Working with Experiments

Andre Gray

Contents

1 Parts of an Experiment

- **Motivation:** What is the "big think" question we want to address?
- **Causal Research Question:** What is the specific research question you want to answer. What is the hypothesis you are testing, or the causal effect you intend to measure (Effect of A on B)
  - A good research question should be (1) motivated by prior qualitative or quantitative research, such as customer feedback (2) identify a cause and effect (3) answerable with measurable outcomes (4) relevant to a business/policy decision (ie. finding or not finding an effect may change a business/policy decision).
  - A research question should imply a specific hypothesis to be tested. For example, the research question "does wine consumption have an impact on the incidence of heart disease?" can lead you to a specific hypothesis H_1: "individuals that consume 1 or more glasses of wine per day on average are more likely to contract heart disease after 10 years relative to individuals who drink less than that amount".
- **Sample Population / Setting**: Who are you recruiting to your experiment, in what setting are you running the experiment. Here you want to outline the "who" of your experiment, including most crucially the sample size (the N of the experiment), which sets the parameters for your statistical power. We also want to know characteristics of your target population (representative sample of population, only adults, a particular ethnic group), we want to know the units you're interested in (households, neighborhoods, cities, individuals), and the geography (a particular county, a particular school, a particular region).

- **Treatments:** What are the treatments, what do treated subjects receive, what do control subjects receive. If you're running an A/B test, then there won't be a control, but rather 2 treatments compared to each other. Part of this should be a discussion of logistics – what needs to happen to implement these treatments. Do you need to hire someone, do you need to purchase something, do you need to produce some materials? Specificity is key.
- **Randomization (Treatment Assignment):** What is the mechanism you are using to ensure that treatment and control groups are the same at baseline. Are you worried about contamination across treatment and control groups? Do you expect spillovers to be a factor in your setting?
  - Here you need to define the level of randomization. Is randomization at the individual level, classroom level, neighborhood, IP address, web page? This affects how much you should worry about spillovers.
- **Timeline:** How many survey waves will you run, when will you run them? How spread apart is your measurement of treatment and outcomes? Are there time varying factors you need to worry about? The logistics of how you will measure outcomes is important here.
  - If your assignment of treatment happens at a different time than the outcome measurement, then you have to think about attrition. Does your sample need to be invited back to measure outcomes, will there propensity to come back be affected by your treatment?
- **Outcomes:** What will you measure, how will you measure it. Are there primary and secondary outcomes? Do you expect to see variation in the outcomes you measure, how can you be sure of this?
  - Stated Vs. Revealed Preference: If a surveyor comes up to you and asks, "how much do you like product X on a scale of 1 to 5?" they are collecting a stated preference. Suppose you know you value the product at 3 (this is the ground truth). But if you have no incentive to report your real value, you could say any number without consequence. For example, if you suspect that the person is going to try to sell you the product, you may strategically under-report your value ("I only value it 1"), this is often what people do when haggling. As the surveyor, we have no reason to trust that your response reflects your real beliefs – this problem is present in many kinds of consumer survey questionnaires. As social scientists, we'd like to design our outcome measures to get at "revealed preference", or the true preferences of our subjects. To do this we usually like our outcome to be "behavioral" – that is, a subject is performing some real action with costs. A real purchase decision, a link click, a list sign-up, a test outcome, a dictator game with real payouts; all of these are behavioral measures.
  - Incentivizing Revealed Preference: In some contexts we need incentives to make sure people tell us the truth about their preferences. I provide some examples of strategies below.
- **Measurement/Analysis:** How will you measure your effect? Are you running a t-test comparison of treatment and control, are you looking at an average difference?

2 What Randomization Does

Randomization creates balance between groups on both observable and non-observable characteristics. If I randomly assign everyone in class to a treatment and control by flipping a coin, I can be

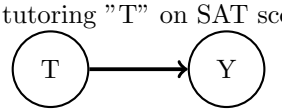
sure that my groups are similar to each other in terms of things we can measure like age, gender, experience. Not only that, but they will also be balanced on things I can't or haven't measured, like family background, patience, IQ, health.

This result is only true for **large enough sample N**. Suppose I have two classrooms, and I randomize at the classroom level. One receives treatment, one receives control. I have not created any balance, because I only have 2 randomized units. The characteristics of one class could be very different from another, and bias the results. For example, if one class is earlier in the day, these students may be more patient or punctual or busier than the other class. My treatment measurement will now partially reflect these socioeconomic differences.

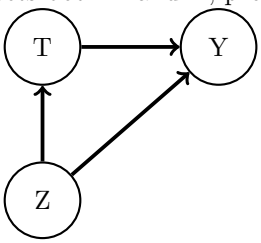
Randomization of treatment T to units allows me to identify a causal effect of T on some outcome Y. It resolves (1) ommitted variable bias and (2) reverse causality.

2.1 Ommitted Variable Bias

Also described as "selection bias" this is bias that occurs when a third, unobserved variable is driving the correlation between your outcome Y and treatment T. Suppose I want to know the effect of tutoring "T" on SAT scores "Y". I'm interested in the direction of causality T on Y:

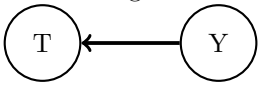


But suppose I just observe some non-randomized data of students who are tutored or not (T=0,1) and their SAT scores. Students who pay for tutoring may come from richer families, and may therefore have many other advantages like more study time at home, more extracurriculars, more parental help that could explain their SAT scores. We could call this bundle of "other stuff" that may be driving SAT scores family background, or Z. If rich family background predicts tutoring take-up AND high SAT scores, then the correlation between tutoring and SAT scores may be coincidental to the true causal link between family background and SAT scores. Z is an ommitted variable that affects both T and Y, producing a spurious correlation between T and Y.



2.2 Reverse Causality

What if some kids choose to get tutored BECAUSE they got low SAT scores? If this is the case, we would observe a correlation between low SAT scores and tutoring. But the causality is the opposite direction, low SAT scores cause kids to get tutoring. With just the observational data, we could infer the wrong direction of causality.



2.3 Potential Outcomes Framework

Formally, we can describe the benefits of randomization using the "Potential Outcomes framework". Suppose I want to study the effect of a treatment T on Y. For example, what is the effect of tutoring on SAT scores? If I have data on many kids, their SAT scores and whether they were tutored, I could calculate the average difference between tutored and non-tutored kids:

$$E(Y_i^T|\text{tutored}) - E(Y_i^C|\text{not tutored}) = E(Y_i^T|T=1) - E(Y_i^C|T=0) \tag{1}$$

Where $E(Y_i^T)$ is the expected (average) SAT score for a person i who was "treated" T with tutoring, and $E(Y_i^C)$ is the expected SAT score of someone who was not treated C. The conditionals $T=1$ and $T=0$ represent what actually happened to this person, whether they were treated or not treated.

Suppose we add and subtract another quantity to this equation, $E(Y_i^C|T=1)$ which represents the expected outcome for a treated person, had they NOT been treated. This is an un-observable, hypothetical quantity. We don't actually observe what would have happened to a treated person in the case they were not treated. Hence, it's a "potential outcome".

$$E(Y_i^T|T=1) - E(Y_i^C|T=0) + E(Y_i^C|T=1) - E(Y_i^C|T=1) \tag{2}$$

If we combine the first and fourth quantities in this equation, we get:

$$E(Y_i^T - Y_i^C|T=1) + E(Y_i^C|T=1) - E(Y_i^C|T=0) \tag{3}$$

The first term is our treatment effect that we want to estimate, the effect of being treated $E(Y_i^T - Y_i^C|T=1)$. It represents the effect of being treated, relative to what would have been the outcome if the person wasn't treated. The second term is a selection bias $E(Y_i^C|T=1) - E(Y_i^C|T=0)$, and it represents what a treated person would have achieved if they had not been treated, relative to a control person that was not treated. Kids with tutoring are different from kids without tutoring, and likely would have performed better than control kids even without the tutoring. This means the selection bias is non-zero, if we are simply observing non-randomized data.

Suppose we decide to control for a bunch of covariates, like family income, race, gender etc. Whatever we control X, we still can't say $E(Y_i^C|X,T=1) = E(Y_i^C|X,T=0)$. We fundamentally don't observe what would have happened to a treated person without treatment $E(Y_i^C|T=1)$. There are always unobserved covariates we can't control for.

Randomization solves this problem. When treatment is randomly assigned, no observed or unobserved covariate of the treatment group is correlated with treatment. So we can say the selction bias is zero, ie. $E(Y_i^C|T=1) = E(Y_i^C|T=0)$

A better more complete explanation of Potential Outcomes can be found in the book "Causal inference in statistics, social, and biomedical sciences" by Rudin & Imbens (Imbens and Rubin, 2015).

3 Types of Randomization

3.1 Unit Level Randomization

Suppose we have a sample of N people that we want to assign to treatment and control. We can flip a coin and assign each person a treatment , $T=1$ or $T=0$.

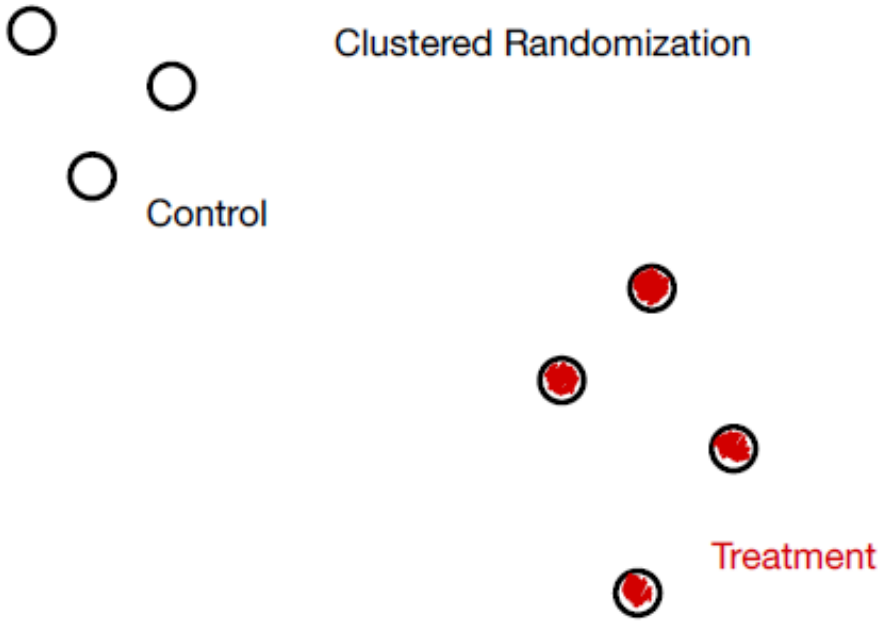


Figure 1: Cluster Randomization Example

3.1.1 Cluster Randomization

In some cases, we might be really worried about spillovers or contamination between people (see a description of this below). For example, if we are offering a tutoring treatment to kids, we might be worried that the kids will interact with non-tutored kids in the same class. Instead, we can randomize at a *cluster* level. This means we randomize a group of units together, if we think they will be interacting a lot. For example we can randomize our tutoring treatment at the classroom level, so that everyone in the same class gets the treatment.

The trade-off here is sample size – if we randomize at a bigger unit, we need more sample to observe effects.

3.1.2 Stratified Randomization

Suppose each person in our sample N has a gender G such that $G = 1$ or $G = 0$. Suppose these groups are very unequal, so we have many more $G = 1$ than $G = 0$. This makes us worried that a simple coin toss might randomly end up sorting all our $G = 0$ people into either treatment or control, so we can't do any subgroup analysis on these people. Instead we can stratify – so we split our sample into all the $G = 1$ and all the $G = 0$ separately, then we randomize to treatment within each of these groups. This guarantees that we have some $G = 0$ people represented in both treatment and control groups.

In large-scale experiments, we often stratify by geography. Suppose I'm rolling out a treatment across many locations, I might first stratify by county, city or urban/rural status, and then randomize within each of these units. This means that if I zoom into any given city or county, I'm guaranteed randomization to treatment, regardless of relative sample size differences across counties or regions.

3.2 Session and Time Level Randomization

In many contexts like e-commerce, we don't randomize treatments by individual, but rather by session or time. Given some background traffic flow to our site, we choose to randomize what version a user sees either by changing the version every new session, or changing the version at some random time Z. This leads to a few concerns. If we choose to expose our traffic flow to a new treatment by switching at some time, we need to be concerned about how the time of switching correlates with baseline traffic flow trends. We could have time of day effects, day of week effects, or even "time of day-week effects". The paper by Andreoni "Avoiding the Ask" provides an example of how to deal with this by making sure you randomize across multiple times/days.

The other concern is contamination. The same person might visit my site multiple times before making a purchase. If I randomize at the session level, then every time they open the site they may experience a different treatment version. When they eventually make a purchase, I don't know if I should I ascribe that decision to being treated or control. If I accidentally label this person as "control" even though they were treated in an earlier session, I will underestimate my treatment effect. See Bojinov and Shephard (2019) for more information on inference for time varying experiments.

Time-blocked randomization is also called "switchback" experimentation in industry. See this blog post for details about how platforms like Lyft use it in large markets.

3.2.1 Sample Size Issues in Session Level Experiments

In practice running session-level experiments can make it difficult to produce enough treatment and control units to analyze. Suppose you want to test the effects of wearing casual vs. formal clothing on in-person solicitation rates. To control for time-of-day and day-of-week effects, you might implement a design like this:

Table 1: Clothing Randomization Design

	Monday	Tuesday	Wednesday
9-12pm	Casual	Formal	Casual
12-3pm	Formal	Casual	Formal
3-6pm	Casual	Formal	Casual

This kind of design gives me variation I can leverage both within day-of-week (compare all

Monday slots to each other), and within time (compare all 9am slots to each other). But if I collect an outcome like "total successful solicitations" or "response rate", then I basically only have 9 observations to use, and confidence intervals will be very wide.

To get more data, we could try splitting up our randomization blocks. So for example, in my "Casual 9-12pm Monday" block, I split it into 6 sessions of 30 minutes, where I record response rates at 30min intervals. The randomization is still the same, but now I have 6 observations rather than 1. I just have to cluster my standard errors at the level of randomization (the 3 hour block).

Another alternative is to collect individual level data. Suppose as part of my solicitation I collect donations, or estimates of willingness to pay, or a purchase decision. I could keep track of every person I approached or interacted with, so that my outcome is the measurement of person i during a given block. If we assume the people walking past are independent, we now have a dataset of N people and their donation rates, or their 1/0 purchase decision. Their treatment is assigned depending on which block they walked by during.

4 Problems in Experiments

Suppose a chain of grocery stores is interested in rolling out a new shop layout, with a more optimized store layout that puts high-value items near the front of the store. Before the chain commits to implementing the layout everywhere, they first want to estimate the impact of the new layout on gross sales. They choose to implement a randomized design at the store level, randomly assigning some stores to receive the new layout, and some stores to remain as is. Then they'll compare the average change in gross sales between treatment and control stores.

What could go wrong with this experimental design? Suppose there's a treatment store really close to a control group store. People who usually go to the control group store decide they love the new layout, and begin to instead shop at the nearby treatment store. The control store has now lost sales simply because they happen to be near a treatment store. Now if I estimate a treatment effect comparing treatment to control group sales, my effect will be biased upward, because the difference in sales will reflect not only the impact of the new store on sales, but also the cannibalization effect that took customers away from my control stores. This is an example of a "spillover" from treatment to control, also called "contamination".

What if the chain is only made up of 2 stores, so they randomly select one to implement the layout, and one as control? Now I have a "sample size" issue. Because I have only randomized across two units, my sample is equal to 2, and I have not achieved the balance between treatment and control units that randomization is meant to do. Any systematic differences between my treatment and control store (their location, local business cycle trends, changes in local foot traffic) can bias my estimate because I don't have enough units to be confident that these characteristics are balanced across my groups.

What if the chain has sufficiently many stores, but a recession hits during my study and 50% of the control group stores close because they did not have the sales bump from the new layout? Now when I compare treatment and control group sales, my control group only consists of surviving control stores, which are the most successful subgroup of control stores. My treatment effect will be biased downward, because I'm only comparing my treatment group to a selected subset of surviving control group stores. This is an example of bias via "differential attrition".

Below I'll outline the different ways that even well-implemented randomization strategies can run into issues.

4.1 Sample Size

An experiment is the testing of a hypothesis. Often we are asking "does X affect Y", which amounts to testing the null hypothesis "there is no effect of X on Y". Testing this hypothesis we are subject to Type 1 error (finding an effect when there isn't one, aka false positive) and Type 2 error (finding no effect when there is one, aka false negative).

Our ability to detect a significant difference between a treatment and control group, assuming a treatment effect *exists* depends on the power of our study, which is a function of our sample size.

Before running an experiment, we do power calculations to see how much sample size we need for a given "minimum detectable effect size" or MDE, which is the minimum effect size of treatment we want to be able to detect. Suppose I expect my treatment to move my outcome by 5%, then I want to make sure my MDE is set to under 5. Sample size can be calculated as

$$N = \frac{\sigma^2(z_{1-\alpha/2} + z_{1-\beta})^2}{\text{MDE}^2} \tag{4}$$

Where σ^2 is the standard deviation of the outcome in the population we're studying. α represents our tolerance of Type 1 errors, which we set usually at 0.05. And β is our tolerance for Type 2 error, which we set at 20% (for 80% power). The z-scores for these scenarios are: $z_{1-0.05/2} = 1.96$ and $z_{0.8} = 0.84$

For example, suppose I want to study the effect of tutoring on SAT scores, where I will assign half my sample randomly to receive a tutoring course. The standard deviation of SAT scores in the US is 195. Suppose I want to be able to detect an effect size of 50 SAT points (this is the minimum effect of tutoring I want to be able to observe). Then the sample size I need is:

$$N = \frac{195^2 * (1.96 + 0.8)^2}{50^2} = 119 \tag{5}$$

4.1.1 Power in Clustered Experiments

When our treatment is clustered, we have to deal with the fact that people in the same cluster are not independent of each other (their outcomes are correlated). I won't walk through the proof here, but according to Duflo et al. (2007) the minimum detectable effect size when J groups are randomized, and there are n people in each group is:

$$MDE = \frac{(z_{1-\alpha/2} + z_{1-\beta})}{\sqrt{P(1-P) * J}} * \sqrt{\rho + \frac{1-\rho}{n}\sigma} \tag{6}$$

Where P is the probablity of treatment, and ρ is the intracluster correlation coefficient (how correlated are outcomes within cluster). The more correlation within-cluster, the less information we get from each individaul in a cluster. This is how clustering might balloon our sample size needs, as we really need to increase our number of clusters J to lower our MDE.

4.2 Contamination/Spillovers

Contamination occurs when a treated individual impacts a person that is not treated. Suppose I provide some new anti-biotic to a person, and I want to study the impact on their health. If this person is friends with a "control group" person, they could talk to them and give them some of the pills. This control person is now "contaminated", because they have essentially been treated. This produces a downward bias, I underestimate the effect of the antibiotic because both people received the treatment in reality.

This is a BIG concern in webpage level randomization. If I randomize people to an A or B version of a webpage when they visit my website, I ignore the fact that many people might visit or refresh my page multiple times. If the person is being re-randomized every time they visit my site, how can I say whether or not they're an "A" or "B" version person, as their behavior is a function of all their past exposures to an A or B version.

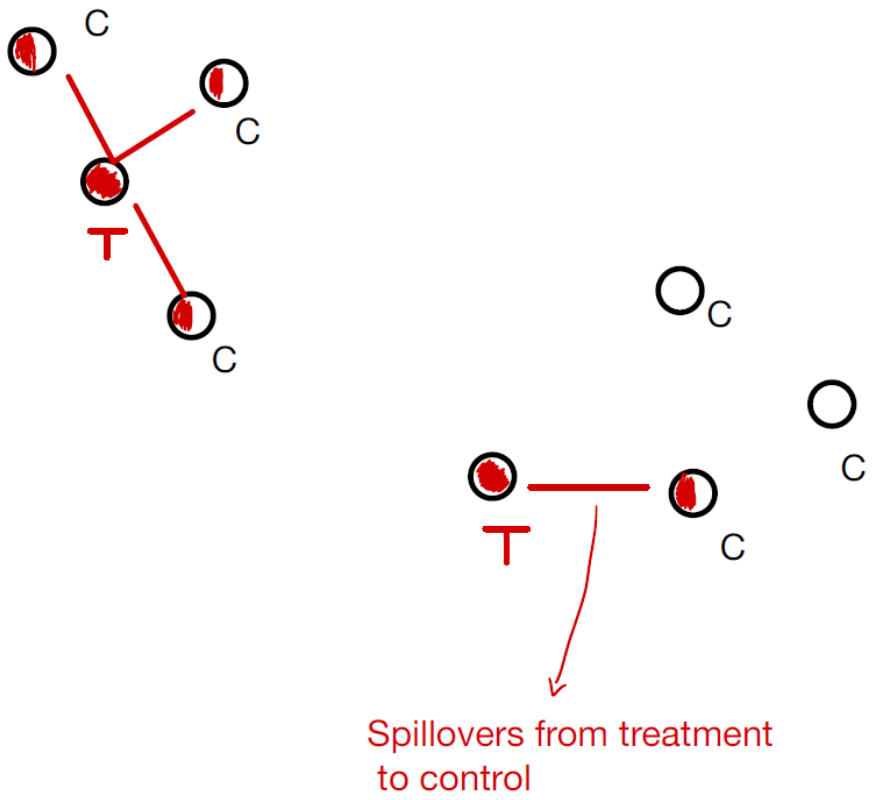


Figure 2: Ex. of Treatment Spillover

This is also a big problem for A/B testing in online platforms. If Uber wants to test a new fee raise on the demand for Uber drivers, it might randomize some drivers to receive the raise, and leave others untreated. But suppose the treated drivers start driving much more, because they're taking advantage of the raise. This has a congestion effect, because it increases the total number of Uber drivers on the road, which indirectly impacts how many rides non-treated drivers get. This is an indirect channel through which treated units affect the outcomes non-treated units.

4.3 Direct vs. Indirect Spillovers

Spillovers can be direct. A kid receiving tutoring may directly interact with a control child that did not receive tutoring. A household that received a cash transfer may share that cash with a control household. There are also indirect spillovers, especially "market spillovers" that operate through

economic forces.

If a treatment affects an aspect of the economic environment that control units participate in, this could be a contamination. Suppose I randomly give certain villages in rural Kenya a cash transfer, and want to study effects on long-term health and wealth, relative to control villages. Even if the villages don't personally interact, they may be connected through markets. It's possible that if a cash transfer is large enough, the cash inflow creates price inflation that affects all villages through the price of food. See Egger et al. (2022) for a direct example of this.

Now let me lay out an example of indirect market spillovers using the tutoring example we discussed above. Suppose I randomly give a new online tutoring program to some schools, and don't give it to other schools. Suppose 10% of students, both treatment and control, use private tutoring services. When online tutoring is implemented, suppose this service loses a lot of business and collapses, harming control students who no longer have access to the private tutoring.

Of course, the treatment kids may also be harmed by this, because they also no longer have access to this private tutoring. So what's the problem if everyone is affected? Well experiments are an attempt to capture a fundamentally "unknowable" quantity – the effect of treatment on an individual's outcome, compared to a hypothetical world where that same individual was not treated. The reason this is unknowable is if I give John the treatment, I will never observe the counterfactual world in which John was never treated. Instead, by using randomization I construct a control group that is balanced on all of John's features, such that I can use the average outcome of the control group as my "counterfactual John" in a world where John was never treated.

Market spillovers, like other spillovers, break this logic. Now I'm no longer comparing the effect of online tutoring compared to a counterfactual world where the same class or school never received treatment. Instead the collapse of the private tutoring market means I'm comparing treatment to a counterfactual world where there is no public NOR private tutoring, which was NOT the status quo pre-treatment. In this way the market spillover has changed the environment AND changed our counterfactual. So the effect I'm measuring doesn't accurately reflect what the policymaker wants: what would happen in a world where I treated everyone, compared to doing nothing.

Suppose I was able to run my experiment across many school districts across the country, each with their own independent private tutoring market that exist far apart. I could randomize some districts to receive online tutoring, others to remain as control. In this case the private tutoring market still collapses in treatment districts, but I have pure control districts who's private tutoring is unaffected. Now my treatment effect on grades would reflect the total impact on grades, including the potential negative effect of the private tutoring collapse, and I would have a pure counterfactual control to compare to, which accurately depicts what grades look like in a world without treatment. (This is of course assuming that I have enough districts to randomize to treatment and control in a balanced way).

4.3.1 Dealing with Spillovers

When I'm worried about spillovers between treatment and control units, one thing I can do is just randomize at a higher level. If I'm implementing a tutoring program, and I'm worried about treatment kids influencing non-treatment kids, I can instead randomize at a school level, assuming that kids don't talk to each other across schools.

To actually study the spillover effect, I could instead implement a *randomized saturation design*. I randomize in two levels. First, I randomize my clusters to either treatment or control (so some schools/classrooms are treated, some are not). Then, within treatment clusters, I randomize the number of units I treat. For example, I randomize where 25% of kids get the treatment, 50% or 70%. The total causal effect is the different between treatment and control clusters, weighted by

treatment saturation. But now I can also measure the spillover effect on non-treated individuals by comparing non-treated units in a treated cluster of a given saturation, to the non-treated units in a pure control cluster (ex. non-treated student performance in a treated classroom, compared to students from a non-treated classroom).

Interference is an active area of study for analysis of A/B testing. A recent popular topic that is being implemented in online platforms like Uber is the idea of "multiple randomization designs". See for example Johari et al. (2022).

4.4 Attrition

In many experiments, you have 2 or 3 survey waves to measure outcomes. In wave 1, you recruit people to your study, and you randomize them to a treatment. The treatment takes place (for ex., administering an antibiotic), and then you survey these people again at a "midline" or "endline" where you measure the outcome you're interested in. This creates a potential problem of "non-random attrition". Inevitably, some people in your experiment will not show up to be measured post-treatment. If certain kinds of people in your control group are LESS LIKELY to show up for endline measurement, this could introduce bias into your causal estimation. For example, if I want to measure the impact of an anti-biotic, and the people who received the placebo KNOW they received the placebo, they may decide that it's not worth showing up to the medical center for the endline survey. Differentially, less healthy people or lower income people in the control group may choose not to come back, while in the treatment group, everyone decides to show up for endline. Now my control group is a "selected sample" of particularly healthy control group people.

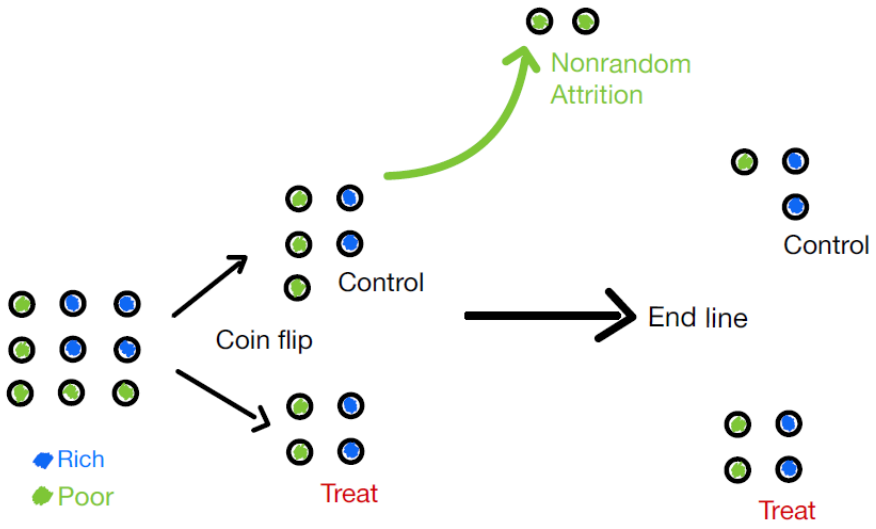


Figure 3: Attrition Example with Wealth Covariate

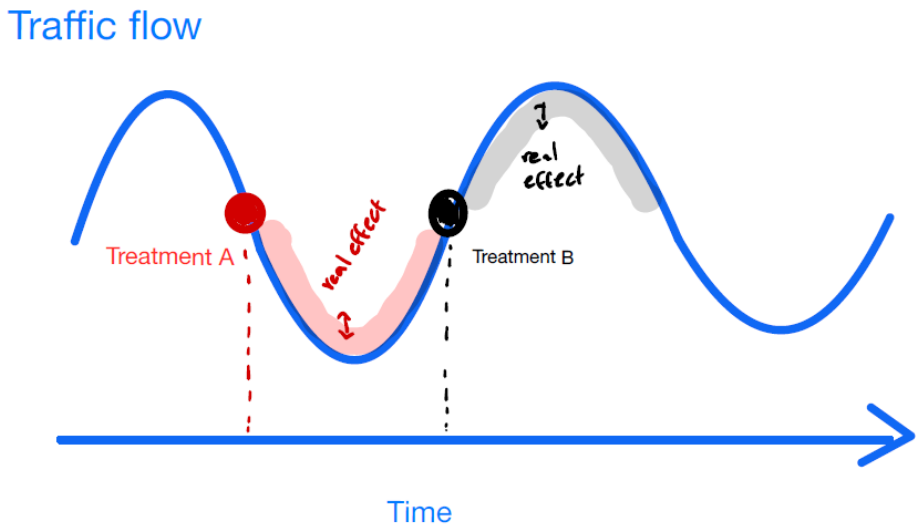


Figure 4: Time Effects in Randomization Design

5 Other Things to Think about in Experiments

5.1 Time Effects

Some designs randomize over time, rather than directly over individuals. For example, a Youtuber might present one video thumbnail to their audience for X amount of time, and then A/B test by switching to another thumbnail at time Y. Here individuals aren't being randomize to seeing one or the other. The randomization relies on the fact that the switch happened at some random time. The assumption here is that the type/flow of click-through would have remained constant across this random change, if not for the thumbnail change. For this type of design, we have to worry a lot of about time trends, cycles, etc. If you decide to change the thumbnail during a particularly low traffic period, or during a trough in the clickthrough cycle, you might misattribute a jump in clickthrough as a causal effect, when it's actually just due to the time of day at which you changed the thumbnail.

5.2 Compliance and Intent to Treat (ITT)

Suppose I want to estimate the effect of a new drug relative to a placebo. I implement a randomized experiment offering some people the drug, and some people the placebo. Suppose some of the treatment individuals simply don't take the drug. They don't comply with my experiment. Maybe these non-compliers are a particularly selected subgroup of my treatment group – for example, maybe low-income people were less likely to adhere to the treatment.

Does this affect my casual estimate? Intuitively, our treatment effect is still a quantity of interest. It reflects the effect of our treatment, given the fact that not everybody actually takes it up, which

lowers the average treatment impact. If our real world implementation of treatment will inevitably have non-compliers, this is exactly the effect we want to estimate.

In an experiment we can have 4 types of people:

1. Compliers: People who follow their treatment assignment (do what they're told)
2. Always Takers: People who always take treatment regardless of their treatment assignment.
3. Never takers: People who never take up treatment regardless of treatment assignment.
4. Defiers: People who do the opposite of their treatment assignment (if treated, they don't take it up, if control, they take treatment).

We think of non-compliance as some combination of always takers and never takers in our sample. In the case of non-compliance, what we estimate is an Intent to Treat (ITT) effect. This is the treatment effect for people who were ASSIGNED to treatment, regardless of whether or not they actually followed through. As the number of compliers goes to 100%, the ITT approaches the Average Treatment Effect (ATE).

We can also look at the treatment effect JUST for people that did comply with treatment; that did actually take the pill. This sample of people may be selected, but it still reflects a causal estimate since randomization insures the selection is uncorrelated with treatment. This estimate is called the local average treatment effect (LATE), because its specifically the treatment effect for the subgroup that was really treated.

5.3 Long term vs short term outcomes

Do you expect your treatment to impact outcomes in the short or long term? This effects your decision of when you evaluate the results of your treatment. If you survey people to get outcomes too early, you may miss the causal impact that comes later. For example, suppose you want to study the impact of preschool on child development. You randomize kids to receive free preschool, and study the impact by running a test at the end of the school year. This just measures a short term impact of preschool on learning. We might be missing the real effect, that could come from the cumulative impact on kids after several years past the treatment.

5.4 External Validity

Does my experimental sample actually speak to a broader effect that would occur in other populations? For example, suppose I want to study the effects of a particular ad campaign on purchase rates, but I choose as my sample population a group of college students at a university. Suppose I find a strong effect of my ad, relative some control ad. Does this mean that I will see the same effects if I try the experiment with a population of 50-60 year old suburban households? Probably not. Even if you randomize, it does not mean that your research finding will replicate in other settings or populations. The more specific or niche your experiment environment is, the less likely it is to be "externally valid".

5.5 Baseline Outcomes or Controls

In a randomized design we don't actually need to control for things if randomization has been implemented correctly. In a regression of treatment on outcome, controlling for demographics like age or gender won't affect bias, but it may help us with precision. You can think of this as essentially

aggregating T-tests of treatment vs. control within just men, just women, just people aged 30, etc. This can help us with our standard errors, even if we don't need them for an unbiased estimate.

The same principle holds for collecting baseline outcomes. For example, if I'm studying effect of a treatment on performance on a test. I don't need to know people's pre-treatment test scores in a randomized experiment. I can just compare their post-treatment test scores, because I know their pre-treatment characteristics are balanced via the randomization. However, If I do collect a pre-treatment test score, then conditioning on this baseline will help me with precision, just like controlling for demographics would.

5.6 Ethics

All experiments we run must pass standards of ethics. A big rule in economics experiments is **do not deceive**. We are not allowed to lie to subject participants about what they will receive. This does not mean we have to completely explain everything we're trying to learn from an experiment, nor do we have to disclose to everyone that they are part of an experiment. It just means that if I promise you a chance to win a raffle, a raffle must in fact take place. If I say that money they donate will go to an NGO, it must go to the NGO. Experiments should also not actively harm subject participants.

5.6.1 Dealing with Fairness/Inequality

Sometimes when you run an experiment with an organization, they might insist that everyone is treated. An example would be a de-worming campaign of school children in Kenya. The government might say that it's unfair to make some schools "control schools" that receive no medicine. The way we deal with this is to use "staggered adoption" designs. Here we say everone will *eventually* be treated, but for the first wave we will randomly select units to be treated first. This gives us treatment and control units in the short term where we can estimate outcomes before eventually treating the rest of the units at endline.

6 Measuring what's difficult to measure

Some "revealed preference" outcomes are easy to measure. You might be looking at performance on a comprehension test, click-rates on a product, time spent on an app or service, money donated. Other outcomes are abstract. Social scientists have a variety of methods to measure abstract things.

6.1 Willingness to Pay

Suppose I want to understand how much people value a product. That is, what is the max price they would be willing to pay for the product, regardless of market price. If I simply survey someone and ask "how much are you willing to pay?", they have an incentive to lie and undersell their desire. If I suspect you're going to try to sell me the product, I want to keep my cards close to my chest, in order to get the best price. The surveyed person may also simply not really know their "maximum willingness to pay", since this concept isn't necessarily how non-economists think about goods. To fix this problem of "revealing true signals", we need to align incentives so that the optimal strategy of the surveyed person is to be truthful about their valuation of the product. There are many variations to these and other methods, and marketing researchers are always developing new methods (He et al., 2024).

1. Auction Formats: Sealed-bid auctions can reveal information about how people value a good. Suppose I run a Vickrey auction, which is a "second-price sealed bid" auction. I tell everyone in my sample that I am auctioning off the product, and they should submit a bid. Highest bid wins, and the highest bidder pays the second-highest bid price for the product. Now people want to submit only truthful bids – if they bid higher than their true value, they risk winning the product and having to pay more than they want. If they bid lower than their true value, they may miss out on winning the product at a good price for them.

2. Multiple Price Lists: Suppose I tell you to read each line of this table, and make a decision about which item you would rather have. Once you've made your selections, I will randomly choose one row of this table that will be "implemented". That is, whatever you picked on that row, you will receive.

Table 2: Multiple Price List

Choice A	Choice B
Product	1$ cash
Product	5$ cash
Product	10$ cash
Product	15$ cash
Product	20$ cash

You decide you'd take the product over 1 or 5 dollars, but you'd rather have 10 dollars than the product. Because you know one of these options will really be implemented, you have an incentive to be truthful about your decisions. I now know you have willingness to pay (WTP) between 5 and 10 dollars.

3. Becker-DeGroot-Marschak (BDM) method: Suppose I tell you to submit a bid for the product. But after you do, I tell you that I will run a random number generator. If the number that comes up is below the bid you gave, then you will pay for the product at the random number price. If the random number is above your bid, you will not receive the product, and pay nothing. Similar to the auction scenario, you have an incentive to say your truthful bid. For an example of this used in practice, see this paper by Chowdurhy et al. on air purifiers and perceptions of pollution.

4. Given enough experimental control and sample size, you could directly estimate an "average willingness to pay" among a group of people by randomly offering the product at different price points and seeing how many people buy it. This amounts to "price discovery", which some firms use to try to estimate demand curves for their products.

6.2 Measurement with Games

There's a zoo of behavioral parameters that economists and psychometricians have been interested in estimating. You've likely heard of the "marshmallow test" to measure patience or restraint in children. This is an example of using a game to measure the abstract idea of "patience". In the world of "social preferences", social scientists try to measure "pro-sociality" via dictator games. Suppose I offer you 10 dollars, and I say that it's up to you to decide how much to keep for yourself, vs. give to another person. These are real dollars that you will keep. Will you split it 50/50? Does your answer change if you can actually see the other person, the race/gender/age of that other person? The

more you choose to give to the other person, the more "pro-social" we might say you are. Haushofer et al. (2023) is a great example of using games to study co-ethnic attitudes in Africa. Another great example for studying social norms and moral values is De La Sierra et al. (2025) found here. For a broader summary of how economists approach the measurement of beliefs and norms, see this paper by Gneezy and co-authors (2025).

Here are some other examples of using games:

1. Ultimatum Games – This is a bargaining version of the dictator game where the receiving player can choose to accept the offer, or blow up the game. Again we use this as a measure of pro-sociality, or as a measure of "norm-enforcement", a person's propensity to punish unfair allocations.

2. Trust Game – This bargaining game forces players to cooperate to increase the total pot of winnings. We often use this to measure cooperativeness, for example between a husband and wife, or two friends or two strangers.

3. Bomb/Balloon Game – This is an example of measuring risk aversion, or risk attitudes. Suppose I want to know how risk taking a person is. Suppose I show you a panel of squares or "boxes".

Table 3: Bomb Game

Box	Box	Box
Box	Box	Box
Box	Box	Box
Box	Box	Box

For each box you collect, you will receive 1 dollar. One of the boxes contains a bomb. If you choose the bomb box, you lose all your earnings. The number of boxes tells me how much money you're willing to risk in exchange for a given amount of risk (probability of bomb selection).

6.3 Real Effort Tasks

I might be interested in studying how some treatment affects a person's willingness to work, or speed of work. This is often the case in managerial research where we're interested in what kinds of incentives might make people work better or harder. If I don't have access to any real performance metrics, I can construct a performance metric by some artificial task (Charness et al., 2018).

Usually we opt for something repetitive and relatively mundane, where we can measure some unit-level completion rate. Stuffing envelopes might be an example. Here are some other examples:

1. Solving a series of small mazes

2. Solving a sheet of simple math summations in a given timeframe

3. Fill a page with random numbers. Have subjects count the number of even numbers on each row of the page.

4. Give a subject 7 letters, and have them come up with as many words as possible in a given time frame.

5. Transcribe a series of cursive letters on a page

These tasks vary in how cognitively demanding they are, how unpleasant they are, and also how much a person's baseline skill matters. In a randomized experiment we're generally comparing performance on the task across two randomized groups, so the particular ease or difficulty of a task doesn't matter for within-task comparisons. The exception is if too many people actually complete the task, or hit a "max value". Then we have no variation in our outcome variable.

6.3.1 Learning Effects

One issue with effort tasks comes when we have the same people complete the task more than once. For example, we want to measure performance on a real-effort task before-after people receive caffeine. But if the task is something that people can get better at over time, then part of their improvement before-after will be due to learning effects. Having a control group helps us partial this effect out, as the control group will also experience the learning effect, so the treatment gain in effort is just the effort change in treatment minus effort change in control.

6.3.2 Incentivizing Tasks

Of course, we ALWAYS need to incentivize our tasks, otherwise there's no reason why someone would bother doing the task. We usually do this with some baseline piece-rate incentive, like paying 5 cents per math problem solved. We could also implement a raffle as follows: for each math problem you solve, you receive 1 raffle ticket to win a prize of a 25$ amazon gift card. Therefore, the more you solve, the higher probability you have of winning. This guarantees that people will have some reason to be working on the task, and then you can layer your particular treatment on top of that general incentive.

6.4 Not all "stated preference" is bad

Many fields rely on stated preferences. A clear example are mental health studies, which often rely on asking subjects a series of standardized questions about how they feel. The best we can do here is to use survey questions that have been in some way validated by the public health community, or shown to be correlated with depression diagnoses, or if the responses have been shown to change in response to anti-depressants.

Another example is the valuation of non-market goods, like clean air. Often policymakers might rely on "contingent valuation" surveys, which in effect ask households their willingness to pay for a reduction X% of air pollution, for example. Many economists don't necessarily trust this method, but it pops up often in litigation around environmental impact.

Another example is consumer confidence or expectation surveys, where people are asked to forecast inflation or to give their opinion on the state of the economy in 6 months. The Fed takes these seriously, even though they're not incentivized. Again, validation here is key. We want to be able to justify our use of stated preference questions by showing that those questions have been shown to correlate to some real behavior, or to track some metric. So for example an investor may have no incentive to report his realistic expected future inflation rate, but if I aggregate his measure with others, and it seems to react to market factors in expected ways, I'm more confident that people are telling me the truth about their beliefs on average.

7 Survey Techniques

Firms and governments rely on surveys to observe economic, social and health behaviors. You might be interested in trying to collect information from participants like income, occupation, wealth, health. One of my favorite webpages is this list of survey and measurement techniques at Poverty Action lab: https://www.povertyactionlab.org/resource/repository-measurement-and-survey-design-resources. While geared towards developing countires, it's a great general resource on how we construct questions to capture household, individual and firm characteristics. You can also explore the US Census for examples of how economic and household questions are often phrased in the state of the art. I'll highlight just a few details about survey design below.

7.1 Attention in Surveys

Many of the surveys we run are uninteresting to the survey taker. This can lead to measurement error if a bored survey taker begins to randomly select options, or continuously select the same option, especially towards the end of longer surveys. When we have many observations, we might excuse away inattention as random measurement error or "white noise". If a minority of individuals randomly choose options, this will make our estimates a bit noisier when we aggregate across people, but it won't produce any systematic biases in our results.

In other contexts the inattention can create big problems. Suppose I ask a battery of question using a Likert scale like the following: "how much do you like pizza on a scale of 1 to 10". I ask many in a row, cycling through pizza, burgers, fries etc. After a few, suppose people get bored and begin selecting 5 for everything. This is a systematic error, and could render our outcome measurement meaningless.

Here are some strategies we use to try to fix or control problems of inattention.

7.1.1 Time Measurement

In online surveys, we can try to capture the amount of time survey takers spend on different parts of a survey. If a survey is implemented in a series of webpages, we can measure the time the survey taker spent on each page. We can then know if anyone is speeding through the survey in a way that suggests inattention, and try adjust our estimates to consider those bad actors.

7.1.2 Validation and Attention Questions

We might use trick or sneaky questions to test if people are actually reading instructions properly. For example, after a battery of multiple choice question I might add a question like "Please ignore the options below and select 'Somewhat agree'." Or I might ask a super obvious question, like "Which of the following is furniture? a) chair b) elephant c) cloud". Last, if I'm asking a bunch of scale questions like "how much do you like ice cream, pie, cake" I might switch the direction in the middle and ask "how much do you dislike scones?"

7.1.3 Incentives and Gamification

Rather than testing for attention and detecting inattention patterns, we might try to make our survey more interesting. We could do this with colorful images, fun button functionality, or include a variety of question types that involve games, competitions, or other types of user interactions (see Duolingo as an example of extreme gamification).

Alternatively we might try to pay or incentivize people into paying attention. This is related to our discussion above about revealed vs. stated preferences. If I tell you I'm going to pay you for taking my survey, or even better if your payment is conditional on certain decisions in the survey, people may pay better attention.

7.1.4 Survey Order

If I have a long survey, or many repetitive questions, I might be more worried about survey fatigue. By the end, maybe survey takers get tired and stop paying attention. To make sure that the last questions in my survey aren't differentialy affected by inattention, I might randomize my survey order. So every survey taker sees my survey questions in a different order. This at least spreads the inattention evenly across survey questions.

7.2 Sensitive Questions

Some questions are difficult to answer because they are about traumatic topics, or because answering them truthfully may be uncomfortable for the respondent. For example, trying to learn about a respondent's experience with domestic violence, or their feelings about a dangerous dictator may be too sensitive to ask directly.

7.2.1 List Randomization

This technique is one way to try to capture aggregate presence of a sensitive behavior. Suppose we're trying to learn what percentage of people in our sample have tried an illicit substance, like fentanyl. We might present a random subset of our sample with a list of statements. The respondent simply tells us how many of the statements are true, not which ones are true. Ex. "I have been skiing", "I have shoplifted", "I have had sex", etc. For another subset of respondents, we ask for the same list, except we add our sensitive question of interest: "I have used fentanyl". The difference in the average number respondents give tells us something about the number of people for whom the fentanyl question is true. We have discovered this without having to force people to reveal any individual data about fentanyl use. Karlan and Zinman (2012) is an example of this technique.

7.2.2 Indirect Vignettes or Opinions

Another technique is to ask people to give their opinion on other people's behavior. For example, instead of asking whether or not someone has experienced domestic violence, you could ask "what percent of the households in your neighborhood do you think have experienced domestic violence". Or you might ask "do you know someone personally that has experienced domestic violence". Again, we learn something about the aggregate presence of a behavior, without having to force individual disclosure.

7.3 Time Use and Recall

In time use, we're trying to get a sense of an individual's behaviors in the past. In a normal time use survey we might ask for a breakdown of what a person did every hour yesterday, and who they did it with. For ex. "Which of the following did you do between 8-9am yesterday?" Sleep/Wake Up/Work/Brush Teeth etc. "

In other recall studies, we might ask for a food diary – "which of the following did you eat in the last week?". Or "how many times did you visit the doctor in the past month". We often use these

types of recall questions when we can't gather pre-treatment data on individuals. Measurement error is a big problem here – the longer into the past we go, the more we should worry about memory slips causing measurement error in what we're trying to capture.

8 Examples of Experiments

8.1 User Interface

1. Halperin, Basil, Benjamin Ho, John A. List, and Ian Muir. "Toward an understanding of the economics of apologies: evidence from a large-scale natural field experiment." The Economic Journal 132, no. 641 (2022): 273-298.

2. Chandar, Bharat, Uri Gneezy, John A. List, and Ian Muir. The drivers of social preferences: Evidence from a nationwide tipping field experiment. No. w26380. National Bureau of Economic Research, 2019.

3. Lindon, Michael, Chris Sanden, and Vaché Shirikian. "Rapid regression detection in software deployments through sequential testing." In Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, pp. 3336-3346. 2022.

8.2 Marketing

1. Simester, Duncan. "Field experiments in marketing." In Handbook of economic field experiments, vol. 1, pp. 465-497. North-Holland, 2017.

2. Gordon, Brett R., Florian Zettelmeyer, Neha Bhargava, and Dan Chapsky. "A comparison of approaches to advertising measurement: Evidence from big field experiments at Facebook." Marketing Science 38, no. 2 (2019): 193-225.

3. Allcott, Hunt, Juan Camilo Castillo, Matthew Gentzkow, Leon Musolff, and Tobias Salz. Sources of market power in web search: Evidence from a field experiment. No. w33410. National Bureau of Economic Research, 2025.

4. Boegershausen, Johannes, Yann Cornil, Shangwen Yi, and David J. Hardisty. "On the persistent mischaracterization of Google and Facebook A/B tests: How to conduct and report online platform studies." International Journal of Research in Marketing (2025).

8.3 Pricing

1. Cooprider, Joe, and Shima Nassiri. "Science of price experimentation at Amazon." Business Economics 58, no. 1 (2023): 34-41.

2. Holtz, David, Ruben Lobel, Inessa Liskovich, and Sinan Aral. "Reducing interference bias in online marketplace pricing experiments." arXiv preprint arXiv:2004.12489 (2020).

8.4 Firms and Management

1. Friebel, Guido, Matthias Heinz, Miriam Krueger, and Nikolay Zubanov. "Team incentives and performance: Evidence from a retail chain." American Economic Review 107, no. 8 (2017): 2168-2203.

2. Bloom, Nicholas, Benn Eifert, Aprajit Mahajan, David McKenzie, and John Roberts. "Does management matter? Evidence from India." The Quarterly journal of economics 128, no. 1 (2013): 1-51.

3. Bandiera, Oriana, Iwan Barankay, and Imran Rasul. "Incentives for managers and inequality among workers: Evidence from a firm-level experiment." The Quarterly Journal of Economics 122, no. 2 (2007): 729-773.

4. Colonnelli, Emanuele, Tim McQuade, Gabriel Ramos, Thomas Rauter, and Olivia Xiong. "ESG Is the Most Polarizing Nonwage Amenity: Evidence from a Field Experiment in Brazil." In AEA Papers and Proceedings, vol. 115, pp. 146-152. 2014 Broadway, Suite 305, Nashville, TN 37203: American Economic Association, 2025.

8.5 Charitable Giving/Solicitation

1. Andreoni, James, Justin M. Rao, and Hannah Trachtman. "Avoiding the ask: A field experiment on altruism, empathy, and charitable giving." Journal of political Economy 125, no. 3 (2017): 625-653.

2. Samek, Anya, and Chuck Longfield. "Do thank-you calls increase charitable giving? expert forecasts and field experimental evidence." American Economic Journal: Applied Economics 15, no. 2 (2023): 103-124.

8.6 Social Preferences

1. Niederle, Muriel, and Lise Vesterlund. "Gender and competition." Annu. Rev. Econ. 3, no. 1 (2011): 601-630.

2. Cappelen, Alexander, John List, Anya Samek, and Bertil Tungodden. "The effect of early-childhood education on social preferences." Journal of Political Economy 128, no. 7 (2020): 2739-2758.

8.7 Expectations

1. Botelho, Anabela, and Ligia Costa Pinto. "Students' expectations of the economic returns to college education: results of a controlled experiment." Economics of education review 23, no. 6 (2004): 645-653.

2. Armantier, Olivier, Scott Nelson, Giorgio Topa, Wilbert Van der Klaauw, and Basit Zafar. "The price is right: Updating inflation expectations in a randomized price information experiment." Review of Economics and Statistics 98, no. 3 (2016): 503-523.

8.8 Learning

1. Oreopoulos, Philip, Richard W. Patterson, Uros Petronijevic, and Nolan G. Pope. "Low-touch attempts to improve time management among traditional and online college students." Journal of Human Resources 57, no. 1 (2022): 1-43.

2. Felkey, Amanda J., Eva Dziadula, Eric P. Chiang, and Jose Vazquez. "Microcommitments: The effect of small commitments on student success." In AEA Papers and Proceedings, vol. 111, pp. 92-96. 2014 Broadway, Suite 305, Nashville, TN 37203: American Economic Association, 2021.

References

Bojinov, I. and Shephard, N. (2019). Time series experiments and causal estimands: exact randomization tests and trading. *Journal of the American Statistical Association*.

Charness, G., Gneezy, U., and Henderson, A. (2018). Experimental methods: Measuring effort in economics experiments. *Journal of Economic Behavior & Organization*, 149:74–87.

Duflo, E., Glennerster, R., and Kremer, M. (2007). Using randomization in development economics research: A toolkit. *Handbook of development economics*, 4:3895–3962.

Egger, D., Haushofer, J., Miguel, E., Niehaus, P., and Walker, M. (2022). General equilibrium effects of cash transfers: experimental evidence from kenya. *Econometrica*, 90(6):2603–2643.

Haushofer, J., Lowes, S., Musau, A., Ndetei, D., Nunn, N., Poll, M., and Qian, N. (2023). Stress, ethnicity, and prosocial behavior. *Journal of Political Economy Microeconomics*, 1(2):225–269.

He, S., Anderson, E. T., and Rucker, D. D. (2024). Measuring willingness to pay: A comparative method of valuation. *Journal of Marketing*, 88(3):50–68.

Imbens, G. W. and Rubin, D. B. (2015). *Causal inference in statistics, social, and biomedical sciences*. Cambridge university press.

Johari, R., Li, H., Liskovich, I., and Weintraub, G. Y. (2022). Experimental design in two-sided platforms: An analysis of bias. *Management Science*, 68(10):7069–7089.

Karlan, D. S. and Zinman, J. (2012). List randomization for sensitive behavior: An application for measuring use of loan proceeds. *Journal of Development Economics*, 98(1):71–75.