

公共管理与政策研究中的实地实验： 因果推断与影响评估的视角

王思琦*

【摘要】实验方法作为统计学因果推断的重要方法，在公共管理与公共政策研究中得到了越来越多的应用。实地实验作为实验方法的新发展，相比传统的实验室实验和调查实验具有更高的外在效度和现实应用价值，被广泛用于变量之间的因果推断及政策与项目的影响评估中。论文在解释实验方法的统计学因果推断基础、并比较各类实验优劣势之后，举例分析了实地实验的基本原理、程序和可能出现的各种现实问题，并以公共政策影响评估为切入点讨论了中国公共管理与政策研究中使用实地实验方法面临的挑战与机遇。

【关键词】因果推断 实地实验 准实验 自然实验 影响评估

【中图分类号】 D63

【文献标识码】 A

【文章编号】 1674 - 2486 (2018) 01 - 0087 - 21

一、引言

实地实验 (Field Experiments, 也称为田野实验或现场实验) 方法是近年来社会科学实验方法发展的一个新阶段。相比于传统的观察研究 (Observational Studies), 如社会科学研究方法中最常用的问卷调查与管理数据分析等, 实验研究 (Experimental Studies) 对模型设定 (Specification) 和变量控制的依赖程度较小, 因此不必限定过多的假设条件和想方设法找出各种控制变量。同时, 实验方法通过随机分配 (Random Assignment) 和进行干预 (Treatment, 也称为处理), 能够有效克服内生性 (Endogeneity)、遗漏变量 (Omitted Variable)、混淆变量

* 王思琦, 西南交通大学公共管理与政法学院, 副教授。感谢复旦大学“第二届比较公共管理/政策研究与教学研讨会”与会者的意见和建议。感谢匿名评审人的意见。

基金项目: 教育部人文社会科学研究青年基金项目“基于 IAT 实验方法的政府信任测量与影响机制研究”(17XJC630009)、四川省软科学项目“风险技术公众态度形成机制研究”(2016ZR0019)。

◆ 专栏：因果推理

(Confounder)、样本选择性偏误 (Selection Bias) 等统计推断中的常见问题。作为目前社会科学因果推断 (Causal Inference) 的前沿方法, 实验方法受到了越来越多的关注与应用。

在实验方法内部, 不同于经典的实验室实验 (Lab Experiments) 通常招募学生作为实验被试 (Subjects)、采用抽象的激励手段 (金钱或点数) 来研究诸如拍卖、决策、偏好等心理学和博弈论的理论问题。实地实验通过随机分配现实中的被试 (包括个人、家庭、机构、村庄、社区、地区等不同单位), 在真实的社会经济环境 (Real Setting) 下, 使用真实的干预手段, 具有更高的外在效度 (External Validity) 和解释力。正是由于实地实验的结论能够推广到更大的范围中去, 除了一般学术研究中的作为因果推断方法的黄金标准 (Golden Standard) 外, 还被世界银行、世界卫生组织等国际组织和各国各地方政府广泛用于公共政策与项目的影响评估 (Impact Evaluation, 或称为结果评估、效果评估) 中, 成为非常具有现实应用价值的一种研究方法 (耿曙等, 2016)。

到目前为止, 实地实验目前应用最多领域是经济学 (罗俊等, 2015), 顶级经济学期刊 (AER、JPE、QJE 等) 几乎每期都刊有采用实地实验方法进行研究的学术论文,^① 涉及行为经济学、发展经济学和劳动经济学等各个领域 (Harrison & List, 2004; Angrist & Pischke, 2008)。近年来, 实地实验在政治学、管理学、传播学等学科的应用也越来越广泛。

虽然在公共管理与公共政策研究中, 实验方法尤其是实地实验方法的应用相对整个实证研究论文的数量来说还较少 (Bækgaard et al., 2015; Doberstein, 2017; Margetts, 2011), 但是越来越多的实验研究范式出现在国际顶级的公共管理刊物上, 如最近几年, JPART、PAR、PMR、PA 等刊物经常都有实验研究, 其中越来越多的研究采用实地实验方法 (Gregg et al., 2017)。有学者认为 (Grimmelikhuijsen et al., 2017; James et al., 2017), 这一方法的发展趋势与“行为公共管理学” (Behavioral Public Administration) 研究范式的兴起是高度一致的。

这些公共管理实地实验研究涉及的研究主题非常丰富,^② 包括: 公民对绩效

① 需要说明的是, 出于控制文献综述篇幅的考虑, 很多经济学研究中具有公共管理和公共政策含义的文献没有被纳入分析当中, 本文中的文献基本上限定在传统公共管理与公共政策研究方面, 感谢审稿人提出的这一点意见。

② 这里列举的仅仅是最近几年国际一流公共管理刊物中的采用实地实验方法的部分论文, 可以看出研究范围与研究主题的多样性和广泛性, 也充分说明了实地实验方法在公共管理与公共政策研究中的巨大发展空间。

信息的感知实验 (Olsen, 2017), 公共服务中的风险偏好 (Bækgaard, 2017), 领导方式对绩效的影响 (Bellé, 2013), 地方政府开放性和透明性的研究 (Ben - Aaron et al., 2017), 透明度、知识对公民政府信任的研究 (Grimmelikhuijsen, 2012), 司法透明度对公共信任的影响研究 (Grimmelikhuijsen & Klijn, 2015), 使用电话来提高公民会议参与率的研究 (Hock & Potoski, 2013), 公共服务的合作生产 (Jakobsen, 2013), 公共组织中社会交换关系研究 (Jakobsen & Andersen, 2013), 如何补偿公民接受的公共服务质量 (Thomassen et al., 2017), 公共部门领导的伦理与操守是否会影响雇员态度和行为 (Thaler & Helmig, 2015), 政府繁文缛节 (Red Tape) 对公民满意度的影响 (Kaufmann & Tummers, 2017; Tummers et al., 2016), 透明是否有利于反腐败 (Peisakhin & Pinto, 2010), 协商实验中的问责与联盟 (Bouwman et al., 2017), 等等。

与国外相比, 国内的公共管理与政策研究中实验方法的使用无论从选题还是数量都明显滞后, 但也逐渐开始出现呼吁加强实验研究的文献 (马亮, 2015), 这些国内外的研究趋势均表明, 越来越多的公共管理与政策研究者希望采用实验方法来增加研究在因果关系建立方面的可信度, 可以预期, 公共管理与政策研究即将迎来方法论上的“实验转型” (景怀斌, 2015; Bouwman & Grimmelikhuijsen, 2016; Morton, 2015)。虽然国内已经有一些文献对实验方法进行了介绍和综述 (周佳、景怀斌, 2015; 余莎等, 2015), 但是实证研究仍然非常少 (李明, 2013), 其中采用实地实验方法的更为缺乏。

因此, 本文尝试回答以下的问题, 来促进实地实验方法在中国公共管理与政策研究中的应用, 这些问题包括: 为什么实地实验方法在因果推断中的意义如此重要? 与其他实验方法相比, 具有什么样的优势? 公共管理实地实验的基本程序和通常面临的挑战是什么? 如何在中国的公共管理制度和文化背景下使用实地实验进行因果推断和政策影响评估?

本文的结构安排如下, 第二部分对统计学与计量经济学中的因果推断范式进行解释和评述, 并且比较实地实验方法与其他准实验、自然实验 (陈少威等, 2016; 马超等, 2015) 以及调查实验、实验室实验之间的差异 (李明, 2013; 徐彪, 2014)。第三部分举例说明了实地实验的基本原则和程序, 并对实地实验中常见的困难和问题进行了分析。第四部分结合中国的制度与文化环境, 对实地实验与公共政策与项目影响评估的关系进行了思考。最后的结论部分进一步对实地实验方法的意义进行了反思。

二、实地实验方法的统计原理：因果推断

（一）因果推断的反事实框架

社会科学的实验方法，其因果推断的统计学基础建立在唐纳德·鲁宾（Donald B. Rubin）（Rubin, 1974, 1980）提出的反事实框架（Counterfactual Framework）之上，在这个统计学框架中，我们可以用 $Y_i(1)$ 表示如果实验被试 i 接受干预（Treatment）的可能结果， $Y_i(0)$ 表示被试不接受干预时的结果。每个被试 i 接受干预 τ_i 的因果效应，被定义为这两种潜在结果（Potential Outcome）之间的差，即表达式（1）：

$$\tau_i = Y_i(1) - Y_i(0) \quad (1)$$

例如，如果我们想计算某个人 i 读大学（放在公共政策背景下，即是接受高等教育的干预）对其将来收入的因果效应，可将其读大学后的年收入 $Y_i(1)$ 假设为 10 万元，没有读大学直接工作的年收入 $Y_i(0)$ 假设为 4 万元，那么读大学对于个体 i 年收入的因果效应 τ_i 就是 $10 - 4 = 6$ 万元。

现实世界中，研究者面临的挑战是，在任何给定的时间点，只能观察到 $Y_i(1)$ 或 $Y_i(0)$ 两者之一，而不能同时观察到两者。如前面的例子那样，在真实的社会中，一个人只能处于读大学和不读大学这两种状态的其中之一，不可能同时处于两种状态，只要选择了一种人生道路，另外一种道路就成为了“反事实”状态，即两者相互构成对方的反事实结果。两种状态不能同时存在的条件下，每个被试的结果可以使用表达式（2）来表示：

$$Y_i = d_i Y_i(1) + (1 - d_i) Y_i(0) \quad (2)$$

一般来说，我们规定干预变量 d_i 取值 0 或者 1，因此等号右边总有一项为零。如果实施了干预（ $d_i = 1$ ），我们将观察到干预导致的潜在结果 $Y_i(1)$ 。如果没有实施干预（ $d_i = 0$ ），我们将观察到没有干预时的潜在结果 $Y_i(0)$ 。

但在之前的例子中，我们关注的是个体意义上的因果效应，而在社会科学研究，尤其是公共管理与政策研究中，通常我们并不关心某个个体的问题，研究者真正想要知道的是某项干预产生的群体意义上的因果效应，即所谓的“平均因果效应”。在潜在结果框架下，这种因果效应也被称为平均干预效应（Average Treatment Effects, ATE），我们可以将平均干预效应用表达式（3）来表示。

在实验方法中，我们将接受政策干预的研究对象命名为干预组（Treatment

Group)，而没有接受干预的研究对象，则构成控制组（Control Group）或者对照组（Comparison Group），两者互为对方的反事实比较组，即干预组是接受了干预的（Treated）控制组，控制组是没有接受干预（Untreated）的干预组。与此对应，这里的 $\mu_{Y(1)}$ 是接受干预的群体（干预组）中所有被试干预效应 $Y_i(1)$ 的平均值，而 $\mu_{Y(0)}$ 是未接受干预的群体（控制组）中所有被试 $Y_i(0)$ 的平均值。

$$ATE = \mu_{Y(1)} - \mu_{Y(0)} \quad (3)$$

在实验方法的文献中（Gerber & Green, 2012），这里的ATE计算公式也被称为均值差估计量（Difference in Means Estimator），^①仍然采用之前读大学的例子，如果我们把一定数量的被试分配到干预组（读大学）以及控制组（不读大学），经过几年大学教育的干预之后，可以通过计算干预组被试与控制组被试的平均收入之差，即得到接受高等教育政策干预的平均干预效应ATE。此外，ATE在不同的统计设定（主要是后文介绍的遵从性、缩减的不同状态）下，还有各种各样的变体，如ATT、ITT、CACE等。ATE估计量与传统统计方法，如回归分析等得到的各种系数相比（除非满足具体的条件，大多数情况下，这些系数体现的只是一种相关性，并不能做因果解释），其简单性、清晰性的优势非常明显，并且与人们的直觉最为相符，不用进行各种各样的转换（如Logistic回归中Odds Ratio那样）。

由于具有随机分配和干预这两个基本程序，与传统的观察性研究相比，实验研究在建立反事实对照组的特殊优势非常明显。整个社会科学，包括部分自然科学，如医学中的随机对照试验（Randomized Controlled Trials, RCT），实验研究方法的目标，就是尽可能通过随机化（Randomization），将实验对象随机分配到干预组和控制组中，从而形成协变量高度平衡（Perfect Balanced）、在统计学意义上各组完全等价、并且可互换的反事实比较组，从而消除混淆变量、不可观察因素等通常因果推断中的障碍，得到某项政策干预产生的因果效应（Causal Effects），即排除了其他所有干扰因素，唯一由政策干预（原因）本身导致的结果。这正是公共政策的影响评估之所以采用实地实验方法为基准的理由。

^① 与个体因果效应能够直接相减不同，由于存在统计概率误差，一般情况下，我们采用方差分析、t检验等统计方法来计算平均干预效应，但如果被试的数量太少（只有几个），不满足传统统计检验的要求，则需要采用置换检验（Permutation Test）来进行。在复杂实验设计中，研究者还经常使用包括众多协变量的传统回归方法分析实验数据，以提高平均干预效应的估计精度。

（二）准实验、自然实验与实地实验的比较

虽然实验方法在因果推断方面具有无可比拟的优势（余莎、游宇，2017）。但是由于研究成本、研究观念，实施难度等问题，相比于问卷调查这类观察性研究，实验方法尤其是实地实验方法并没有在社会科学中得到广泛的应用，但已处于迅速发展的过程当中。社会科学的学术研究文献中，尤其是经济学研究中，我们最常看到的因果推断的统计与计量方法，仍然是“准实验”（Quasi - Experiments）或者“自然实验”（Natural Experiments）^①的方法（Kamkhaji，2015），一般来说，包括工具变量方法 IV（Instrumental Variable）、双重差分方法（Difference in Differences，DID）、回归间断设计方法（Regression Discontinuity Design，RDD）以及倾向值匹配方法（Propensity Score Matching，PSM）四种。

这些准实验或者自然实验的方法，在一定程度上，可以理解为将观察性研究得到的数据，通过某种统计学的分析，模拟成为实验方法的数据（这种逻辑在倾向值匹配方法中尤其明显），从而得到因果效应。但是，由于其数据生成过程（Data Generating Process）本身的局限性，观察性研究的本质并没有改变，很多具体的研究数据集，其内在的随机化过程，只能看成近似（As If）随机化，而非真正意义上的随机化，因此在因果推断很容易遇到各种各样的难题。^②

例如，众所周知，工具变量方法中，如何选择合适的工具变量，以满足排除限制（Exclusion Restriction）和弱工具变量检验的难度非常大，公共管理与政策研究中由于学科特点，很多观察变量都是内生于制度本身的，要想找到一个合适的工具变量，难度会更大。双重差分虽然对于数据形式的要求较为宽松，但是其两个政策影响组具有共同趋势（Common Trends）的要求并不容易满足。回归间断设计中，间断点（Cut Offs）附近的带宽（Bandwidth）选择对分析结果有高度的敏感性，因此这种方法计算得到的因果效应，基本上是局部平均干预效应（Local Average Treatment Effect，LATE），很难评估大规模（时间和空间）政策的效果。倾向值匹配方法虽然解决了传统协变量匹配的困境，而且在经济学以外的领域中用得更加广泛，但匹配对于样本量大小要求较高，而且匹配程

① 自然实验与准实验在很多研究文献中是等同的，但二者也存在区别。自然实验通常指在真实社会环境下，研究者观察到的某种由社会、地理、气候等因素自发形成的随机分配机制。准实验则更强调通过统计分析将观察数据转换为实验数据的逻辑。因此，根本上来说，两者都不是真正的实验，仍然属于观察性研究。

② 经济学等学科已经初步实现了统计方法上的“可信性革命”（Angrist & Pischke，2008），相比之下，目前公共管理中采用准实验方法进行因果推断仍然较少，因此公共管理与政策研究中的实地实验发展会面临更多的机遇与挑战。

序仍然基于可观察的协变量进行，难以控制不可观察因素的影响，因此本质上仍然属于传统回归方法的变体，很难完全解决各种因果推断中出现的问题。

因此，虽然有越来越多的研究者希望将这些准实验和自然实验方法建立在基于设计的（Design Based）模式下来进行，避免纯粹的数据驱动的统计研究，以提高因果推断的效力，但是，在研究者的研究条件允许使用实验方法情况下，建议尽量采用实验设计，以避免上面讨论的各种困境。当然，准实验和自然实验研究在某些方面的优势，也是实地实验方法无法取代的，比如，实地实验由于必须采用随机化和干预程序，因此研究问题的层次不可能太高，因为过于大范围和高层次的主体很难被研究者控制与操纵。^① 一般来说，较长历史时期的、宏观的社会经济现象以及国家与国际层面的制度与文化问题，仍然只能采用传统的统计方法或者准实验、自然实验研究设计。^②

（三）实验方法：实验室、调查还是实地？

实验方法本身也是在不断的发展过程当中，目前来看，可以将社会科学中的实验方法按照实验样本类型、实施环境、干预类型大致上分为三类，即实验室实验（Lab Experiments）、调查实验（Survey Experiments）和实地实验（Field Experiments）。必须指出的是，这种分类方法是一种相对意义上的^③，现实研究中采用的实验可能是兼具三种实验的特征及其混合的方法（Belle & Cantarelli, 2017；Harrison & List, 2004）。

表 1 三种社会科学实验方法的比较

实验类型	样本特征	实施环境	干预类型
实验室实验	学生被试为主	高度封闭	诱导价值与信息
调查实验	学生与公众兼有	封闭或开放	文字图像等信息
实地实验	公众与机构被试	开放实地环境	高度多样化

资料来源：作者自制。

① 虽然在理论上，我们可以将较大的地区甚至国家作为实验单位，进行随机分配和干预，来研究跨国的长时段、宏观的制度与文化问题，例如，将全世界两百多个国家和地区随机分配到两个组（民主或专制），研究制度的长期历史结果（GDP），但却不可能实施这种实验。

② 例如，在很多公共政策和项目的影响评估中，出于评估难度（时间、经费）的考虑，相当多的评估仍然采用准实验及自然实验设计。

③ 具体可以参见（Harrison & List, 2004）对于实验室实验和实地实验相互关系的类型划分。

传统的实验室实验起源于心理学，并在实验经济学（尤其是行为实验经济学与博弈论）早期研究中具有重要的价值和作用，例如对于市场机制的各种研究。在实验室实验当中，由于处于相对封闭的环境中，研究者对于实验的环境以及干预有着高度的控制，大多采用专门的计算机系统和软件（如 Z - Tree），甚至专门的心理学和神经科学设备，能够获得较高的实验数据质量，避免突发和意外因素的干扰，因此在三类实验中，内部效度（Internal Validity）是最高的。

但是，实验室实验通常为了成本的考虑和研究便利性的考虑（很多研究者都是高校教师），经常采用学生样本；并且由于研究主题的理论化，而惯于使用抽象框架（Abstract Framing）来进行干预设计；研究者在此过程中，为了保证所研究理论的简洁性，往往加入了一些规则（条件或假设）。在实验室实验中，由于样本的特殊性和研究的抽象性（Anderson & Stritch, 2016），从而降低了研究的外部效度（External Validity），即结论的外推能力（Bozeman & Scott, 1992; Edwards, 2015）。可见实验方法在外部效度与内部效度之间存在一种权衡（Trade off）关系。

调查实验，综合了问卷调查方法与实验方法的设计（孟天广等，2015）。一方面，调查实验从外在形式上，类似于传统的问卷调查方法，都是由问题及对应选项构成的，因此在纸质问卷和网络问卷环境下都可以进行实验，调查样本与问卷调查也基本一致，除了学生之外还包括普通公众。但另一方面，又具有一定的实验设计特征，即随机化与干预。随机化（随机分配）体现在问卷按照不同的干预内容被分为多个版本，按照某种随机分配方法（如问卷编号的单双号交替）分配给调查的受访者（被试），多个版本的问卷对应了多个实验组（如 n 个干预组，1 个控制组， $n \geq 1$ ）。通过问卷中嵌入的干预文字段落或者填答问卷时呈现给受访者的图片、声音或图像信息等，来实施干预（James & Ryzin, 2017）。

由于调查实验的干预实施和结果测量呈现都基于传统的问卷调查，因此成本与实施难度都比较低（Charbonneau & Ryzin, 2015）。此外，由于很多大规模问卷调查是从一个特定的研究总体中随机抽样，因此基于总体的（Population Based）调查实验，其结论的代表性和推广性较高，兼具了实验和问卷调查的优势。但是，由于问卷载体本身的限制，这种信息干预的强度往往较弱，并且研究主题大多限于问卷调查通常测量的态度、认知等方面（Hvidman & Andersen, 2016），采用自我报告的形式，也容易受到诸如启动效应（Priming Effect）和情景效应（Context Effect）的影响（Hjortskov, 2017），并且难以测量行为等客观

维度的变量。

实地实验方法与实验室实验相比，研究采用的样本基本上都是现实样本（具体可以依据研究主题来确定，包括公众和机构）。如果经费和研究资源充足，可以先从某个研究总体中进行随机抽样，然后再对抽取的样本进行随机分配（类似调查实验），这样内在效度和外在效度都可以得到保证。更重要的是，进行实地实验的环境是真实的社会环境（Field），大多数实地实验中，研究被试对于自己正在参与实验是无察觉的，即非介入性（Unobtrusive）的，避免了某些诸如霍桑效应（Hawthorn Effect）和约翰·亨利效应（John Henry Effect）^①等问题。实地实验采用的干预除了实验室实验和调查实验通常的信息干预外，还可以采用多种多样的现实干预（物理的、生理的、心理的、经济的、政治的、社会的），如果实地实验用于公共政策与项目的影响评估，这些政策与项目的内容就是干预本身。

当然，实地实验相比实验室实验和调查实验，也存在一些不足，例如，由于在真实的社会、经济和政治环境中进行实验，实验方法和结果的估计较为复杂，实验的设计、管理、实施和测量的经费成本较高，需要解决很多交通、通讯、协调和后勤问题。由于对公众的工作和生活有直接影响，也容易产生研究伦理问题，例如某种政策和项目会给实验对象带来实际的利益（现金或者实物补助、免费医疗、免费教育），如果被分配到控制组，就意味着无法获得这些资源，从而可能陷入更加困难的境地。尤其是在实地环境下的实验，相对于实验室实验的控制程度较低，很容易遇到各种现实问题，如不遵从、样本缩减、相互干扰等。下面将在基于实地实验的程序和原理的基础上，举例来进行具体分析。

三、实地实验研究的基本步骤：一项实例

（一）开展实地实验的三个阶段

为了更清晰和具体地说明在公共管理与政策研究中进行实地实验的基本过

^① 约翰·亨利效应是指分配到控制组的被试，可能会刻意改变自己的行为（如提高生产效率）来表示对没有被分配到干预组的某种不满，从而导致干预效应有偏，即降低了干预效应的数值。

程，^① 下面我们将使用一项公共管理实地实验的研究实例，来帮助读者理解。^②

这个例子（Ben - Aaron et al. , 2017）是发表在 PAR（*Public Administration Review*）上的论文《遵从导致的透明：一项评估地方政府信息公开的实地实验》。^③ 这项研究旨在通过实地实验理解美国地方政府（县政府）信息公开程度的影响因素之一：群体遵从性（Conformity），即政府组织间的相互影响（Peer Compliance）是否会导致地方政府信息公开程度的变化。研究发现，遵从性不仅影响了地方政府履行（Fulfill）信息公开要求的初始响应时间，还影响了地方政府履行要求的比率。

研究者向北卡罗莱纳州的 100 个县政府提出公开政府内部电子邮件档案（E - Mail Archives）的要求（Ben - Aaron et al. , 2017）。研究者之所以选择电子邮件档案作为信息公开的内容，有两个原因。首先，电子邮件是大多数政府机构的标准传播媒介。第二，电子邮件是许多行政区域公共记录的一部分。研究者采用了一个顺序要求（Sequential Request）过程，来检验县政府履行信息公开要求的行为是否会受到其他县政府（Peers）已履行同样要求的影响。^④

1. 随机分配阶段

随机分配是实验研究的关键步骤，正是通过随机化，实验方法才能将干预组和控制组被试在各种可以观察以及不可观察特征上的差异平均分配。在实地实验中，由于随机分配的对象都是真实的社会单位（个人、机构、区域），这些实验单位通常呈现分层和聚类的空间分布模式，因此除了实验室实验通常采用的简单和完全随机化以外，更多地采用两种随机分配方法，即区块（Block）随机化与整群（Cluster）随机化（类似随机抽样中的分层和整群抽样）。区块随机化中，被试被划分为多个亚组（或称为区块或层），每一个区块内采用完全随机

① 在有的文献中，将实验数据的分析阶段也作为实验研究的程序，参见陆方文（2017）。

② 由于社会科学各学科包括公共管理与政策研究的实地实验方法在不断发展过程中，各种研究在具体实验设计上不尽相同，因此本例中的实验操作化并不能作为标准来看待。但是众多实地实验研究必须满足随机化与干预两个必要条件。

③ 之所以选择这篇论文作为实例，一方面是因为该文的研究主题属于公共管理传统领域“地方政府透明度”，便于公共管理领域的读者理解概念和术语，另一方面，论文本身的设计与目标也有公共政策影响评估的含义，即评估地方政府信息公开政策的效果。此外，由于本文属于方法论性质的综述，所以主要将关注点集中在这项研究的实验方法方面，而非实质性的理论和观点。

④ 这种政府组织之间影响机制在社会学文献中也被称为参照群体（Peer Reference Group）。

分配。例如，一项实验中我们有 200 个被试，其中 100 位男性和 100 位女性。实验设计要求将 100 个被试分配到干预条件，如果使用完全随机分配，偶然性会导致干预组中的男性与女性数量不同。区块随机化，则可以保证分配到每一实验条件的男性与女性数量相同。整群随机化中，同一个群的所有被试被随机分配到一个组，要么干预组要么控制组。整群分配可以消除多个被试在同一个群里但却被分配到不同实验条件的可能性。例如，某个村庄的所有农民被作为一个群被一起分配到干预或控制条件，这样，他们共同拥有的某种属性被置于同一种实验条件下。实施良好的区块与整群随机分配，可以在节约研究资源的同时，保证实验结果的可靠性与精度。

在这项信息公开实地实验中，随机分配的被试单位是政府机构，即北卡罗来纳的县政府（County Governments）。这些县政府属于同一个州，因此具有共同的属性，即受同一部信息公开法（Sunshine Law）的约束，此外电子邮件是各地区政府共同使用的政府公务沟通手段，因此各个实验单位之间具有可比性。在实验研究的可行性方面，由于州法律规定政府公务电子邮件被视为一种公共记录（Public Records），并且编制非保密信息的成本不能由信息公开要求者承担（降低了实验经费成本），给实地实验的进行提供了基本条件（Ben - Aaron et al. , 2017）。

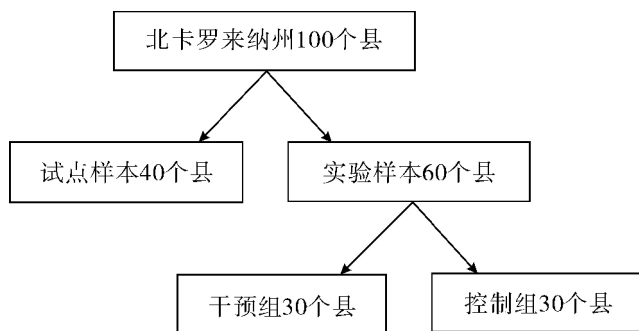


图1 北卡地方政府信息公开研究中的随机分配

资料来源：作者根据（Ben - Aaron et al. , 2017）的相关内容绘制。

在随机分配设计方面，由于一个州包括的县政府数量并不很大，研究者最终采用的是简单随机分配（Ben - Aaron et al. , 2017）。研究者将北卡罗来纳州的县随机分为两个样本：一个由 40 个县组成的试点样本（Pilot Sample）和一个由 60 个县组成的实验样本（Experimental Sample），见图 1 所示。然后将后一组样本分为干预组和控制组，随机分配 30 个县到每组。与很多实地实验研究直接进行一次随机分配不同，该研究之所以在正式随机分配前还进行了一次随机分

配，是因为这些试点的县政府样本刚好构成了干预本身，即遵从信息公开的县政府（以便在后续的实验中，将其信息透露给实验县政府样本）。

2. 干预实施阶段

在进行了随机分配之后，就可以开展实验干预了，实地实验干预的内容和形式非常多元化，若实地实验研究者具有足够的创造性和想象力，几乎所有行为和政策都可以被用于实地实验的干预，但其选择从根本上仍然要依据文献和理论假设，即必须充分反映研究概念和原因变量本身的内在特质。干预的数量越多，越能够反映一个变量的多个维度和多个研究变量（在政策和评估的背景下，意味着各种不同的政策方案比较），通过因素设计（Factorial Design）等，还可以探索变量之间的交互作用。不过随着干预变量数量的增加，为了保证实验结果的统计效力，实验中干预组数量和被试数量也会急剧增加，所以必须基于经费、时间和资源等进行仔细的考虑与权衡。

在信息公开研究中，研究采用的干预是对县政府的信息公开要求，要求每个县公开所有非私人的电子邮件。这些电子邮件包括县政府管理者（County Managers）和各部门管理者（Department Managers）发件箱与收件箱中的电子邮件，随机选择三个月作为邮件覆盖的时间范围。^①研究者通过检查每个县政府的网站，找出最有可能满足要求管理者的电子邮箱，并给其发送一份包含信息公开要求的邮件。研究者首先对试点样本中的40个县政府实施了这种干预，然后对实验样本中的60个县政府进行了同样干预。

需要指出的是，这项研究的干预并非一次性的，在等待60天后，研究者发出了第二份信息公开邮件（只对干预样本，即实验组中30个县），邮件内容如下：“为了供你参考，我们想让你知道我们已经向北卡罗来纳州的其他县政府发出了这一请求。它们有些已经满足了我们的要求，包括xxx县等。”从研究中看到，邮件中满足了信息公开要求的县政府一共8个（这8个正是试点样本中满足了要求的县政府）。所以第二份邮件实际上才是真正的干预（体现研究理论假设中的群体遵从压力信息）。

3. 结果测量阶段

实地实验研究中的测量，分为对结果变量的测量和协变量的测量。测量的工具和干预手段一样，是高度多样化的。例如在很多态度、认知相关的实地实验研究中，测量采用的是传统的问卷调查与访谈方法，可见，实地实验方法与

^① 之所以随机选择，可能是研究者为了避免政府电子邮件某种周期性对实验结果产生的系统性偏差。

传统社会科学研究方法之间的关系不是互斥的，而是相互配合使用的。测量有时候会在随机化之后、干预实施之前进行，即所谓的前测（Pre - Test），在政策评估背景下，有时也称为基线测试（Base Line Test）。在干预之后的一段时期，可以进行后测（Post - Test）或者终线测量（End Line Test）。测量也可以在实验前、实验中和实验后多次进行，以便测量变量的变化趋势乃至长期实验结果。

这项信息公开研究中，由于实验单位是机构（县政府）而非个人，^①并且没有面对面的交流，采用的结果测量实际上是一种行为测量。研究者测量了实验县的履行信息公开要求行为的两个方面：第一，干预组县政府在初步答复前已过去了多少天；第二，如果县政府做出答复，他们最终是否满足了要求。

结果发现，60个县中有44个在60天内做出了初步答复，其中干预组20个，控制组24个。平均来说，干预组县政府做出答复的天数更少，最后履行信息公开要求的比例也更高。为了在后面的数据分析（回归）中，提高平均干预效应的估计精度，减少偏误，研究者还对协变量（Covariates）进行了测量，包括每个县的人口数量、每个县的政府工作人员数量和透明度得分^②。

（二）实地实验中的常见问题

如前所述，实地实验由于在真实的社会环境中，采用真实的干预，对实际的社会主体进行操作，会遇到一些在实验室实验中较为少见的问题（Gerber & Green, 2012）。这些问题中，比较突出的问题包括被试的不遵从（Noncompliance）、实验被试的缩减（Attrition）和实验单位的相互干扰（Interference）等。

不遵从，是由于某些被试没有接受分配的干预而产生的问题（Gerber & Green, 2012）。一种情况称为“单边不遵从”（One - Sided Noncompliance）或“干预失败”（Failure - to - Treat），即某些分配到干预组的被试并没有实际接受干预。另一种更复杂的情况是“双边不遵从”（Two - Sided Noncompliance），即某些分配到干预组的被试没有接受干预，同时某些分配到控制组的被试却接受了干预。

被试样本缩减，指的是实验中结果变量的数据缺失（Missing Data），可能是由于被试的迁徙流动，也可能是被试拒绝接受实验测量导致的（Gerber & Green, 2012）。某些类型的（往往是非随机的）缩减会对因果效应的无偏推断

① 虽然实施干预（发送邮件）的对象是个人（县政府管理者），但是这些个人实际上是视为机构代表。

② 这些透明度得分数据来自于约翰·洛克基金会（John Locke Foundation），分为ABCD四个档次。

产生严重威胁。如果只是简单的从数据集中删除没有结果测量的观察值，意味着干预组或控制组中剩余被试不再是最初被试集的随机样本，对各组平均值的比较不再是平均干预效应的无偏估计量。

实验单位之间相互干扰，会对潜在结果框架的基本假定产生威胁，在因果推断的统计学术语中，这个假定被称为“稳定的单位干预值假定”（Stable Unit Treatment Value Assumption, SUTVA）即该被试的潜在结果只受到其自身是否接受干预的影响，与其他单位接受的干预无关。在具体的实证研究文献中，这种干扰问题有时候用溢出效应（Spillover Effect）来形象描述。

在这项信息公开研究中，由于研究样本的数量不多，实验单位都是政府机构，而且干预（电子邮件）本身是受到高度控制的，因此几乎没有出现不遵从以及样本缩减问题。^①但样本间的相互干扰却可能成为一种较严重的问题，包括试点样本县政府与实验样本县政府之间的相互影响，以及实验样本内部干预组县政府与控制组县政府之间的相互影响。这些县均位于北卡罗来纳州内，区域邻近性（Proximity）以及政府管理者之间人际网络均可能导致干预的溢出效应。不过，在本研究中，由于自变量（干预）本身恰好就是这种网络扩散效应（遵从压力），从而避免了对研究结果的干扰。

在实地实验方法中，为了处理各种各样的问题对实验数据因果推断的威胁，往往会采用各种复杂的统计和计量方法来对统计结果进行修正（例如，不采用最简洁清晰的均值差估计量，而采用包括众多协变量的回归模型来估计因果效应，以提高估计值的精度），以获得无偏性和一致性的估计值（Gerber & Green, 2012）。本研究同样使用了包括协变量、交互项的多种回归模型以及置换检验（Permutation Testing）和调整模型参数（如时间）的模拟等统计方法，来增强研究结果的稳健性。

这些修正虽然一定程度上能够降低这些问题的危害性，但归根结底还是应当在实验设计一开始，就想方设法预计到可能出现的问题，并巧妙地用各种经验和诀窍来尽量避免，而不是实验结束之后对数据缺陷进行补救，否则实验数据相对于传统观察性数据的优势就会丧失殆尽。

四、实地实验的重要应用：政策影响评估

在公共管理与公共政策研究中，实地实验方法除了用于变量的因果推断之

^① 研究者考虑到了可能出现的不遵从问题，如政府管理者虽然收到邮件，却没有查看干预信息。这样会低估干预效应，实验结果可能是真正干预效应的下界。

外，一项非常具有现实意义和长远前景的应用领域，就是公共政策与项目的影响评估。虽然在影响评估实践领域中，各种准实验的推断方法，如工具变量、双重差分、回归间断设计等也经常被采用，传统的问卷调查、观察、访谈等资料收集方法也是重要的组成部分，但是实地实验方法作为因果推断的黄金标准，具有不可替代的价值和作用（Gertler et al.，2016）。其他方法都是在随机实地实验方法不可行的时候，才会采用，或者作为实地实验方法的补充。^①

近年来，公共政策研究的一个重要发展趋势，是基于证据的政策制定或称为循证决策（Evidence - Based Policy Making）越来越被研究者和实践者（政府与非政府组织）作为政策制定的核心标准（Gertler et al.，2016），例如许多发展经济学家广泛参与到世界银行、世界卫生组织在全世界发展中国家进行的援助项目的影响评估中（Alatas et al.，2012）。大规模的评估项目包括阿富汗国家重建项目（Afghanistan's National Solidarity Program）评估，研究者采用随机对照试验（RCT）评估了项目对数百个（250 个干预组村庄，250 个控制组村庄）阿富汗村庄的公共服务、基础设施、社区治理、政治态度的影响（Beath et al.，2013）。著名的影响评估研究机构包括隶属于麻省理工学院的“贫困行动实验室”（The Abdul Latif Jameel Poverty Action Lab，J - PAL）和“影响评估国际行动”（International Initiative for Impact Evaluation，3IE）等。

虽然政策和项目影响评估的全过程非常复杂，包括建立变迁理论（Theory of Change）、确定结果链（Results Chain）、说明评估问题、选择结果指标、开展随机实验（或准实验方法）、政策结果测量、数据整理和分析、评估报告写作等（Gertler et al.，2016）。但是其原则却非常简单，即致力于建立某项公共政策或项目（原因、干预）与某种影响（效果、结果）之间可信的因果关系。

传统的政策与项目评估方法，由于其在因果推断方面存在不足，往往难以建立政策（项目）和某个结果之间的确定关系（Khagram & Thomas，2010）。这些不足主要体现在两个方面。其一，许多传统政策评估方法，关注于政策的实施过程评估而忽视了政策的最终效果，这些评估方法，不管采用专家评估还是第三方评估的模式，不管使用的方法是访谈、问卷调查还是数据分析，都非常强调政策过程中的投入和产出测量，例如本项目花费了多少经费、投入了多少人力物力、文件档案是否齐备、设施和场所修建情况、有没有浪费等，但往往却忽视了项目本身是否真正对干预对象的某种结果产生了实质影响，是否改善

^① 审稿人提出，虽然实地实验与影响评估之间存在一些共同部分，但是各自存在自己的话语体系。因此这里仅对二者的交集进行分析，限于篇幅，影响评估本身的理论和方法体系不做进一步讨论。

◆ 专栏：因果推理

了学生的短期入学率、到校率，健康状况，乃至长期的收入、心理与婚姻，而这些才是政策的根本目标。

其二，即使某些政策与项目评估方法，关注到了结果的重要性，但是由于采用方法本身难以进行因果推断，得出了有偏的估计值。例如，很多政策与项目的效果评估采用将项目参与者的某种结果与非参与者的某种结果进行对比，即所谓的“内外比较”，有时候还采用将参与者参加项目之前与之后的某种结果对比，即所谓的“前后比较”，但这些比较方法都没有解决样本选择偏误的问题。因此，可能会将项目之外其他因素造成的结果，归结于项目本身，从而高估或者低估了政策和项目的因果效应。例如，某项政府提供的再就业培训项目，比较了参加再就业培训之前参与者的就业状态与培训后的就业状态，发现就业率有了显著的提高，于是政府大力宣传再就业项目的成功，但是这些参与培训者的就业提高，可能与宏观经济本身的改善导致失业率下降有关，而非政府培训的作用。

如果不进行培训前后比较，而进行参加培训的人员与没有参加培训人员的比较，同样也存在因果推断的困难。因为，参加培训者的某种隐藏特征（就业技能、再就业积极性、能够动用的社会关系）与非参加者有系统性的差异，正是这种差异而非政府培训项目导致了他们在再就业结果方面的不同。换句话说，参不参加培训，这些培训对象都很有可能再就业。

这种情况下，如果我们通过基于随机实地实验的方法，随机将需要再就业的人群分配到培训（干预组）与不培训（控制组）中，然后使用培训来进行干预，最后（长期或短期）进行再就业的结果测量（通过调查或者就业统计数据），才能有效避免这些问题，建立培训项目与再就业结果之间的因果关系。

因此，在中国的现实背景下，影响评估具有特别的意义，通过将政府公共政策与项目的评估建立在影响评估的基础上，在某些情况下，就可以避免所谓的形象工程与政绩工程。通过建立在影响评估之上的问责机制，可以促使政府部门将精力花在确实有效的政策工具上，而非选择一些只有宣传效果的政策手段。

此外，通过政策影响评估，还可以对不同的政策方案效果进行比较，基于成本数据，能够得到哪种方案和哪些政策工具的组合具有最高的性价比。政策实验本身还有试点（Trial and Error）的价值：如果某种政策在某个地区或人群中被实验证明有效，那么就可以逐渐将政策用于更多地区，即扩大政策的实施范围；如果某种政策被发现缺乏实际效果或者性价比太低，也可以取消或用其他政策来替代。可以预期，通过公共政策与项目的影响评估，长期来看，可以

有效提升政府的治理能力与治理水平。

五、结论

本文对实地实验方法在因果推断中的价值，实验方法的不同类型及各自的优缺点进行了深入分析，在此基础上，进一步举例说明了实地实验方法的原理、程序和常见问题，并分析了其在公共政策影响评估中的重要意义。

基于以上的分析，我们可以得到一些启示。首先，实地实验在中国的制度与文化背景下可能会面临一些挑战与机遇。例如，在公共管理与政策研究中采用实地实验方法，需要得到政府部门（或者非政府组织）的大力配合与协助，无论是抽样框架的获取，随机分配的实施，干预的进行，还是结果测量，都需要动员相当大的人力物力资源，政府部门可能出于敏感性以及对研究结果不利的担心，而拒绝配合甚至制造障碍。此外，由于实验方法在中国公众与政府工作人员中的认知与了解程度较低，使得他们更容易接受传统的问卷调查和访谈方法，不太愿意参与实验研究或者觉得多此一举或浪费资源。

当然，实地实验方法在中国公共管理与政策研究中也拥有一些独特机遇。中国快速的经济发展和 社会变迁，为实验方法提供了大量的有理论意义和现实意义的研究问题，可能出现具有重大国际比较研究价值的选题。同时，如果能够取得政府部门（非政府组织）的信任，将实验研究作为其常规工作内容、政策试点和实施的一部分，通过巧妙的设计自然地融入其中，将干扰性和负面影响降低到最小程度，并且能够为公共管理与政策提供有价值的信息和建议，那么实地实验也是有很多研究机会的。

总之，实地实验在中国公共管理与公共政策的研究中发展空间是巨大的，可以预期，将会有越来越多的研究采用实地实验以及其他实验方法。但是，与最早采用随机对照试验的医学研究相比较，如果社会科学中的实验研究，缺乏医学研究那样对于研究结论的不断验证（Arceneaux & Nickerson, 2009; Walker et al., 2017）与复制（Replication），以及将同类研究聚合起来，进行各种元分析（Meta Analysis），那么，通过实验研究得到的可能是只是大量分散的、琐碎的、甚至相互矛盾的知识（Jilke et al., 2016, 2017），而非对社会现象中的因果机制具有深刻洞察与理解的自治理论。

参考文献

陈少威、王文芹、施养正(2016). 公共管理研究中的实验设计——自然实验与田野实验. 国

◆ 专栏：因果推理

外理论动态, 5: 76 – 84.

耿曙、余莎、孔晏(2016). 实验方法在公共政策研究中的应用: 以纳税遵从为例. 公共管理与政策评论, 3: 45 – 52.

景怀斌(2015). 专栏导语: 强化公共管理实验研究的几个问题. 公共行政评论, 3: 120 – 125.

李明(2013). 拥挤的动机: 公益投资中的公共服务动机与外部激励. 心理科学, 5: 1181 – 1186.

陆方文(2017). 随机实地实验: 方法、趋势和展望. 经济评论, 4: 149 – 160.

罗俊、汪丁丁、叶航、陈叶烽(2015). 走向真实世界的实验经济学——田野实验研究综述. 经济学(季刊), 3: 853 – 884.

马超、顾海、孙徐辉(2015). 参合更高档次的医疗保险能促进健康吗? ——来自城乡医保统筹自然实验的证据. 公共管理学报, 2: 157 – 15.

马亮(2015). 公共管理实验研究何以可能: 一项方法学回顾. 甘肃行政学院学报, 4: 13 – 23.

孟天广、杨平、苏政(2015). 转型中国的公民意见与地方财政决策——基于对地方政府的调查实验. 公共管理学报, 3: 57 – 68.

徐彪(2014). 公共危机事件后政府信任受损及修复机理——基于归因理论的分析 and 情景实验. 公共管理学报, 2: 27 – 38.

余莎、耿曙、孔晏(2015). 如何有效征税: 来自纳税遵从实验的启发. 公共行政评论, 3: 151 – 175.

余莎、游宇(2017). 不操纵无因果: 实验政治学几个核心的方法论问题. 甘肃行政学院学报, 2: 53 – 66.

周佳、景怀斌(2015). 观念形态激活对公共问题决策的影响与机制. 公共行政评论, 3: 126 – 150.

Alatas, V., Banerjee, A., Hanna, R., Olken, B. & Tobias, J. (2012). Targeting the Poor: Evidence from a Field Experiment in Indonesia. *American Economic Review*, 102(4): 1206 – 1240.

Anderson, D. M. & Stritch, J. M. (2016). Goal Clarity, Task Significance, and Performance: Evidence from a Laboratory Experiment. *Journal of Public Administration Research and Theory*, 26(2): 211 – 225.

Angrist, J. D. & Pischke, J. S. (2008). *Mostly Harmless Econometrics: An Empiricist's Companion*. Princeton: Princeton University Press.

Arceneaux, K. & Nickerson, D. W. (2009). Who Is Mobilized to Vote? A Re-analysis of 11 Field Experiments. *American Journal of Political Science*, 53(1): 1 – 16.

Bækgaard, M. (2017). Prospect Theory and Public Service Outcomes: Examining Risk Preferences in Relation to Public Sector Reforms. *Public Administration*, 3: 1 – 16.

- Bækgaard, M. , Baethge, C. , Blom – Hansen, J. , Dunlop, C. A. , Esteve, M. , Jakobsen, M. , Kisida, B. , Marvel, J. , Moseley, A. , Serritzlew, S. , Stewart, P. , Thomsen, M. K. & Wolf, P. J. (2015). Conducting Experiments in Public Management Research: A Practical Guide. *International Public Management Journal*, 18 (2) : 323 – 342.
- Beath, A. , Christia, F. & Enikolopov, R. (2013). *Do Elected Councils Improve Governance? Experimental Evidence on Local Institutions in Afghanistan*. World Bank Policy Research Working Paper, No. 6510.
- Belle, N. & Cantarelli, P. (2017). Do Ethical Leadership, Visibility, External Regulation, and Prosocial Impact Affect Unethical Behavior? Evidence from a Laboratory and a Field Experiment. *Review of Public Personnel Administration*, 81 : 1 – 23.
- Bellé, N. (2013). Leading To Make a Difference: A Field Experiment on the Performance Effects of Transformational Leadership, Perceived Social Impact, and Public Service Motivation. *Journal of Public Administration Research & Theory*, 24 (1) : 109 – 136.
- Ben – Aaron, J. , Denny, M. , Desmarais, B. & Wallach, H. (2017). Transparency by Conformity: A Field Experiment Evaluating Openness in Local Governments. *Public Administration Review*, 77 (1) : 68 – 77.
- Bouwman, R. & Grimmelikhuijsen, S. (2016). Experimental Public Administration from 1992 to 2014. *International Journal of Public Sector Management*, 29 (2) : 110 – 131.
- Bouwman, R. , Van Thiel, S. , Van Deemen, A. & Rouwette, E. (2017). Accountability and Coalitions: Evidence from a Negotiation Experiment. *Public Administration Review*, 78 (1) : 37 – 47.
- Bozeman, B. & Scott, P. (1992). Laboratory Experiments in Public Policy and Management. *Journal of Public Administration Research and Theory*, 2 (3) : 293 – 313.
- Charbonneau, É. & Van Ryzin, G. G. (2015). Benchmarks and Citizen Judgments of Local Government Performance: Findings from a Survey Experiment. *Public Management Review*, 17 (2) : 288 – 304.
- Doberstein, C. (2017). An Emerging Experimental Approach to Public Administration Research in Canada. *Canadian Public Administration*, 60 (1) : 135 – 139.
- Edwards, B. C. (2015). Unfulfilled Promise: Laboratory Experiments in Public Management Research. *Public Management Review*, 17 (10) : 1518 – 1542.
- Gerber, A. S. & Green, D. P. (2012). *Field Experiments: Design, Analysis, and Interpretation*. New York: W. W. Norton.
- Gertler, P. J. , Martinez, S. , Premand, P. , Rawlings, L. B. & Vermeersch, C. M. J. (2016). *Impact Evaluation in Practice*. New York: World Bank Publications.
- Gregg, G. , Van Ryzin, G. , James, O. & Jilke, S. (2017). *Experiments in Public Management Research: Challenges and Contributions*. Cambridge: Cambridge University Press.
- Grimmelikhuijsen, S. (2012). Linking Transparency, Knowledge and Citizen Trust in Government: An Experiment. *International Review of Administrative Sciences*, 78 (1) : 50 – 73.

◆ 专栏：因果推理

- Grimmelikhuijsen, S. & Klijn, A. (2015). The Effects of Judicial Transparency on Public Trust: Evidence from a Field Experiment. *Public Administration*, 93(4):1623 – 1630.
- Grimmelikhuijsen, S., Jilke, S., Olsen, A. L. & Tummers, L. (2017). Behavioral Public Administration: Combining Insights from Public Administration and Psychology. *Public Administration Review*, 77(1):45 – 56.
- Harrison, G. W. & List, J. A. (2004). Field Experiments. *Journal of Economic Literature*, 42(4):1009 – 1055.
- Hjortskov, M. (2017). Priming and Context Effects in Citizen Satisfaction Surveys. *Public Administration*, 26:1 – 15.
- Hock, S., Anderson, S. & Potoski, M. (2013). Invitation Phone Calls Increase Attendance at Civic Meetings: Evidence from a Field Experiment. *Public Administration Review*, 73(2):221 – 228.
- Hvidman, U. & Andersen, S. C. (2016). Perceptions of Public and Private Performance: Evidence from a Survey Experiment. *Public Administration Review*, 76(1):111 – 120.
- Jakobsen, M. (2013). Can Government Initiatives Increase Citizen Coproduction? Results of a Randomized Field Experiment. *Journal of Public Administration Research & Theory*, 23(1):27 – 54.
- Jakobsen, M. & Andersen, S. C. (2013). Intensifying Social Exchange Relationships in Public Organizations: Evidence from a Randomized Field Experiment. *Journal of Policy Analysis & Management*, 32(1):60 – 82.
- James, O. & Van Ryzin, G. G. (2017). Motivated Reasoning about Public Performance: An Experimental Study of How Citizens Judge the Affordable Care Act. *Journal of Public Administration Research and Theory*, 27(1):197 – 209.
- James, O., Jilke, S. & Van Ryzin, G. G. (2017). Behavioural and Experimental Public Administration: Emerging Contributions and New Directions. *Public Administration*, 95:865 – 873.
- Jilke, S., Petrovsky, N., Meuleman, B. & James, O. (2016). Measurement Equivalence in Replications of Experiments: When and Why It Matters and Guidance on How to Determine Equivalence. *Public Management Review*, 19(9):1293 – 1310.
- Jilke, S., Walle, S. V. D. & Kim, S. (2016). Generating Usable Knowledge through an Experimental Approach to Public Administration. *Public Administration Review*, 76(1):69 – 72.
- Kamkhaji, J. C. (2015). Regulators and Reform: A Quasi – Experimental Assessment of the Effects of Training Inspectors. *International Public Management Journal*, 18(2):304 – 322.
- Kaufmann, W. & Tummers, L. (2017). The Negative Effect of Red Tape on Procedural Satisfaction. *Public Management Review*, 19(9):1311 – 1327.
- Khagram, S. & Thomas, C. W. (2010). Toward A Platinum Standard for Evidence – Based Assessment by 2020. *Public Administration Review*, 70(s1):S100 – S106.
- Margetts, H. Z. (2011). Experiments for Public Management Research. *Public Management Review*, 13

(2):189 – 208.

Morton,R. (2015). Experiments in Public Management Research. *International Public Management Journal*,18(2):151 – 170.

Olsen, A. L. (2015). Human Interest or Hard Numbers? Experiments on Citizens' Selection, Exposure, and Recall of Performance Information. *Public Administration Review*,44(5):61 – 67.

Peisakhin,L. & Pinto,P. (2010). Is Transparency an Effective Anti – Corruption Strategy? Evidence from a Field Experiment in India. *Regulation & Governance*, 4(3):261 – 280.

Rubin,D. B. (1974). Estimating Causal Effects of Treatments in Randomized and Nonrandomized Studies. *Journal of Educational Psychology*,66(1):688 – 701.

Rubin,D. B. (1980). Randomization Analysis of Experimental Data:The Fisher Randomization Test Comment. *Journal of the American Statistical Association*, 75:591 – 93.

Thaler,J. & Helmig, B. (2015). Do Codes of Conduct and Ethical Leadership Influence Public Employees' Attitudes and Behaviours? An experimental analysis. *Public Management Review*, 18(9):1 – 35.

Thomassen,J. P. ,Leliveld,M. C. ,Steven,V. D. W. & Ahaus,K. (2017). Compensating Citizens for Poor Service Delivery:Experimental Research in Public and Private Settings. *Public Administration* 7:1 – 17.

Tummers,L. ,Weske,U. ,Bouwman,R. & Grimmelikhuijsen,S. (2016). The Impact of Red Tape on Citizen Satisfaction: An Experimental Study. *International Public Management Journal*, 19(3): 320 – 341.

Walker,R. M. ,James,O. & Brewer,G. A. (2017). Replication,Experiments and Knowledge in Public Management Research. *Public Management Review*,3:1 – 14.

Field Experiments in Research of Public Administration and Public Policy: Causal Inference and Impact Evaluation

..... Siqi Wang

Abstract As an important method of statistical causal inference, experimental methods have been used more and more in the study of public administration and public policy. Field experiments, as a new development of experimental methods, have higher external validity and broader practical implications than traditional laboratory experiments. They are widely used in policy and program impact evaluations. After explaining the statistical principle of experimental methods and comparing various sorts of experiments, this paper analyzes the potential problems of field experiments. Using public policy impact evaluation as an example, the paper discusses the challenges and opportunities of field experiment methods in China.

Key Words Causal Inference; Field Experiments; Quasi – Experiments; Natural Experiments; Impact Evaluation

Impact Assessment Based on a Result Chain and Its Experimental Method

..... Tangzhe Cao

Abstract In a result chain, an impact assessment focuses on the relationship between the intervention and the net effect. The purpose is to separate the net impact from the total effect and to prove the causal relationship between the intervention and the net effect. An impact assessment involves almost all the issues of public management research. In this study, experimental methods were used for Ex Ante and Ex Post impact evaluations. Study results can further promote the development of impact assessments from the perspective of methodology.

Key Words Result Chain; Impact Assessment; Experimental Method

The Potential Outcomes Model in the Causal inference: Origins, Logics and Implications

..... Wenzhao Li

Abstract The Potential Outcomes Model is a revolution in causal inference in statistics but it also provides a new perspective on causal inference. The simple idea of the potential outcomes model focuses on the comparison of different treatments on the same object. The differences between the treatments are the causal effects of one treatment compared with the other. The two hypotheses of the potential outcomes model are the Stable Unit