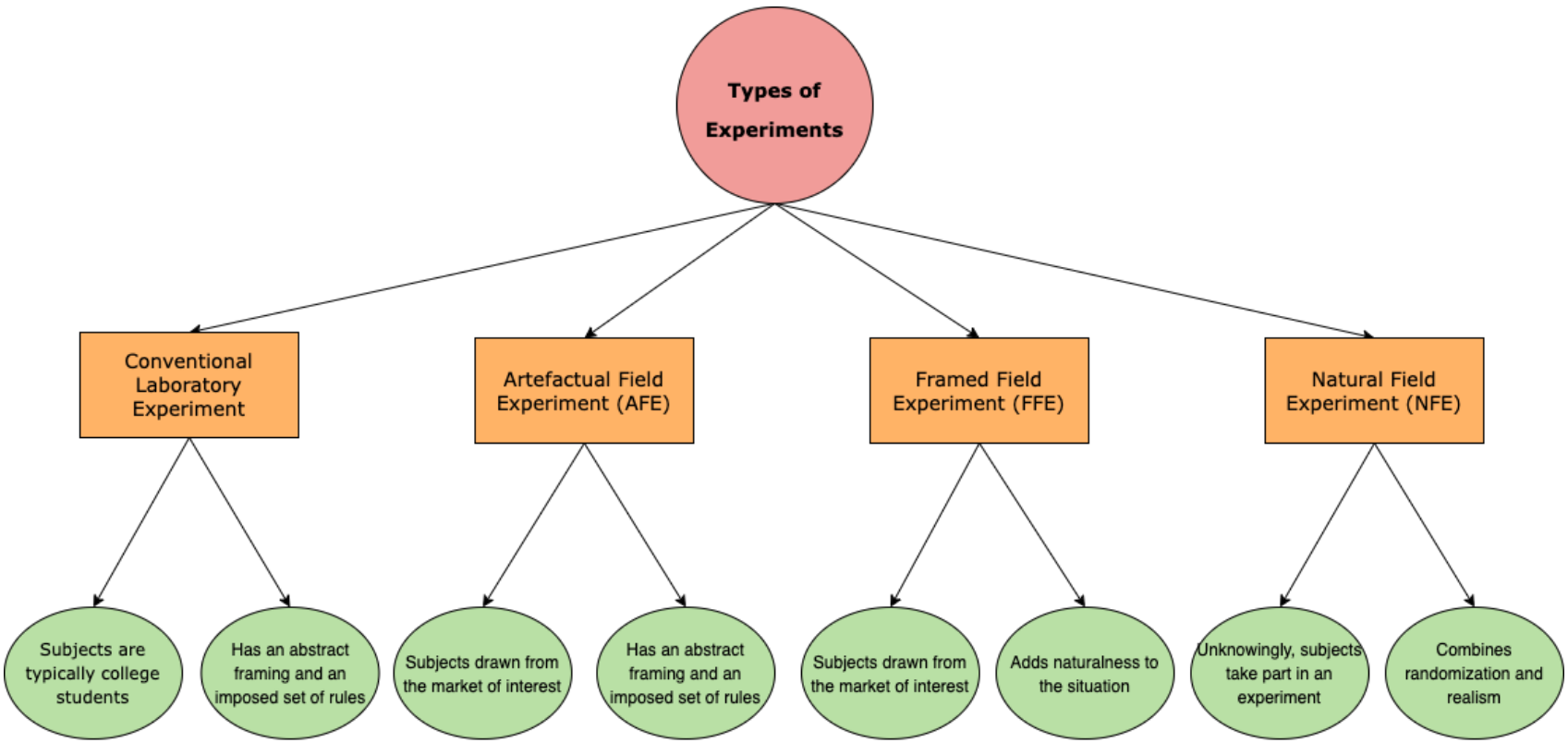
1. The use of experiments in economics has steadily grown over the past several decades, with both lab and field experiments now playing an important role in causal inference.

2. Economic experiments are unique in that the assignment mechanism is known and controlled by the researcher.

3. Field experiments, which can be categorized into three types (artefactual, framed, and natural), represent a unique bridge between laboratory and naturally occurring data.

4. Each type of economic experiment has strengths and weaknesses. Viewed as complements, they provide a blend of control and realism that is unparalleled, allowing social scientists a unique opportunity to provide fresh insights.

Mind's Eye


Types of Experiments
Conventional Laboratory Experiment
Artefactual Field Experiment (AFE)
Framed Field Experiment (FFE)
Natural Field Experiment (NFE)
Subjects are typically college students
Has an abstract framing and an imposed set of rules
Subjects drawn from the market of interest
Has an abstract framing and an imposed set of rules
Subjects drawn from the market of interest
Adds naturalness to the situation
Unknowingly, subjects take part in an experiment
Combines randomization and realism

Introduction

- Economists have turned to the experimental model of the physical sciences as a method to understand human behavior
- For example, in a laboratory, Thurstone (1931) asked respondents their preferences for a single or multiple sets of options in controlled experimental settings
  - In essence, he used **randomization** to show that it is possible to obtain information about a person's (stated or hypothetical) preference ordering that may not be available in revealed preference data
- Chamberlin (1948) tested competitive equilibrium in a lab setting but used randomly determined **induced values** to provide the necessary control to test the relevant theory
- These studies, combined with the seminal work of Fisher (1936), Neyman (1923), and others, set the wheels in motion for modern causal inference within experimental economics

Our Four Running Examples

1. Exley, Niederle, and Vesterlund (2020) recruit undergraduate students in a **conventional laboratory experiment** to explore negotiation behaviours between men and women

2. Callen, Isaqzadeh, Long, and Sprenger (2014) measure risk preferences on violence-exposed individuals with multiple price listing in an **artefactual field experiment (AFE)**

3. Bursztyn and Jensen (2015) employ a **framed field experiment (FFE)** to measure sign-up rates by randomizing whether choices were private or publicised

4. Pallais (2014) utilizes a **natural field experiment (NFE)** by randomly hiring workers and altering the depth of feedback to measure employment outcomes

---

Textbook Section 2.1

Our Four Running Examples - Summary

Type	Units	Control Condition	Treatment Conditions	Outcome
Exley, Niederle and Vesterlund (2020)	Lab	Undergraduate students at Stanford	Subjects choose whether to negotiate	Subjects are forced to negotiate
Callen et al. (2014)	AFE	Afghan civilians likely exposed to violence	Describing an event from past year	Describing a fear-inducing event Describing a happy event
Bursztyn and Jensen (2015)	FFE	Students from four large public high schools in Los Angeles	Sign-up decision is private	Sign-up decision is public
Pallais (2014)	NFE	oDesk workers	Not hired by experimenter	Hired by experimenter; detailed or course feedback Subsequent employment

Outline

The Empirical Approach in Economics

- Let's begin by assuming that individual i has covariates x_i, and the measurement approach consists of the following stages:
  - Z is the assignment mechanism and let z_i denote assignment to treatment
  - D is the unit-specific treatment and let d_i be the realized treatment status, which is the treatment individual i actually receives (e.g. $d_i = 1$ if unit i actually takes up treatment)
  - Y is the observed outcome and let y_{i1} be the outcome of interest when treatment status is $d_i = 1$, and y_{i0} when treatment status is $d_i = 0$
  - P_i is an indicator variable for the participation decision, such that $P_i = 1$ if subject i chooses to participate in the experiment, and $P_i = 0$ otherwise

---

Textbook Section 2.2

The Empirical Approach in Economics

- We follow the potential outcomes framework, in which an individual i has outcome y_{i1} in the treated group and y_{i0} in the control group
- Ideally, when conducting an experiment or examining naturally-occurring data, researchers could measure individual treatment effects for each individual i, $y_{i1} - y_{i0}$, which is the difference in outcomes for individual i being in the treated versus the control group
- In practice, we cannot observe both of these outcomes – we can only observe individual outcomes in one of the treated states, and the counterfactual outcome in the other state remains unobserved
- Thus, instead of individual treatment effects, researchers usually consider the **Average Treatment Effect (ATE)**, given by:

$$\tau^* = \mathbb{E}[y_{i1} - y_{i0}]$$

The Empirical Approach in Economics

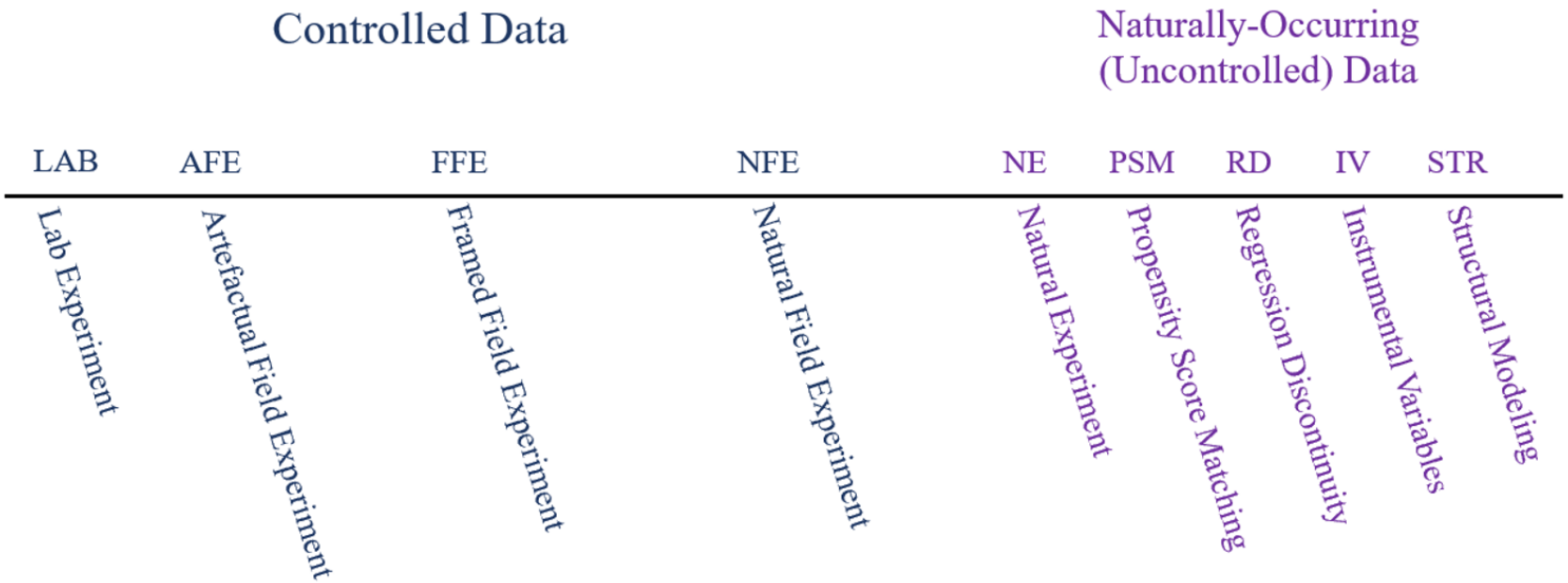
- The ATE measures the average difference in the outcomes for the population

- The ATE, τ^*, is not directly observable either; instead, researchers recover τ, defined as:

$$\tau = \mathbb{E}[y_{i1}|d_i = 1] - \mathbb{E}[y_{i0}|d_i = 0]$$

- We return to this topic in greater detail in chapters 3 and 4, but use it here to illustrate how the various empirical approaches achieve identification and the various assumptions underlying causal inference and estimation

Data Generation and Modeling for Causal Inference



[Figure 2.1 in the Textbook]

The Empirical Approach in Economics

- Figure 2.1 provides an empirical spectrum from approaches where the researcher knows and controls the assignment mechanism (Lab, AFE, FFE, and NFE), to empirical approaches where the assignment mechanism is neither under the control nor known by the researcher (NE, PSM, RD, IV, and STR)

- We categorize these as "controlled" and "uncontrolled" because the researcher knows and controls the assignment mechanism in one case but not in the others

  - Each of these approaches considers a unit i that has a potential outcome $Y_i(D_i)$ under treatment $D \in 0, 1$. The treatment effect for unit i can then be measured as $\tau_i = Y_i(1) - Y_i(0)$

  - The major problem, however, is one of a missing counterfactual: τ_i is unknown because for any individual, we can only observe $Y_i(1)$ or $Y_i(0)$. If we could observe the outcome for an untreated observation, had it been treated, then there is no evaluation problem

The Empirical Approach in Economics

- Natural experiments consider policies that are administered for some group, but not another. The treatment effect on the group on which the policy is administered is given by the change in the outcome for that group over some time period relative to the group without the policy
- Assuming two time periods, $T_i \in \{0, 1\}$, this can be expressed more formally as:

$$\mathbb{E}[\tau_i | D_i = 1] = \mathbb{E}[Y_{i1} - Y_{i0} | D_i = 1] - \mathbb{E}[Y_{i1} - Y_{i0} | D_i = 0]$$

- The major identifying assumption in difference-in-differences estimation is that there are no time-varying, unit-specific shocks to the outcome variable that are correlated with treatment status, and that selection into treatment is independent of temporary individual-specific effect
- If these assumption do not hold, difference-in-differences estimation does not recover a causal effect

The Empirical Approach in Economics

- The intuition behind propensity-score matching (PSM) is that if the researcher can select observable factors so that any two individuals with the same value for these factors will display homogeneous responses to the treatment, then the treatment effect can be measured without bias
  - One can use statistical methods to identify which two individuals are "more homogeneous lab rats" for the purposes of measuring the treatment effect
- The advocated solution is to find a vector of pre-determined covariates, X_i, such that the treatment of interest is as good as randomly assigned conditional on the propensity score given those covariates

The Empirical Approach in Economics

- The regression discontinuity method applies to settings where there is some continuous "running variable" which changes the administration of treatment discontinuously at some cutoff

- Under an assumption that potential outcomes change continuously at the cutoff, the regression approach is able to identify the treatment effect for those near the cutoff

- Problems of statistical power and the possibility of other factors changing at the cutoff can frustrate the researchers ability to uncover causal estimates of treatment effects

The Empirical Approach in Economics

- Instrumental variables (IV) draws from approaches to structural econometric methods
- The crux of the IV approach is to find a variable that is excluded from the outcome equation, but which is related to treatment status and has no direct association with the outcome
- The weakness of the IV approach is that such variables do not often exist, or that implausible assumptions must be maintained for them to be used to identify the treatment effect of interest
- When considering field experiments, one can view randomization itself as the IV

The Empirical Approach in Economics

- Structural modeling often entails a heavy mix of identifying restrictions, impose structure on technology and preferences, and simplifying assumptions about equilibrium outcomes

- Perhaps the best-known class of such structural models is computable general equilibrium models, which have been extensively applied to evaluate trade policies, for example

- Such structural models typically rely on complex estimation strategies, but yields structural parameters that are well-suited for ex-ante policy simulation, provided one undertakes systematic sensitivity analysis of those parameters

Outline

Experiments in Economics

- In this section, we discuss laboratory experiments as well as field experiments, including:
  - artefactual field experiments,
  - framed field experiment, and
  - natural field experiment

Textbook section 2.3

Laboratory Experiments

- Within economics, a **conventional laboratory experiment** is one that employs a standard subject pool of students, an abstract framing, and an imposed set of rules, such as Exley et al. (2020)
- The experimenter's goal is to create a small-scale environment in the laboratory where adequate **control** is maintained
- In the traditional lab experimental literature, an economic environment consists of a set of agents whereby each agent is described by a utility function, a technology or knowledge endowment, and a commodity endowment
- To complete the microeconomic environment, the experimenter specifies the institutional setting which includes the appropriate message space, the allocation rules, and other relevant characteristics of the specific institution of interest

Laboratory Experiments

- The experimental system thus is composed of the microeconomic environment and the institution
- Agents, who are assumed to possess consistent preferences and to make decisions to maximize their own well-being, choose messages, and the institution determines allocations via the governing rules
- To measure reliably the behavioral principles between preferences, institutions, and outcomes, experimenters have proposed a set of sufficient conditions for a valid controlled microeconomic experiment (Wilde, 1981 Smith, 1982):

    1. **Nonsatiation**: more money is preferred to less
    2. **Salience**: actions are linked to rewards
    3. **Dominance**: the reward structure dominates subjective costs
    4. **Privacy**: each subject is given his/her own payoff structure
    5. **Parallelism**: behavior observed in the lab applies in the extra-lab world under ceteris paribus conditions

Laboratory Experiments

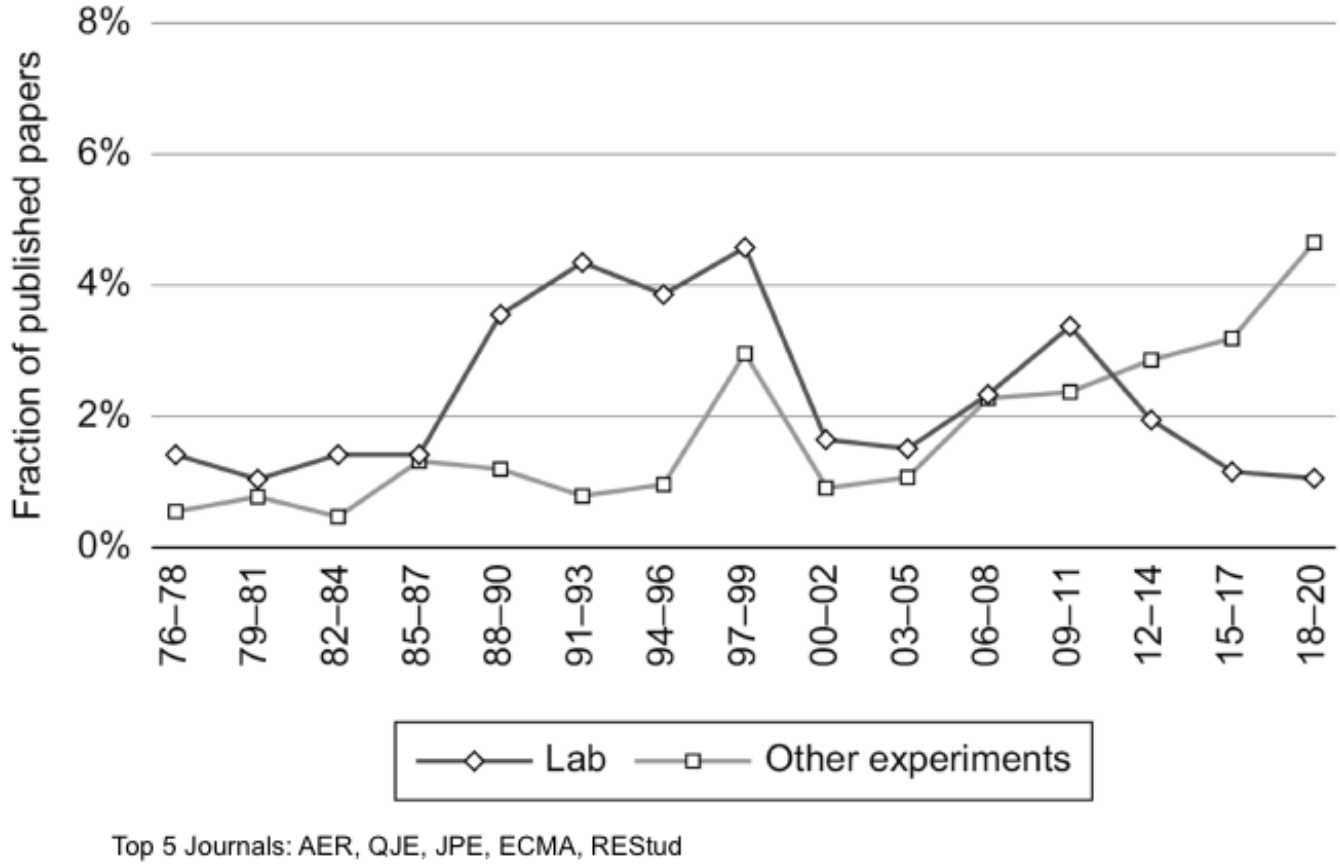
- Lab experiments provide the investigator with a means to directly influence the set of prices, budget sets, information set, and actions available to actors, and thus measure the impact of these factors on behavior within the context of the laboratory

- Individual demand and supply curves tend to be invisible in the real world, but they come alive when the lab experimentalist uses induced values to construct a market to test if Adam Smith's invisible hand is as powerful as most surmise

- In this manner, randomization of induced values acts as an instrumental variable and therefore allows the analyst to make strong causal statements within the domain of study

- This particular attractiveness helped to make laboratory experiments a "boom industry" in the late 20th century

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



[Figure 2.2 in the Textbook]

Laboratory Experiments

- As Holt (2006) documents, academic publications using the methodology were almost non-existent until the mid-1960s, surpassed 50 annually for the first time in 1982, and by 1998 exceeded 200 per year

- Figure 2.2 shows similar growth over this period within the Top 5 Economics journals

- While the growth in this set of journals has slowed since the early years, there remains important representativeness of lab experiments in the first few decades of the 21st century

Laboratory Experiments

- Early on, there were many explorations of theory testing based on whether neoclassical theory organized choice and data across various institutions, such as a Chamberlin market, double-oral auction, or various one-sided auctions

- The late 1980s and 1990s ushered in many laboratory approaches measuring preferences and exploring novel tests of more behavioral theories than the early years witnessed

- In this manner, the literature took a decided turn from a primary focus on markets to now include more attention on preference elicitation and measurement; for example using a class of experiments to measure pro-social preferences

Summary of Experimental Approaches to Measure Social Preferences

Name of game	Summary	Typical finding	Social preference interpretation
Ultimatum game	A two-stage game where two people, a proposer and a responder, bargain over a fixed amount of money. In the first stage, the proposer offers a split of the money, and in the second stage, the responder decides to accept or reject the offer. If accepted, each player receives money according to the offer; if rejected, each player receives nothing.	Proposer: Majority of offers in the range of 25–50% of fixed endowment. Few offers below 5%. Responder: Frequently reject offers below 20% of fixed endowment.	Proposer: Fairness preferences Responder: Inequity aversion, negative reciprocity
Dictator game	A variant of the ultimatum game: strategic concerns are absent as the proposer simply states what the split will be and the proposer has no veto power, rendering the proposed split as effective.	Usually more than 60% of subjects offer a positive amount, though the average transfer is only about 20% of the endowment.	Altruism; fairness preferences, such as inequity aversion.
Trust game	A sequential prisoner's dilemma game wherein the first mover decides how much money to pass to the second mover. All money passed is increased by a factor, $f > 1$, and the second mover then decides how much money to return to the first mover. In this light, the second mover is a dictator who has been given his endowment by the first mover.	Proposer: Typically transfer about 50% of endowment. Responder: Typically return about 50% of received transfer. Average repayment rate is nearly 50% of transfer.	Proposer: Trust; positive reciprocity Responder: Trustworthiness, positive reciprocity

[Table 2.2 in the Textbook]

Summary of Experimental Approaches to Measure Social Preferences (Cont.)

Name of game	Summary	Typical finding	Social preference interpretation
Gift exchange game	Similar to the trust game, but the money passed by the first mover (often labeled the "wage" or "price" offer) is not increased by a factor; rather it represents a pure lump-sum transfer. Also, the first mover requests a desired effort, or quality, level in return for the "wage" or "price" offer. The second mover then chooses an effort or quality level that is costly to provide, but increases the first mover's payoff.	Proposer: "Wage" or "price" offer is typically greater than the minimum allowed. Responder: Effort or quality increases in response to "wage" or "price" offer.	Proposer: Trust; future positive reciprocity Responder: Trustworthiness, positive reciprocity
Public goods game	Generalization of the prisoner's dilemma game in that group members decide simultaneously how much to invest in the public good. The payoff function is given by $p_i = e_i + b\sum e_j$, where e_i is the initial endowment, g_i is the level of tokens that subject i places in the group account, b is the marginal payoff of the public good, and $\sum g_j$ is the sum of the individual contributions to the public good. By making $0 < b < 1 < nb$, the dilemma follows.	Players' contribution to public good is roughly 50% of endowment in one-shot games. Many players' contributions unravel to approach 0% in latter rounds of multi-period games.	Altruism; fairness preferences, conditional reciprocity

[Table 2.2 in the Textbook]

Field Experiments

- Similar to the spirit of the early laboratory experiments, field experiments typically use randomization to achieve identification

- However, different from laboratory experiments, field experiments vary the subject pool, oftentimes occur in the *natural environment* of the agent being observed, and in many cases examine behaviors that cannot be reasonably distinguished from the tasks the agent has entered the marketplace to complete

- Levitt and List (2009) document three distinct waves of field experimentation within economics

Field Experiments

1. The first wave, the dawn of field experimentation, is rarely considered to be part of the field experimental genre in economics

   - Considering that none of these studies were experiments with human subjects, and few were published in economics journals, this is understandable

   - Nevertheless, the work of Fisher and Neyman in the 1920s and 1930s is worthwhile for three reasons:

     1.1 Answered important economic questions regarding agricultural productivity

     1.2 They are generally believed to be the first to conceptualize randomization as a key element of the experimental method

     1.3 Neyman's (1923) "potential yields" framework set the stage for the Rubin Causal Model, a framework that considerably improved the language and reasoning of causal inference

Field Experiments

2. The second wave is the latter half of the 20th century, during which government agencies conducted a series of large-scale social experiments

   - This second stage was important as the early 1970s ushered in the work of Donald Rubin (1974), who put the language and framework of potential outcomes front and center for causal inference

   - Whereas Neyman (1923) focused on completely randomized experiments, Rubin (1974) made it the centerpiece of all empirical work, experimental or non-experimental

   - Rubin (1974) also clarified the import of the assignment mechanism in the potential outcomes framework

Field Experiments

3. The third wave of field experimentation is the surge of field experiments in economics in the past few decades

   - This most recent movement approaches field experiments by taking the tight controls of the lab to the field
   - The analyst bridges laboratory and naturally-occurring data by systematically relaxing the controls inherent in a laboratory experiment
   - Harrison and List (2004) propose experimental features that can be used to determine the field context of an experiment. Leveraging these features, we can discuss a broad classification scheme that helps to organize one's thoughts about the factors that might be important when moving from the lab to the field

Seven Criteria that Define Field Experiments

- The seven criteria are:
  - *Experimental Subjects: Population and Selection*
    1. The nature of the subject pool
    2. The nature of the information that subjects bring to the experimental task
    3. Selection into the experiment
  - *Experimental Environment*
    4. The nature of the commodity in the experiment
    5. The nature of the experimental task
    6. The nature of the stakes
    7. Experimental proclamation

Seven Criteria that Define Field Experiments

- Using these criteria, Harrison and List (2004) propose the following taxonomy to classify the field context of experiments:

1. An **artefactual field experiment (AFE)** is a conventional lab experiment, but with a non-standard subject pool

2. A **framed field experiment (FFE)** is an artefactual field experiment, but with field context in either the commodity, task, or information that the subjects can use

3. A **natural field experiment (NFE)** is a framed field experiment, but where the environment is one where the subjects naturally undertake these tasks and where the subjects do not know that they are taking part in an experiment

What Parameters do the Various Experimental Types Recover?

- We can collapse the seven experimental features described above into four factors to describe the distinct parameters that the experimental types recover
  - Let S=1 denote the population of interest and S=0 be standard student samples
  - Let E=1 denote a natural environment and E=0 denote an artificial environment
  - Let A=1 denote that subjects are aware they are part of an experiment and A=0 denote that subjects are not aware of being an experimental participant
- We use the above notation, along with the participation notation introduced above ($P = 0, 1$), to summarize the differences between each type of experiment in Table 2.3

Summary of four key characteristics of each experimental type

	Laboratory	**AFE**	**FFE**	**NFE**
Population	College Students $S = 0$	Population of Interest $S = 1$	Population of Interest $S = 1$	Population of Interest $S = 1$
Environment	Artificial $E = 0$	Artificial $E = 0$	Natural $E = 1$	Natural $E = 1$
Awareness	Overt $A = 1$	Overt $A = 1$	Overt $A = 1$	Covert $A = 0$
Selection/Who do we observe?	Those Sorting into Experiment $P = 1$	Those Sorting into Experiment $P = 1$	Those Sorting into Experiment $P = 1$	All in Market $P = 0$ and $P = 1$

[Table 2.3 in the Textbook]

What Parameters do the Various Experimental Types Recover?

- In terms of the notation introduced in Table 2.3, we can further summarize the parameters estimated from each type of overt experiment as follows:

$$\tau^{lab} \equiv \mathbb{E}[\tau_i \mid S = 0, E = 0, A = 1, P = 1]$$
$$\tau^{AFE} \equiv \mathbb{E}[\tau_i \mid S = 1, E = 0, A = 1, P = 1]$$
$$\tau^{FFE} \equiv \mathbb{E}[\tau_i \mid S = 1, E = 1, A = 1, P = 1]$$

- The defining distinction of NFEs compared to the other three experimental types is that there is no experimental selection
- Subjects are unaware that they are participating in an experiment
- In other words, there is no $P = 1$ and $P = 0$ groups because subjects in NFEs are drawn from the population of interest and are randomized into treatment and control without their knowledge

What Parameters do the Various Experimental Types Recover?

- While, NFEs rid us of the $P = 0$ issue, similar to the other experimental types, NFEs still have one source of selection: the decision to be in the naturally-occurring market where the NFE is conducted
- Depending on the characteristics of the market and the goals of the researchers and policy makers, the concerns that stem from this type of selection can vary
- The lack of an experimental proclamation also avoids the interference of the behavioral effects of the experimental proclamation (behavioral changes due to knowing one is in an experiment)
  - Behavioral biases in experiments include the **Hawthorne Effects** (or **Pygmalion Effects**), the **John Henry Effects** and the **resentful demoralization**

What Parameters do the Various Experimental Types Recover?

- The Hawthorne effect, sometimes referred to as experimenter effects, alludes to the notion that subjects might alter their behavior as a result of being aware that they are being observed
- In other words, subjects might modify their behavior because they know that their actions are being recorded and analyzed as part of an experiment
- The John Henry effect is the bias introduced into experiments when members of the control group are aware that they are being compared to the experimental group and react by trying harder
- On the flip side of this argument is resentful demoralization. In this case, subjects do the opposite: emotions resulting from the knowledge of being disadvantaged spill over into a negative performance change

What Parameters do the Various Experimental Types Recover?

- With all of this in mind, we can write the parameter estimated from NFEs as:

$$\tau^{NFE} \equiv \mathbb{E}[\tau_i \mid S = 1, E = 1, A = 0]$$

where the treatment effect is now independent of the participation decision and behavioral effects of taking part in an experiment are muted

- This means that NFEs avoid biases arising both from selection into treatment and selection into the experiment

What Parameters do the Various Experimental Types Recover?

- However, it should not be assumed that these characteristics are always desirable
- The type of research or policy questions should inform which type of estimator provides the most relevant test in the field
- Yet, NFEs do provide what we are after from the beginning: the average treatment effect for the entire population

$$\tau^{NFE} \equiv \mathbb{E}[\tau_i \mid S = 1, E = 1, A = 0] = \tau = \mathbb{E}[Y_i(1) - Y_i(0)] \quad (2.5)$$

Outline

What Experimental Type to Choose?

- The problem for the individual researcher revolves around comparing marginal benefits and marginal costs of the various empirical approaches
- On the benefit side, researchers need to balance two key aspects that determine the value of their contribution to science and policy
- These two aspects follow naturally from EP1 and EP2 discussed in chapter 1

1. Making correct statistical inference, which includes internal validity (the property of being able to identify the parameters of interest in a given design), measurement, understanding mediators and moderators, and informativeness (how much a result can change the prior of the scientific community)
2. Generalizability (external validity): whether a causal relationship continues to hold when subjects, context, location, or treatment details are modified

---

Textbook section 2.4

Control Across the Experimental Spectrum for Identification Purposes

- In chapter 3 we will discuss the four exclusion restrictions necessary to ensure internal validity in randomized experiments, but for reference, they are:

1. **The Stable Unit Treatment Value Assumption (SUTVA):** any unit's potential outcomes do not vary with the treatments assigned to or undertaken by any other unit and that there is no variation in the form of treatment that leads to different potential outcomes
2. **Observability (non-attrition):** all participants in the study remain in the study, and for each, we observe their outcome, treatment assignment, actual treatment, and unit-level characteristics
3. **Compliance:** assumes that every unit assigned to $Z_i = z$ ends up taking $D_i = z$
4. **Statistical independence:** the assignment mechanism governing the D_i is independent of potential outcomes (i.e., the researcher controls the assignment mechanism)

Control Across the Experimental Spectrum for Identification Purposes

- An initial consideration is the properties of the different types of experiments from the point of view of ensuring internal validity
  1. Viewed through the lens of the four exclusion restrictions, all of the experimental types have utility over their naturally-occurring data counterparts because the researcher knows and controls the assignment mechanism. This key feature is shared, and therefore isomorphic, across the four experimental types
  2. Lab and AFEs typically, but not always, have the best control over SUTVA, Observability, and Compliance
     - This is because, by design, they are set up in a tightly controlled environment where interference, attrition, and non-compliance are easier to minimize
     - The short-run substitution effect estimates recovered within a tightly controlled environment, which most lab and AFEs have a comparative advantage to approximate, are the resulting yield
     - This is a keen advantage for most lab and AFEs, as will be discussed in chapter 19

Control Across the Experimental Spectrum for Measurement Purposes

- A second major consideration across the empirical spectrum relates to measurement
- As discussed in chapters 4 and 5, the ability to generate more precise estimates from a given sample size is a key concern in experimental design
- While some view the traditional subject pools of undergraduate students in lab experiments as a distinct drawback, in terms of experimental power it is a distinct feature. This is because of the ability of the researcher to make more precise empirical statements due to the lower variance, discussed further in chapter 5
- Researchers also need to consider the resource cost of conducting each type of experiment, including the monetary and logistical costs and the opportunity cost of doing other types of research

Control Across the Experimental Spectrum for Measurement Purposes

- As discussed in chapter 19, lab experiments typically cost less per unit of observation than field experiments
- In fact, without organizational partnerships, experimental cost will tend to monotonically increase when moving from West to East in Figure 2.1
  - There are exceptions. For instance, FFEs and NFEs can be much more cost effective when researchers partner with governmental agencies, firms, non-profits, and NGOs, creating win–win partnerships
  - Moreover, the unit cost per subject can be reduced in field experiments due to economies of scale
- Consequently, researchers often begin by exploring questions using lower-cost laboratory experiments, and later move into the field to replicate their initial results in a more diverse environment and population
  - This is the "low cost wind tunnel" approach

The Ability to Replicate Across the Experimental Spectrum

- A third consideration across the empirical spectrum relating to the benefits of general knowledge creation is the ability to replicate the received results
- As discussed in chapter 15, there are several types of replication:
  1. **Pure replication:** examine the same question and model using the underlying original data set
  2. **Robustness Analysis:** use the exact same data as the original analysis but modify the data or the empirical methods to examine if the results are robust to reasonable changes to the original analysis
  3. **Same population replication:** follow the original protocol and draw a new sample from the same subject population
  4. **Similar population replication:** target a similar population, such as the $P = 0$ group or uses a closely related population
  5. **Disparate population replication:** examine the same question and model using a population dissimilar from the original experiment
  6. **Conceptual replication**: test the hypotheses of the original study using a new research design

The Ability to Replicate Across the Experimental Spectrum

- Under these definitions, four of the replication types are roughly isomorphic in terms of difficulty in achieving across the experimental types (pure replication, robustness analysis, disparate population replication, and conceptual replication)

- Alternatively, for the other two replication types, laboratory experiments and AFEs usually have a distinct advantage in that their ease of replicability is a keen source of inferential prowess

- This follows from the fact that lab experiments, for example, can offer a more portable protocol than many NFEs, and it is usually easier to maintain consistent experimental conditions with more precision

Control Across the Experimental Spectrum for Inferential Purposes

- When taken to the lab, the experimentalist usually invokes the assumptions of the theory and tests various predictions
- Economists have used induced values, for example, to test theory. In this sense, NFEs present a trade-off to the control of a laboratory experiment or an AFE in that induced-valuations or robots guaranteed to play equilibrium strategies against human subjects are by definition not present in NFEs
- FFEs and NFEs, therefore, may not be "clean theory testers" compared to laboratory experiments or AFEs on this dimension, where researchers have superior control by inducing preferences to accord with theoretical assumptions and excluding other complicating factors
- In this manner, FFEs and NFEs often represent a joint test of the theoretical predictions and whether the theoretical assumptions were in place naturally

Control Across the Experimental Spectrum for Inferential Purposes

- Thus, FFEs and NFEs can present a useful check on the robustness of laboratory results in a natural setting, where the mathematical assumptions of the theory cannot necessarily be guaranteed to hold
- A related inferential point revolves around the fact that the tools social scientists use to measure parameters of interest are constructed to explore human choice and are therefore subject to human vagaries not admitted in the typical physical science experiment
- Thus, we face a universal question of construct validity – are we measuring what we think we are measuring?
- When we run an experiment in a chosen context, we have theory about how people behave in that context and believe that the variables we measure correspond to their namesakes in our theoretical constructs. In testing our theories, we need to be sure that we have the right theory for the context

Control Across the Experimental Spectrum for Inferential Purposes

- An aspect of the lab over which experimenters have incomplete control is that subjects may not be playing the game that the experimenter intends
- On the other end of the empirical spectrum, in an NFE the market has provided the ecological validity, and randomization is used to identify the treatment effect of interest
- In this way, the researcher takes the field environment as given and does not construct a new setting or develop a new interpretation of the existing market. In this case, it is less likely that the experimenter's labels are incorrect or that she did not have a representative design
- Thus, results from the laboratory can serve as an important complement to field insights when construct validity is a concern

Control Across the Experimental Spectrum for Inferential Purposes

- One simple, informal, approach to detect potential construct validity issues in their own work as well as in others' studies is to first place yourself in the shoes of the experimental subject and immediately turn to the experimental instructions
  - We call this the "construct validity pre-read."
- That is, before reading even the abstract (or title) of the study, we advise reading the experimental instructions
- Then, we urge a thought experiment to determine how you would interpret the received data. If your interpretation after reading the instructions matches the scenario the experimenter advocates, then you can be more confident about this aspect of the study

Control Across the Experimental Spectrum to Ensure External Validity

- Beyond EP1, it is also important to understand external validity (EP2): predicting whether the causal impacts of treatments implemented in one environment transfer to other environments, be them spatially, temporally, or scale differentiated
- AFEs, FFEs, and lab experiments have a two-stage selection process whereby subjects first decide what market or what university they will attend, if any. Then they then decide whether to participate in the study
- In NFEs, there is one stage of selection: selection into the market of interest. There is no second-stage selection, so the ATE is the ATE of the population of interest
- NFEs therefore offer an advantage in that they better control the experimental sample, therefore avoiding **randomization bias** and **participation bias**

Control Across the Experimental Spectrum to Ensure External Validity

- This issue can be especially important if researchers want to generalize to a population of others or scale up their proposed program. For lab, AFEs, and FFEs, one must take caution that the recovered parameter applies to the underlying population due to selection concerns (i.e., does $\tau_{p_1=1} = \tau_{p_1=0}$?)
- In chapter 16 we identify four potential threats to generalizability that vary across the experimental spectrum in Figure 2.1
- The general conclusion is that overt experiments, in which subjects are aware of being part of an experiment, have characteristics that make them particularly prone to the various threats to generalizability
- In this sense, contrary to conventional wisdom, NFEs have more control, not less, compared to lab, AFEs, and FFEs when it comes to understanding the subject pool and selection

Outline

What have we learned?

- This chapter highlights that when the researcher moves the exploration from the lab to an NFE, the potential loss in some types of control comes in exchange for increased realism and gains in other types of control

- NFEs match the real-world settings that theory attempts to explain: players competing for real goods rather than explicit cash values, they are not told explicitly the distributions of other's valuations, and they are likely to have previous experience

- That is, NFEs have the virtue of resembling natural economic phenomena as closely as possible

- Alternatively, the lab, AFEs, and FFEs are quite useful to test the efficacy of a policy or a program that is voluntary in nature

What have we learned?

- Ultimately, each type of experiment serves different purposes, and each recovers a different parameter due to both selection and behavioral effects
- In this manner, the various experimental types offer keen complementarities in the production of scientific knowledge
- This chapter identifies several features that the researcher should consider when deciding on where they want to augment the economic science using the experimental approach
- The decision naturally comes down to comparing marginal benefits and marginal costs
- Yet, in the end, aggregating results from qualitatively different empirical methods across different settings, subject pools, and selection rules will increase our confidence in building the knowledge framework necessary for testing theories and providing empirical advice

References I

Al-Ubaydli, O. and J. A. List (2013). On the generalizability of experimental results in economics: With a response to camerer. NBER Working Paper 19666, National Bureau of Economic Research.

Al-Ubaydli, O. and J. A. List (2015). Do natural field experiments afford researchers more or less control than laboratory experiments? *American Economic Review 105*(5), 462–466.

Andreoni, J. and J. Miller (2002). Giving according to GARP: An experimental test of the consistency of preferences for altruism. *Econometrica 70*(2), 737–753.

Angrist, J. D. and A. B. Krueger (1992). The effect of age at school entry on educational attainment: An application of instrumental variables with moments from two samples. *Journal of the American Statistical Association 87*(418), 328–336.

Bacon, F. (1620). *Novum Organum*. Oxford: Clarendon Press.

Berg, J., J. Dickhaut, and K. McCabe (1995). Trust, reciprocity, and social history. *Games and Economic Behavior 10*(1), 122–142.

Bohm, P. (1972). Estimating the demand for public goods: An experiment. *European Economic Review 3*(2), 111–130.

Bohm, P. (1979). Estimating willingness to pay: Why and how? *The Scandinavian Journal of Economics 81*(2), 142–153.

Bohm, P. (1984a). Are there practicable demand-revealing mechanisms? *Public Finance and the Quest for Efficiency 19*, 185–200.

Bohm, P. (1984b). Revealing demand for an actual public good. *Journal of Public Economics 24*(2), 135–151.

Bohm, P. (1994). Behavior under uncertainty without preference reversal: A field experiment. In *Experimental Economics*. Heidelberg: Physica-Verlag.

Bolton, G. E. and A. Ockenfels (2000). A theory of equity, reciprocity, and competition. *The American Economic Review 90*(1), 166–193.

Brown, M., A. Falk, and E. Fehr (2004). Relational contracts and the nature of market interactions. *Econometrica 72*(3), 747–780.

Brunswik, E. (1952). *The Conceptual Framework of Psychology*. International Encyclopedia of Unified Science. Chicago: University of Chicago Press.

Brunswik, E. (1956). *Perception and the Representative Design of Psychological Experiments* (2 ed.). Berkeley, CA: University of California Press.

Bursztyn, L. and R. Jensen (2015). How does peer pressure affect educational investments? *The Quarterly Journal of Economics 130*(3), 1329–1367.

Callen, M., M. Isaqzadeh, J. D. Long, and C. Sprenger (2014). Violence and risk preference: Experimental evidence from Afghanistan. *American Economic Review 104*(1), 123–148.

Camerer, C. (2011). The promise and success of lab-field generalizability in experimental economics: A critical reply to Levitt and List. Available at SSRN.

Camerer, C. and K. Weigelt (1988). Experimental tests of a sequential equilibrium reputation model. *Econometrica 56*(1), 1–36.

References II

Cárdenas, J.-C. and E. Ostrom (2004). What do people bring into the game? Experiments in the field about cooperation in the commons. *Agricultural Systems 82*(3), 307–326.

Chamberlin, E. H. (1948). An experimental imperfect market. *Journal of Political Economy 56*(2), 95–108.

Charness, G. (1997). Attribution and reciprocity in a simulated labor market: An experimental investigation. Technical report, Universitat Pompeu Fabra, Department of Economics and Business.

Charness, G., G. R. Frechette, and J. H. Kagel (2004). How robust is laboratory gift exchange? *Experimental Economics 7*(2), 189–205.

Charness, G. and M. Rabin (2002). Understanding social preferences with simple tests. *The Quarterly Journal of Economics 117*(3), 817–869.

Chetty, R., N. Hendren, and L. F. Katz (2016). The effects of exposure to better neighborhoods on children: New evidence from the moving to opportunity experiment. *American Economic Review 106*(4), 855–902.

Cook, T. D. and D. T. Campbell (1979). The design and conduct of true experiments and quasi-experiments in field settings. In R. T. Mowday and R. M. Steers (Eds.), *Research in Organizations: Issues and Controversies*. Santa Monica, CA: Goodyear Publishing Company.

Cunningham, S. (2021). *Causal Inference: The Mixtape*. New Haven: Yale University Press.

Exley, C. L., M. Niederle, and L. Vesterlund (2020). Knowing when to ask: The cost of leaning in. *Journal of Political Economy 128*(3), 816–854.

Fehr, E. and A. Falk (1999). Wage rigidity in a competitive incomplete contract market. *Journal of Political Economy 107*(1), 106–134.

Fehr, E., S. Gächter, and G. Kirchsteiger (1997). Reciprocity as a contract enforcement device: Experimental evidence. *Econometrica 65*, 833–860.

Fehr, E., G. Kirchsteiger, and A. Riedl (1993). Does fairness prevent market clearing? An experimental investigation. *The Quarterly Journal of Economics 108*(2), 437–459.

Fehr, E. and J. A. List (2004). The hidden costs and returns of incentives—trust and trustworthiness among CEOs. *Journal of the European Economic Association 2*, 743–771.

Fehr, E. and K. M. Schmidt (1999). A theory of fairness, competition, and cooperation. *The Quarterly Journal of Economics 114*, 817–868.

Fisher, R. A. (1935). *The Design of Experiments*. Edinburgh: Oliver and Boyd.

Friedman, M. (1953). The methodology of positive economics. In *Essays In Positive Economics*, pp. 3–16. Chicago: University of Chicago Press.

Gächter, S. and A. Falk (2002). Reputation and reciprocity: Consequences for the labour relation. *The Scandinavian Journal of Economics 104*(1), 1–26.

References III

Gallen, T. (2020). Broken instruments. Available at SSRN.

Grether, D. M. and C. R. Plott (1979). Economic theory of choice and the preference reversal phenomenon. *The American Economic Review 69*(4), 623–638.

Güth, W., R. Schmittberger, and B. Schwarze (1982). An experimental analysis of ultimatum bargaining. *Journal of Economic Behavior & Organization 3*(4), 367–388.

Hannan, R. L., J. H. Kagel, and D. V. Moser (2002). Partial gift exchange in an experimental labor market: Impact of subject population differences, productivity differences, and effort requests on behavior. *Journal of Labor Economics 20*(4), 923–951.

Harrison, G. W. and J. A. List (2004). Field experiments. *Journal of Economic Literature 42*(4), 1009–1055.

Holt, C. A. (2006). *Markets, Games, & Strategic Behavior: Recipes for Interactive Learning*. Boston: Pearson Addison Wesley.

Imbens, G. W. and D. B. Rubin (2015). *Causal Inference for Statistics, Social, and Biomedical Sciences: An Introduction*. Cambridge: Cambridge University Press.

Kahneman, D., J. L. Knetsch, and R. Thaler (1986). Fairness as a constraint on profit seeking: Entitlements in the market. *The American Economic Review 76*(4), 728–741.

Ledyard, J. O. (1995). Public goods: A survey of experimental research. In J. Kagel and A. Roth (Eds.), *Handbook of Experimental Economics*, Volume 2. NJ: Princeton University Press.

Levitt, S. D. and J. A. List (2007). What do laboratory experiments measuring social preferences reveal about the real world? *Journal of Economic Perspectives 21*(2), 153–174.

Levitt, S. D. and J. A. List (2009). Field experiments in economics: The past, the present, and the future. *European Economic Review 53*(1), 1–18.

List, J. A. (2006). The behavioralist meets the market: Measuring social preferences and reputation effects in actual transactions. *Journal of Political Economy 114*(1), 1–37.

Mill, J. S. (1844). On the definition of political economy; and on the method of investigation proper to it. In *Essays on Some Unsettled Questions of Political Economy*.

Neyman, J. (1990). On the application of probability theory to agricultural experiments. Essay on principles. Section 9. *Statistical Science 5*(4), 465–472. Originally published 1923; translated by Dorota M. Dabrowska and T. P. Speed.

Nobel Prize Outreach AB (2009). The Sveriges Riksbank prize in economic sciences in memory of Alfred Nobel 2009. `https://www.nobelprize.org/prizes/economic-sciences/2009/summary/`.

Nobel Prize Outreach AB (2012). The Sveriges Riksbank prize in economic sciences in memory of Alfred Nobel 2012. `https://www.nobelprize.org/prizes/economic-sciences/2012/summary/`.

References IV

Ortmann, A. (2016). Episodes from the early history of experimentation in economics. In A. Svorenčík and H. Maas (Eds.), *The Making of Experimental Economics: Witness Seminar on the Emergence of a Field*, pp. 195–217. Cham: Springer International Publishing.

Ostrom, E. and H. Nagendra (2006). Insights on linking forests, trees, and people from the air, on the ground, and in the laboratory. *Proceedings of the National Academy of Sciences 103*(51), 19224–19231.

Pallais, A. (2014). Inefficient hiring in entry-level labor markets. *American Economic Review 104*(11), 3565–3599.

Peirce, C. S. and J. Jastrow (1885). On small differences in sensation. *Memoirs of the National Academy of Sciences 3*(1), 75–83.

Rabin, M. (1993). Incorporating fairness into game theory and economics. *American Economic Review 83*(5), 1281–1302.

Reuben, E., S. X. Li, S. Suetens, A. Svorenčík, T. Turocy, and V. Kotsidis (2022). Trends in the publication of experimental economics articles. *Journal of the Economic Science Association 8*(1), 1–15.

Rigdon, M. L. (2002). Efficiency wages in an experimental labor market. *Proceedings of the National Academy of Sciences 99*(20), 13348–13351.

Robinson, J. (1977). What are the questions? *Journal of Economic Literature 15*(4), 1318–1339.

Roth, A. E. (2008). What have we learned from market design? *The Economic Journal 118*(527), 285–310.

Roth, A. E. and I. Erev (1995). Learning in extensive-form games: Experimental data and simple dynamic models in the intermediate term. *Games and Economic Behavior 8*(1), 164–212.

Roth, A. E., V. Prasnikar, M. Okuno-Fujiwara, and S. Zamir (1991). Bargaining and market behavior in Jerusalem, Ljubljana, Pittsburgh, and Tokyo: An experimental study. *The American Economic Review 81*(5), 1068–1095.

Rubin, D. B. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology 66*(5), 688–701.

Rubin, D. B. (1975). Bayesian inference for causality: The importance of randomization. *The Proceedings of the Social Statistics Section of the American Statistical Association*, 233–239.

Rubin, D. B. (1978). Bayesian inference for causal effects: The role of randomization. *The Annals of Statistics 6*(1), 34–58.

Samuelson, P. A. and W. D. Nordhaus (1985). *Economics* (12 ed.). New York: McGraw-Hill.

Sanbonmatsu, L., J. R. Kling, G. J. Duncan, and J. Brooks-Gunn (2006). Neighborhoods and academic achievement: Results from the moving to opportunity experiment. *Journal of Human Resources 41*(4), 649–691.

Siegel, S. and L. E. Fouraker (1960). *Bargaining and Group Decision Making: Experiments in Bilateral Monopoly*. New York, NY: McGraw-Hill.

Slonim, R., C. Wang, E. Garbarino, and D. Merrett (2013). Opting-in: Participation bias in economic experiments. *Journal of Economic Behavior & Organization 90*, 43–70.

References V

Smith, V. L. (1962). An experimental study of competitive market behavior. *Journal of Political Economy 70*(2), 111–137.

Smith, V. L. (1965). Experimental auction markets and the Walrasian hypothesis. *Journal of Political Economy 73*(4), 387–393.

Smith, V. L. (1982). Microeconomic systems as an experimental science. *The American Economic Review 72*(5), 923–955.

Smith, V. L. and B. J. Wilson (2019). *Humanomics: Moral Sentiments and the Wealth of Nations for the Twenty-First Century*. Cambridge, UK: Cambridge University Press.

Stigler, S. M. (1978). Mathematical statistics in the early states. *The Annals of Statistics 6*(2), 239–265.

Thurstone, L. L. (1931). The measurement of social attitudes. *The Journal of Abnormal and Social Psychology 26*(3), 249–269.

Viceisza, A. C. G. (2012). Treating the field as a lab: A basic guide to conducting economics experiments for policymaking. Food Security in Practice Technical Guide Series 7, International Food Policy Research Institute, Washington, DC.

Wallis, W. A. and M. Friedman (1942). The empirical derivation of indifference functions. In *Studies in Mathematical Economics and Econometrics in Memory of Henry Schultz*, pp. 175–189.

Wilde, L. L. (1981). On the use of laboratory experiments in economics. In J. C. Pitt (Ed.), *Philosophy in Economics*, pp. 137–148. Dordrecht: Springer Netherlands.