第八讲 调查实验

孟天广 副教授

清华大学政治学系

2021年4月19日

内容提要

- 调查实验的原理
- 调查实验的设计
- 调查实验的分析
- 调查实验举例

为什么调查实验设计？

- 实验（室）情景vs社会现实
- 方便样本~总体推论
- 被试样本少——统计功效低
- 调查方法因果推论困境：时序问题、内生性问题、虚假相关spurious correlation
- ✓ 警察数量与犯罪
- ✓ 果树与共和党偏好

参照类：实验室实验

- 在实验室内进行，干预变量与环境受到严格控制，实验结果可以清楚观测的实验。
- 米尔格莱姆的权威—服从实验
- ✓ 优势：更好控制；更精准测量；更丰富的被设计的实验“世界”
- ✓ 局限：人造环境、社会动物、外部效度（方便样本）、情景异质性
- ✓ 应用：心理学、行为经济学

调查实验

- 社会调查+实验设计

- ✓ experiment-embeded survey：社会调查中嵌入实验设计
- ✓ survey-based experiment：讲实验研究从实验室搬到社会现实场景
- ✓ 调查实验是跨对象系统地改变调查的一个或多个要素并且评估这种改变对于一个或多个结果变量的作用（Nock & Guterbock，2010）
- ✓ 内嵌于调查中个人决策实验（Morton & Williams，2010）
- ✓ 基于总体的调查实验（Mutz，2011）
- ✓ 方便样本；代表性样本；特定群体样本

调查实验的要素

- 实验要素
  - ✓ 随机化分配干预
  - ✓ 被试实验干预（操作）
  - ✓ 比较实验组与控制组的实验结果
- 调查要素
  - ✓ 概率或非概率抽样
  - ✓ 结构化提问
  - ✓ 非实验室场景（实地、电话、网络）

调查实验的优势与局限

- 优势：Field experiment + Lab experiment；因果验证；外部效度；
- 局限：缺乏实验环境控制；多个实验嵌套引起外溢效应（spillover）；外溢（污染）效应

调查实验的发展

- Roper实验（1940）

  1（a）你赞同萨姆纳•韦尔斯出访欧洲吗？

  1（b）你赞同罗斯福总统派遣萨姆纳•韦尔斯出访欧洲吗？

  2（a）你认为美国是否应该比目前更多援助英国和法国？

  2（b）你认为美国是否应该比目前更多援助英国和法国抗击希特勒？

  - ✓ 提及罗斯福引起不赞同比例从25%上升到31%
  - ✓ 提及希特勒引起支援英法率从13%上升到22%

- 对半投票设计（Rosen,1973）：一个因子（两个状态）；随机分组；回答不同版本问卷；比较两组结果

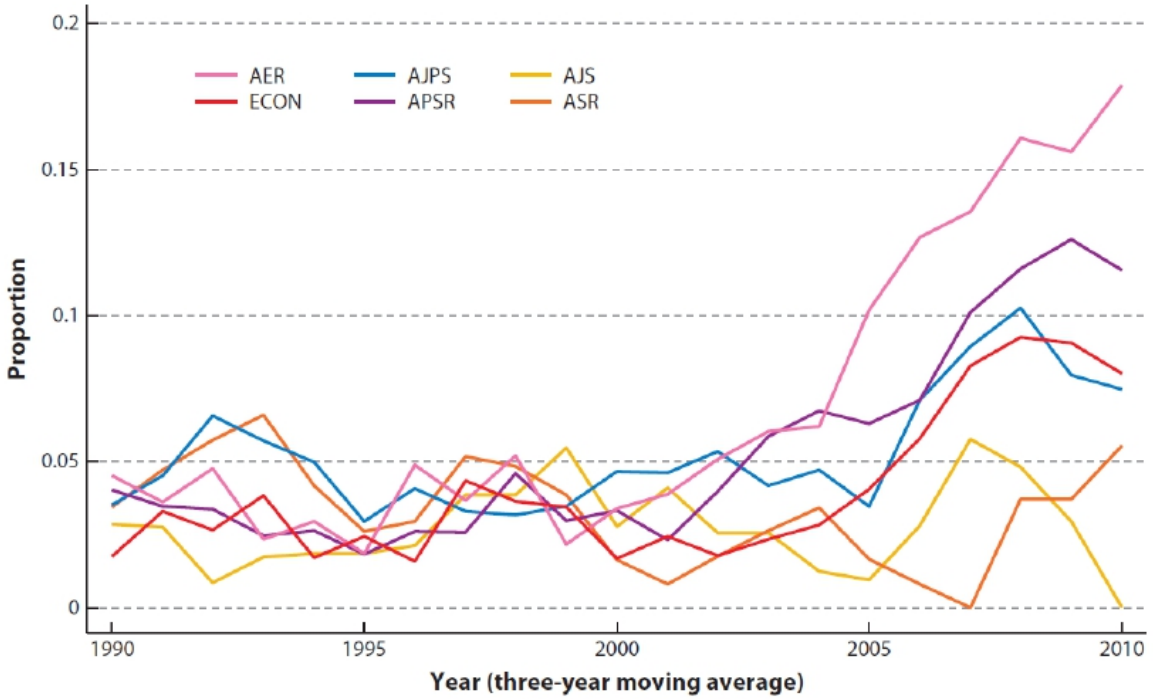


Figure 1

The proportion of research articles employing experimental design in top-ranked journals in economics, political science, and sociology (three-year moving average). Abbreviations: AER, *American Economic Review*; ECON, *Econometrica*; AJPS, *American Journal of Political Science*; APSR, *American Political Science Review*, AJS, *American Journal of Sociology*; ASR, *American Sociological Review*.

调查实验的发展

- 实验技术的发展
  - ✓ 对问题语句和顺序的操作
  - ✓ 嵌入图片、视频作为干预
  - ✓ 应对敏感问题的测量实验（列举实验、背书实验）
  - ✓ 析因实验（多个因子，虚拟情景）
  - ✓ 多阶段调查实验（within subject design)
  - ✓ 自填调查、面访调查、电话调查、网络自填调查
  - ✓ 随机化分配干预（印制不同版本问卷——计算机辅助调查）

调查实验的应用场景

- 调查方法：无应答酬金实验；调查模式；顺序效应；访问员效应
- 测量实验：社会期望偏差；人际不可比
- 实质性议题：
  - ✓ 决策行为：Priming effect，信任与合作，投票
  - ✓ 社会偏好：种族/移民歧视、收入再分配、政治支持
  - ✓ 冲突解决：民主和平论、集体行动、国际冲突
  - ✓ 制度与政策变革：政策偏好、政府回应、决策模式

调查实验的设计与执行

- 调查实验设计的要素
  - ✓ 实验结果（Outcome）
  - ✓ 影响因子（Factors）：实验因子+观察因子
  - ✓ 实验干预（Treatments）
  - ✓ 被试对象（Subjects）
  - ✓ 测量工具（Measurement）
  - ✓ 成本与统计功效（Power)

调查实验的设计与执行

- 准备阶段
- ✓ 调查实验？
- ✓ 哪种实验处理技术？
- ✓ 哪种分组设计？
- ✓ 如何抽取样本？
- ✓ 何种调查模式？
- ✓ 何种统计方法？

调查实验的设计与执行

- 实验处理技术
  - ✓ 常规调查实验（直接处理）
  - ✓ 启动效应实验（Priming effects experiment）
  - ✓ 析因调查实验
  - ✓ 列举实验/背书实验

调查实验的设计与执行

- 分组设计
  - ✓ Between subject design
  - ✓ Within subject design
  - ✓ mixed design
  - ✓ 完全随机设计
  - ✓ 随机化区组设计（配额设计）
  - ✓ 析因设计

调查实验的设计与执行

- 抽样方法
  - ✓ 非概率抽样
  - ✓ 概率抽样
- 调查模式
  - ✓ CAPI计算机辅助面访
  - ✓ CATI计算机辅助电话调查
  - ✓ CAWI网络调查
  - ✓ F2F面对面访谈
  - ✓ 自填问卷调查

调查实验的设计与执行

- 统计分析方法
  - ✓ 均值差Differnce in Means（T检验）
  - ✓ 方差分析（ANOVA）
  - ✓ 协方差分析（ANCOVA）
  - ✓ 回归分析
  - ✓ 倾向值匹配分析（不同组样本不均衡时）

调查实验的执行

- 调查过程
  - ✓ 设计调查问卷和实验题目
  - ✓ 预调查
  - ✓ 招募和培训调查员
  - ✓ 调查实施（入户、电话、网络）
  - ✓ 调查质量控制
  - ✓ 数据清洗

虚拟情景锚定实验的设计与分析

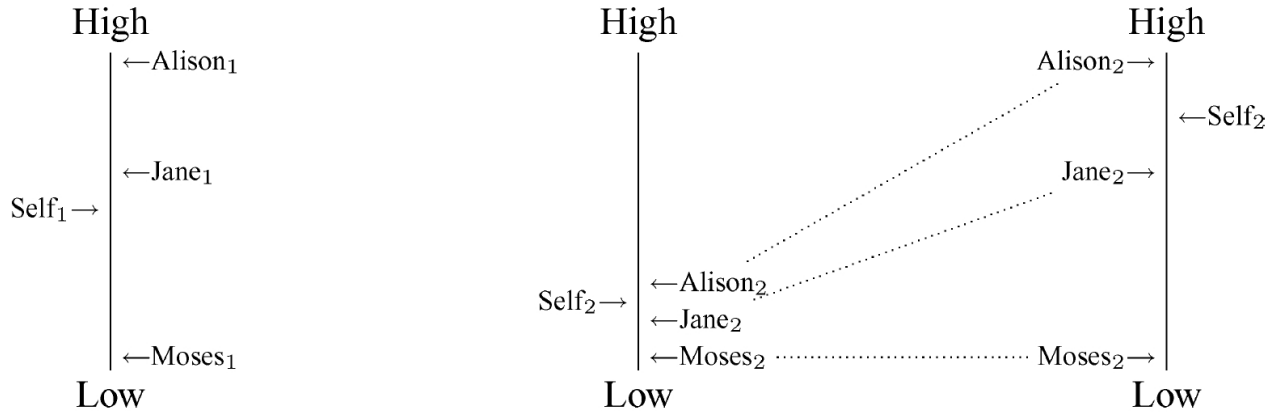
- 人际可比性与虚拟情景锚定实验
  - ✓ 社会科学研究对象的人际异质性
  - ✓ 跨文化、跨区域研究的挑战
  - ✓ 传统方法：受访者自举法(Cantril,1965)；参照类法(Aldrich,1977)
  - ✓ 虚拟情境锚定实验(King etc., 2004) 测量不同群体间真实、可比的变量
  - ✓ 虚拟情境的设置可以最大限度削除个人经验、生活情境等带来的测量误差
  - ✓ 根据受访人在虚拟情景的应答来计算corrected *differential item functioning*(DIF)

虚拟情景锚定实验的设计与分析

- 中国与墨西哥的政治效能感(King etc.,2004)
- How much say [does 'name'/do you] have in getting the government to address issues that interest [him/her/you]?

1. "[Alison] lacks clean drinking water. She and her neighbors are supporting an opposition candidate in the forthcoming elections that has promised to address the issue. It appears that so many people in her area feel the same way that the opposition candidate will defeat the incumbent representative."
2. "[Imelda] lacks clean drinking water. She and her neighbors are drawing attention to the issue by collecting signatures on a petition. They plan to present the petition to each of the political parties before the upcoming election."
3. "[Jane] lacks clean drinking water because the government is pursuing an industrial development plan. In the campaign for an upcoming election, an opposition party has promised to address the issue, but she feels it would be futile to vote for the opposition since the government is certain to win."
4. "[Toshiro] lacks clean drinking water. There is a group of local leaders who could do something about the problem, but they have said that industrial development is the most important policy right now instead of clean water."
5. "[Moses] lacks clean drinking water. He would like to change this, but he can't vote, and feels that no one in the government cares about this issue. So he suffers in silence, hoping something will be done in the future."

FIGURE 1. Comparing Preferences



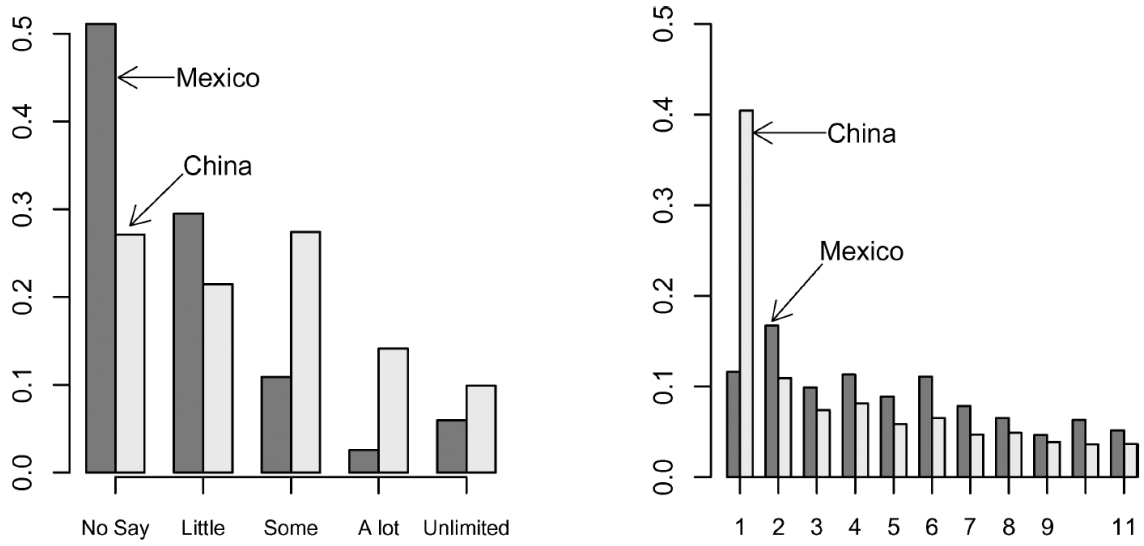
Note: Respondent 1, on the left, reported a higher self-assessment of political efficacy than respondent 2, in the middle. On the right, Respondent 2's reported scale is deformed into one comparable to 1's scale: Now 2's vignette assessments match those for Respondent 1, revealing that Respondent 2 has a higher actual level of political efficacy than Respondent 1.

A SIMPLE (NONPARAMETRIC) APPROACH

To define this idea more generally, let y_i be the categorical survey self-assessment for respondent i $(i = 1, \ldots, n)$ and z_{ij} be the categorical survey response for respondent i on vignette j $(j = 1, \ldots, J)$. Then for respondents with identical ordinal rankings on all vignettes ($z_{i,j-1} < z_{ij}$, for all i, j), the DIF-corrected variable is

$$C_i = \begin{cases} 1 & \text{if} \quad y_i < z_{i1}, \\ 2 & \text{if} \quad y_i = z_{i1}, \\ 3 & \text{if} \quad z_{i1} < y_{is} < z_{i2}, \\ \vdots & \vdots \\ 2J+1 & \text{if} \quad y_i > z_{iJ}. \end{cases}$$

FIGURE 2. Nonparametric Estimates of an Electoral Dimension of Political Efficacy



Note: The left graph is a histogram of the observed categorical self-assessments. The right graph is a histogram of *C*, our nonparametric DIF-corrected estimate of the same distribution.

TABLE 2. Comparing Political Efficacy in Mexico and China

		Ordered Probit		Our Method	
Eq.	Variable	Coeff.	(SE)	Coeff.	(SE)
μ	**China**	**0.670**	(**0.082**)	−**0.364**	(**0.090**)
	Age	0.004	(0.003)	0.006	(0.003)
	Male	0.087	(0.076)	0.114	(0.081)
	Education	0.020	(0.008)	0.020	(0.008)
τ^1	**China**			−**1.059**	(**0.059**)
	Age			0.002	(0.001)
	Male			0.044	(0.036)
	Education			−0.001	(0.004)
	Constant	0.425	(0.147)	0.431	(0.151)
τ^2	**China**			−**0.162**	(**0.071**)
	Age			−0.002	(0.002)
	Male			−0.059	(0.051)
	Education			0.001	(0.006)
	Constant	−0.320	(0.059)	−0.245	(0.114)
τ^3	**China**			**0.345**	(**0.053**)
	Age			−0.001	(0.002)
	Male			0.044	(0.047)
	Education			−0.003	(0.005)
	Constant	−0.449	(0.074)	−0.476	(0.105)
τ^4	**China**			**0.631**	(**0.083**)
	Age			0.004	(0.002)
	Male			−0.097	(0.072)
	Education			0.027	(0.007)
	Constant	−0.898	(0.119)	−1.621	(0.149)
Vignettes	θ_1			1.284	(0.161)
	θ_2			1.196	(0.160)
	θ_3			0.845	(0.159)
	θ_4			0.795	(0.159)
	θ_5			0.621	(0.159)
$\ln \sigma$				−0.239	(0.042)

Note: Ordered probit indicates counterintuitively and probably incorrectly that the Chinese have higher political efficacy than the Mexicans, whereas our approach reveals that this is because the Chinese have comparatively lower standards (τ's) for moving from one categorical response into the next highest category. The result is that although the Chinese give higher reported levels of political efficacy than the Mexicans, it is the Mexicans who are in fact more politically efficacious.

设计要求与注意事项

- 两个测量假定
  - ✓ *response consistency* assumption，受访人评估虚拟情景的方式与评估自己的方式雷同
  - ✓ *vignette equivalence* assumption，受访人对虚拟情景中变量的水平具有一致理解
- 先提问自评问题，然后在随机提问虚拟情景问题

析因设计(Factorial design)

- 多个自变量对因变量的同时效应
- 多个干预（或干预水平）的结合
- 主效应(Main effects)：单个自变量的效应
- 交互效应(Interaction effects)：两个或三个自变量的结合之效应

实验举例

- 政治资源与集体行动对政府回应的影响
- 两个自变量（干预）+两个干预水平
- 具体自变量组合参见下表：

		B （集体行动）	
		b_1(有)	b_2(无)
A (政治资源)	a_1(有)	a_1b_1	a_1b_2
	a_2(无)	a_2b_1	a_2b_2

实验操作

- 研究人员分析自变量分为a1b1、a2b1、a1b2、a2b2时政府回应的程度。
- **被试** 政府官员
- **程序** 随机把被试分为4个小组，引入4个自变量，观察、记录因变量的结果。
- **假设** 依据理论假设推论出工作假设，按照作用小大预测干预效果（结果变量变化）的顺序：a1b1>a1b2>a2b1>a2b2。
- **结果？？？**

Top-Down Accountability as Propaganda: Evidence from an Online Experiment

Why would states publicize political accountability?

Our theory

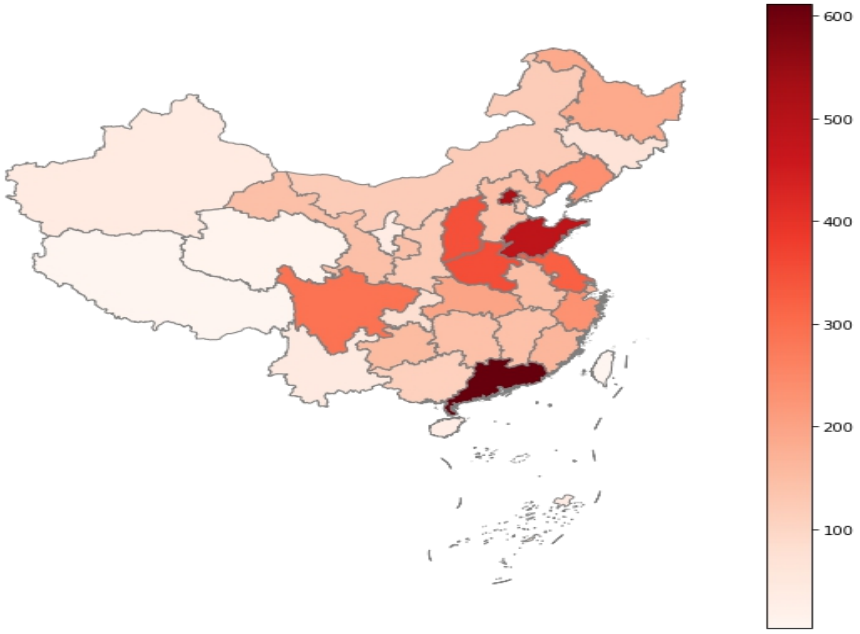
- Information: Top-down Accountability reveals information about incapacity of local officials, lowering citizens' support for local state (reduce the perceived responsibility for the center).
- Monitoring: Top-down Accountability implies the signal central government is play an active role in supervising local officials.
- Punishment: Top-down Accountability is often accompanied with punishment of incapable officials, compensating the information effect.

Hypothesis

		Support for central government	Support for local government
Investigation		+	?
Revealing	Negative Information	?	-
	Positive Information	-	+
Punishing		?	+(When decision is made by the local government)

DATA

- Survey experiment by Tsinghua Data and Governance Research Center (Meng, 2020)
- National online survey on Digital Technology and Modernization of Public Health Governance
- 6046 respondents, 5838 valid samples
- Covered 300+ prefectures in all province.
- Except Xizang, Ningxia, Qinghai and Xinjiang, 50-611 samples per province. 188 on average.



Experiment Design

请阅读关于M市的资料，然后回答以下问题：
M市位于东南沿海某省，距离湖北较远，人口约500万，经济比较发达，流动人口较多，近日受新型冠状病毒肺炎疫情影响。

	Oversight	Pos Info	Neg Info	Punishment
Control: 据悉，M市政府召开了全市防控新型冠状病毒感染的肺炎疫情工作调度会，对我市疫情防控工作再调度、再安排、再部署。				
T1: 中央督导组一行来到M市新型冠状病毒感染的肺炎防控工作指挥部，详细了解当地疫情防控情况和疑似病例检测工作。	√			
T2: 中央督导组一行来到M市新型冠状病毒感染的肺炎防控工作指挥部，详细了解当地疫情防控情况和疑似病例检测工作。在督查核查过程中，督导组对M市疫情防控工作取得的成效给予充分肯定。	√	√		
T3: 中央督导组一行来到M市新型冠状病毒感染的肺炎防控工作指挥部，详细了解当地疫情防控情况和疑似病例检测工作。在督查核查过程中，M市有关门负责人含糊其辞，对于定点医院的收治能力和具体床位数量表示不太清楚。	√		√	
T4: 中央督导组一行来到M市新型冠状病毒感染的肺炎防控工作指挥部，详细了解当地疫情防控情况和疑似病例检测工作。在督查核查过程中，M市有关部门负责人含糊其辞，对于定点医院的收治能力和具体床位数量表示不太清楚。据悉，市委已提名免去相关部门负责人职务。	√		√	√

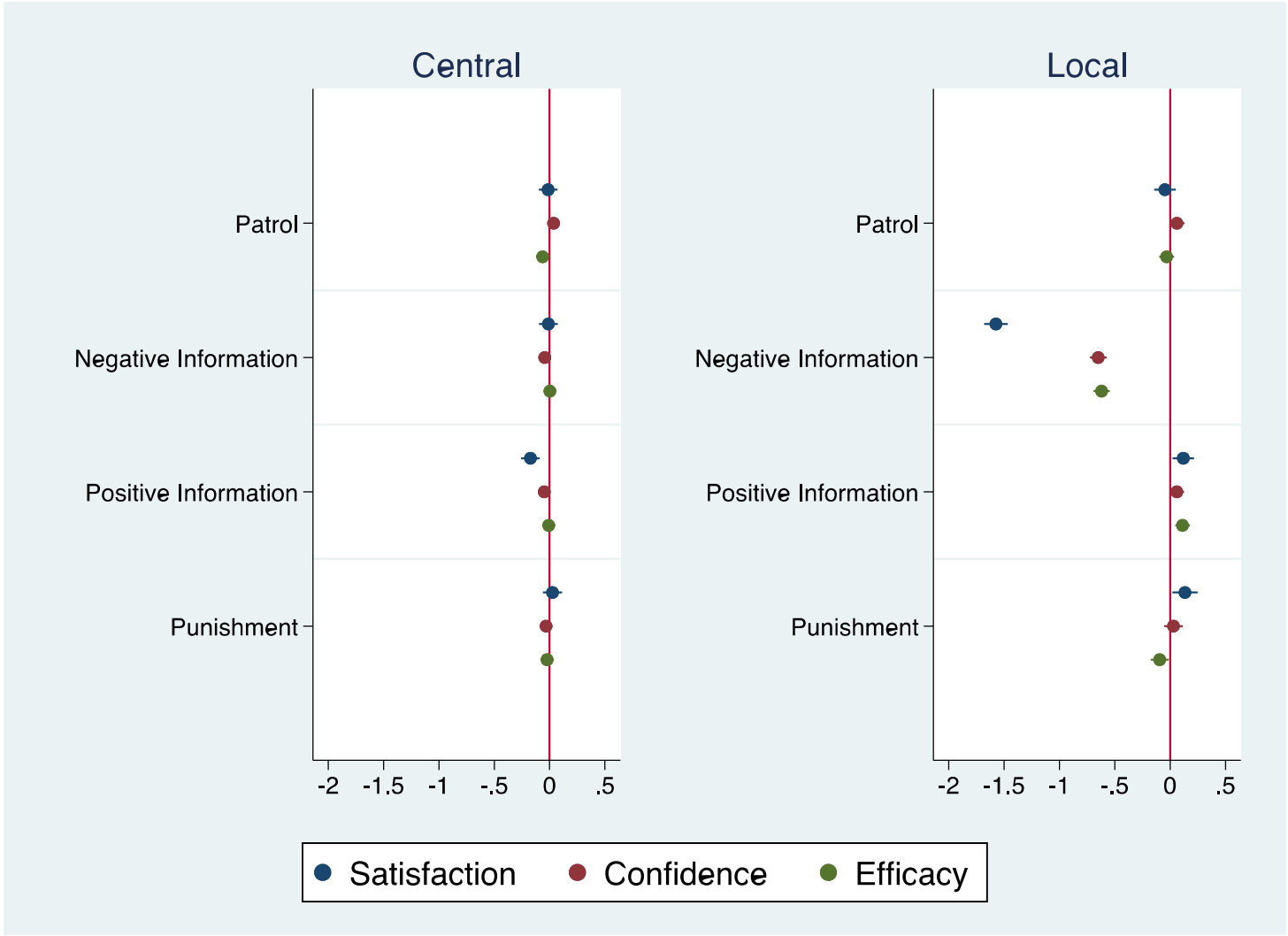
Dependent variable

- Political Support
- ✓ Satisfaction, 您对中央/M市政府部门在本次疫情防治中的总体表现是否满意？
- ✓ Efficacy, 整体而言，您认为中央/M市政府和卫计委采取的疫情应对措施在多大程度上有效？
- ✓ Confidence, 您对中央/M市能够有效地防控此次疫情的信心是？
- Responsiveness, 在本次疫情防控中，您认为人民群众的意见和诉求能被中央/M市政府及时听取吗？
- Propaganda, 我认为中央/地方媒体披露的信息是真实情况

Baseline regression

	Satisfaction		Confidence		Efficacy	
	Central	Local	Central	Local	Central	Local
Patrol	-0.01	-0.05	0.04	0.06+	-0.06*	-0.03
	(0.04)	(0.05)	(0.03)	(0.03)	(0.03)	(0.03)
Patrol& Positive Information	-0.19***	0.07	-0.01	0.12***	-0.07*	0.08*
	(0.04)	(0.05)	(0.03)	(0.03)	(0.03)	(0.03)
Patrol& Negative Information	-0.02	-1.62***	-0.01	-0.59***	-0.06+	-0.65***
	(0.04)	(0.06)	(0.03)	(0.04)	(0.03)	(0.04)
Patrol& Negative Information& Punishment	0.00	-1.49***	-0.04	-0.56***	-0.08**	-0.75***
	(0.04)	(0.05)	(0.03)	(0.04)	(0.03)	(0.04)
External Efficacy	-0.19***	-0.13***	-0.16***	-0.11***	-0.12***	-0.08***
	(0.02)	(0.02)	(0.01)	(0.01)	(0.01)	(0.01)
Internal Efficacy	-0.08***	-0.04+	-0.04***	-0.01	-0.02*	0.01
	(0.02)	(0.02)	(0.01)	(0.01)	(0.01)	(0.01)
Knowledge of COVID	-0.09***	0.01	-0.05***	-0.05*	-0.08***	-0.02
	(0.02)	(0.03)	(0.02)	(0.02)	(0.02)	(0.02)
Party Member	0.08*	0.06	0.04+	0.02	0.01	0.02
	(0.03)	(0.04)	(0.02)	(0.03)	(0.02)	(0.03)
Confirmed Cases	-0.05***	-0.04*	-0.04***	-0.03*	-0.03**	-0.02
	(0.01)	(0.02)	(0.01)	(0.01)	(0.01)	(0.01)
Control	Y	Y	Y	Y	Y	Y
Observations	5838	5838	5838	5838	5838	5838
Adjusted R-squared	0.143	0.283	0.156	0.152	0.108	0.166
AIC	17153.10	19532.65	12513.79	15542.75	12998.43	15196.13

Decomposition of Treatment effect



小结：什么是好的调查实验？

- 无偏(Unbiased)
  - Randomization
  - Balance Check
  - Blinding
  - Rondom sample
- 精度(High precision)
  - Uniform measure
  - Replication
  - Sensitive
- 适用性(Applicability)
  - List exp
  - Priming design
  - Factorial design
- 不确定性(uncertainty)
  - Replication
  - Randomization
  - Large sample and Statistical power

Q&A