Experimental methodology[1]

Ondřej Krčál

DXH_MET1, MUNI
23.–24. 11. 2023

[1] *Some sources*: James Tremewan's course thought at MUNI in May 2022, *Handbook of Research Methods and Applications in Experimental Economics* and *Handbook of Experimental Economic Methodology*

Are you using experimental methods in your research?

Are you using experimental methods in your research?

Some kind of experimental design necessary for causal inference.

Are you using experimental methods in your research?

Some kind of experimental design necessary for causal inference.

You can look for experimental designs in observational data (Štěpán Mikula).

Are you using experimental methods in your research?

Some kind of experimental design necessary for causal inference.

You can look for experimental designs in observational data (Štěpán Mikula).

Or you can produce your own experimental data – popular in many fields:

- **Gender and labor market** – huge literature in the lab and field
- **Enviromental economics** – common lab and field studies
- **Management/adaptation** – common experiments: (1, 2)
- **DSGE** – lab studies: (1, 2)

Overview

Economic experiments (as compared to surveys) have two main features:

- **randomization**
- incentivization (exceptions: survey experiment, some lab studies)

Overview

Economic experiments (as compared to surveys) have two main features:

- **randomization**
- incentivization (exceptions: survey experiment, some lab studies)

Running **experiments** (as opposed to using experiments in the data):

- Design
  - informed by **theory**
  - to be able to **analyze data efficiently** (power, avoid assumptions)
- Admin
  - **Pre-registration**
  - **Ethics approval**
- Data collection (depends on the type of the experiment)
- Data analysis

MU Experimental Economics Laboratory (MUEEL)

We offer research infrastructure:

- **experimental laboratory** – hroot, otree programmers, support (lab manager, payments)
- access to **software** and **equipment** (Qualtrics, Veyon, Air Quality, ...)
- **experience** with lab/field/survey experiments and agent-based models
- **network of experts** – local (CERGE-EI, VŠE, WU) and broader

Questionnaires in the lab

Scientific disciplines are as precise as the measurements of basic concepts.

Main example in economics: **preferences** (consistency, stability, values/distributions)

- risk preferences
- time preferences: discount rate, present bias
- social preferences: e.g. altruism, trust, reciprocity

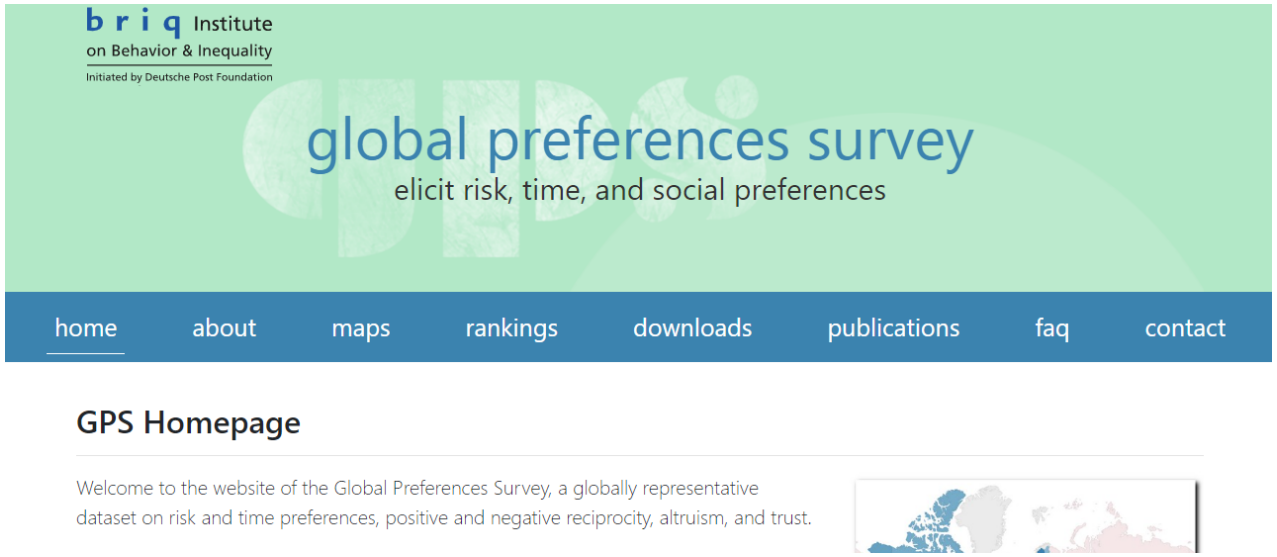
Questionnaires in the lab

Scientific disciplines are as precise as the measurements of basic concepts.

Main example in economics: **preferences** (consistency, stability, values/distributions)

- risk preferences
- time preferences: discount rate, present bias
- social preferences: e.g. altruism, trust, reciprocity

Global preference survey:



Theory and experiments (1/2)

Economic theory provides **structure for the examination** of how people behave in economic situations.

Experiments not motivated by theory may lead to **duplication of effort**.

Theory and experiments (1/2)

Economic theory provides **structure for the examination** of how people behave in economic situations.

Experiments not motivated by theory may lead to **duplication of effort**.

Experiments are best at:

- **testing theories** – make comparative-static predictions (about direction and not about size/parameters of the model)
- **testing assumptions** field work is based on

Theory and experiments (2/2)

Economic theory

- **is strong and works**: *e.g. market experiments*
- **is strong and does not work**: *e.g. expected utility theory*
- **is weak**: *e.g. ultimatum game*
  - selfish preferences are not theory
  - experience

Theory and experiments (2/2)

Economic theory

- **is strong and works**: *e.g. market experiments*
- **is strong and does not work**: *e.g. expected utility theory*
- **is weak**: *e.g. ultimatum game*
  - selfish preferences are not theory
  - experience

Adjustments to make theory **more useful**, e.g.

- **other-regarding preferences** (Fehr & Schmidt, 1999; Bolton & Ockenfels, 2000)
- **quantal response equilibrium** (McKelvey & Palfrey, 1995)
- one-one shot interactions – **level-k theory** (Stahl & Wilson, 1995)

Design choices: treatments

Random assignment to treatment $\rightarrow$ **causal inference**
Treatments should differ **only in one** thing!

Design choices: treatments

Random assignment to treatment $\rightarrow$ **causal inference**
Treatments should differ **only in one** thing!

Sometimes experiments involve multiple factors, e.g. communication (with/without) and stake-size (low/high) in a dictator game.

Full factorial design would lead to 2 x 2 = 4 designs.

Design choices: treatments

Random assignment to treatment $\rightarrow$ **causal inference**
Treatments should differ **only in one** thing!

Sometimes experiments involve multiple factors, e.g. communication (with/without) and stake-size (low/high) in a dictator game.

Full factorial design would lead to 2 x 2 = 4 designs.

Rule of thumb: Use the minimum number of treatments possible to test your hypothesis!

- Why?
  - More treatments – problem of **multiple testing**.
  - Fewer treatments = more observations per treatment
- Avoid
  - **intermediate treatment** unless interested in **non-linearities**.
  - **full factorial design** unless interested in the **interaction term**.

Design choices: within/between subject design

Within-subject design: Each subject in more than one treatments

- Advantages:
  - Effect on individual
  - Greater power
- Disadvantages:
  - Subjects may want to choose the same to be **consistent**
  - **Experimental demand effect** = changes in behavior due to cues about what constitutes appropriate behavior
  - Potential **order effects** = order of tasks affects decisions

Design choices: within/between subject design

Within-subject design: Each subject in more than one treatments

- Advantages:
  - Effect on individual
  - Greater power
- Disadvantages:
  - Subjects may want to choose the same to be **consistent**
  - **Experimental demand effect** = changes in behavior due to cues about what constitutes appropriate behavior
  - Potential **order effects** = order of tasks affects decisions

Between-subject design: Each subject only in one treatment

- Advantages – solves problems of demand effects and consistency
- Disadvantages – power and subjective judgment (Birnbaum, 1999)

Design choices: incentives

- Transparent (*minimum payment* in lab experiments)
- Incentive compatible
- Pay all vs. random incentive system (RIS) (Clot et al., 2018)

Other design choices

- **One-shot vs. repeated** play of a game/situation
  - advantages: allow learning, interested in learning effects
  - distadvantages: dilutes incentives (random lottery incentive mechanism to avoid wealth effects)

Other design choices

- **One-shot vs. repeated** play of a game/situation
  - advantages: allow learning, interested in learning effects
  - distadvantages: dilutes incentives (random lottery incentive mechanism to avoid wealth effects)
- **Partner vs. stranger matching**
  - **Partner matching** – play with the same subjects each time
  - **Stranger matching** – randomly rematched every time in a given **matching group** (**perfect stranger matching** if played always with someone else)

Other design choices

- **One-shot vs. repeated** play of a game/situation
  - advantages: allow learning, interested in learning effects
  - distadvantages: dilutes incentives (random lottery incentive mechanism to avoid wealth effects)
- **Partner vs. stranger matching**
  - **Partner matching** – play with the same subjects each time
  - **Stranger matching** – randomly rematched every time in a given **matching group** (**perfect stranger matching** if played always with someone else)
- **Direct response vs. strategy method** in sequential games:
  - **direct-response**: P1 makes a decision, P2 is informed of the decision then makes their own decision.
  - **strategy method**: P1 makes a decision, and simultaneously P2 decides for all possible choices of P1. The strategies of the two players are combined and the outcome determined.

Power calculations

Power is the probability of finding an effect that is really there.

Underpowered studies (too few subjects):

- less likely to detect real treatment effects.
- more likely to have a large effect, even when there is no real effect

Power calculations

Power is the probability of finding an effect that is really there.

Underpowered studies (too few subjects):

- less likely to detect real treatment effects.
- more likely to have a large effect, even when there is no real effect

Useful exercise. **Problem**: you need to know the effect size.

Especially useful to get some idea about number of subjects when interested in more complicated effects (e.g. interaction effect)

Replication crisis

- **Psychology**: 36/100 studies replicated, effect sizes 50% of original effect (Open Science Collaboration, Science, 2015)
- **Experimental economics:** 11/18 (61%) studies in AER or QJE replicated, effect sizes 66% of original effect (Camerer et al, *Science*, 2016)
- **Social sciences:** 13/21 (62%) studies in Nature or Science, effect sizes 50% of original effect (Camerer et al, *Science*, 2018)

Replication crisis

- **Psychology**: 36/100 studies replicated, effect sizes 50% of original effect (Open Science Collaboration, Science, 2015)
- **Experimental economics:** 11/18 (61%) studies in AER or QJE replicated, effect sizes 66% of original effect (Camerer et al, *Science*, 2016)
- **Social sciences:** 13/21 (62%) studies in Nature or Science, effect sizes 50% of original effect (Camerer et al, *Science*, 2018)

Reasons:

- Publication bias
- The way tests work (sample size, α, reasonable hypotheses)
- P-hacking
- Testing multiple hypotheses
- ...

P-hacking

- Selectively removing outliers
- Running multiple tests – reporting lowest p-value
- Running multiple regressions specifications and reporting only the one that works.
- Including pilot data to the analysis
- Stopping gathering data as soon as $p < 0.05$

Multiple testing

Problem when using multiple simultaneous tests (common in experimental economics – multiple outcomes/subgroups/treatments).

Family-wise error rate (FWER): Probability of at least one false rejection for k statistically independent tests at α level:

$$\text{FWER} = 1 - (1 - \alpha)^k$$

For $\alpha = 0.05$ and $k = 2, 3, 4, \ldots$, FWER $= 0.098, 0.143, 0.185, \ldots$

Multiple testing

Problem when using multiple simultaneous tests (common in experimental economics – multiple outcomes/subgroups/treatments).

Family-wise error rate (FWER): Probability of at least one false rejection for k statistically independent tests at α level:

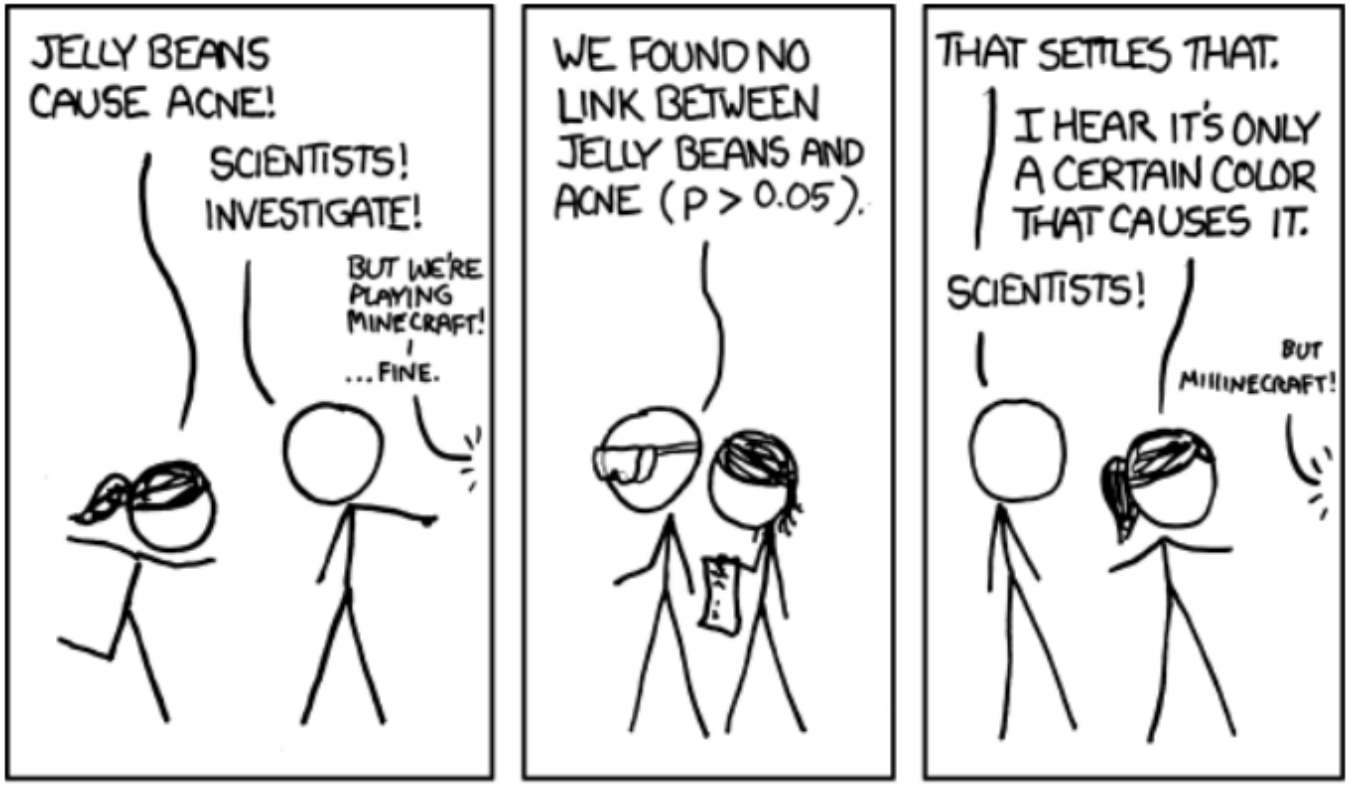
$$\text{FWER} = 1 - (1 - \alpha)^k$$

For $\alpha = 0.05$ and $k = 2, 3, 4, \ldots$, FWER $= 0.098, 0.143, 0.185, \ldots$

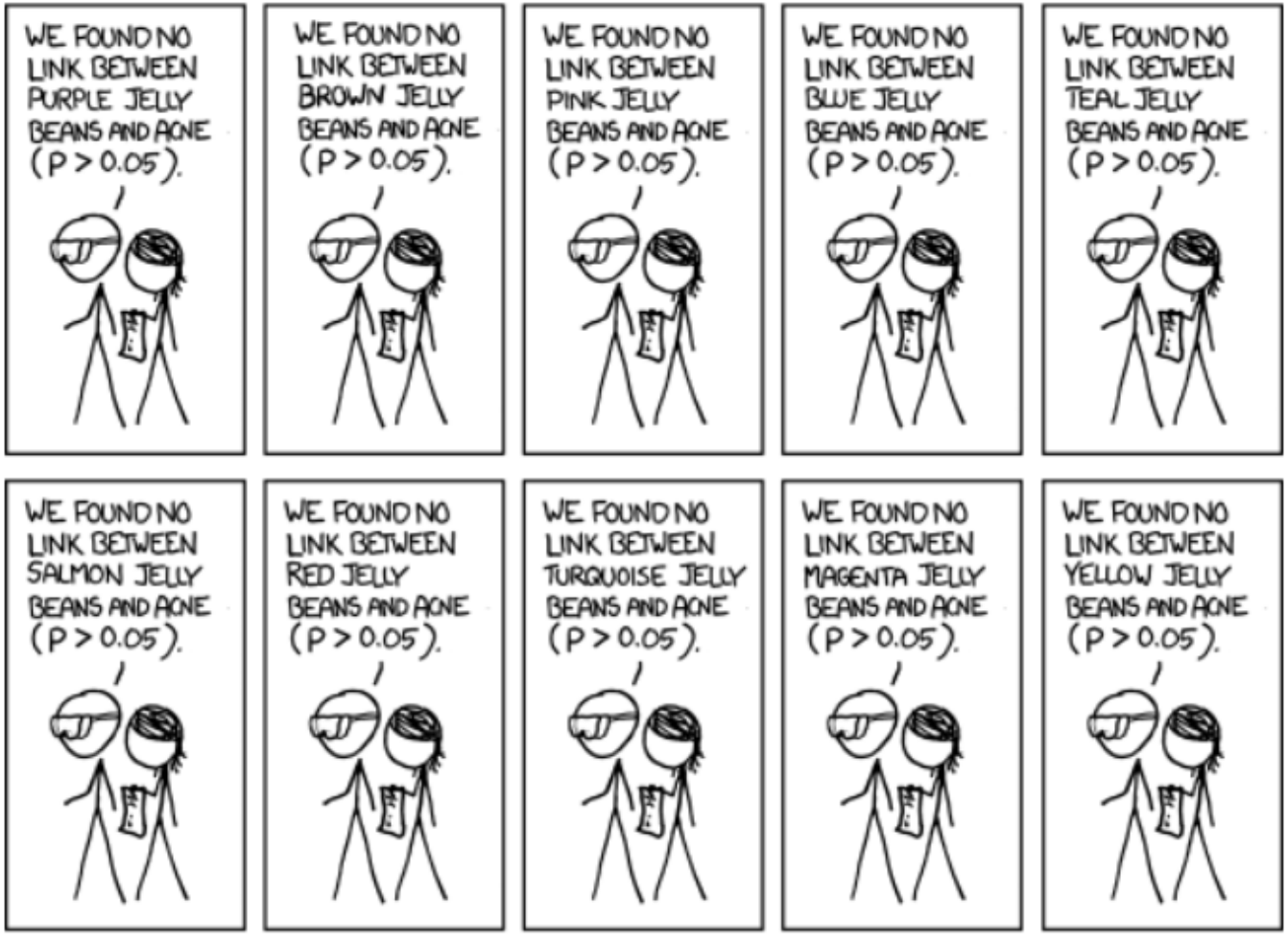
There are methods used for adjusting for multiple hypotheses:

- Bonferroni
- Holm-Bonferonni
- List (2019)
- ...

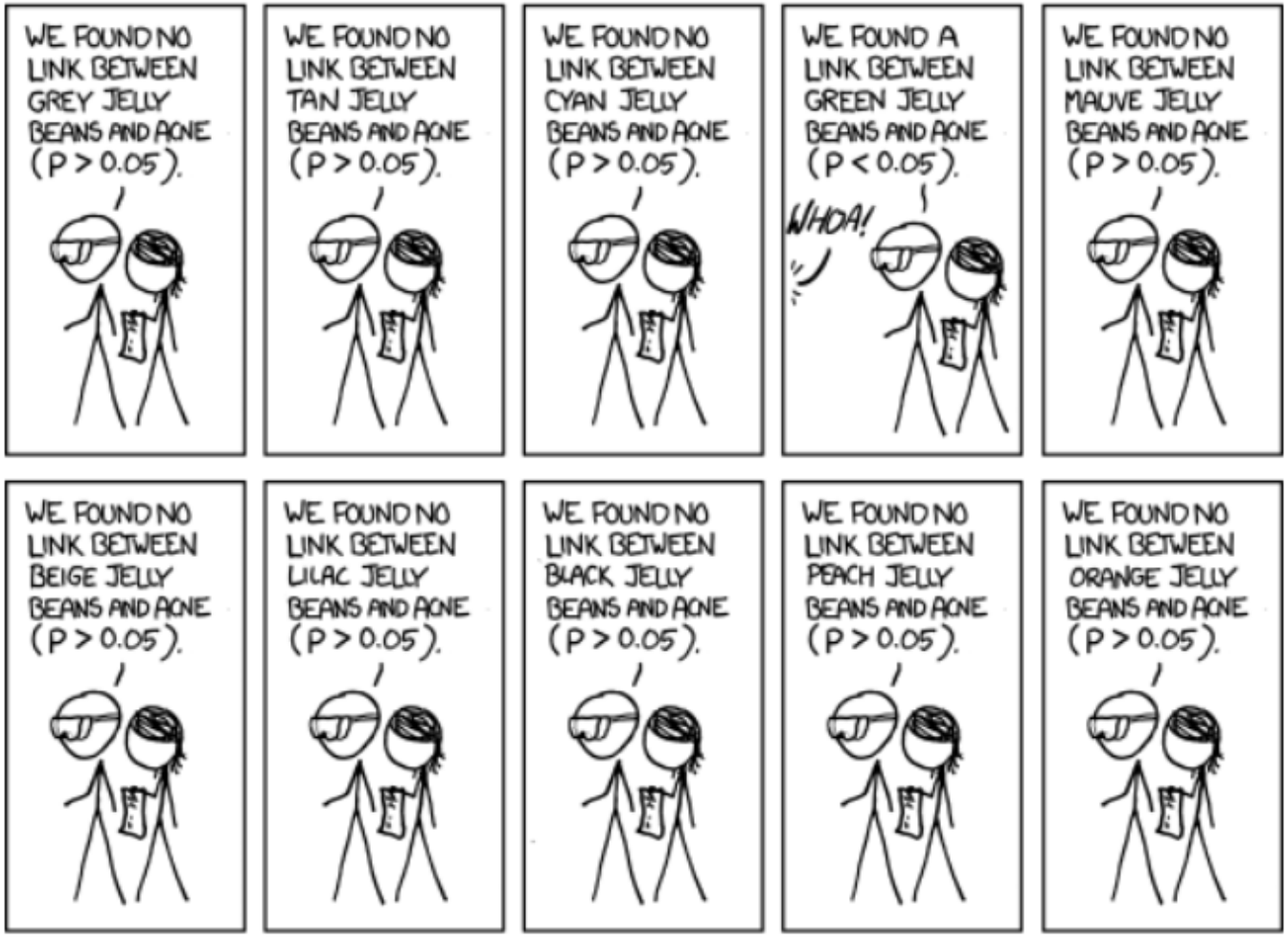
When to adjust?



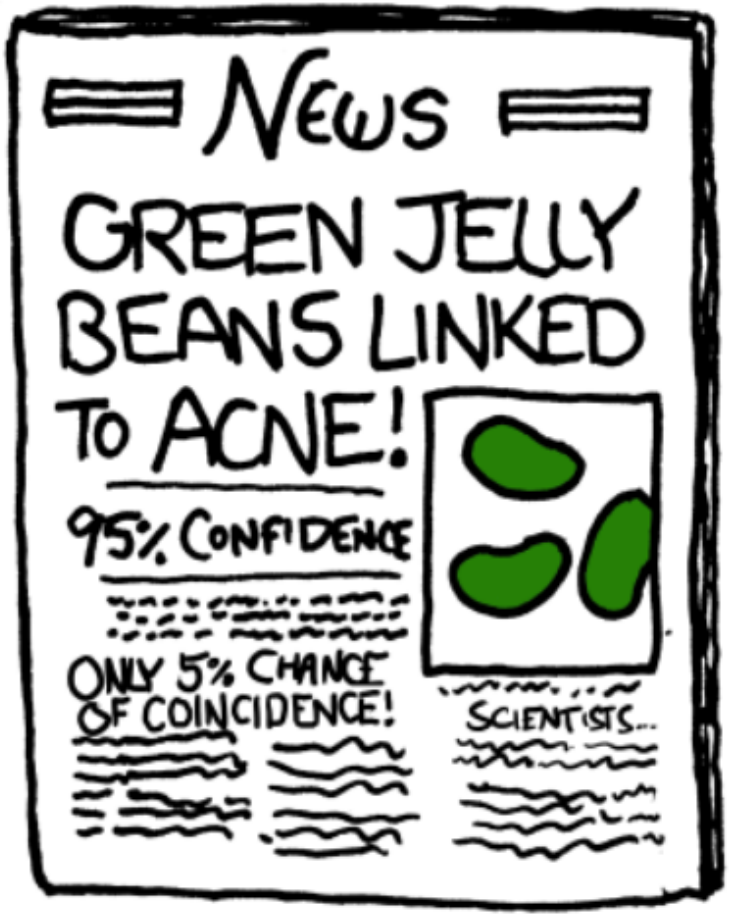
When to adjust?



When to adjust?



When to adjust?



https://xkcd.com/882/

When to adjust?

Should we adjust for multiple hypotheses?

- Yes, if the null hypothesis is that jelly beans cause acne. Why test it separately by color, then?
- Perhaps not, if you want to test the link between colour and acne (if you have a theory why each tested color might cause acne)
- Always important to report the tests to all hypotheses (and to pre-register)

Pre-registration

You register your desing prior to data collection (AEA RCT Registry, AsPredicted).

What do you register?

- number of observations – problem with stopping
- hypotheses – multiple comparisons

Pre-registration

You register your desing prior to data collection (AEA RCT Registry, AsPredicted).

What do you register?

- number of observations – problem with stopping
- hypotheses – multiple comparisons

When important/necessary?

- design leads to many potential hypotheses (problematic for exploratory studies)
- expected by journals – field experiment, now increasingly common in lab experiments

Ethics approval

- Design (social importance)
- Procedures
- Hiring
- Ethics questions

Ethics approval

- Design (social importance)
- Procedures
- Hiring
- Ethics questions

Necessary/more important when

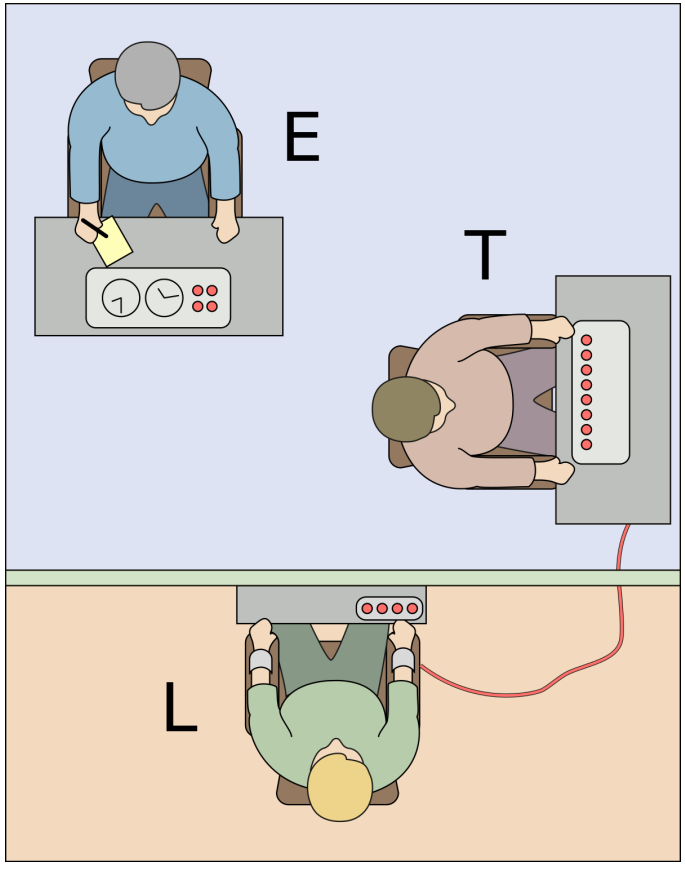
- **expected by the journal** (in some countries obligatory whenever research on people)
- we collect personal information
- subjects can be **harmed** (e.g. we induce stress or strong emotions)/use of vulnerable population
- we take biological material (measuring hormone levels)
- we use **deception**

Deception

- widespread in (social) psychology, factually **banned** in laboratory experimental economics
- **reputational rather than ethical** concerns (*ex ante* informed consent, *ex post* debriefing)
- acts of **commission** (e.g. lying to subjects) vs. acts of **omission** (not telling subjects things)
- evidence (Ortmann and Hertwig, 2002) shows that economists should avoid acts of commission and acts of omission that violate default assumptions
- and what about field experiments? difficult to get informed consent - ethically problematic (audit studies)

Examples: Milgram experiment

Milgram, S. (1963). Behavioral study of obedience. *The Journal of abnormal and social psychology*



Public Announcement

WE WILL PAY YOU $4.00 FOR ONE HOUR OF YOUR TIME

Persons Needed for a Study of Memory

*We will pay five hundred New Haven men to help us complete a scientific study of memory and learning. The study is being done at Yale University.

*Each person who participates will be paid $4.00 (plus 50c carfare) for approximately 1 hour's time. We need you for only one hour: there are no further obligations. You may choose the time you would like to come (evenings, weekdays, or weekends).

***No special training, education, or experience is needed. We want:**

Factory workers	Businessmen	Construction workers
City employees	Clerks	Salespeople
Laborers	Professional people	White-collar workers
Barbers	Telephone workers	Others

All persons must be between the ages of 20 and 50. High school and college students cannot be used.

*If you meet these qualifications, fill out the coupon below and mail it now to Professor Stanley Milgram, Department of Psychology, Yale University, New Haven. You will be notified later of the specific time and place of the study. We reserve the right to decline any application.

*You will be paid $4.00 (plus 50c carfare) as soon as you arrive at the laboratory.

TO:
PROF. STANLEY MILGRAM, DEPARTMENT OF PSYCHOLOGY, YALE UNIVERSITY, NEW HAVEN, CONN. I want to take part in this study of memory and learning. I am between the ages of 20 and 50. I will be paid $4.00 (plus 50c carfare) if I participate.

NAME (Please Print). . . . . .

ADDRESS . . . . . .

TELEPHONE NO. . . . . Best time to call you . . . . . .

AGE . . . . . OCCUPATION . . . . . SEX . . . . . .
CAN YOU COME:

WEEKDAYS . . . . . EVENINGS . . . . . WEEKENDS . . . . . .

Examples: Color of a free ride

Mujcic, R., Frijters, P. (2021). The colour of a free ride. *The Economic Journal*

Controversy/opinion:

OPINION Bus driver research by UQ's Frijters and Mujcic deserves publication

04 MAR 2015

Research by University of Queensland economists Paul Frijters and Redzo Mujcic revealing discrimination by Brisbane bus drivers deserves to be published, writes Andreas Ortmann.

Types of experiments

- **laboratory experiments**
- **field experiments** (experiment run outside of laboratory)
  - **artefactual** *(lab-in-the-field)* – non-standard subjects (Henrich et al. 2001)
  - **framed** – the same as an artefactual, except that it incorporates important elements of the context of the naturally-occurring environment
  - **natural** – natural environment and subjects do not know that they are participants in an experiment
  - *strange category*: **survey experiments** – often population-based, typically without incentives

Laboratory experiments

Subject pool management (hroot, Orsee, Sona, Prolific, mTurk, ...)

- ethics (clear rules, ability to stop any time, anonymity)
- motivation (monetary incentives, regular invitations to experiments, good/fair experience)

Experimental laboratory (or online environment)

- privacy during choices (dividers)
- privacy during payments
- experimental environment (zTree, oTree, Qualtrics, ...)

Field experiments

- useful when the "right" population
- natural experiments are more valuable, but ethical concerns
- cooperation with authorities often needed

Hypotheses and p-values

Tests:

- *null (H_0)* vs. *alternative hypothesis (H_1)*
- **one-sided vs. two-sided test**
  - strong theoretical reasons $\rightarrow$ one-sided
  - competing theories (direction is unclear) $\rightarrow$ two-sided

Hypotheses and p-values

Tests:

- *null (H_0)* vs. *alternative hypothesis (H_1)*
- **one-sided vs. two-sided test**
  - strong theoretical reasons $\rightarrow$ one-sided
  - competing theories (direction is unclear) $\rightarrow$ two-sided

Interpret p-values carefully:

- It is not the **probability the null hypothesis is true**
- **Never "accept the null"**: $p > 0.1$ is not evidence the null is true
- A small p-value does not mean an effect is **important**

Types of tests

Treatment tests:

- **Non-parametric tests** do not assume that data comes from a particular probability distribution, e.g. normal distribution.
  - important because experimental data often has small sample size and is non-normally distributed
- At the cost of stronger (typically untestable) assumptions about true distribution of data, **parametric methods** allow us to do more interesting stuff, e.g.:
  - regression analysis (controls, non-independent data)
  - treating with heterogenous types and zeros
  - structural models (estimating utility functions)

Types of tests

Treatment tests:

- **Non-parametric tests** do not assume that data comes from a particular probability distribution, e.g. normal distribution.
  - important because experimental data often has small sample size and is non-normally distributed
- At the cost of stronger (typically untestable) assumptions about true distribution of data, **parametric methods** allow us to do more interesting stuff, e.g.:
  - regression analysis (controls, non-independent data)
  - treating with heterogenous types and zeros
  - structural models (estimating utility functions)

- **Exact tests** calculate the p-value exactly (true asymptotically)
- Non-exact tests are OK for a big-enough sample (but "big enough" depends on the true distribution of the data, which is unknown)

Overview of statistical methods used[2]

- Treatment tests
  - looking for least assumptions (non-parametric) and most power
  - depends on type of data (binary/multi-valued) and comparison (within/between subject)

[2]Moffatt, P. (2020). Experimetrics: Econometrics for experimental economics. Bloomsbury Publishing.

Overview of statistical methods used[2]

- Treatment tests
  - looking for least assumptions (non-parametric) and most power
  - depends on type of data (binary/multi-valued) and comparison (within/between subject)
- Regression analysis (parametric)
  - testing effect of more than one treatment simultaneously
  - controlling for additional variables
  - accounting for dependence between observations

[2] Moffatt, P. (2020). Experimetrics: Econometrics for experimental economics. Bloomsbury Publishing.

Overview of statistical methods used[2]

- Treatment tests
  - looking for least assumptions (non-parametric) and most power
  - depends on type of data (binary/multi-valued) and comparison (within/between subject)
- Regression analysis (parametric)
  - testing effect of more than one treatment simultaneously
  - controlling for additional variables
  - accounting for dependence between observations
- Structural modelling = estimating parameters of a utility function

[2] Moffatt, P. (2020). Experimetrics: Econometrics for experimental economics. Bloomsbury Publishing.

What is external validity

External validity (ecological validity, generalizability) – validity of applying the conclusions of a scientific study outside the context of that study (across other situations, people, stimuli, and times).

Internal validity – reflects how well the study is conducted and hence to what extent do the observed results describe in the population we are studying.

Levitt and List (2007)

Levitt and List (2007) – behavior in the lab (when measuring **social preferences**) is influenced by factors other than monetary incentives:

- the presence of **moral and ethical considerations**
- the nature and extent of **scrutiny of one's actions** by others
- the **context** in which the decision is embedded
- **self-selection** of the individuals making the decisions
- the **stakes** of the game

Camerer's reply to Levitt and List

Camerer (2011) comes with three arguments against the criticism:

1. The goal of experimental economics is to establish a general theory linking economic factors to behaviour. **Generalizability** from the lab to the field **is not a primary concern** in a typical experiment.
2. The factors listed by Levitt and List are not essential for all lab experiments (except for obtrusiveness, because of human subjects protection) and there is **little evidence that typical lab features undermine generalizability**.
3. Economics experiments designed to test lab-field generalizability show that **laboratory findings could be generalized to comparable field settings**.

Camerer's reply to Levitt and List

Camerer (2011) comes with three arguments against the criticism:

1. The goal of experimental economics is to establish a general theory linking economic factors to behaviour. **Generalizability** from the lab to the field **is not a primary concern** in a typical experiment.
2. The factors listed by Levitt and List are not essential for all lab experiments (except for obtrusiveness, because of human subjects protection) and there is **little evidence that typical lab features undermine generalizability**.
3. Economics experiments designed to test lab-field generalizability show that **laboratory findings could be generalized to comparable field settings**.

Comparing lab and field: Laboratory experiments are more easily replicated whereas field experiments are less obtrusive.

Lab experiments focus on qualitative effect

Kessler and Vesterlund (2015) argue that

- the debate concentrates around a **straw-man version of external validity** that quantitative results are externally valid
- for most laboratory studies it is only relevant to ask whether the **qualitative results** (= direction or the sign of the estimated effect) are externally valid
- laboratory studies are conducted to identify general principles of behavior and therefore promise to generalize

Lab experiments focus on qualitative effect

Kessler and Vesterlund (2015) argue that

- the debate concentrates around a **straw-man version of external validity** that quantitative results are externally valid
- for most laboratory studies it is only relevant to ask whether the **qualitative results** (= direction or the sign of the estimated effect) are externally valid
- laboratory studies are conducted to identify general principles of behavior and therefore promise to generalize

Different methodologies are not in competition. They are complementary.

Case: Are bankers cheaters? (1/6)

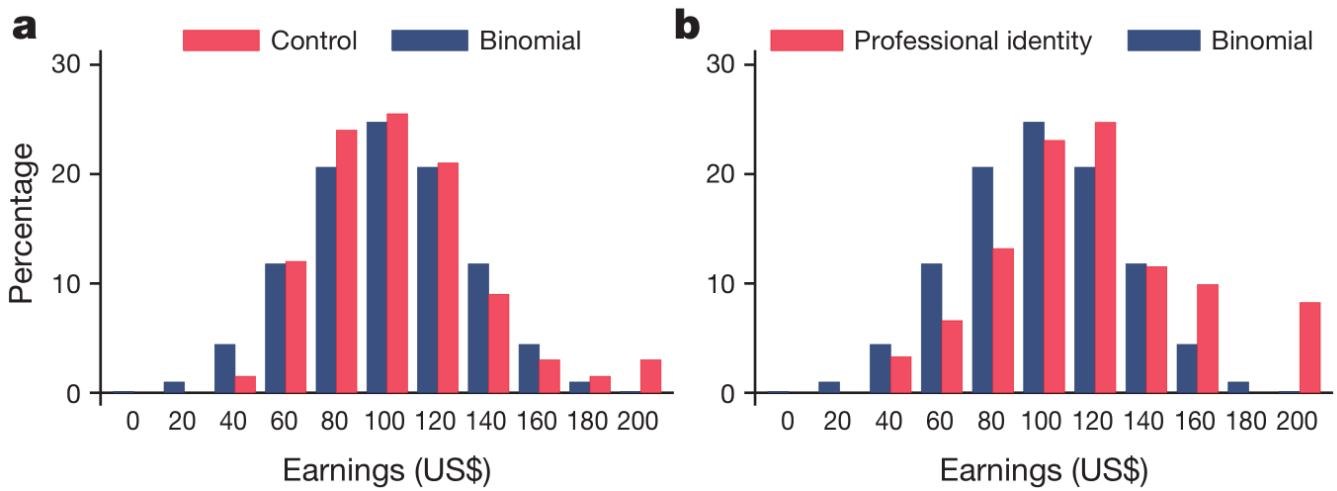
Cohn et al. (*Nature*, 2014) run a lab-in-the-field experiment with 128 bankers

- **Priming** professional identity with seven questions
  - **T:** about their professional background (e.g. "At which bank are you presently employed?" or "What is your function at this bank?")
  - **C:** that were unrelated tot heir profession (e.g. "How many hours per week do you watch television on average?")
- Playing a **cheating game** (10 coin throws earning \$20 or 0 each) + paid only if higher than earnings of a randomly drawn person from a pilot
- **Manipulation check** – converting word fragments into meaningful words (__oker_ = *broker* or *smoker*)

Case: Are bankers cheaters? (1/6)

Cohn et al. (*Nature*, 2014) run a lab-in-the-field experiment with 128 bankers

- **Priming** professional identity with seven questions
  - **T:** about their professional background (e.g. "At which bank are you presently employed?" or "What is your function at this bank?")
  - **C:** that were unrelated tot heir profession (e.g. "How many hours per week do you watch television on average?")
- Playing a **cheating game** (10 coin throws earning \$20 or 0 each) + paid only if higher than earnings of a randomly drawn person from a pilot
- **Manipulation check** – converting word fragments into meaningful words (*__oker* = *broker* or *smoker*)



Case: Are bankers cheaters? (2/6)

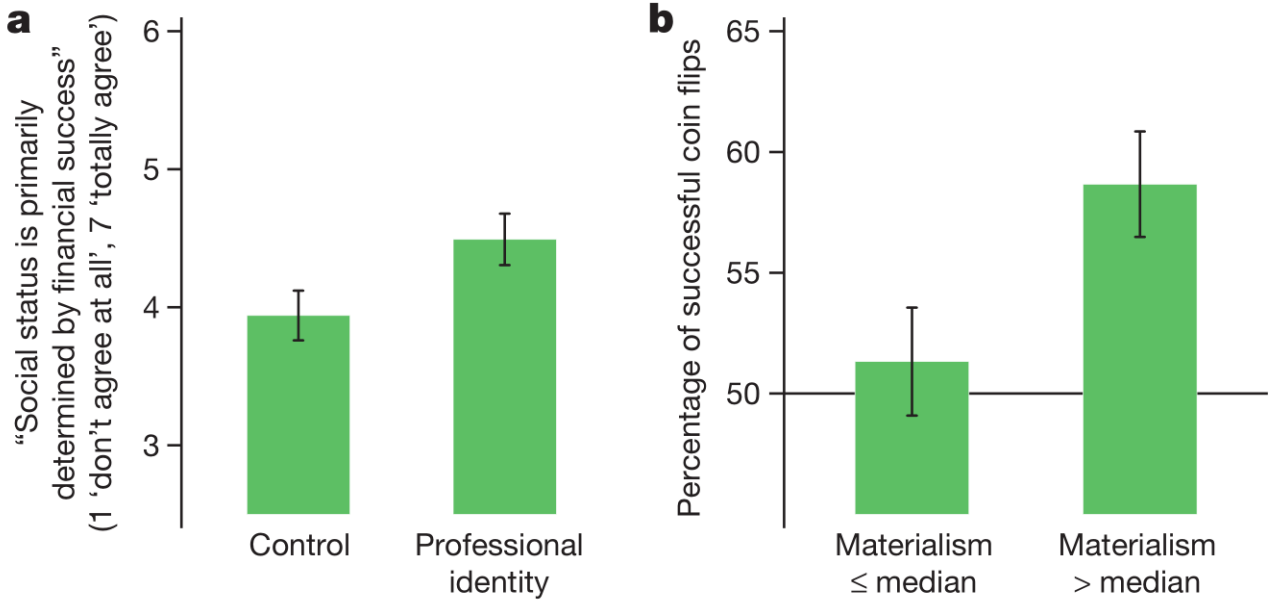
Several channels though which priming may have worked:

- **Competitive behaviour** in banking is intrinsically desirable (no evidence)
- Salience of **competitive incentive schemes** (no difference between *core* and *support* units)
- **Norm obedience** – what I should do/what others do (no difference in beliefs what others will do)
- Evidence on **materialistic values** below

Case: Are bankers cheaters? (2/6)

Several channels though which priming may have worked:

- **Competitive behaviour** in banking is intrinsically desirable (no evidence)
- Salience of **competitive incentive schemes** (no difference between *core* and *support* units)
- **Norm obedience** – what I should do/what others do (no difference in beliefs what others will do)
- Evidence on **materialistic values** below



Case: Are bankers cheaters? (3/6)

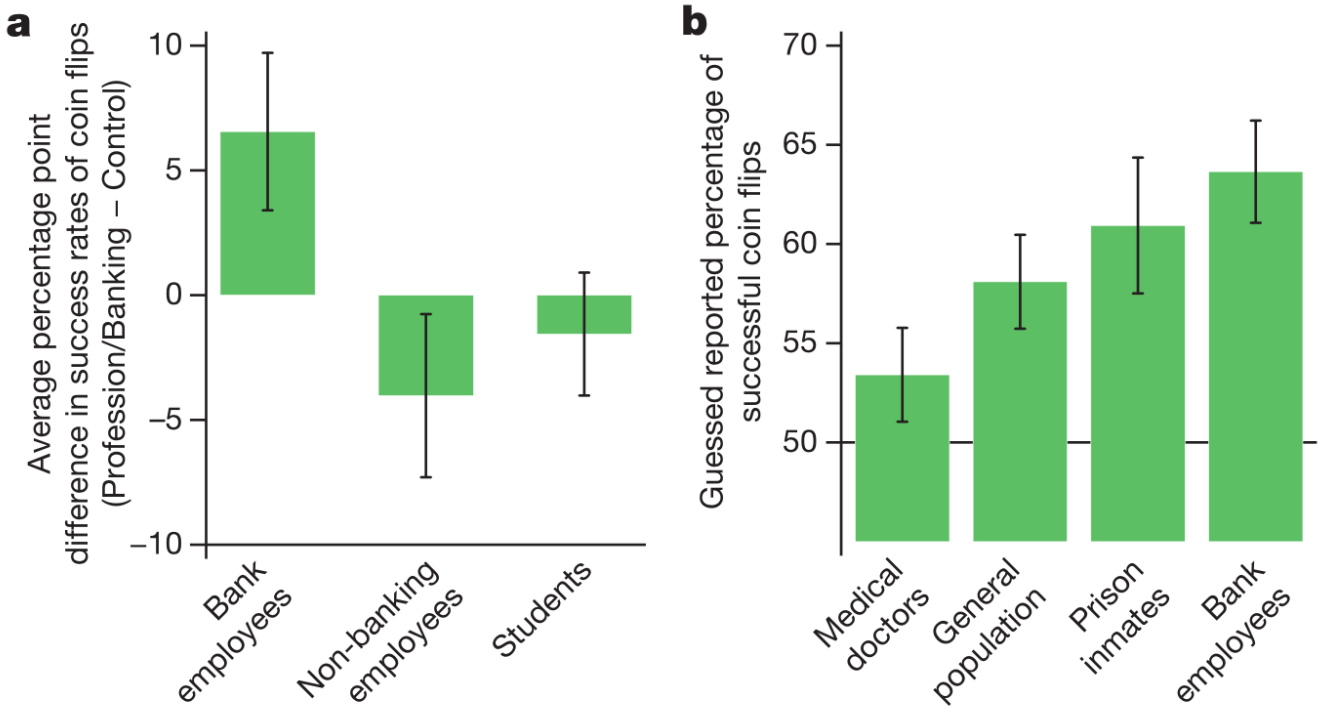
Other tests and robustness:

- other industries
- beliefs

Case: Are bankers cheaters? (3/6)

Other tests and robustness:

- other industries
- beliefs



Case: Are bankers cheaters? (4/6)

Criticism:

- Vranka and Houdek (*FrontPsy*, 2015) criticize interpretation:
  - **stereotype threat** instead of social norms
  - what is treatment? control group **primed by leisure activities** – call for **multiple control groups**
  - priming bankers **with money** instead of professional identity
  - why not **replicate** when they had bankers giving expectations?
- Hupe (2018) criticize the analysis (not published)
  - propose a different H0: *cheating the same in all six groups*
  - criticizes the use of MW test (ties)

Case: Are bankers cheaters? (5/6)

Replications:

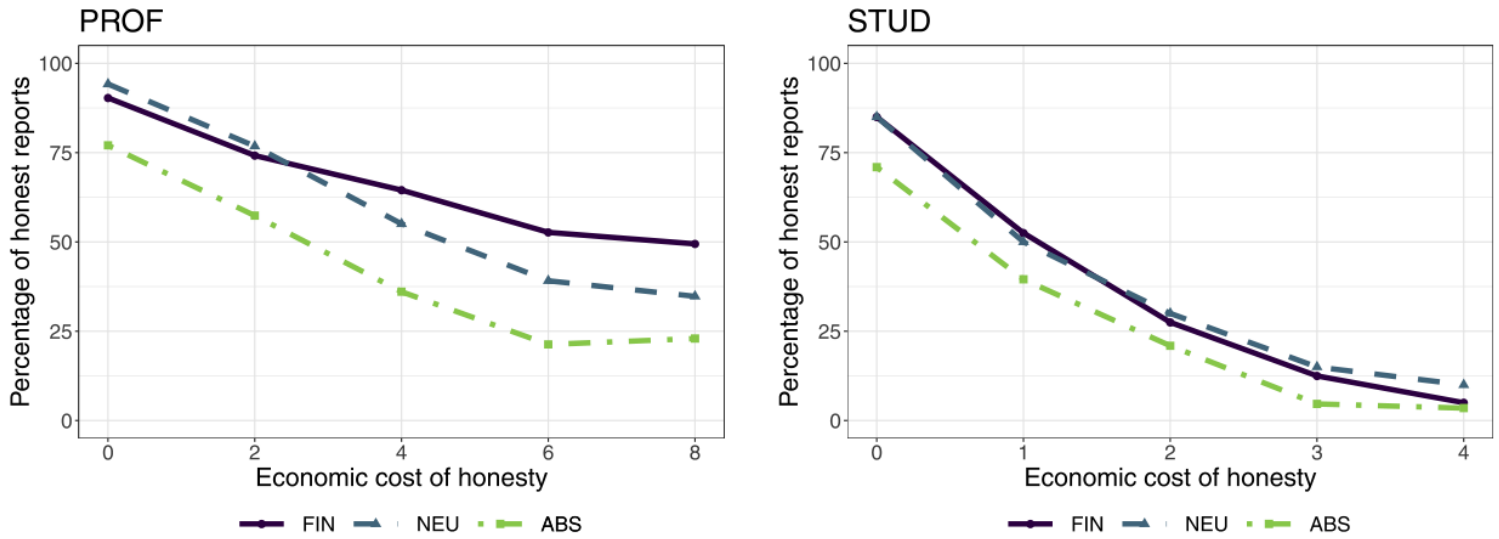
- Rahnwan et al. (*Nature*, 2019) do not replicate the result ($n = 768$)
- Cohn et al. (*Nature*, 2019) why the replication by Rahnwan not valid:
  - Media attention $\rightarrow$ selection (2 out of 27 banks)
  - Disclosed purpose at the beginning (demand effect)
  - Manipulation check did not work in Middle East
  - Basic retail services vs. investment/private banking
  - Study tests whether "*prevailing scandals involving fraud and unethical behaviour in the banking industry were partly the result of a problematic business culture, rather than, for example, the employment of dishonest people in the banking industry.*" instead of *"the prevailing business culture in the banking industry weakens and undermines the honesty norm*"
- Huber and Huber (*JEBO*, 2020)

Case: Are bankers cheaters? (6/6)

- Huber and Huber (*JEBO*, 2020):
  - Lying game (T: 31; F: 37) in which individual lying is observed
  - Framing:
    - **Abstract:** *Imagine there are two possible states of nature and one of your tasks is to report the current state.*
    - **Neutral:** *Imagine you are a security clerk at a museum and one of your tasks is to inform the manager each week about the average number of visitors in the preceding week.*
    - **Financial:** *Imagine you are the Chief Executive Officer (CEO) of a publicly listed company and one of your tasks is to inform shareholders each quarter about the course of business and the earnings per share.*

Case: Are bankers cheaters? (6/6)

- Huber and Huber (*JEBO*, 2020):
  - Lying game (T: 31; F: 37) in which individual lying is observed
  - Framing:
    - **Abstract:** *Imagine there are two possible states of nature and one of your tasks is to report the current state.*
    - **Neutral:** *Imagine you are a security clerk at a museum and one of your tasks is to inform the manager each week about the average number of visitors in the preceding week.*
    - **Financial:** *Imagine you are the Chief Executive Officer (CEO) of a publicly listed company and one of your tasks is to inform shareholders each quarter about the course of business and the earnings per share.*



Some more examples

A good strategy is to **combine experiments with observational data**:

- Bursztyn, Leonardo, Thomas Fujiwara, and Amanda Pallais. "'Acting Wife': Marriage Market Incentives and Labor Market Investments." *American Economic Review* 107.11 (2017): 3288-3319.
- Kuziemko, I., Buell, R. W., Reich, T., Norton, M. I. (2014). "Last-place aversion": Evidence and redistributive implications. *The Quarterly Journal of Economics*, 129(1), 105-149.

Some more examples

A good strategy is to **combine experiments with observational data**:

- Bursztyn, Leonardo, Thomas Fujiwara, and Amanda Pallais. "'Acting Wife': Marriage Market Incentives and Labor Market Investments." *American Economic Review* 107.11 (2017): 3288-3319.
- Kuziemko, I., Buell, R. W., Reich, T., Norton, M. I. (2014). "Last-place aversion": Evidence and redistributive implications. *The Quarterly Journal of Economics*, 129(1), 105-149.

Some Czech examples:

- Bartoš, V., Bauer, M., Cahlíková, J., Chytilová, J. (2022). Communicating doctors' consensus persistently increases COVID-19 vaccinations. *Nature*, 1-8.
- Vranka, M., Frollová, N., Pour, M., Novakova, J., Houdek, P. (2019). Cheating customers in grocery stores: A field study on dishonesty. *Journal of Behavioral and Experimental Economics*, 83, 101484.
- Krčál, O., Peer, S., Staněk, R., Karlinova, B. (2019). Real consequences matter: Why hypothetical biases in the valuation of time persist even in controlled lab experiments. *Economics of Transportation*, 20, 10013

Literature (1/3)

- Birnbaum, M. H. (1999). How to show that 9>221: Collect judgments in a between-subjects design. *Psychological Methods*, 4(3), 243.
- Bolton, G. E., & Ockenfels, A. (2000). ERC: A theory of equity, reciprocity, and competition. *American economic review*, 90(1), 166-193.
- Butler, D. M., & Desposato, S. (2022). Proposing a Compensation Requirement for Audit Studies. *Political Studies Review*, 20(2), 201-208.
- Camerer, C. (2011). The promise and success of lab-field generalizability in experimental economics: A critical reply to Levitt and List. Available at SSRN 1977749.
- Camerer, C. F., Dreber, A., Forsell, E., Ho, T. H., Huber, J., Johannesson, M., ... & Wu, H. (2016). Evaluating replicability of laboratory experiments in economics. *Science*, 351(6280), 1433-1436.
- Camerer, C. F., Dreber, A., Holzmeister, F., Ho, T. H., Huber, J., Johannesson, M., ... & Wu, H. (2018). Evaluating the replicability of social science experiments in Nature and Science between 2010 and 2015. *Nature Human Behaviour*, 2(9), 637-644.
- Clot, S., Grolleau, G., & Ibanez, L. (2018). Shall we pay all? An experimental test of Random Incentivized Systems. *Journal of Behavioral and Experimental Economics*, 73, 93-98.
- Cohn, A., Fehr, E., & Maréchal, M. A. (2014). Business culture and dishonesty in the banking industry. *Nature*, 516(7529), 86-89.
- Cohn, A., Fehr, E., & Maréchal, M. A. (2019). Selective participation may undermine replication attempts. *Nature*, 575(7782), E1-E2.

Literature (2/3)

- Fehr, E., & Schmidt, K. M. (1999). A theory of fairness, competition, and cooperation. *The quarterly journal of economics*, 114(3), 817-868.
- Fréchette, G. R., & Schotter, A. (Eds.). (2015). Handbook of experimental economic methodology. Handbooks of Economic Methodol.
- Henrich, J., Boyd, R., Bowles, S., Camerer, C., Fehr, E., Gintis, H., ... & Tracer, D. (2001). Economic man in cross-cultural perspective: Behavioral experiments in fifteen small-scale societies (No. 01-11-063).
- Huber, C., & Huber, J. (2020). Bad bankers no more? Truth-telling and (dis) honesty in the finance industry. *Journal of Economic Behavior & Organization*, 180, 472-493.
- Hupé, J. M. (2018). Shortcomings of experimental economics to study human behavior: a reanalysis of Cohn et al. 2014, Nature 516, 86–89,"Business culture and dishonesty in the banking industry".
- Levitt, S. D., & List, J. A. (2007). What do laboratory experiments measuring social preferences reveal about the real world? *Journal of Economic perspectives*, 21(2), 153-174.
- List, J. A., Shaikh, A. M., & Xu, Y. (2019). Multiple hypothesis testing in experimental economics. *Experimental Economics*, 22(4), 773-793.
- List, J. A. (2020). Non est disputandum de generalizability? A glimpse into the external validity trial (No. w27535). *National Bureau of Economic Research*.
- Milgram, S. (1963). Behavioral study of obedience. *The Journal of abnormal and social psychology*, 67(4), 371.
- McKelvey, R. D., & Palfrey, T. R. (1995). Quantal response equilibria for normal form games. *Games and economic behavior*, 10(1), 6-38.

Literature (3/3)

- Mujcic, R., & Frijters, P. (2021). The colour of a free ride. *The Economic Journal*, 131(634), 970-999.
- Open Science Collaboration. (2015). Estimating the reproducibility of psychological science. *Science*, 349(6251), aac4716.
- Ortmann, A., & Hertwig, R. (2002). The costs of deception: Evidence from psychology. *Experimental Economics*, 5(2), 111-131.
- Radner, R. (1980). Collusive behavior in noncooperative epsilon-equilibria of oligopolies with long but finite lives. *Journal of economic theory*, 22(2), 136-154.
- Rahwan, Z., Yoeli, E., & Fasolo, B. (2019). Heterogeneity in banker culture and its influence on dishonesty. *Nature*, 575(7782), 345-349.
- Schram, A., & Ule, A. (Eds.). (2019). Handbook of research methods and applications in experimental economics.
- Stahl, D. O., & Wilson, P. W. (1995). On players models of other players: Theory and experimental evidence. *Games and Economic Behavior*, 10(1), 218-254.
- Vranka, M. A., & Houdek, P. (2015). Many faces of bankers' identity: How (not) to study dishonesty. *Frontiers in Psychology*, 6, 302.