第六讲 实地实验

张开平 助理教授

2021年3月29日

实地实验的概念

- 使用**实验操纵和随机机制**，在**自然发生的真实场景**下研究偏好和行为的方法
  - "Identify causal effect via randomization while studying people and groups in their <u>naturally occurring contexts</u>" (Baldassari & Abascal, 2017)
  - "field experiment refers to the implementation and evaluation, by comparing different treatment groups chosen at random, of an intervention or a set of interventions specifically designed to test a hypothesis or a set of hypotheses (Banerjee & Duflo, 2009).
  - 区别于实验室实验高度控制和人造的环境
  - 田野工作与实验两种方法的结合

  Experiment happen in the real-world!

实地实验的特征

- 在真实环境下进行**随机**研究
  - 随机化分配发生在实际场景而非实验室场景
  - 个人、群体或机构被随机分配至干预组和对照组以保证每个观察接受干预的概率相同

- **真实场景**：尽可能接近地模拟一个因果过程发生的条件
- **自然干预**：实验者发挥尽可能无干扰、不引注目的作用，干预自然而不突兀
- **真实影响**：产生实际效果
- **真实测量**：自然而不突兀的测量方式；在某些情况下，受试者并不知道自己是实验参与者，使研究者更能直接观测人们的真实行为，减小自我汇报偏差
- **真实人群**：学生群体之外的一般群体

实地实验举例

- Piano Staircase Initiative Experiment

- 2004年，在美国圣地亚哥某社区，研究者将字条随机挂在1000户的门把手上，
  鼓励居民节约能源。字条共有五个版本：
  - 更经济
  - 更环保
  - 对下一代的社会责任
  - 你的邻居也在使用节能方法
  - 无理由

- 发展经济学：外部援助、教育
  - 减少辍学率、提高学生成绩、减少贫困、卫生习惯、接种疫苗

Nolan, J. M., Schultz, P. W., Cialdini, R. B., Goldstein, N. J., & Griskevicius, V. (2008). Normative social influence is underdetected. *Personality and social psychology bulletin, 34*(7), 913-923.

Banerjee, A. V., Cole, S., Duflo, E., & Linden, L. (2007). Remedying education: Evidence from two randomized experiments in India. *The Quarterly Journal of Economics, 122*(3), 1235-1264.

实地实验的优势

- 通过随机机制进行**因果推断**，并在**真实而自然的场景**中研究个人和群体

  - 比观察性研究更能推断因果关系，比实验室实验更真实

  - 结合实验室实验的内部效度和实地场景的外部效度，在实验室和自然发生的数据间架起了桥梁，具有更强的可推广性（List 2007）

Causal Inference on Real-world Behavior	**Advance Theory**	**Inform Policy**

Let the evidence talk!

实地试验适合研究什么问题?

- **政策领域**：经济和教育政策的干预效果
  - 减少辍学率、提高学生成绩、减少贫困、提高卫生习惯、经济发展
- **行为领域**：社会规范、动机和激励等如何影响行为
  - Nudge、bystander effect、亲社会行为、健康习惯
- **政治学领域**
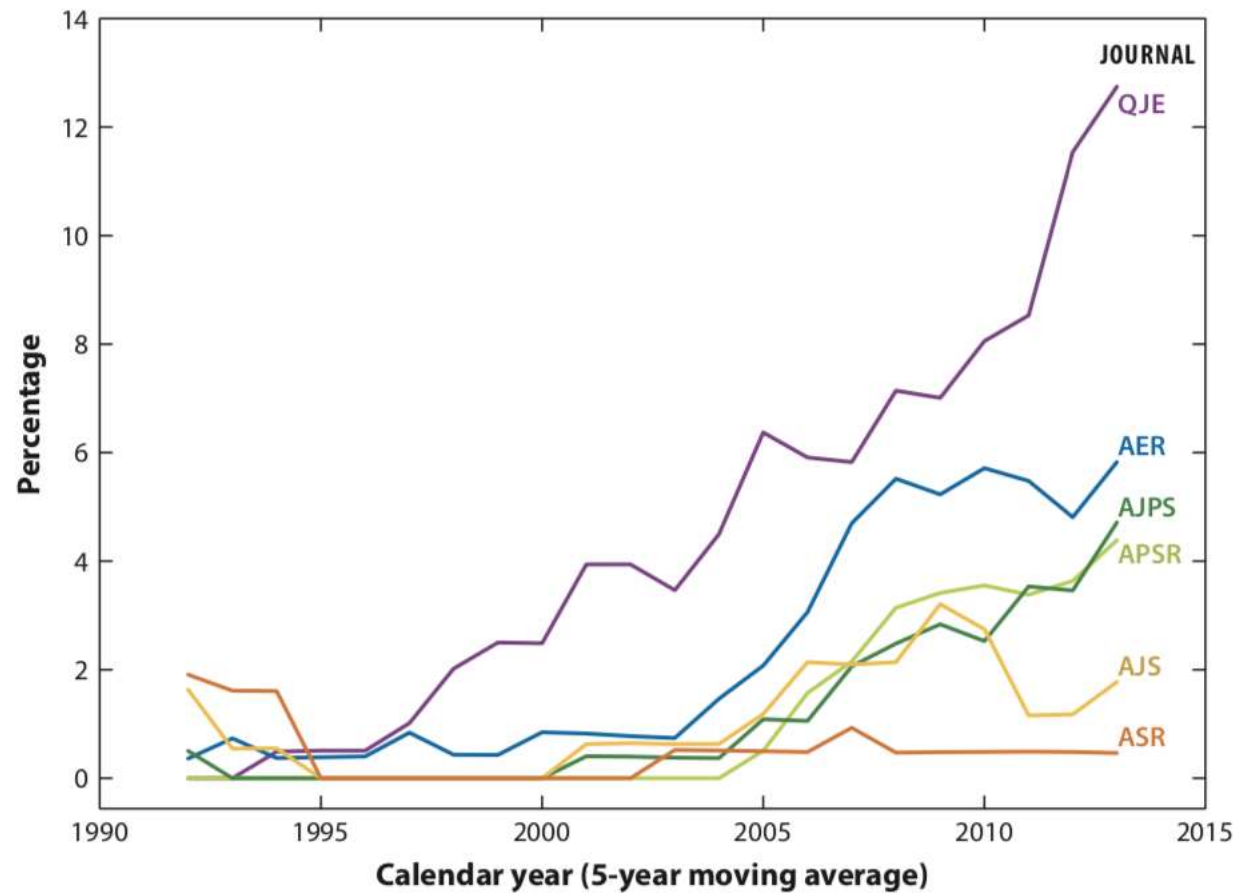 - 政治态度、立法行为、选举投票、社交媒体、社群影响、族群与身份
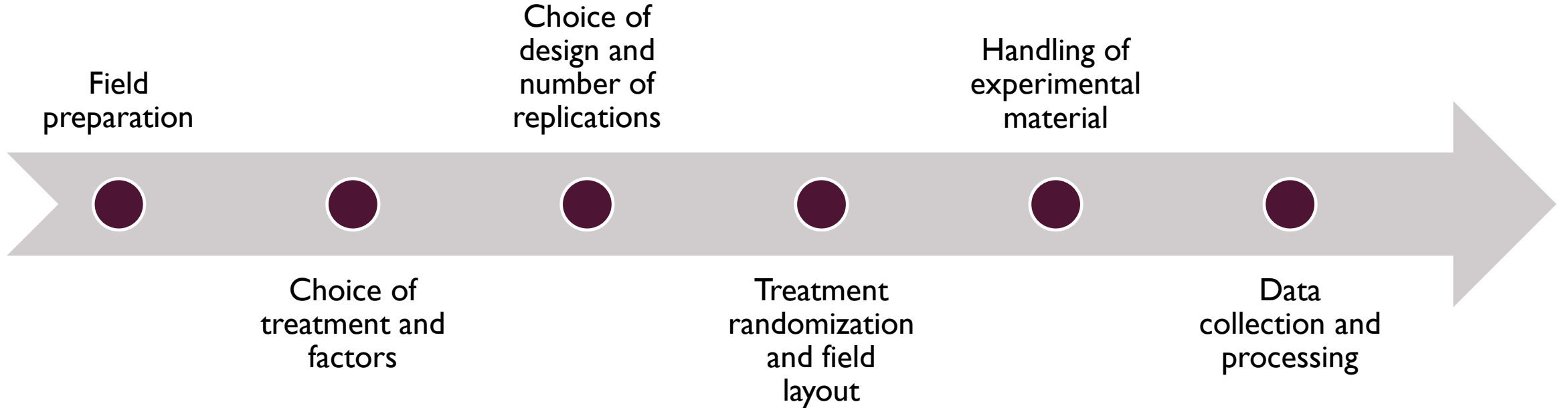- **心理与态度研究**
  - 偏见与歧视

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, D., & Abascal, M. (2017). Field experiments across the social sciences. *Annual Review of Sociology, 43*, 41-73.

实地实验的操作

Field preparation

Choice of treatment and factors

Choice of design and number of replications

Treatment randomization and field layout

Handling of experimental material

Data collection and processing

实地实验的设计

2020   2021   2022

Group A   Group A   Group A
Group B   **Group B**   Group B
Group C   Group C   **Group C**

- 寻求合作伙伴：政府部门、国际组织、非营利组织、学校、医院等
- 随机方法：
  - 完全随机
  - 超额随机 (oversubscription)（抽签）：结果阐释只针对干预为完全随机的群体
  - 分阶段随机 (Randomized order of Phase-in)：在伦理或实践上不允许对控制组不提供任何支持/干预
    - 分组分阶段接受干预，减少样本流失但较难评估长期效果。
    - 不同阶段期间预留充足的时间以观察干预效果
    - 当对照组对干预有预期时可能改变行为，从而影响其作为counterfactual的有效性
  - 组内随机 (Within-group randomization)：当合作方不允许研究者接触或采集全体数据时
    - 为每个区域中的部分群体(subgroup A, subgroup B) 随机提供干预
  - 随机鼓励 (Encouragement Design)：
    - randomly assign treatment over subjects → randomly assign subjects an encouragement to receive the treatment
    - 随机鼓励作为外生的工具变量，通过2SLS估计干预效果
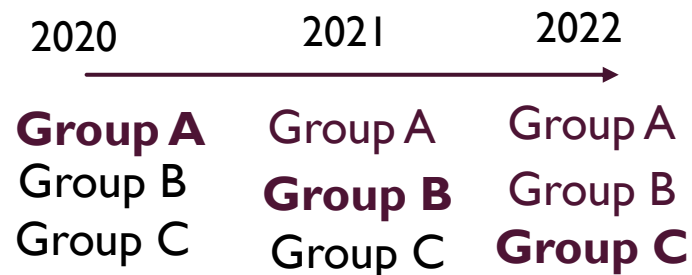
实地实验的实施

- **计算样本量**: Power analysis (the probability that for a given effect size, and a given statistical significance level, we are able to reject the null hypothesis if the alternative is true)
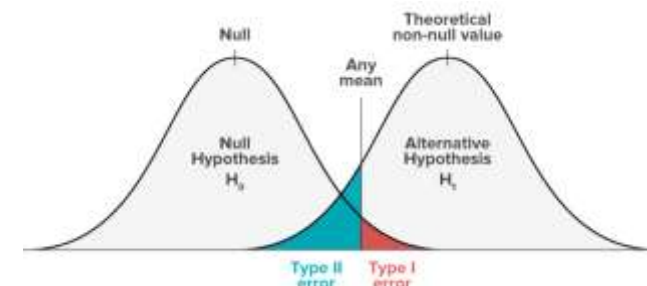
- **决定随机单元** (Level of randomization)

  - 个体、家庭、区、学校、群组
  - 随机的单位越大，达到同样power所需要的总样本量越大，因此个体层面的随机样本量更节省，影响预算。
  - Group-level randomization 能减少个体层面溢出效应影响，减少对照组对预期干预的影响，操作更方便，减少样本抵触和样本流失
  - When randomization happens at the group level，errors are not iid → Grouped errors
  - 当干预昂贵但数据收集成本低时，实验组与对照组的最佳比例为

$$\frac{p}{1-p} = \sqrt{\frac{c_C}{c_T}}$$

$$Y_{ij} = \alpha + \beta T + v_j + \omega_{ij}$$

Standard error of $\hat{\beta}$ for group – level randomization

组内方差

$$\sqrt{\frac{1}{p(1-p)}} \sqrt{\frac{n\tau^2 + \sigma^2}{nJ}}$$

实验组样本比例 → ，组间方差 →，J组，每组N个样本

S.E. of $\hat{\beta}$ for individual – level randomization

$$\sqrt{\frac{1}{p(1-p)}} \sqrt{\frac{\tau^2 + \sigma^2}{nJ}}$$

Design effect：
S.E. of $\hat{\beta}$ for group – level randomization/ S.E. of $\hat{\beta}$ for individual – level randomization →

$$D = \sqrt{1 + (n-1)\rho}$$

组内方差占总方差的比例 →

$$\rho = \tau^2 / \tau^2 + \sigma^2$$

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

PARTIAL COMPLIANCE 部分服从

- 实地实验往往有较高的不服从率，是否意味其因果估计存在偏差（biased)?
  - 常见原因：联系不到实验组的人(low contact rate)
  - 解决方法：
    - 估计试图干预效应（ITT）
    - 在满足前提假设的情况下（Angrist, Imbens & Rubin, 1996），转为估计实际接受干预者/服从者平均干预效应，即LATE/CATE (Local/Complier ATE)

$$LATE = \frac{E(Y_1 - Y_0)}{p(D_1|T_i = 1)}$$

$$ITT = E(Y_{i1}|Z_i = 1) - E(Y_{io}|Z_i = 0)$$

Angrist, Imbens, & Rubin (1996)

估计LATE的假设条件

- **干预分配机制的可忽略性 Exclusion Restriction**
  - 每个单元的结果只与干预（发生或不发生）有关
  - 无安慰剂效应 (No placebo effect)
- **稳定的单元处理值假设 Stable Unit Treatment Value Assumption (SUTVA)**
  - 对某个单元的干预只影响该单元的效果
  - 无溢出效应(No spillover effect)

- **单向性 Monotonicity**
  - 实验组接受干预的概率不小于对照组接受干预的概率
- **随机分配的非零效果 Nonzero causal effect of assignment on treatment**
  - 至少一些被随机分配到干预组的人实际接受了干预

部分服从问题

- 当存在不服从时，对于干预效应的估计只限于服从者人群，推广至一般人群需要假设服从者与一般人群的干预效应同质性（很难证实）。当干预效应同质时，LATE=ATT=ATE

- 较低的服从率使LATE估计具有较大的标准误（SE）

- 提高的服从率能够减小ATE与LATE之间的差别，但无法完全消除

$$\sqrt{\frac{1}{p(1-p)}}\sqrt{\frac{\tau^2+\sigma^2}{nJ}}$$

- 人群中complier的比例取决于人群特征和实验设计

- 部分服从极大程度上影响statistical power，需要更大样本量

样本流失 ATTRITION

- 常见原因：被试退出或无法回访

- 解决方法
  - A. 忽略缺失值
    - 0.5/(0.5 + 0.2) – 0.4/(0.4+0.2) = 0.05
  - B. 假设缺失值为 Y = 0
    - 0.5/(0.5 + 0.2 + 0.3) – 0.4/(0.4+0.2+0.4) = 0.1
  - C. 计算干预效果的上下边界值
    - Upper bound: 0.8 – 0.4 = 0.4
    - Lower bound: 0.5 – 0.8 = -0.3
    - $\{-0.3, 0.4\}$
  - D. 插补 Imputation
    - 如缺失值随机出现，则插补方法更优
    - 如缺失值由于系统性的无法观测的变量，则插补加剧bias.

|	Y = 1	Y = 0	Missing
实验组 ($Z_i = 1$)	50%	20%	30%
对照组 ($Z_i = 0$)	40%	20%	40%

溢出效应

- 溢出效应，即实验干预不仅影响了实验组，还影响了对照组，违反稳定的单元值假设（SUTVA），对照组不再提供准确的counterfactual，使干预效应估计偏小。

- 方法

  - Randomization at the group level

  - Multilevel experimental design

  - Relaxing the non-interference assumption, redefine potential outcomes, use inverse-probability weighted regression (IPW) in statistical analyses

Sinclair, B., McConnell, M., & Green, D. P. (2012). Detecting spillover effects: Design and analysis of multilevel experiments. *American Journal of Political Science, 56*(4), 1055-1069.

- 实地实验不控制背景性活动，是否会带来(bias)偏差?
  - 背景活动影响实验结果的解读，但不产生偏差。
    - 因为干预是随机分配的，背景情况同时影响实验组和对照组的 Y_0, Y_1。
    - 然而，如果实验干预与背景性活动相关，且背景性活动影响 Y_0, Y_1，那么就违反了干预机制的可忽略性假设(ignorability)
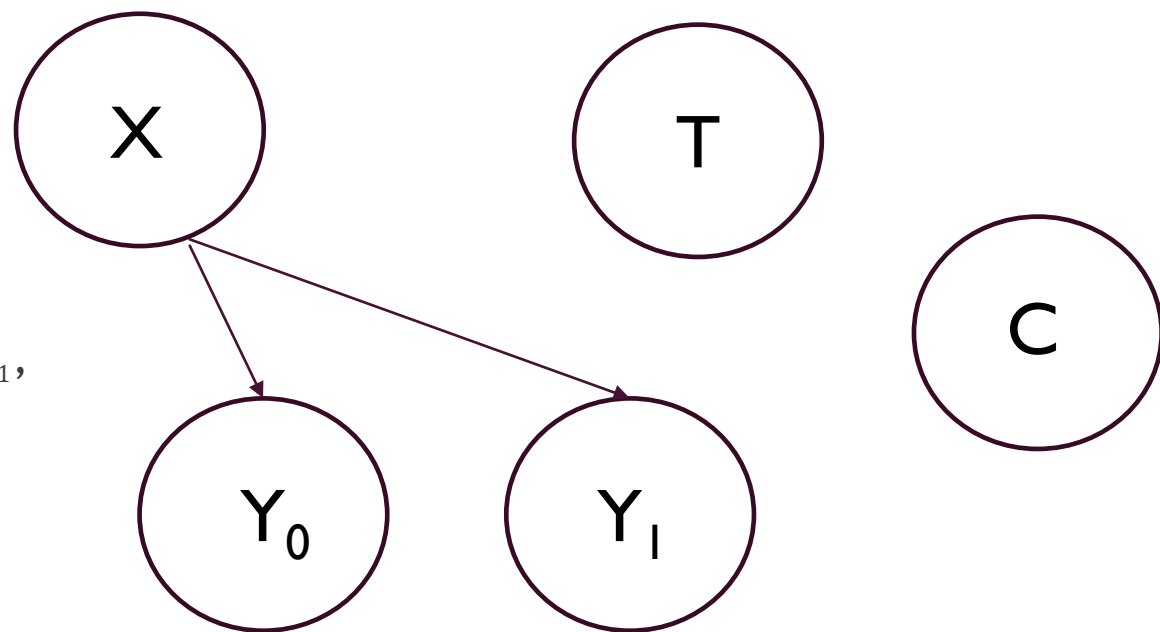
$$(Y_0, Y_1 \perp T \mid X, C)$$

实地实验的挑战：内部效度

- 在实地场景易缺乏对干预实施的严格控制，进而影响内部效度 (McDermott 2011, Gerber & Green 2012)
  - 不服从 no-compliance
  - 样本流失 attrition
  - 溢出效应 spillover

实地实验一定具有更好的推广性？

- Policy-driven instead of theory driven
  - Knowing it works but not why it works limits the generalizability of the findings

- Scalability of the effect
  - 能推广到不同的人群、情境么？
  - 可推广性是指在不同的人群、情境、干预、测量进行**重复实验**并得到一致的结果
    - **Replication** across population, contexts, treatments, outcome measurements（Banerjee & Duflo 2011, Keizer et al. 2008, McDermott 2011）
  - Random assignment + random sampling

实地实验的伦理问题

Emotional states can be transferred to others via emotional contagion, leading people to experience the same emotions without their awareness. Emotional contagion is well established in laboratory experiments, with people transferring positive and negative emotions to others. Data from a large real-world social network, collected over a 20-y period suggests that longer-lasting moods (e.g., depression, happiness) can be transferred through networks [Fowler JH, Christakis NA (2008) *BMJ* 337:a2338], although the results are controversial. In an experiment with people who use Facebook, we test whether emotional contagion occurs outside of in-person interaction between individuals by reducing the amount of emotional content in the News Feed. When positive expressions were reduced, people produced fewer positive posts and more negative posts; when negative expressions were reduced, the opposite pattern occurred. These results indicate that emotions expressed by others on Facebook influence our own emotions, constituting experimental evidence for massive-scale contagion via social networks. This work also suggests that, in contrast to prevailing assumptions, in-person interaction and non-verbal cues are not strictly necessary for emotional contagion, and that the observation of others' positive experiences constitutes a positive experience for people.

computer-mediated communication | social media | big data

- 通过操纵算法改变人的心情？
- 读取20年情绪数据推测用户抑郁程度?
- 同意Facebook使用条款等于允许使用其个人数据并进行实验干预?
- IRB是否通过?

PSYCHOLOGICAL AND COGNITIVE SCIENCES

PNAS is publishing an Editorial Expression of Concern regarding the following article: "Experimental evidence of massive-scale emotional contagion through social networks," by Adam D. I. Kramer, Jamie E. Guillory, and Jeffrey T. Hancock, which appeared in issue 24, June 17, 2014, of *Proc Natl Acad Sci USA* (111:8788–8790; first published June 2, 2014; 10.1073/pnas.1320040111). This paper represents an important and emerging area of social science research that needs to be approached with sensitivity and with vigilance regarding personal privacy issues.

Questions have been raised about the principles of informed consent and opportunity to opt out in connection with the research in this paper. The authors noted in their paper, "[The work] was consistent with Facebook's Data Use Policy, to which all users agree prior to creating an account on Facebook, constituting informed consent for this research." When the authors prepared their paper for publication in PNAS, they stated that: "Because this experiment was conducted by Facebook, Inc. for internal purposes, the Cornell University IRB [Institutional Review Board] determined that the project did not fall under Cornell's Human Research Protection Program." This statement has since been confirmed by Cornell University.

Obtaining informed consent and allowing participants to opt out are best practices in most instances under the US Department of Health and Human Services Policy for the Protection of Human Research Subjects (the "Common Rule"). Adherence to the Common Rule is PNAS policy, but as a private company Facebook was under no obligation to conform to the provisions of the Common Rule when it collected the data used by the authors, and the Common Rule does not preclude their use of the data. Based on the information provided by the authors, PNAS editors deemed it appropriate to publish the paper. It is nevertheless a matter of concern that the collection of the data by Facebook may have involved practices that were not fully consistent with the principles of obtaining informed consent and allowing participants to opt out.

Inder M. Verma
Editor-in-Chief

Kramer, A. D., Guillory, J. E., & Hancock, J. T. (2014). Experimental evidence of massive-scale emotional contagion through social networks. *Proceedings of the National Academy of Sciences, 111*(24), 8788-8790.

实地实验的伦理问题

- 一部分人得不到治疗方案或援助
  - encouragement design: 所有人都得到实验干预,得到的激励和成本进行随机化 (Thornton 2008)
  - Phase-in design: 不同人在不同时段分批接受干预（Miguel & Kremer 2004）
- 难以预料的其他负面影响 (Schuler et al., 1998)
  - Example: Facebook study (Kramer et al., 2014)
  - 影响政治与社会走向 (Cantoni et al., 2019)
- 方法
  - pilot study
  - deeply local knowledge
  - informed consent and debriefing,
  - option for opting out
  - IRB approval

实地实验的伦理原则

- **Respect for persons** -- requires that research subjects are not coerced into participating in a study and requires the protection of research subjects who have diminished autonomy

- **Beneficence** -- requires that experiments do not harm research subjects, and that researchers minimize the risks for subjects while maximizing the benefits for them

- **Justice** -- requires that all forms of differential treatment among research subjects be justified
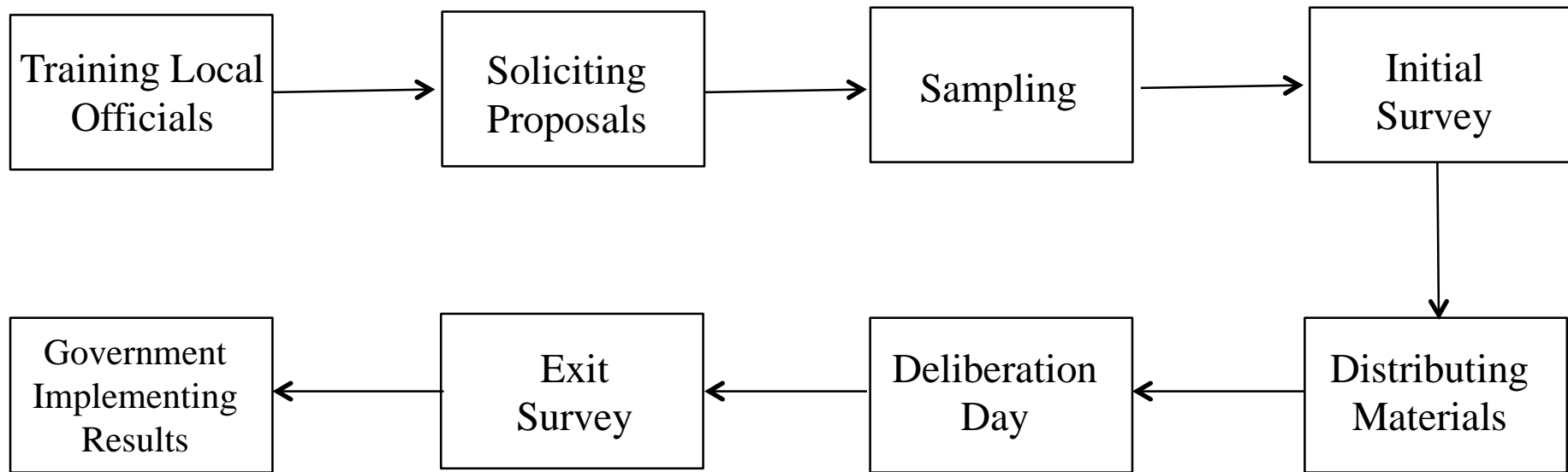
准实验/类实验 QUASI-EXPERIMENT

- 出于伦理或操作性的原因，RCT难以实现，可采用准实验方法
- 实验与准实验的相似点
  - 被试分为实验组和对照组
  - 测量因变量
  - 比较实验组和对照组得到实验干预的效果
- 不同之处
  - 准实验中的**分配并不一定是随机的**，缺少随机生成的对照组
  - 对一些因素缺乏严格控制**缺乏严格的控制**，因此可能在实验干预之外存在其他可能的解释
- 尽量选取相似的对照组和参与组
  - 比如：协商式民意调查

上海浦兴路街道协商民意调查（2015）

Deliberative Participative Budgeting in Shanghai

- 从全街道18万居民（6.9万户）中随机抽取400户，每户随机抽取一名成年公民

```
┌──────────────┐     ┌──────────────┐     ┌──────────────┐     ┌──────────────┐
│Training Local│ ──> │  Soliciting  │ ──> │   Sampling   │ ──> │   Initial    │
│  Officials   │     │  Proposals   │     │              │     │   Survey     │
└──────────────┘     └──────────────┘     └──────────────┘     └──────────────┘
                                                                       │
                                                                       v
┌──────────────┐     ┌──────────────┐     ┌──────────────┐     ┌──────────────┐
│  Government  │ <── │     Exit     │ <── │ Deliberation │ <── │ Distributing │
│ Implementing │     │    Survey    │     │     Day      │     │  Materials   │
│   Results    │     │              │     │              │     │              │
└──────────────┘     └──────────────┘     └──────────────┘     └──────────────┘
```

协商民意测验大会材料

项目说明材料

浦兴路街道
2015 年 5 月

TAMALE DELIBERATIVE POLL (2015)

Topics

- Food security and livelihood

- Water, sanitation, and hygiene

Procedure

- A scientific sample of 243 individuals: A random selection of households, then a random selection within the households

208 participants:

- 47.6% male, 52.4% female

- 33.7 years old on average

- 27.9% had never been to school

Source: Center for Deliberative Democracy, Stanford University

解剖一只麻雀：GERBER & GREEN (2000)