试点与实验:社会实验法及其对试点机制的启示

刘军强　胡国鹏　李　振

内容提要:试点是中国政策创新与制度建设的一个重要机制。试点有助于在可控范围内降低政策创新和制度变迁的风险,但对试点效果的评估容易受到选择性偏差和霍桑效应的影响。社会实验技术的运用为我们提供了一种改善的可能性,它比自然实验设计和统计控制分析方法对观察数据的处理更能有效地减少各类偏误的影响。本文在讨论社会科学偏误类型和控制机制的基础上,总结了社会实验法的分类、设计原则、实施程序及其在社会科学中的应用。社会实验可以对偶然因素、偏误和混淆因素实现有效控制,为政策评估和因果推理提供可靠依据。当然,社会实验在适用范围和实施方面也面临很多限制和挑战。从实践层面看,试点机制如果能吸收社会实验法的相关要素,可以将政策创新与制度建设建立在更为坚实的基础之上。

关键词:随机控制实验　因果关系　政策评估　偏差　统计控制

一、作为治理机制的试点:优势与局限[①]

近年来,试验或试点作为一种治理机制,成为众多中国研究学者关注的热点议题。来自经济学[②]、政治学[③]等不同学科的研究主要是在政策过程的框架下展开讨论的,并将试验或试点视为政策创新或扩散的有效机制,进而将其视为解释中国渐进式经济改革取得成功或经济获得发展的原因之一。也有一些学者从学习和适应能力的角度,将这一机制归为政体适应能力的具体体现之一。[④] 当然,试验或试点机制并非仅出现在经济政策领域,也出现在诸如医疗卫生[⑤]、社会保

① 一些英文文献将试点翻译为 experiment 是不严谨的,其对应的英文术语应该是 pilot 或 demonstration,因为实验(experiment)往往需要有严格的随机分组,但大多数试点并不具备这个要素。

② G. H. Jefferson and T. G. Rawski, "Enterprise Reform in Chinese Industry", *The Journal of Economic Perspectives*, Vol. 8, No. 2, 1994; B. Naughton, *Growing Out of the Plan: Chinese Economic Reform, 1978 ~ 1993*, New York: Cambridge University Press, 1995.

③ S. Heilmann, "From Local Experiments to National Policy: The Origins of China's Distinctive Policy Process", *The China Journal*, No. 59, 2008; S. Heilmann, "Policy Experimentation in China's Economic Rise", *Studies in Comparative International Development*, Vol. 43, No. 1, 2008; 梅赐琪、汪笑男、廖露、刘志林:《政策试点的特征:基于〈人民日报〉1992 ~ 2003 年试点报道的研究》,《公共行政评论》,2015 年第 3 期。

④ S. Wang, "Adapting by Learning: The Evolution of China's Rural Health Care Financing", *Modern China*, Vol. 35, No. 4, 2009.

⑤ 李丽辉:《中央百亿元支持公立医院改革》,《人民日报》,2015 年 10 月 19 日。

障[①],乃至于行政改革[②]等领域。

也有学者对试验或试点机制的有效性提出了质疑,主要有:第一,从试点过程中不同层级政府间的互动特点来看,有学者认为中央政府在整个过程中的控制性角色过于显著;中央政府通过掌握试点的选择、实施和纠正等权限,从而能够对地方试点形成整体性控制。[③] 在不同类型的政策试验中,只有开展了对比性试验(comparative trial)的试点选择会考虑单位的代表性,并投入较多精力来评估试点的效果。[④] 还有学者指出,试点的选择容易受到领导个人意志支配,一些不具备典型性且缺乏试点条件的领导干部的"联系点"往往容易成为试点。[⑤]

第二,很多来自试点单位内部的问题可能会影响到政策试验的效果,而这些因素与政策本身相关性不大。例如,有学者认为地方政府的态度是影响试点成败的关键因素。[⑥] 还有学者发现,财政转移支付是中央政府对试点单位提供财政激励的有效手段,如果没有来自上级政府的财政支持,那么下级单位执行试点工作的积极性往往容易成问题。[⑦] 另外,某些试点单位因为承担了某一项试点任务之后,可能随之而来的是另外更多的试点任务,被选为试点的地方政府往往缺乏足够的时间、资源和权限来开展政策创新和试行。[⑧] 还有学者指出,下级试点单位官员还有可能将获取试点资格本身视为一种政绩;而单位领导的更替可能会造成试点工作的中断或更替,从而出现试点工作内容"翻烧饼"的状况,层出不穷的政策创新可能只是在间或地重复着之前的方案。[⑨]

第三,从对试点效果的评估过程来看,有学者认为,中央政府往往通过领导人或专家的调研来评估试点的效果,其信息的收集也往往有赖于试点所在地方政府的反馈,这有可能提高地方政府在央地互动过程中的议价权,[⑩]地方政府也可能会通过不同的方式来扭曲信息的传递过程。[⑪] 地方的试点情况到底怎样才算成功,标准掌握在中央手上,中央也有最终权力决定是否在全国推广地方经验,甚至会在地方试点没有取得一定成效之前就在全国推进所试的政策。[⑫]

由此可以看出,对试点机制中承担试点任务的单位开展研究,有利于完善试点机制的理论研究和实践运作。在本文中,我们的讨论也将聚焦于影响试点运行成败的因果机制的识别,以寻

---

① 郑文换:《地方试点与国家政策:以新农保为例》,《中国行政管理》,2013 年第 2 期。

② W. Tsai and N. Dean, "Experimentation under Hierarchy in Local Conditions: Cases of Political Reform in Guangdong and Sichuan, China", *The China Quarterly*, Vol. 218, 2014.

③ 刘培伟:《基于中央选择性控制的试验:中国改革"实践"机制的一种新解释》,《开放时代》,2010 年第 4 期。

④ X. Zhu and H. Zhao, "Experimentalist Governance with Interactive Central-Local Relations: Making New Pension Policies in China", *Policy Studies Journal*, forthcoming.

⑤ 杨雪冬:《制度运行的逻辑》,第 154 页,社会科学文献出版社,2017 年版。

⑥ W. Tsai and N. Dean, "Experimentation under Hierarchy in Local Conditions: Cases of Political Reform in Guangdong and Sichuan, China" *The China Quarterly*, Vol. 218, 2014.

⑦ X. Zhu and H. Zhao, "Experimentalist Governance with Interactive Central-Local Relations: Making New Pension Policies in China". *Policy Studies Journal*, forthcoming.

⑧ C. Mei and Z. Liu, "Experiment-based Policy Making or Conscious Policy Design? The Case of Urban Housing Reform in China", *Policy Sciences*, Vol. 47, No. 3, 2014; 陈那波、蔡荣:《"试点"何以失败? ——A 市生活垃圾"计量收费"政策试行过程研究》,《社会学研究》,2017 年第 2 期。

⑨ 杨雪冬:《制度运行的逻辑》,第 158 页。

⑩ X. Zhu and H. Zhao, "Experimentalist Governance with Interactive Central-Local Relations: Making New Pension Policies in China". *Policy Studies Journal*, forthcoming.

⑪ 周雪光:《基层政府间的"共谋现象":一个政府行为的制度逻辑》,《社会学研究》,2008 年第 6 期。

⑫ C. Mei and Z. Liu, "Experiment-based Policy Making or Conscious Policy Design? The Case of Urban Housing Reform in China". *Policy Sciences*, Vol. 47, No. 3.

求提高试点机制实际运行有效性的可能方向。从社会科学的学理上讲,试点的运行过程可能受到两大效应的影响,从而使得其因果关系的推理受到质疑,这两个效应是选择性偏差(selection bias)和霍桑效应(Hawthorn effects),它们都会使得政策试行的结果被高估或被低估。

具体说来,首先,试点地区的选择和自我选择容易对结果产生影响。如果试点选在了条件比较好的地方,那么政策本身的成功就会被高估(例如在经济发达地区试点就业培训项目);如果试点选择了问题比较严重的地区,那么政策效应则会被低估(例如在困难国企聚集的地区试点养老保险改革)。在允许地方自行申报试点的情况下问题也很严重,那些积极参加试点的地区和其他地区很可能存在系统性的差异,例如积极试点的地区可能问题比较严重,迫切需要做出政策回应。而这种选择和自我选择带来的差异会严重干扰结果,使得政策干预和结果之间的因果关系变得极不可靠(即内部效度)。① 其次,进行试点的地区在资源分配、组织管理方面,往往对试点的项目进行特殊的照顾,甚至有的地区因为是在试点,所以有"只能成功、不能失败"的想法。② 这些干扰会使得试点地区很难成为典型,因为一旦这个项目铺开,其他地区很难获得试点地区类似规格的"待遇"。这也是为什么有些试行政策在试点地区很成功,一旦推广后立刻褪色的原因,即试点带来的霍桑效应削弱了效度。③

那么,有没有一种克服上述问题的方法呢?从20世纪60年代开始兴起的社会实验,作为一种新型的政策评估方式为我们提供了一个可能。到现在为止,无论是发达国家、发展中国家,还是国际组织,都日益将随机控制实验(randomized control trials)作为重要的政策发展和项目评估工具。例如墨西哥政府曾经聘请哈佛大学研究人员对其庞大的医疗保险、营养计划进行设计、评估,④世界银行、世界卫生组织等机构对其扶贫开发项目也日益采用随机控制实验作为评估工具。⑤ 在学术研究中,实验法在社会科学各个学科中(包括政治学⑥、政治经济学⑦、社会学⑧、经济学⑨、管理学⑩等)都呈现复兴之势。无论应用性的政策倡议还是基础性的学术研究,复杂的现实世界对因果关系识别提出了更高的要求,而传统的以观察性数据为基础的方法很难有效地控制干扰变量,这也是实验法重新兴起的主要原因。

与试点机制相比,实验法通过随机分配实验组和控制组的方式,可以有效控制干扰因素或变

① 杨雪冬:《改革试点的走样变形》,《北京日报》,2014年5月12日。

② 有的地区为了凸显政绩,更是不惜代价确保试点成功,这就使得试点本身不再"典型",不再具有可复制性。

③ 陈那波、蔡荣:《"试点"何以失败?——A市生活垃圾"计量收费"政策试行过程研究》,《社会学研究》,2017年第2期。

④ G. King, et al, "Public Policy for the Poor? A Randomised Assessment of the Mexican Universal Health Insurance Programme", *The Lancet*, Vol. 373, No. 9673, 2009.

⑤ 阿比吉特·班纳吉、埃斯特·迪弗洛:《贫穷的本质:我们为什么摆脱不了贫穷》,第13~15页,中信出版社,2013年版。

⑥ R. McDermott, "Experimental Methods in Political Science", *Annual Review of Political Science*, Vol. 5, 2002.

⑦ T. R. Palfrey, "Laboratory Experiments in Political Economy", *Annual Review of Political Science*, Vol. 12, 2009.

⑧ M. Jackson and D. R. Cox, "The Principles of Experimental Design and their Application in Sociology", *Annual Review of Sociology*, Vol. 39, 2013.

⑨ A. V. Banerjee and E. Duflo, "The Experimental Approach to Development Economics", *Annual Review of Economics*, Vol. 1, No. 1, 2009; F. Guala and L. Mittone, "Experiments in Economics: External Validity and the Robustness of Phenomena", *Journal of Economic Methodology*, Vol. 12, No. 4, 2005.

⑩ R. Croson, J. Anand and R. Agarwal, "Using Experiments in Corporate Strategy Research", *European Management Review*, Vol. 4, No. 3, 2007.

量,从而使得出的因果推理具有比较可靠的基础,保证了结论的内在效度。[①] 虽然社会实验往往旷日持久且费用高昂,但不经过实验检验而导致的决策失误代价更高。因此,从长远来看,只要条件允许,科学决策最好将随机控制的实验思路纳入其中,以提高其因果推理的品质。

当下,中国的改革已经进入深水区,试点作为一种检验政策或方案有效性的常用工具,将可能被长期继续使用下去。试点有效性的评估来源于对问题发生过程及因果关系的可靠认识,这种认识又会影响到全国性政策回应的质量高低。因此,如何提高政策试点的有效性成为非常关键的实践问题。本文将系统介绍实验方法在改善因果推理和政策评估的作用,以期能带来更多的讨论和关注。

二、影响因果机制确立的偏误类型与控制机制

为了辨识事物间的因果机制,社会科学研究的基础策略是比较与控制(compare and control):通过比较变异来获得变量间的相关关系(association or correlation),通过控制其他变量来获得相关基础上的因果关系(causal relationship)。[②] 因果关系的确定需要满足三个条件:(1)变量间存在关系;(2)时间上有先后;(3)排除其它竞争性假设。下文将讨论社会科学中影响因果机制确立的偏误类型及现有的控制机制。

(一)社会科学中的偏误类型

当一个变量取值发生变化时,我们会考察其他变量是否也会变化,这称之为共变(co-variation),它是社会科学研究的重要对象,也是统计关系的基础。统计上发现相关是非常容易的,然而从相关到因果的鉴定却极其苛刻。[③] 要精确确定一个因素与另外一个因素具有因果关系,研究者需要分离三类因素:(1)随机扰动(chance/random variation);(2)偏误(bias);(3)混杂因素(confounder)。[④]

随机扰动如一些小概率事件,通常可以随着样本量的增大而减小,因而对结果的影响是可控的。统计检验可以帮助我们判断随机扰动在因果推理中的角色。在统计分析的实践中,如果统计模型的系数通过了某一门槛的统计检验(通常是 P 值小于 0.01 或 0.05),就可以视同排除了

---

① N. Cartwright, "Causal Powers: What are they? Why Do We Need them? What Can Be Done with Them and What Cannot?", London: Center for Philosophy of Natural and Social Science, 2007, p. 58; N. M. Castillo and D. A. Wagner, "Gold Standard? The Use of Randomized Controlled Trials for International Educational Policy", *Comparative Education Review*, Vol. 58, No. 1, 2014.

② F. Bechhofer and L. Paterson, *Principles of Research Design in the Social Sciences*, London; New York: Routledge, 2000, p. 1.

③ 大数据分析兴起后,有学者认为有了大数据,因果关系就不再重要了,找到相关就可以了(参见维克托·迈尔-舍恩伯格、肯尼思·库克耶:《大数据时代:生活、工作与思维的大变革》,第 213 页,浙江人民出版社,2013 年版),但笔者不同意这样的看法。

④ 我们并没有以内生性作为讨论的主题,虽然对内生性的关注已经成为整个社会科学实证研究的核心问题。内生性是一个广泛提及但是定义不那么统一的概念。经典的《社会科学中的研究设计》将内生性视为因果关系方向无法确定,参见 G. King, R. O. Keohane and S. Verba, *Designing Social Inquiry: Scientific Inference in Qualitative Research*, Princeton, N. J.: Princeton University Press, 1994, p. 185;而陈云松则将回归分析中的若干问题统一纳入到内生性的范畴,参见陈云松、范晓光:《社会学定量分析中的内生性问题:测估社会互动的因果效应研究综述》,《社会》,2010 年第 4 期。内生性问题主要有反向因果关系(reverse causality)和遗漏变量(omitted variables),参见 M. Baekgaard, et al, "Conducting Experiments in Public Management Research: A Practical Guide", *International Public Management Journal*, Vol. 18, No. 2, 2015;也有学者认为是反向因果关系和自选择偏误(self-selection bias),参见 J. Blom-Hansen, R. Morton and S. Serritzlew, "Experiments in Public Management Research", *International Public Management Journal*, Vol. 18, No. 2, 2015。本文讨论的各类偏误实际上与内生性有较大重叠部分。

随机扰动的作用。当然,学术界已经兴起对 P 值的反思。[①] 偏误是我们很难准确测量的变异,它通常跟研究设计和资料收集方式有关,很难用统计手段完全控制或分离。科学推理中的偏误是如此之多,因此充分的讨论需要一篇文章甚至一本书的篇幅,[②]这里只能做简要的梳理。对因果关系产生干扰的偏差大致可以分为两类:选择性偏误(selection bias)和信息性偏误(information bias)。混杂偏误有时也被视为一种偏误,但由于它本身的一些特质,一般作单独处理。

1. 选择性偏误。如果我们比较的对象在某些方面存在显著的不同,这些不同会干扰解释变量和结果变量的关系。例如,如果一项减肥计划通过自由报名来征集参与者,积极报名者和其他人群在主动性、自律性、肥胖程度、健康意识方面可能存在显著的差异。如果不做控制,直接比较这两类群体会得到有偏差的结果。这种偏误称为自选择偏误(self-selection bias,也称为 publicity bias)。自选择偏误的另外一个例子是"健康工人效应"或称为"在职者效应"(healthy worker effect):当我们以在职人口作为研究对象时,他们的特征是没法推论到其他群体的,因为在职人口很可能是更健康的,其他群体则因为年龄、工伤、健康等原因离开了劳动力市场。

在社会科学的调查中,还有一种比较常见的偏误是退出偏误(withdraw bias 或者 follow up bias)。这种偏误在大型跟踪性的调查中最为常见。如果有大量的样本丢失,研究者需要判断样本损失的原因是否与关键的解释变量有关系。例如,我们跟踪调查家庭的收入变动状况,如果第二次、第三次跟踪调查发现流动人口的样本损失率远高于整体样本的平均损耗率,那么这个样本就是有偏差的。因为流动人口的工作稳定性更差,他们的大量流失会导致后两次样本跟第一次样本不可比,进而带来结论的偏差。

2. 信息性偏误。这类偏误通常是指因为资料收集的方法不恰当,以至于产生系统性偏差。例如我们在调查居民收入时,如果问卷的问题仅仅围绕工资所得,那么对某些高资产群体会产生低估。因为这些群体的相当一部分收入并不来自于工资,而是来自资产收益所得。信息性偏误通常根源于受访者差异(subject variation)、信息采集者差异(observer variation)、工具缺陷(deficiency of tools)、测量的技术错误(technical errors in measurement)等因素。比较常见的信息性偏误包括回忆偏差(recall bias)和报告偏差(reporting bias)。

回忆偏差是指受访对象受到当前或过去因素的影响而提供了不够准确的信息。例如我们调查受访者对过往婚姻状况的感受,有过离婚经历的往往会放大过去不愉快的记忆,因此会使得对这一类受访者的结果会产生系统偏差。报告偏差与霍桑效应有相似之处,也是一类需要注意的问题。如果受访者因参与调查而改变了自己的行为和回答,那么结果就是有偏差的。例如我们在进行居民消费行为的调查时,受访者因为需要记账而对自己的花销变得更敏感,那么他的消费行为就是有偏差的。我们访谈时如果录音,有些受访对象说话就会更加谨慎和保守。

3. 混杂偏误。现实中的各个因素往往混杂在一起,使得辨识因果关系非常困难。更为致命的是,与关键解释变量混杂在一起的因素还会对因果关系产生误导性的影响。混杂偏误中混淆因子的存在,使得我们观测到的变量间关联并不能代表变量间的真实关联。[③] 例如,当我们考察吸烟(关键自变量,IV)与肺癌发病率(因变量,DV)关系时,年龄就是一个很典型的混淆因子:年老的人跟年轻的人相比,吸烟的比例更大、烟龄更久;同时,年龄越大,各种癌症的发病率越高。

① R. Nuzzo, "Scientific Method: Statistical Errors", *Nature*, Vol. 506, 2014.

② D. L. Sackett, "Bias in Analytic Research", *Journal of Chronic Diseases*, Vol. 32, 1979; D. A. Grimes and K. F. Schulz, "Bias and Causal Associations in Observational Research", *Lancet*, Vol. 359, No. 9302, 2002.

③ S. Greenland and H. Morgenstern, "Confounding in Health Research", *Annual Review of Public Health*, Vol. 22, 2001.

由于年龄与肺癌存在正向关系,如果不控制,很容易高估吸烟对肺癌发病率的影响。一个变量如果是一个混淆因子,需要满足三个条件:(1)这个变量必须在观察对象中有不同的分布,尤其是关键自变量所对应的不同子群体之间(在上述例子中,吸烟和不吸烟的人,平均年龄是不一样的);(2)即使没有自变量,混淆因子对因变量仍然有直接的影响(年龄越大,患癌概率越大);(3)混淆因子并不在自变量和因变量的因果链条上(年龄并不是吸烟—肺癌的中介变量)。三者的关系如图 1 所示。

图 1　混淆因子的典型示例

处理混杂偏误时需要将各个因素的作用予以区分,以辨清它们各自的净效应(partial effect)。通常的处理方式分为两个阶段:在研究设计阶段,研究者可以通过限制、配对和随机分配的方式控制;在数据分析阶段,研究者只能通过配对分析、分层分析和多变量分析的方式予以控制。[①] 所有的处理方式中,最为彻底和可靠的就是随机分配了,也就是本文所讲的随机控制实验。

(二)统计分析阶段的控制方法及其适用性

当资料收集完毕进入分析阶段,就只能通过统计手段去控制和降低偏误对因果识别的影响了。统计分析策略多种多样,限于篇幅我们仅介绍两类:匹配法和自然实验。

1. 匹配法。多元回归中的统计控制和倾向值匹配,是最为常用的统计分析方法,然而这些方法也都存在种种限制和缺陷,无法代替受控实验。

(1)多元回归的控制。在讨论回归分析结果时,我们常常需要加上“其他条件不变”的限定,这就是所谓的统计控制,即通过在回归方程中加入控制变量,努力使得需要比较的对象在各方面尽量一致。统计控制的原理是:在分析一个自变量与因变量的关系时,尽量使因变量在其他自变量涉及的维度相等或相近。例如,当我们分析教育对收入的影响时,需要控制性别、年龄等因素,控制变量所起的作用是:比较学历为本科和高中的人的收入时,我们尽量只比较性别相同、年龄接近的人。

然而回归无法取代受控实验。[②] 这是因为:首先,很多变量没法测量,因此无从控制。这对社会科学而言尤其严重。无法测量的概念只能寻找替代变量。例如,人的聪明程度是很难测量的,智商测验可以作为替代变量,但其效度、信度还有待商榷。因此,最终能纳入控制变量的都是那些比较方便测量、也较少争议的变量。真正需要控制的因素不见得一定会控制得住。其次,测量误差的存在。许多变量即使可以测量,也难以逃避测量误差的存在。社会科学概念的测量问题尤其严重。[③] 测量误差不仅来源于概念本身,指标设计、资料收集方法都会影响测量的精度,进而影响控制的效果。[④] 实际上,回归方程对测量误差的处理不太充分。再次,统计控制的有限

① V. J. Schoenbach and W. D. Rosamond, *Understanding the Fundamentals of Epidemiology: An Evolving Text*, Chapel Hill: North Carolina, 2000, pp. 335 ~ 380.

② P. D. Allison, *Multiple Regression: A Primer*, Thousand Oaks, California: Pine Forge Press, 1998, pp. 16 ~ 19.

③ 赵鼎新:《社会与政治运动理论:框架与反思》,《学海》,2006 年第 2 期。

④ M. Blackwell, J. Honaker and G. King, “A Unified Approach to Measurement Error and Missing Data: Details and Extensions”, *Sociological Methods & Research*, Vol. 46, No. 3, 2017.

性。虽然统计上尽量匹配比较对象,但受限于变量取值的分布和样本量,实际分析中很难做到完全匹配。例如我们很难找到这样两个比较对象:家庭背景、年龄、工作经验、所在行业、智商等变量都一致,仅仅教育程度不一样。所以,控制变量只是名义上在控制,充其量也就是将高学历、低学历等组内的差异变小而已,但组内差异很难消失。最后,统计控制的尺度很难把握。如果我们担心其他变量的影响,希望尽可能地考虑其他因素,因此把大量的变量放入回归方程,这样做会造成模型失去简约性,直接的后果是方差过大,统计系数的显著性会受到影响;而且放进来的变量间的相互关系、测量误差等很可能成为额外的问题,即所谓的过度控制。

(2)倾向值匹配法(propensity score matching)。倾向值匹配法可视为统计控制的升级版,它的主要设计是首先计算控制组的人参与实验的概率(倾向值),然后将控制组和实验组里倾向值相同或相近的样本进行比较。这一方法将统计控制的理念升华为更为精致、系统的方法,使得比较的对象更具有可比性。在社会科学研究中,由于变量观测的诸多限制,倾向值匹配法是较为贴近研究实际的一种重要的因果识别方法,广受社会科学研究者的欢迎。与其他方法相比,倾向值匹配法更适合用于调查研究数据的分析。

倾向值匹配法的缺陷在于:在计算倾向值时,它仅能涵盖那些可观察到的变量;对于无法观测或获得的变量,就不能运用这一方法了。因此它的有效性依赖于一个很强的假设:比较组参与实验的倾向值只受那些观测到的变量影响。但实际研究中,这一假设往往有问题。并且,该方法只在因果关系方向确定的情况下使用,仍然没能突破因果方向假设前提的限制,无法用于甄别变量间可能存在的互为因果的关系。

2. 自然实验设计。自然实验设计是近些年兴起的一组研究方法,主要包括工具变量、断点回归等方法。[①] 这些方法本质上都是回归分析,不过与普通的回归相比,它们在因果推理方面具备一定的优势。但这些方法都有其适用范围,也存在许多假设。离开了这些范围和假设,其因果推理的力量将会大大削弱。

(1)工具变量(instrument variable)。工具变量主要解决内生性问题,即解释变量与误差项的相关问题。[②] 工具变量本质上是二阶最小二乘法。工具变量的选择面临苛刻的条件:其一,工具变量与解释变量的相关性大小。太小的话,那么这个工具变量就会非常弱。其二,工具变量须与因变量无关。在很多情况下这两个变量总会有些微关系。实际上,只有随机分配实验才能完美的符合上述两个条件。因此,随机实验实际上是工具变量的一个特例。[③]

(2)断点回归(regression discontinuity)。断点回归是一个非常精妙的设计,[④]它利用经济、社会、政治中的一些特殊设置形成的分界线来推论因果关系。例如用淮河南北这一供暖政策的分界线来推断空气污染对预期寿命的影响。[⑤] 断点回归的问题在于有意义的断点并不多见,并且

① M. R. Rosenzweig and K. I. Wolpin, "Natural 'Natural Experiments' in Economics", *Journal of Economic Literature*, Vol. 38, No. 4, 2000.

② J. P. Newhouse and M. McClellan, "Econometrics in Outcomes Research: The Use of Instrumental Variables", *Annual Review of Public Health*, Vol. 19, No. 1, 1998。详细的讨论参见陈云松:《逻辑、想象和诠释:工具变量在社会科学因果推断中的应用》,《社会学研究》,2012 年第 6 期。

③ A. Finkelstein, et al, "The Oregon Health Insurance Experiment: Evidence from the First Year", *The Quarterly Journal of Economics*, Vol. 127, No. 3, 2012.

④ D. S. Lee and T. Lemieux, "Regression Discontinuity Designs in Economics", *Journal of Economic Literature*, Vol. 48, No. 2, 2010.

⑤ Y. Chen, A. Ebenstein, M. Greenstone and H. Li, "Evidence On the Impact of Sustained Exposure to Air Pollution On Life Expectancy From China's Huai River Policy", *Proceedings of the National Academy of Sciences*, Vol. 110, No. 32, 2013.

断点回归外推性不好,所得结论可能仅对断点附近的取值有参考价值。

总之,自然实验设计通过巧妙的设计能够克服部分因果推理的难题,但由于缺少真正的随机分配过程,无论是统计控制还是自然实验设计仍然属于观察性研究(observational studies)。观察性研究不管采取什么样的分析手段,最理想的情况也就是能获得与实验研究可比的结果。[①] 因此,因果推理的"金标准"还是随机控制实验。

三、社会实验的分类与设计原则

面对如此众多的偏差和干扰因素,怎样才能识别可靠的因果关系呢?统计学家的解决方案是用随机分配的方法。[②] 按照此逻辑设计的随机控制实验成为生物学、农学等学科的主要研究范式。随机控制实验法(以下简称实验法)可界定为"研究者控制干预并能主动操纵的一种试验"。[③] 而观察性研究中,无论是统计控制还是自然实验,都不是研究者可以实际参与操作的,前者本质上是一种数学方法上的设定与计算;[④]后者是依据某个事件发生后,研究对象是否受到影响来分组得到两组对比数据。[⑤] 因此,有无实验者参与操作的"外生的干预",是实验法与观察性研究最本质的区别标准。了解到这一不同,研究者需要判断何种问题适合使用实验法解决,这需要对实验法有更为具体而细致的认识。接下来,我们对社会实验法进行分类介绍。

(一)实验法的分类

一些学者将实验法分为以下四类:田野实验(field experiment)、实验室实验(lab experiment)、调查实验(survey experiment)、自然实验(natural experiment)。[⑥] 其中,尽管自然实验获得了日益广泛的应用,但其"外生的干预"是自然发生的,实验者无法参与操作,[⑦]不能将其归到真正意义上的实验法的范畴。下文将讨论前三种实验类型。

1. 田野实验。田野实验是在真实的世界中开展的实验,可控程度较低。田野实验的要素有:(1)研究的主题能在自然条件下进行;(2)受试者对实验任务的信息接收是在自然状态下;(3)外生干预都是在自然条件下发生的;(4)研究的环境也都是自然的。[⑧] 田野实验的每个环节均能反映出自然状态下的真实性,因此其外部效度比较高。[⑨] 然而,田野实验的真实性也带来了

---

① T. D. Cook, W. R. Shadish and V. C. Wong, "Three Conditions under Which Experiments and Observational Studies Produce Comparable Causal Estimates: New Findings from Within-Study Comparisons", *Journal of Policy Analysis and Management*, Vol. 27, No. 4, 2008.

② 萨尔斯伯格:《女士品茶:20 世纪统计怎样变革了科学》,第 47 页,中国统计出版社,2004 年版。

③ Rebecca B. Morton and Kenneth C. Williams, "Experimental Political Science and the Study of Causality: From Nature to the Lab", New York: *Cambridge University Press*, 2010.

④ A. Lijphart, "Comparative Politics and the Comparative Method", *American Political Science Review*, Vol. 65, No. 3, 1971.

⑤ J. S. Sekhon and R. Titiunik, "When Natural Experiments are Neither Natural Nor Experiments", *American Political Science Review*, Vol. 106, No. 1, 2012.

⑥ R. Bouwman and S. Grimmelikhuijsen, "Experimental Public Administration From 1992 to 2014: A Systematic Literature Review and Ways Forward", *International Journal of Public Sector Management*, Vol. 29, No. 2, 2016; G. W. Harrison and J. A. List, "Field Experiments", *Journal of Economic Literature*, Vol. 42, No. 4, 2004.

⑦ T. Dunning, *Natural Experiments in the Social Sciences: A Design-Based Approach*, New York: Cambridge University Press, 2012.

⑧ G. W. Harrison and J. A. List, "Field Experiments", *Journal of Economic Literature*, Vol. 42, No. 4, 2004.

⑨ 实验法中的外部效度测量的是研究结论的普遍适用性,内部效度测量的是研究结论的确定性。参见 R. Mcdermott, "Experimental Methods in Political Science", *Annual Review of Political Science*, Vol. 5, No. 1, 2002; D. M. Dimitrov and P. D. Rumrill Jr, "Pretest-Posttest Designs and Measurement of Change", *Work*, Vol. 20, No. 2, 2003.

一些局限性:(1)不少重要的研究问题很难通过田野实验开展;[①](2)研究者的主动性受到一定程度的制约;(3)田野研究的实施成本较高。

鉴于田野实验在外部效度测量上的优势,它在实际应用中备受推崇。加里·金(Gary King)等学者受墨西哥政府委托在墨西哥 13 个州实施的全民医保(SPS)实验是迄今为止最大的政策领域的实验之一,[②]政府的委托和配合使研究者能够克服来自政治和社会方面的制约。国内的试点从政策制定到试点选取,乃至最终的验收,也通常遇到政治或社会压力,如何减少这些外在压力对试点实施环境的"污染",使政策评估的结论具备较好的外推性?后文我们将进行讨论。

2. 实验室实验。实验室实验是在可控环境中开展的实验。实验环境的高度可控性使其具备如下优势:(1)研究者较容易将其他影响实验结果的因素排除在外;(2)研究者可以开展现实中无法进行的研究;(3)研究者可以自主地创造有利的实验机会。尽管如此,实验室实验仍存在一些局限性:(1)实验样本数量较少;(2)实验场景的非真实性。

自然科学研究中,实验室实验对研究对象样本的选择分组过程是一种较为理想的随机分配。然而在社会科学领域,研究者面临的情形则复杂得多,如何尽可能做到近似随机分组,降低选择性偏误,将是首要挑战。同样,就本文所探讨的试点来说,降低试点样本的选择性偏误也极其重要,可作为试点实施前的首要任务,我们将在后文探讨这一选取过程。

3. 调查实验。调查实验是借助调查问卷开展的一类实验研究。调查实验的优势在于:(1)相对成本较低且易于实施;(2)支持的样本量较大;(3)实验者可以控制实验干预的实施。当然,调查实验也存在一些局限性:(1)问卷回收率不高;(2)实验干预较弱。

调查实验受到许多研究者的青睐,有学者进行过统计,在几种主要的实验法中,调查实验使用的频率最高。[③] 近年来,得益于网络平台收集实验数据的便利性,[④]调查实验的实施成本大大降低。国外已有很多基于调查实验法的研究,尤其是依托耶鲁大学的 Breadboard 实验平台开展的大量研究。[⑤] 国内也有学者开始了这一类的研究。[⑥]

---

① J. Blom-Hansen, R. Morton and S. Serritzlew, "Experiments in Public Management Research", *International Public Management Journal*, Vol. 18, No. 2, 2015.

② G. King, et al, "A 'Politically Robust' Experimental Design for Public Policy Evaluation, with Application to the Mexican Universal Health Insurance Program", *Journal of Policy Analysis and Management*, Vol. 26, No. 3, 2007.

③ R. Bouwman and S. Grimmelikhuijsen, "Experimental Public Administration from 1992 to 2014: A Systematic Literature Review and Ways Forward", *International Journal of Public Sector Management*, Vol. 29, No. 2, 2016.

④ Y. Chen, F. M. Harper, J. Konstan and S. X. Li, "Social Comparisons and Contributions to Online Communities: A Field Experiment on Movielens", *The American Economic Review*, Vol. 100, No. 4, 2010; A. Moseley and G. Stoker, "Putting Public Policy Defaults to the Test: The Case of Organ Donor Registration", *International Public Management Journal*, Vol. 18, No. 2, 2015; H. Z. Margetts, "Experiments for Public Management Research", *Public Management Review*, Vol. 13, No. 2, 2011; J. D. Marvel, "Unconscious Bias in Citizens' Evaluations of Public Sector Performance", *Journal of Public Administration Research and Theory*, Vol. 26, No. 1, 2015.

⑤ A. Nishi, H. Shirado, D. G. Rand and N. A. Christakis, "Inequality and Visibility of Wealth in Experimental Social Networks", *Nature*, Vol. 526, No. 7573, 2015; A. Nishi, H. Shirado and N. A. Christakis, "Intermediate Levels of Network Fluidity Amplify Economic Growth and Mitigate Economic Inequality in Experimental Social Networks", *Sociological Science*, Vol. 2, No. 26, 2015; H. Shirado, F. Fu, J. H. Fowler and N. A. Christakis, "Quality Versus Quantity of Social Ties in Experimental Cooperative Networks", *Nature Communications*, Vol. 4, 2013; D. G. Rand, M. A. Nowak, J. H. Fowler and N. A. Christakis, "Static Network Structure Can Stabilize Human Cooperation", *Proceedings of the National Academy of Sciences*, Vol. 111, No. 48, 2014.

⑥ T. Meng, J. Pan and P. Yang, "Conditional Receptivity to Citizen Participation: Evidence From a Survey Experiment in China", *Comparative Political Studies*, Vol. 50, No. 4, 2017.

(二)实验法的理论原理

实验法是一种研究者与研究对象主动对话的研究方法,卡文迪许实验室就鼓励研究者"自己去做",所以实验法也通常被认为是一种人为色彩浓厚的方法。[①] 从早期的物理学家所进行的实验探索,到逐渐蔓延到心理学、管理学、经济学的诸多著名实验,[②]通常都是研究者借助实践中的人为干预将偶然的、次要的因素分离,从而推断出确定的因果作用路径。以下从人为干预逻辑、因果作用路径两方面对实验法理论原理进行讨论:

1. 实验法的人为干预逻辑。如果说观察性研究是在被动地记录和分析经验事实,那么实验法能够让研究者积极主动干预研究对象的发展进程。对实验法原理的把握,关键就在于理解这种人为干预逻辑:对研究对象实施这样的干预依据何在?一方面,通过人为干预,可以得到自然状态下无法形成的环境和条件,从而认识到实验对象在自然情况下无法观测到的一些特征,推断出确定的规律和结论;另一方面,通过人为干预甚至可以模拟自然状态和环境,为一些不可再现和无法直接观测研究对象的研究提供了可行性。

更重要的是,通过观察性研究得出的因果关系尽管是真实可靠的,但只能解释自变量 X 是否能影响因变量 Y,并不能确定地得出 X 是否总能影响到 Y。[③] 穷尽并建立所有 X 对 Y 的假设,并进行一一验证是根本无法做到的,通常是通过研究方法的设计,将研究者所关注的因果联系分离出来,[④]借助统计控制虽然可以近似实现这一点,却无法像实验法通过引入人为干预得到确定的 X 是否总能影响到 Y 这样的结论。

2. 实验法的因果作用路径。实验法的设计包含了一系列的操作性环节:实验对象的分配、"前测"、引入干预、"后测"。这些环节构成了一条因果作用路径,这条因果作用路径是这样勾连起来的:首先,借助随机分配,得到同质的对照组和实验组,并对两组对象实施"前测";接下来,实施干预并进行"后测";最后,通过比较前后测量的差异,得出实验结果。由于干预前的两组对象同质,那么实施干预之后,两组测量结果之间的差异,就是实验干预所导致的。

更进一步,实际研究中复杂因果作用路径中的调节变量(moderator)和中介变量(mediator)的识别,同样可以使用实验法进行解决:针对这类变量提出研究假设,然后以这些假设作为干预因素,从而识别因果作用路径中的调节效应和中介效应是否存在,这种对因果关系作用路径的识别,在研究中比仅仅只是确定因果关系重要得多。[⑤]

(三)实验法的设计原则

好的实验设计致力于尽可能地降低系统偏误,并提高实验结果的可靠性。[⑥] 实验法的设计原则也应立足于这两点:

1. 降低系统偏误。在实验研究设计中,研究者对实验组和控制组的随机分配、对实验干预

---

① G. Burtless, "The Case for Randomized Field Trials in Economic and Policy Research", *Journal of Economic Perspectives*, Vol. 9, No. 2, 1995.

② 近现代社会科学领域比较有代表性的实验有:心理学的人差实验和反应实验,管理学的红珠实验和破窗实验,经济学的风洞实验等,参见刘晓君:《走进实验的殿堂》,第 93 ~ 101 页,上海交通大学出版社,2006 年版。

③ D. G. Mook, "In Defense of External Invalidity", *American Psychologist*, Vol. 38, No. 4, 1983.

④ 彼得·伽里森:《实验是如何终结的》,第 3 ~ 6 页,上海交通大学出版社,2017 年版。

⑤ A. G. Pirlott and D. P. Mackinnon, "Design Approaches to Experimental Mediation", *Journal of Experimental Social Psychology*, Vol. 66, 2016.

⑥ R. A. Fisher, *The Design of Experiments*, Edinburgh: Oliver and Boyd, 1937, pp. 11 ~ 29.

的完全控制常常被认为是科学的“金标准”。[①] 实现随机分配意味着将研究对象以相同概率分配到对照组和实验组,各种因素作用下所产生的组间差异将互相抵消,基本消除了系统偏误。即便对于研究者控制程度较低的实验,也有必要在操作中尽可能做到近似的随机分配,从而使得系统偏误最小化。

2. 提高结果可靠性。一方面,实验者可以设计多次“后测”,以便对测量的结果进行反复的验证,还可以加入重复实验的设计,进行更合理的结果估计;另一方面,实验者可以对结果进行分组比较,分析和比对原因;还有一种更稳妥的做法就是进行多重实验比较,将多个研究问题同时付诸实验,这样能推导出更为可靠的结论。

总而言之,与观察性研究相比,实验设计从源头上大大减少了“内生性”问题和各种偏误的产生。而且,实验过程中的“前测”和“后测”之间存在明确的时间先后关系,通过干预前后的差异比较可以得出确切的因果关系,排除了互为因果的可能。此外,作为实验刺激的“干预”,是完全外生的变量,不存在受到因变量的影响的情况,从而可以杜绝由于反向因果关系引起的内生性问题。[②]

四、实施中的社会实验

尽管存在上述优势,与观察性研究相比,社会实验在实施中还存在若干问题,甚至有的研究并不适合采用实验方法。另外,部分社会实验的实施过程复杂且成本较高。下面我们进一步讨论这些问题。

(一)社会实验的实施

1. 实验的动员与组织。实验室实验通常由专门的实验室来组织实施,并大多使用学生样本;田野实验发生在真实世界,需要实验者具备强有力的组织动员能力和丰富的社会资本;相较之下,调查实验所需的主要工具比如问卷较容易得到,且组织实施的门槛也较低。

2. 实验对象的随机分配。实验室实验的随机分配通常利用随机数对现场实验对象进行指派分组;田野实验的随机分配是研究者通过实验的实施机构进行,随机分配实施过程依赖于相应机构人员的配合;调查实验的问卷回收往往是不完整的,这对随机分配原则是一大挑战,实验者尤其要注意问卷回收率的问题。[③]

3. 资料的收集与分析。实验室实验的资料和数据收集比较直接完整,也相对容易;田野实验的资料和数据收集涉及到的因素众多,这就需要尽可能在研究开展之初、实验干预实施之前,收集到完备的“基线”资料和数据,同时尽可能打通各个环节的授权过程,以便研究者能够接触到足够的资料和信息;[④]调查实验过程中由于研究对象的态度和重视程度不一,其资料和数据可能存在缺失,研究者可以借助在线交互实验平台、激发实验对象的参与积极性来进行改善。

(二)适用范围

对社会实验方法而言,如下问题或情境并不适合使用:其一,存在伦理问题的情况。如果有的研究问题涉及到人身损害等内容,是没法进行随机分配的。其二,大规模系统方面的问题很难

---

①② J. Blom-Hansen, R. Morton and S. Serritzlew, “Experiments in Public Management Research”, *International Public Management Journal*, Vol. 18, No. 2, 2015.

③④ M. Baekgaard, et al, “Conducting Experiments in Public Management Research: A Practical Guide”, *International Public Management Journal*, Vol. 18, No. 2, 2015.

通过实验法去验证,一个折衷的办法是化整为零,将针对大体量的组织或系统的研究问题转化为对中小群体的研究。其三,实验法适合去评估影响和判定因果关系,但它对一个全新的、缺少积累的议题则用处有限。例如我们从头开始设计一项新的政策,这时比较恰当的方式是通过一个小的试点开始积累政策创新和制度设计的经验,在试点开展一段时间后,再运用实验法对政策或制度方案的实施效果进行评估,从而准确判断既有方案的优缺点,为下一步全国性方案的出台奠定基础。

(三)影响因素

一项社会实验设计的优劣,是通过实验结果的效度反映出来的。社会实验实施中的一些因素可能会威胁到其效度。

1."污染"的问题。在社会实验中,由于实验组和控制组是较难做到完全双盲的,因此组间"污染"的现象很难杜绝。例如,参加技能培训的实验组可能会将一些技能传授给认识的控制组成员等,类似情形在田野实验中较为常见。如果很难通过双盲设计来消除,可以考虑用单盲设计来解决,这样做的好处是无需欺骗和控制现场,同时对照组和实验组之间也不会相互影响。①

2. 遵从度的问题。即如果研究对象没有严格遵守实验准则,结果会受到干扰,主要表现为以下两种情况:第一,无法遵守随机分配原则。这个问题多因客观条件限制所致,解决办法是尽可能做到近似随机分组,并对实验结果采取多组比较来提高可靠性。第二,由于研究对象主观因素导致违背实验准则的情形,比如临床医学研究中常见的研究对象因为预料或相信治疗有效而产生的安慰剂效应。应对这种情况的方法除上面提到的双盲或单盲设计外,还可以通过重复实验来提高实验结果的效度。

3. 外部效度问题。与观察性研究相比,实验者的"参与操作"使实验法的外部效度饱受争议:一是实验环境往往与真实的环境不完全一致,尤其是实验室实验,无法有效解释真实世界的一般性社会过程;二是实验对象往往非随机抽样所得,比较突出的是田野实验,所以其代表性也存在缺陷,研究的结果应用到真实场景时会受到质疑。提升外部效度可考虑如下几点:其一,在新的条件下进行复制性实验,比较因变量与自变量之间关系有无改变;②其二,基于已有理论,在因变量与自变量之间构建由一组变量来调节的关系链条,③这种从理论中挖掘出的关系链,使得实验条件具有更广泛的代表性,从而提升实验的外部效度;其三,使实验环境尽可能模拟和接近真实世界,比如让实验参与者佩戴可穿戴传感设备,在虚拟现实(VR)场景中实现沉浸式参与,接受外生刺激,并在接近自然的状态下做出接近真实的反应,一些学者已经将虚拟现实技术用于风险偏好的测量研究中。④

(四)实施成本

总体而言,实验法尤其是田野实验实施的成本较高,⑤除了实验的动员和组织成本外,甚至

---

① 所谓"单盲"是指对于随机分配过程及干预的实施,研究者清楚但实验对象不清楚。采取单盲设计,由于实验对象不了解这些随机分组和干预信息,相互之间就不会产生影响,从而避免了对实验结果的干扰。

②③ F. M. Garcia and L. Wantchekon, "Theory, External Validity, and Experimental Inference: Some Conjectures", *The ANNALS of the American Academy of Political and Social Science*, Vol. 628, No. 1, 2010.

④ V. Dixit, G. W. Harrison and E. E. Rutström, "Estimating the Subjective Risks of Driving Simulator Accidents", *Accident Analysis & Prevention*, Vol. 62, 2014; S. M. Fiore, G. W. Harrison, C. E. Hughes and E. E. Rutström, "Virtual Experiments and Environmental Policy", *Journal of Environmental Economics and Management*, Vol. 57, No. 1, 2009.

⑤ 目前社会科学中盛行的跟踪性调查(longitudinal survey)同样耗时耗力。实验法实际上也属于此类调查,只不过它需要对调研对象进行随机分配。在同样花费巨大的情况下,实验法对因果推理的作用更大一些,不过,它的涵盖面可能要小一点,而普通的跟踪调查往往包罗万象,可以容纳许多变量。

还存在一些意料之外的善后成本。[①] 周期较短的实验成本相对可控,而对周期长的实验而言,改善的对策是加强对行为参数类型实验的研究:通过关注更为基础的人类行为,可以降低政策周期带来的资源浪费。行为参数是所有政策设计的基础,这种实验设计以基础性数据的方式为政策服务。[②] 例如,美国兰德公司耗时八年(1974~1982)完成了一项健康保险的实验,如此长的周期使得这一实验无法回应短期的政策需求,但它对不同健康保险设计下参保者行为参数的研究,例如医疗服务的需求弹性、不同自付比例下参保者的服务使用率等,为后来政策设计提供了重要的参考和学理支持。[③]

五、社会实验法对完善中国的试点机制的启示

本节内容主要集中于回应第一部分提出的试点机制存在问题的解决之道。社会实验方法通过"前测"和"后测"的差异比较,在实践层面可以较为客观地评估政策实施的效果。当然,中国的政策试验的每个阶段,从政策目标的设定,到选择模范试验,再到界定推广的政策选项,"由点到面"的过程一直都是一个充分政治化的过程,这个过程可能会包含了利益竞争、意识形态的分歧、个人间竞争、策略的投机主义或者特定的政策妥协等。[④] 但是,社会实验法的日益广泛应用,为我们完善试点机制的运作,特别是应对前文总结的试点机制存在的三个方面的问题提供了一个方向。

(一)增加试点单位选取的代表性

参照社会实验中有控制的比较的原则,在试点单位的选取阶段,要以试点单位的代表性作为主要选择标准。可以设立由接受社会实验方法训练的专家或工作人员,以及所需试点的政策领域的工作人员组成的试点单位遴选工作小组。该小组根据所试政策的目标,梳理可能影响实施效果的因素,然后根据这些因素对所有单位划分类型,并从每种类型中分别选取一定数量的单位作为试点。针对中国地区间发展差异巨大、各地实际情况复杂多样等特点,要在不同类型的地区分别选取一定数量的试点单位参与试点工作,以降低选择性偏误对试点效果的影响。当然,所选单位要具备开展试点工作的基础条件,以避免本单位对开展试点工作产生"畏难"情绪。

对前文提到的对比性试验,在条件允许的情况下,应尽量选择数量更多的单位同时开展试点工作,以便进行多角度的对比分析。对于其它类型的政策试验,其试点的选取要尽量减少政治化程度,降低上级政府领导人的个人意志对试点选取工作的影响,类似领导人的"联系点"等不应成为试点单位的优先选择。另外,不应以下级单位自行申请作为选取试点的方式,以避免出现为了获取试点工作的配套资源为目的的申请。针对一个单位同时承担多项试点任务的情况,要建立试点单位登记备案制度,在选取试点单位时要事先查询登记备案数据库,同一单位不得同时承担多项试点工作。选择基层一级作为试点时,除非确有必要且该单位有充足的资源,否则,尽量

① B. Kisida, J. P. Greene and D. H. Bowen, "Creating Cultural Consumers: The Dynamics of Cultural Capital Acquisition", *Sociology of Education*, Vol. 87, No. 4, 2014.

② E. Floyd and J. A. List, "Using Field Experiments in Accounting and Finance", *Journal of Accounting Research*, Vol. 54, No. 2, 2016.

③ W. G. Manning, J. P. Newhouse, N. Duan, E. B. Keeler and A. Leibowitz, "Health Insurance and the Demand for Medical Care: Evidence from a Randomized Experiment", *The American Economic Review*, Vol. 77, No. 3, 1987.

④ S. Heilmann, "From Local Experiments to National Policy: The Origins of China's Distinctive Policy Process", *The China Journal*, No. 59, 2008.

避免将那些已承担且尚未完成试点工作的单位再次选为试点。

(二)减少试点单位内生因素对试点效果的影响

我们应借鉴社会实验方法中防止“污染”因素的做法,努力营造一个较为“纯净”的试点实施环境。这样一来,政策试行效果就可以归因于这个外生的政策的执行,而非来自试点单位内部的其它因素。假如这个试点工作实施效果明显,那么得出政策有效的结论将更为可靠,其外推性和延续性就更能经受住考验。

营造“纯净”试点环境的方向包括:首先,建立合理的干部激励机制,使得下级政府的官员不以获得试点资格为政绩,而以完成试点工作为目标,同时,建立健全试点过程中的容错机制,不以试点工作是否成功作为评判官员政绩的标准,鼓励试点单位的工作人员大胆创新;其次,上级政府要尽可能向所有被选定为试点的单位提供平等的资源,特别是财政激励,这一点对那些非经济领域的政策试点工作尤为重要,这样有助于避免承接试点的单位因资源缺乏而导致的试点工作失败;再次,建立试点工作的责任制和交接机制,避免因负责试点工作的人员的岗位变动导致的中断或废止;最后,建立试点工作内容登记备案制度,详细记录试点单位已开展的政策试验,在新授权该单位开展试点工作时进行比对,避免在同一单位的不同时期对相同或相近的政策方案重复开展试点工作的现象。

(三)注重引入多方评审主体和机制

试点既然是“试”,那就已经预设了既可能成功,也可能失败。但是,在既有的试点实践中,即便做到了之前提出的选取具有代表性的试点单位,减少了试点单位内生因素对试点效果的影响,也可能会在效果评估和经验推广阶段出现不符合“科学”规律的现象。假如将存在效度问题的试点经验推广实施,有可能会造成无法估量的损失,从而影响政府的公信力。为了避免出现这种状况,我们有必要借助一套严谨有效的机制,审慎地评估和验证试点结果的外推性。

这些可能的机制包括:首先,借鉴欧盟试验主义治理的做法,采用同行评审机制来评估某个试点单位的试点效果,[①]这里的同行包括但不仅限于开展同类试点工作的单位的工作人员。评审时,在沿用以往实地或现场评估方式的同时,也要引入对试点绩效报告文本的双向匿名评审,以减少领导人的个人意志对试点效果评估的影响。其次,还应纳入更多的评审主体,如第三方研究机构、试点单位辖区的民众等;借鉴“开门立法”的形式,借助网络平台建立民众,特别是所试政策的受众参与评估的机制,减少试点评估过程中上级政府在信息方面对下级政府的依赖和上下级间的信息不对称,从而获得更为客观的评估结果。第三,不仅要注重试点成功经验的总结和推广,也要注重对试点失败原因的分析,这不仅有利于后续政策的修订和改进,也是实践容错机制和激励试点失败单位的工作人员的一种方式。

作者:刘军强,中山大学中国公共管理研究中心、政治与公共事务管理学院(广州市,510275)

胡国鹏,广东外语外贸大学社会与公共管理学院(广州市,510006)、中山大学政治与公共事务管理学院(广州市,510006)

李振,山东大学政治学与公共管理学院(山东省青岛市,266237)

(责任编辑:孟令梅)

① C. F. Sabel and J. Zeitlin, “Learning from Difference: The New Architecture of Experimentalist Governance in the EU”, *European Law Journal*, Vol. 14, No. 3, 2008.