Studying Discrimination (2)

Javier Polavieja

The D-Lab
Discrimination & Inequality Lab

uc3m | Universidad Carlos III de Madrid

Outline

1. What are field experiments?
    1. Basic concepts
    2. The logic of randomization
    3. Field experiments vs observational data
2. Types of field experiments
    1. Audit studies
    2. Correspondence studies
        - Research designs for correspondence
3. Literature Review on LMD
    1. Some recent examples
    2. The GEMM study

What are Field Experiments?

- **Field experiments** combine experimental methods (to improve causal ID) with real-life contexts (to enhance external validity) (Gerber & Green, 2011)
  - "*A data collection strategy that employs manipulation and random assignment to investigate preferences and behaviors in naturally occurring contexts*" (Baldassarri & Abascal 2017:43)
- **Random assignment** of participants into treatment conditions **excludes** the possibility of **unobserved confounders** affecting the outcome, except by calculable chance
- **Randomization allows** for **causal identification** of the effect of the treatment
- **Participants** are **unaware** of the experiment and **this excludes** the possibility of **desirability bias** (no observer effects!)…

…but raises **ethical concerns**…

Glossary of key terms in exp research

- **Experimental units:** Subjects of the experiment
  - In LMD research subjects are employers or their agents (whoever is involved in the recruitment process)
- **Treatment:** the 'variable' of interest, the effect of which we want to assess
  - The treatment is defined by the research question
  - In LMD, treatments are applicants' charact. we hypothesize might trigger D (e.g. gender, ethnicity, race…)
- **Treated group**: Set of subjects that receive the treatment
- **Control group**: Set of subjects that do not receive the treatment
- Observed **outcome**: Outcome of interest as given by the research question
  - $Y_i(1)$→ Outcome if treated; $Y_i(0)$ outcome if not treated

Average Treatment Effect (ATE) = $\frac{1}{N}\sum_{i=1}^{N} Y(1)$ - $\frac{1}{N}\sum_{i=1}^{N} Y(0)$ = $\frac{1}{N}\sum_{i=1}^{N}(Y(1) - Y(0))$

  - In correspondence tests for D, we use Callback Ratios (CBR) to measure ATEs:

Callback Ratio (CBR)= $\frac{\frac{1}{N}\sum_{i=1}^{N} Y(0)}{\frac{1}{N}\sum_{i=1}^{N} Y(1)}$

, where Y(0) is 'majority group' and Y(1) is the treatment we believe might trigger D

The logic of randomization

- Randomization ensures both observed and unobserved factors affecting the outcome of interest are equally likely to be present in the treatment (T) and control (C) groups
  - i.e. if subjects were randomly assigned to T and C and no treatment was actually administered, there would be no reason to expect differences in the outcome
- Randomization provides **accuracy** of the estimate
  - Randomization eliminates the possibility of confounders (except by calculable random chance)
  - When units are randomly assigned, a comparison of average outcomes in T and C groups (the *so-called difference in means estimator*) is an **unbiased** estimator of the true ATE
  - Any given experiment might under/over estimate the effect of the treatment but if experiments are conducted repeatedly *under similar conditions* the average experiment would accurately estimate the true treatment effect
- Estimate's **precision** is a function of sample size
  - Statistical power analysis should be used to establish the minimum N required to detect treatment effects for a given margin of error

Field Experiments vs Observational data

Compared to observational and lab experiments, **field experiments have**:

- Greater internal validity (i.e. greater potential for causal identification) than observational data
- Greater external validity (i.e. greater generalizability) than lab experiments
  - But researchers have lower control over implementation than in lab settings
- But lower external validity than observational data
  - Note no single concrete experiment is generalizable!
  - Generalizability is achieved by replication across settings
  - But note e.g. of a large experiment on LMD involving N different occupation is equivalent on N different experiments involving one single occupation→ i.e. some experiments are more generalizable than others
  - Researchers must always acknowledge/reflect upon the **scope conditions** of their experiments

Fexps have become increasingly popular in the social siences

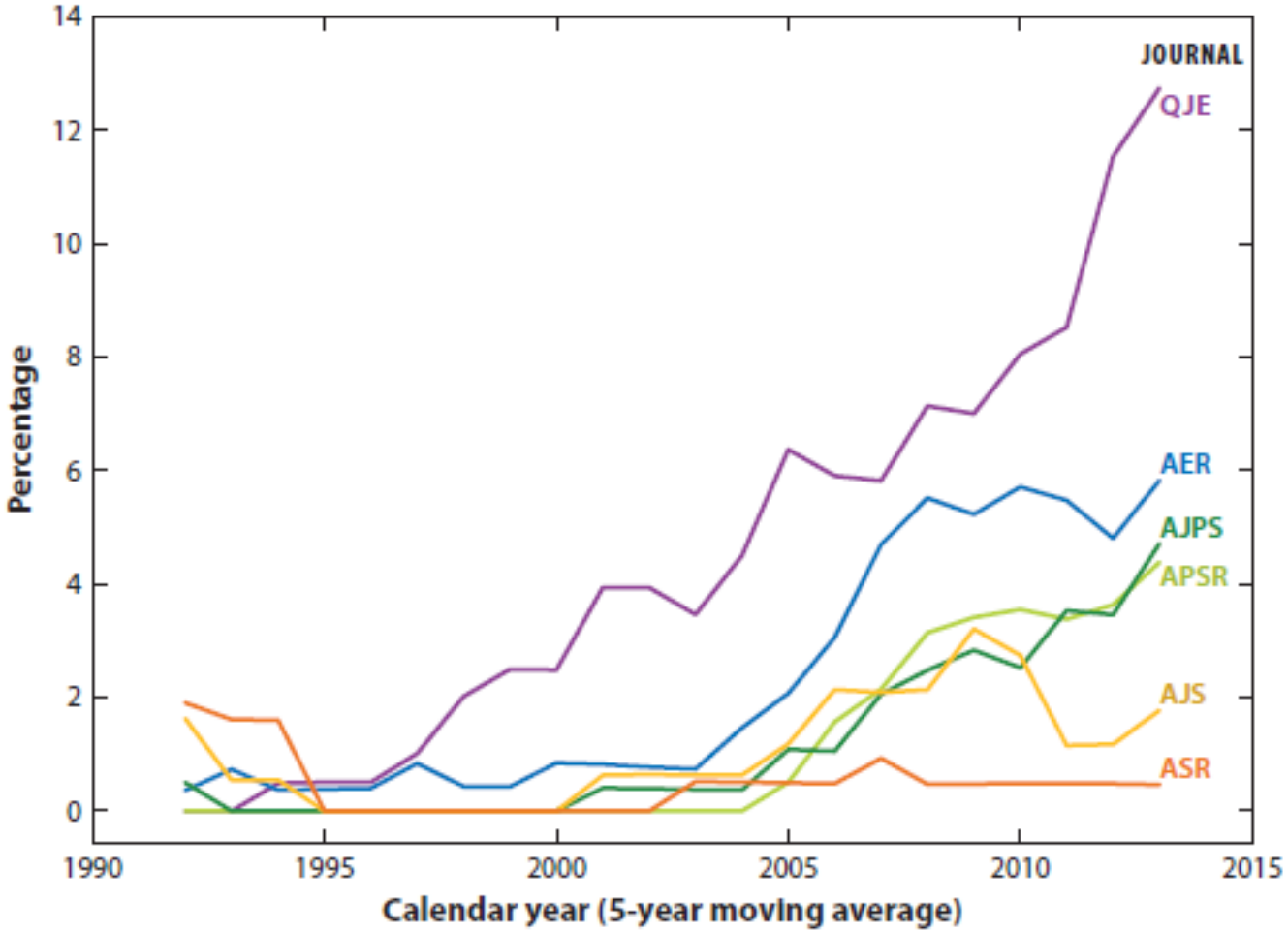


Figure 1

The percentage of research articles reporting field experiments. Abbreviations: AER, *American Economic Review*; AJPS, *American Journal of Political Science*; APSR, *American Political Science Review*; AJS, *American Journal of Sociology*; ASR, *American Sociological Review*; QJE, *Quarterly Journal of Economics*.

Source: Baldassarri & Abascal (2017)

Types of FExs

1. **Randomized Control Trials** (RCTs)→ The gold standard to evaluate policy interventions

   E.g. Perry Preschool Project and the MTO (desegregation) experiments in the US; PROGRESA in Mexico; Anti-poverty experiments in Africa by the MIT Poverty Action Lab, etc

2. **Social Norms Experiments**

   e.g. Broken-Window Experiments on social norms; Lost-Letter Experiments on trust, etc;

3. **Political Mobilization experiments**

   e.g. Get-Out-the Vote Experiments in the US (Green & Gerber 2008)

4. **Behavioral Games in the field**

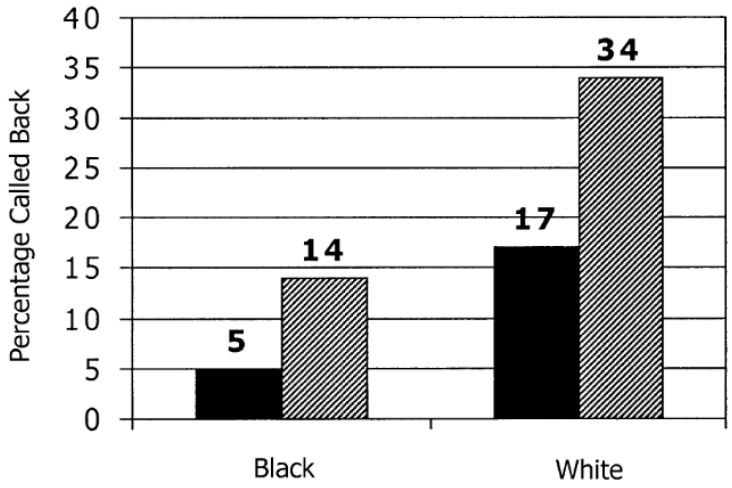
   e.g. cultural differences in trust, cooperation, competitiveness, reciprocity, sanctioning, monitoring, etc

5. **Discrimination experiments**→ The best tool to study market discrimination

   1. Audit (simulation) studies
   2. Correspondence studies

*1. **Audit Studies** for hiring D*

- Two or more trained employees of the researcher -auditors or testers- apply for real entry-level jobs
- Auditors are matched for all relevant personal characteristics other than those tested for discrimination (e.g. gender, race)
  - E.g. Pager (2003) investigates the effect of applicants' race and criminal records on employers' hiring decisions in the US using an audit study



Black bars represent criminal record; striped bars represent no criminal record.
The main effects of race and criminal record are statically significant (P>.01). The interaction between the two is not significant in the full sample.

Limitations of audit studies

- **Testers** from different groups may **not** appear **identical** to employers (Heckman & Siegelman, 1993; Heckman, 1998)
- **Tester bias**→ Audits are not double-blind, this could generate (un)conscious motives to generate data consistent with their beliefs about labour market discrimination
- Due to **high running costs**, audit studies can only deal with **small n** of treatments
- **Ethically questionable** (imply high levels of deception**)**

2. Correspondence studies *for hiring D*

- Involve sending **written applications** of fictitious job applicants to real potential employers, varying only the treatment(s) under study
  - Ethnic origin & gender of the applicant is typically manipulated with the **applicant's name**
- There is now **strict comparability** across groups for all information seen by employers
- But **only** accounts for **discrimination at the initial stage** of the job seeking process
- YET audit tests show about 90% of D takes place at this stage (Riach & Rich 2002:494)

Research designs for correspondence studies

Depending on matching method

- **Paired** (or matched) **design**→ 2 identical CVs but for the treatment are sent to each vacancy
- 3 PROBLEMS:
    1. High detection risks
    2. "Treatment-salience" bias
    3. Higher ethical birden (we impose a higher costs on recruiters)
- **Unpaired** (or unmatched) **design**→ only 1 CV to each vacancy.
    - Captures average LM for the selected occupations (no value in court as prove of organizational-level D)
    - Lower detection risks
    - No "treatment-salience bias"
    - BUT requires larger N

Research designs for correspondence studies

Depending on treatment randomization

- A full **factorial design→** uses 2 or more randomized treatments (e.g. race, gender, past incarceration, religion) and all experimental units take on all possible combinations across all such treatments

  → **each treatment is orthogonal to the others** (e.g. fictitious CVs include all possible combinations of race gender criminal record and religiosity and each combination is represented in the experiment with equal probability due to **randomization**)

- **Fractional factorial design→** some of the possible combinations are omitted for realism or efficiency

  → Not all combinations of treatments are possible—e.g. gender and ethnicity (migrant origin) are orthogonal but e.g. religion cannot be realistically orthogonal to ethnicity

3. Literature Review on LMD

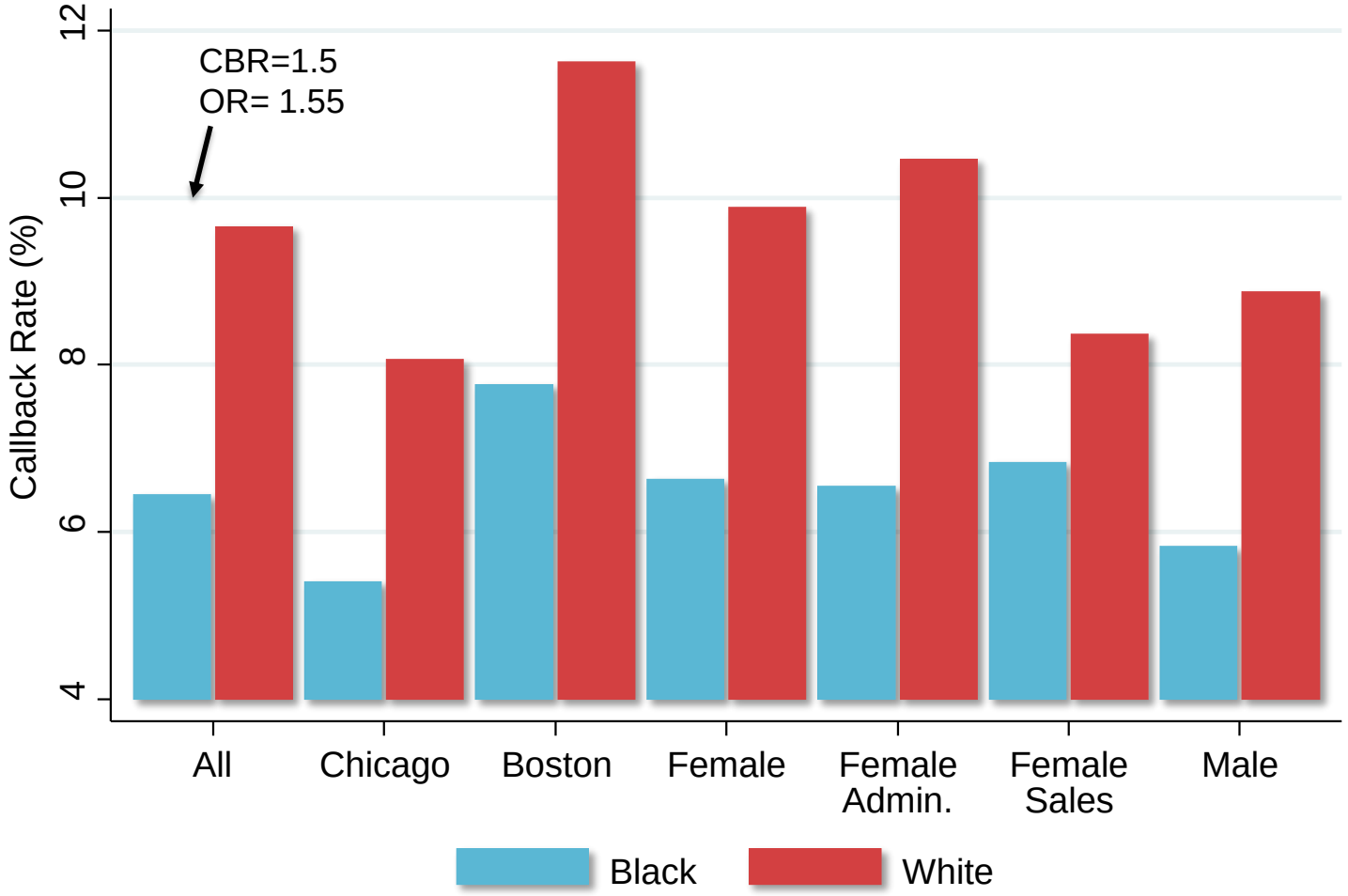
Recent contributions and findings

Some recent examples…

Bertrand and Mullainathan's (2004) send fictitious resumes to newspaper ads in Boston and Chicago

- Applicants' race is signaled with African-American (Lakisha/Jamal) or White (Emily/Greg) sounding names
- White names receive 50 percent more callbacks for interviews
- For White names, a higher quality resume elicits 30 percent more callbacks whereas for African Americans, it elicits a far smaller increase

Job Callback Rates by Race for Resumes with Otherwise Identical Credentials, US



Source: Bertrand and Mullainathan (AER 2004)

2 measures of discrimination

- D estimates measure differences in employers' callback across treatment conditions, i.e. for majority and minority applicants
- 2 measures:

1. **Callback Ratio** (CBR)→ Intuitive, most widely used

$$\text{CBR}= \frac{N\ callbacks_{maj}/N\ applicants_{maj}}{N\ callbacks_{min}/N\ applicants_{min}}$$

, where maj= Majority applicants; min=minority applicants

2. **Odds Ratio** (OR)→ Less intuitive but arguably more suitable for comparing D estimates across contexts with large differences in overall callback rates

$$\text{OR}= \frac{P\ callback_{maj}\Big/_{1-(P\ callbacks_{maj}}}{P\ callback_{min}\Big/_{1-(P\ callbacks_{min}}$$

, where maj= Majority applicants, and min=minority applicants

Some recent examples…

- Riach and Rich (2006) send fictitious resumes to advertised positions in the English labour market to test for gender D
  - They find D in sex-stereotyped occupations: against men in the 'female occupation' secretary, and against women in the 'male occupation' - engineer
  - D against men also found in two 'mixed occupations' - trainee chartered accountant and computer analyst programmer
- Baert et al. (2015) test the relationship between hiring D and labour market tightness at the level of the occupation
  - No D in against candidates with foreign-sounding names in occupations for which vacancies are difficult to fill but sig D for occupations for which labour market tightness is low
- Lancee et al (2019, JEMS Special Issue), present the results of the GEMM study, a harmonized field-experiment involving DE, ES, NL, NO & UK and provide evidence on ethnic and religious D against second generation applicants
- Birkelund et al (2021, R&R), find no evidence of gender D (against women) in none of the countries of the GEMM study (and some signs of D against men in some occupations)
- Polavieja et al. (2021 R&R) find evidence of phenotypical D against "visible" minorities in DE, NL and ES.

Summary of findings on ethnic and gender D

(see meta-analyses by Zschirnt & Ruedin (2016) and Riach and Rich (2002))

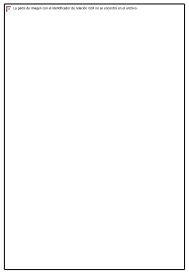
- Widespread discrimination in hiring for ethnic and racial minority groups
  - Equivalent minority candidates need to send around 50 per cent more applications to be invited for an interview than majority candidates
- Strong gender discrimination in sex-stereotyped occupations
- …but no D (against women) in occupations requiring high specific human capital investments where job-interruptions should produce high skill atrophy
- Taste-based (or perhaps implicit) discrimination remains dominant for both ethnic and gender discrimination, although in some instances there is evidence that statistical discrimination also plays a role
- More extensive and standardised procedures of job application seem to reduce statistical discrimination (e.g. labour market in Germany vs other countries' labour markets) but not tasted-based/implicit discrimination
- Suggests importance of

  1) Employers' considerations about consumers' tastes (Baer & De Pauw 2014)

  2) LM tightness at the level of occupations (Baer et al 2015)
- More research is needed!!

The GEMM study

www.gemm2020.eu



- The largest comparative field experiment on hiring D for children of migrants ever carried out in Europe (over 19,000 European firms targeted)
- Conducted **simultaneously** and with a **fully harmonised** design in **5 European countries**: Germany, the Netherlands, Norway, Spain and the UK over a period of approximately 18 months (ES→ Nov2016 until May 2018)
- Unique in scope, complexity and theoretical ambition
  - Involving 6 institutions: **Oxford** University, **Uc3m, WZB**, University of **Olso**, University of **Utrech,** University of **Amsterdam**
  - Large investments in human capital, infrastructure & IT development (e.g. D-Lab in uc3m)
  - Strict ethical clearance procedures (many bodies involved) https://www.d-labsite.com/ethics
  - A host of ancillary validity tests required (photograph ratings on attractiveness and friendliness, phenotype plausibility tests, name recognition surveys, extension of fieldwork, etc..)
- Unpaired fractional design; multiple treatments (origin, phenotype, religion, gender), 6 occupations, 53 different national origin groups

Summary of findings, GEMM study

- Ethnic discrimination in all countries; substantial differences across countries
- Big difference across ethnicities → ethnic hierarchies
- Differences across occupations
- Evidence of phenotypic discrimination
  - Less phenotypic D in Spain than in Germany or the Netherlands
  - Suggestive of ethnicity*phenotype intersections
- Discriminations against Muslim applicants
- Evidence suggests discrimination against men! (in some occupations)
- Evidence predominantly in line with non-rational explanations of discrimination (taste-based or implicit D)

To see further research currently carried out at the D-Lab check:

https://www.d-labsite.com

That's all

Many thanks for your attention!

The D-Lab

Discrimination & Inequality Lab

uc3m | Universidad Carlos III de Madrid