Survey Experiments in Social Science
Session 2 – Treatment and effect

Federico Vegetti
University of Turin

University of Newcastle

Nonresponse and Noncompliance

Nonresponse:

- When we fail to observe the outcome
- E.g. People refuse to tell us their vote choice
- Problematic for inference when it affects some conditions more than others

Noncompliance

- When those assigned to the treatment group do not receive the treatment, or when those assigned to the control group receive the treatment
- E.g. people refuse to watch the debate (or watch it anyway)
- The treatment effect is reduced, although inference is still possible

Missing data

Missing data are problematic if they affect some conditions more than others

- For instance, people who watched the debate might be less likely to reveal their vote choice than people who did not watch the debate

This can be dealt with in the design stage by including incentives for people who complete the study

If there are still missing data, but they are independent from the treatment, then they can be regarded as random

Yet, this will reduce our sample, potentially affecting the power of our experiment

Noncompliance

Noncompliance is a threat to our ability to manipulate the DGP

In the extreme case in which no subject complies with the manipulation, then the manipulation is independent from the treatment

In other words, the manipulation becomes a *weak instrument* for the treatment

Types of noncompliance

- **Pre-treated subjects**: subjects who have already experienced the manipulation before being randomly assigned
  - E.g. people who have watched the debate before (if e.g. we run the experiment in the lab showing a recording)
- **Never-takers**: subjects who choose against the manipulation regardless the random assignment
  - E.g. people who will not watch the debate no matter what
- **Always-takers**: subjects who choose for the manipulation regardless the random assignment
  - E.g. people who will watch the debate no matter what
- **Defiers**: subjects who choose the opposite of the assigned manipulation

Failure to Treat

- Another way to conceptualize noncompliance is to look at our **ability to treat** the subjects
- This is is threatened at three steps:

Delivery

- We fail to deliver the treatment
- E.g. we sample a subject, but the person has no phone

Receipt

- The subject fails to receive the treatment
- E.g. we call the subject, but s/he does not pick up

Adherence

- The subject does not obey the prescribed treatment regime
- E.g. the subject picks up the phone, but refuses to participate

How to deal with it?

To deal with noncompliance we need to work on the design

Dealing with pre-treated subjects:

- Easier in lab settings, where treatments can be fully artificial
- In our case, we could exclude people who have already watched the debate before
- To avoid influencing them, we can ask them questions that they could only answer if they saw the debate (and exclude those who can answer correctly)

In other cases, we can offer subjects financial incentives conditional to them following the instructions

Manipulation checks

- We can check whether the subjects who received the manipulation perceived the treatment in the way we wish
  - For instance, if subjects have to read a text where the tone has been manipulated (e.g. negative in one version, positive in the other), we can ask them after the treatment how would they rate the text (from positive to negative)
  - In our example, we can ask respondents whether they watched the debate (lame), or ask them some questions about the debate that only people who watched it can answer (risky)
- This again depends on what we think the treatment is
- Which depends on the theory
- Manipulation checks are not always possible, but they can persuade the reader when treatments are not intuitive or difficult

Multiple manipulations

Back to our example: we also want to know whether e-mail campaigns affect the voters

We decide to send to half of our subjects an email containing information about the candidates' programs, and to the other half an email containing information about a new TV show

Since we are already planning the experiment about TV debates, we decide to use the same pool of respondents

Thinking about it, we realize that we can also compare the effect of the two, and see how they interact

Factorial Design

We shall call our manipulations **factors**

Each form that one manipulation can take is a **level** of a factor

A **condition** (or “treatment”) in this case is a combination of levels of different factors

When many combinations of factors are of interest, the experiment is called a **factorial experiment**

Factors do not have to be manipulations: we can also create factors based on observables (like in the block design) in case we have theoretical reasons to do so

Designing factorial experiments

It is useful to distinguish factors in two types:

Treatment factors

- Factors based on our manipulation
- They are of direct interest for our theory, or
- They are expected to modify the effect of factors that are of direct interest, or
- They are necessary for the design

Classification factors

- Factors based on levels of observables
- They are included for control, or
- They are expected to modify the effect of factors that are of direct interest (i.e. will allow us to see whether the effects are heterogeneous across subpopulations)

Measurement in factor levels

Qualitative

- There is no natural order between levels, and each level is of interest
- This is by far the most common type

Ordinal

- There is a natural order between levels, but the extent of the increase is not quantifiable

Quantitative

- Levels are ordered, and the increase between two levels is quantifiable
- Normally, some specific values of interest are chosen

Our example

- Our example is a 2 × 2 design
- It is usually represented in one of the 2 following ways:

1) "Stacked" form

Watch the debate YES	E-mail YES	Condition 1
Watch the debate NO	E-mail YES	Condition 2
Watch the debate YES	E-mail NO	Condition 3
Watch the debate NO	E-mail NO	Condition 4

2) Matrix form

	E-mail YES	E-mail NO
Watch the debate YES	Condition 1	Condition 3
Watch the debate NO	Condition 2	Condition 4

Number of Factors, Levels, Conditions

The more factors are in the experiment, and the more levels for each factor, the more conditions there will be

For an experiment with F factors, each of them consisting of N_L levels, the total number of conditions N_C is equal to

$$N_C = N_{L1} \times N_{L2} \times \ \ldots \times N_{LF}$$

- For instance, in our case it is 2 × 2 = 4

This has implications for the number of subjects that you will need to run the experiment!

This is an important aspect to keep in mind before designing factorial experiments with many factors

Random Assignment in Factorial Experiments

The assignment to the levels of one factor should be independent from the assignment to the levels of other factors

In other words, every observation has to have the same chance to get any of the N_C conditions

In our example, we can

- Run the random assignment once for all conditions (with probability ¼)
- Run the random assignment independently for the two factors (with probability ½ and ½)

Main Effects

How do we look at the effect of individual factors on the outcome?

Watch the debate YES	E-mail YES	Condition 1
Watch the debate NO	E-mail YES	Condition 2
Watch the debate YES	E-mail NO	Condition 3
Watch the debate NO	E-mail NO	Condition 4

Main Effects

To find the effect of the debate we compare

- Condition 1 & 3 to Condition 2 & 4

To find the effect of the e-mail we compare

- Condition 1 & 2 to Condition 3 & 4

The assumption is that the effect produced by one factor is the same across all the levels of the other factors

In other words, the effects are additive

Main Effects

How reasonable is this assumption?

Depends on the nature of the treatments and of the phenomenon that we are examining

- In our case, both treatments contain political information about the two candidates running
- Given that all respondents comply, and given that political information has an effect, our assumption implies that decisiveness is a linear function of information
- Every bit of information that we add, affects decisiveness to the same extent

Interaction Effects

If we do not think that this is the case, we can look for **interaction effects** between the manipulations

We observe an interaction effect when the effect of one variable on the outcome depends on the state of another variable

- For instance, the information coming via the debate and the e-mail may be redundant, so the effect of one manipulation on indecision may be weaker for the subjects who were assigned the other manipulation as well

Interaction Effects

Watch the debate YES	E-mail YES	Condition 1
Watch the debate NO	E-mail YES	Condition 2
Watch the debate YES	E-mail NO	Condition 3
Watch the debate NO	E-mail NO	Condition 4

Which comparison do we have to make in this case and what do they mean?

On Precision

- In general, in our experiment we want to maximize the estimated effect of the treatment, hence to minimize the noise
- The more treatment factors we add, and the more levels they have, the harder it gets to achieve this
- Block design gets cumbersome when we want to control for many observables
- In many survey experiments, we may be forced to resort to “simple” (aka non-stratified) random assignment
- In a full factorial design, the number of conditions gets out of hand fast
- One important thing to keep in mind is that, to maximize precision in the estimate, we need many observations

Statistical Power

What is the problem when our estimates of the treatment are *noisy*?

We increase the chance to make a Type II error

That is, to accept the null hypothesis (H_0) when the alternative hypothesis (H_1) is true

In other words, we lose statistical power –the probability to reject H_0 when H_1 is true

In such a case, even if there is an effect, our experiment will fail to detect it

Precision and number of observations

- In statistical applications, the size of the standard error of an estimate is inversely proportional to the number of observations used to obtain the estimate
- For instance, the standard error of the estimated difference between two group means is:

$$\sigma_{M_1 - M_2} = \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$$

- Where σ_1^2 is the estimated variance of Group 1 and σ_2^2 is the estimated variance of Group 2; and n_1 and n_2 are the number of observations in each group
- The more observations we have in our experiment, the more likely it is that, if there is any effect, it will be statistically significant

Precision and number of observations

What is the logic of this?

- When we work on a sample, the goal is often to generalize our finding beyond the sample itself, to the population from which the sample comes
- When we calculate the treatment effect, we want to generalize our finding to the whole target population
- Measures of precision (like the standard error) tell us how much the parameter is expected to *vary from sample to sample*
- Obviously, the larger our sample, the smaller is the variation *by chance*
- Hence, the larger our sample, the more confident we are that the parameter we observed is similar to the true population parameter

Power and effect size

Small effects are harder to detect than large effects

This is intuitive if you think that statistical hypothesis testing often implies dividing the size of the effect by the error of the estimate

So, given a certain magnitude of the error, a large effects are more likely to detect – hence, they bear more statistical power

Effect size and number of observations

Since the error is inversely correlated with the sample size, and

Statistical hypothesis testing are obtained dividing the effect size by the error,

We conclude that:

We need a larger sample to detect a smaller effect

Determining the number of observations

- There are formulas to calculate the number of observations that we should have in our study given the effect size that we expect, the variance in the population, and the power that we want to achieve

However, they require

- Having a quantitative expectation of the effect size
- Knowing population parameters such as the variance

- Moreover, in many cases we will have economic constraints, so we can’t really add as many observations as we like

Determining the number of observations

In fact, there is no “magic” method to tell how many observations are necessary

Consider that the weaker a treatment (because e.g. of noncompliance, or simply because the “real” treatment effect is minimal), and the more disturbance associated to it (because e.g. of heterogeneity in our data), the harder to detect any effect

In such a case, you will need a larger sample to obtain statistical significance

One way to get this information: pre-test

Validity

One important concern of experimental researchers (and reviewers) is about the validity of the experimental findings

What is validity?

Generally speaking, validity is the approximate truth of the inference we are making

How true is the statement that $\underline{T \textit{ causes } Y}$

This implies, whether or not we can prove that our explanation (our theory) is correct

Two types of validity

Applied to any empirical study, this question can be divided in two sub-questions:

1. To what extent our results actually show what we think they show
2. To what extent our results are likely to be the same in populations other than our data

These two questions are at the basis of what we call, respectively internal validity and external validity

Internal Validity

Internal validity is about how valid our experimental results are *per se*

It is made of three components

Statistical Validity

- Is the observed relationship between T and Y statistically significant and sizable?

Causal Validity

- Is the relationship that we observe *causal*?

Construct Validity

- Is the relationship that we observe a good proxy for the relationship that we theorize?

Statistical Validity

Statistical validity is what we strive to ensure when we apply sophisticated methods in our data analysis

It relates to find the model that takes into account:

- The structure of relationship between variables
- The way the variables are measured

In other words, the statistical model that best describes the DGP

Moreover, statistical validity refers to the robustness of our results: whether they take into account the uncertainty of the data appropriately, given alternative sets of assumptions

Causal Validity

Experiments, more than analyses of observational data, maximize the causal validity of our study

This is the case because of manipulation, random assignment, and control

Construct Validity

Construct validity is what most people mean with "internal validity" of experimental results

It refers to the extent to which our empirical analysis matches our theory

- To what extent our variables measure the abstract concepts that we discuss in our theory
- To what extent a treatment reproduces and singles out the causal factor that we hypothesize has an effect on the outcome

These questions are ultimately about the quality of the experiment

Construct Validity and Nonresponsiveness

Construct validity is threatened by noncompliance and by the presence of missing data

Noncompliance

- If many of our subjects do not comply with the treatment, then the observed effect is weakened

Missing data

- If it is more difficult to observe the outcome for subjects in one conditions than for subjects in another, that may produce a biased estimation of the treatment effect

A patch on this problem is claiming that we observe the Intent-to-Treat effect

If missing data are at random, there are formulas to correct for it when calculating the ITT effect

Compliance and experimental design

- A point about compliance that can be generalized about construct validity is that it depends on the subjects taking the experiment seriously
- The more psychologically engaged the subjects are, the more effective (and valid) the treatment will be
- This may force experimenters to put the subjects in situations that are more artificial (e.g. lab settings where they are being monitored)
- A good practice to follow is to take the point of view of the subjects
- How would you react if you were asked a certain question in a certain way? What would you think if you were asked to perform a certain task?

Measurement

Construct validity also relates to the way we measure the abstract concepts described in our theory

The more *stretches* we need to adapt what we measure with what we theorize, the worse for validity

A great deal of research in psychology is about how to measure abstract constructs

- E.g. measures of personality profiles, ideology, etc.

In many cases we can take advantage of pre-validated measures

Self-Reports vs. Behavioral Measures

Let us consider the outcome variable in our examples: indecision, observed asking respondents who they will vote for at the election

This is a **self-report**: subjects stating that they will do (or have done) something

However, what we care ultimately is the actual voting behavior

Using a self-report requires us to trust the subjects that they will indeed do what they said they will do

Self-Reports and stretches

This trust is an example of a *stretch*:

- It puts what we observe is one step further from what we theorize
- It requires that we have faith that the reported intention or behavior corresponds to truth
- In theory, faith is not a good tool for scientific investigation
- In practice the amount of faith required to interpret some social science results is immense

This reduces the construct validity

- We hypothesize that T affects voting behavior
- We find that T affects the reporter voting behavior (or intention)
- Most likely, it is correlated with actual behavior, but not perfectly

Behavioral Measures

In the case of voting behavior, there is no alternative

However, when our theory is about the effect of T on a behavior, and when we can, we should try to observe the behavior itself

Example

- Emails sent to local governors asking for information, signed with different names: some with typical white names, others with typical black names (Butler & Broockman, 2011)
- Behavioral outcome: whether the governors respond to the email requests to people of different races

Other measures not based on self-reports

- Another alternative is to measure responses that are not under the direct control of the subjects' rational decision-making
- For instance, measuring response time to associations between concepts may case a light on the subjects' implicit attitudes
- Another possibility is to observe physiological reactions like skin conductance, to directly measure the subjects' arousal level
- Observe brain activity using fMRI (or EEG, cheaper but less informative) in response to given stimuli or tasks
- Eye-tracking techniques allow to identify where subjects look on a screen – potentially useful as a treatment check as well
- These techniques require more resources, and can hardly be used out of a lab
- However, they provide valid measures that may be of interest

External Validity

- External validity is about whether our results generalize beyond the target population of our study
- This does not only relate to whether the target population is representative or not: one could seek to generalize to other target populations as well
- External validity is hardly achieved with one study only, as it requires the study to be replicated on different target populations or circumstances
- People tend to mix it up with ecological validity

Replicability

- External validity is mostly related to **replicability**
- When an experimental study can be replicated in different populations, different contexts, or with different designs, then the study has a high external validity
- What does it mean?
- That the theory postulated by the study reflects a *generally true* aspect of human behavior (and not limited to a specific case)
- Replications can also cast a light on the conditions under which some results can be replicated
- This helps refining the theory, adding nuances to the mechanism that we describe

Ecological Validity

Ecological validity is the extent to which the conditions in which our experiment is conducted are similar to the real-world conditions where individuals leave

It is also called the “mundane realism” of the experiment

This sometimes comes at the expenses of the “experimental realism”, that is what ensures that subjects get psychologically engaged by the experiment

To sum up

What effect we want to observe and how we can make the best out of it

- Non-response and non-compliance
- More than one single manipulation: experiments can be made more complex using factorial design, however we need to be cautious
- Interaction effects between different manipulations and between manipulations and observable variables
- How many observations? Hard to tell, but we should think about it
- Internal validity, and issues of nonresponse and measurement
- External validity, replicability, artificiality