

基于结果链的影响评估及其实验方法

曹堂哲*

【摘要】在结果链中，影响评估重点关注干预与净影响的关系，目的是将净影响从总影响中分离，并证明干预与净效果之间存在因果关系。公共管理领域的影响评估几乎涉及了公共管理研究的所有议题。从方法论的角度来看，将这些议题的事前、事后影响评估建立在实验方法的基础上，才可能真正推进影响评估的进一步发展。论文以唐纳德·坎贝尔与朱利安·斯坦利的著作《实验与准实验研究》为框架，阐明公共管理影响评估的实验方法体系，并结合相关领域的成果分析了实验方法在公共管理事前和干预后影响评估中的运用。

【关键词】结果链 影响评估 实验法

【中图分类号】D63

【文献标识码】A

【文章编号】1674-2486 (2018) 01-0108-16

一、引言

影响评估重点关注干预与净影响的关系，目的是将净影响从总影响中分离，并证明干预与净效果之间存在因果关系。凡是涉及社会干预和社会变迁的地方就可能开展影响评估。20世纪70年代以来，规制影响评估、环境影响评估和社会影响评估成为了公共管理影响评估研究和实践的重要领域。规制影响评估根据基本的管制目标，追求与确认所有可以达成这些目标的政策干预，利用一致的分析方法，例如成本效益分析法，来检验与测量新的或现行管制的可能效益、成本与效果，有系统地选择最有效果与效率的方案，其对决策质量最重要的贡献不是精确的计算，而是分析的行动，探究、了解真实的影响与探查假设（OECD，2011）。规制影响评估包含了诸如立法评估、行政规范性文件评估、政策以及其他政府制度和行动的影响评估。目前经济合作与发展组织（OECD）国

* 曹堂哲，中央财经大学政府管理学院，副教授。感谢匿名评审人的意见。

基金项目：北京市哲学社会科学基金一般项目“基于整体性治理的北京市预算绩效管理
制度执行能力研究”（16GLB033）、中央财经大学研究生优质课程社会科学研究方法（公共
管理研究方法）阶段性成果、国家社科基金重点项目“公共政策与改革协同研究”
（14AZD016）。

家、欧盟已经将重要战略、政策、工作计划、提案、项目、国际协议等纳入了规制影响评估的范畴 (Allio, 2007)。环境影响评估和社会影响评估是影响评估的成熟领域, 拥有较为成熟的对象、技术、方法和职业队伍, 已经发展成为一个专门的学术、政策和职业领域。

目前, 世界银行、西班牙的世界银行影响评价信托基金、国际粮食政策研究所 (International Food Policy Research Institute, IFPRI)、国际影响评估倡导组织 (International Initiative for Impact Evaluation, 3IE)、国际影响评估协会 (International Association of Impact Assessment, IAIA)、麻省理工的贾米勒减贫行动实验室 (The Abdul Latif Jameel Poverty Action Lab, J-PAL) 等机构在影响评估领域取得了有益的成果。国内学者在环境、大气治理、教育、医疗等领域也开始了影响评估的研究, 但实验方法的应用还非常有限。

影响评估的方法包括定性和定量两类。基于逻辑模型的评价、参与式的评价以及焦点小组访谈法是一些学者 (Hatry, 2007; Patton, 2011) 常用的定性评估方法。定量方法可以分为非实验方法和实验方法两类。非实验方法包括前后对比法、有无对比法、横截面回归法、匹配法、倍差法 (Difference-in-Differences, DID)、断点回归设计法 (Regression Discontinuity Design, RDD) 和工具变量法 (Instrumental Variables, IV) 等, 这些方法采用观察性数据, 依靠对 (参与项目的) 选择过程的不同假设作评估 (亚洲开发银行, 2011; Khandker et al., 2010; Gertler et al., 2011)。由于公共管理的研究传统, 以及很少对博士生进行实验方法的训练等原因, 实验研究方法虽然在自然科学和社会科学中的运用已经发育得相对成熟, 但在公共管理领域中的影响评估研究中, 仅占据很少的分量, 亟待强化公共管理领域实验方法的应用, 以增强影响评估的科学性。

二、结果、结果链与影响

公共管理影响评估与结果链紧密相关。结果 (Results) 是一个与过程相对应的概念, 包括产出 (Output)、成果 (Outcome) 和影响 (Impact)。产出是政府政策、项目、行动等干预形成的直接产品和服务。成果是产出传递 (Deliver) 的短期和中期利益 (Benefits)。影响是一个在技术上含义更加狭窄的概念, 指项目的净效果 (Net Effects)。这些环节环环相扣, 形成结果链, 结果链的每一个环节对应特定层次的目标, 构成目标层级。在理论上学者们对结果 (Results)、结果导向 (Results-Based) 等概念并没有达成共识。有些学者认为结果包括了产出、成果和影响三部分; 有些学者认为成果包括产出和影响两部分。国内亦存在结果 (Results) 与成果 (Outcome) 翻译混用的情况。表面上是名词术语的混

淆，实质上反映了人们对影响评估属性和方法论认知的混乱，进而影响到人们对结果导向管理模式认知的误区。

（一）产出、成果和影响

人们很容易混淆产出与成果，区别两者的简单方式是某项指标是否描述了项目的有效性（Effectiveness），如果没有描述有效性则是产出，如果描述了有效性则属于成果。项目的有效性包括双重含义，一是指项目的归因问题，即项目是否引发了或产生了该观察的结果？另一层含义与实际结果相联系：观察结果是否与预期目标相符合（麦克戴维、霍索恩，2011：16）。总之，“成果衡量”指的是对项目行动的成果对照其目标进行评估（邱霏恩，2004）。

成果与影响在时间范围、效应范围、评估方法等方面存在一定的差异。首先，影响涉及的时间往往比成果涉及的时间更长，就拿一个培训评估项目而言，过程评估可能确认“已经完成了培训的人次”。成果评估则告诉人们“多少人由于培训获得新技能找到了工作”。影响评估则可能关注“培训项目参与者一生的改变，甚至隔代的改变”。其次，成果评估是描述性的，侧重产出与对利益相关者目标的满足程度，并不重点关注如何将干预的净效应分离出来。影响评估则是因果分析式的，目的是将干预的净效应从总效应中分离出来，发现干预与净效应之间的因果关系。因此从方法论上来讲，影响评估通常需要使用独立的鉴别器、建立控制组、测量一个长期的改变。因此影响评估可能需要花费大量的成本，并且一个可信赖的结果往往需要花费几年时间。尽管成果评估和影响评估存在诸多的差异，但是在效应的时间维度上，两者有时候很难截然划出一条明确的界限，在进行影响评估的时候往往也涉及到结果评估的内容。图1是一个结果链的例子：

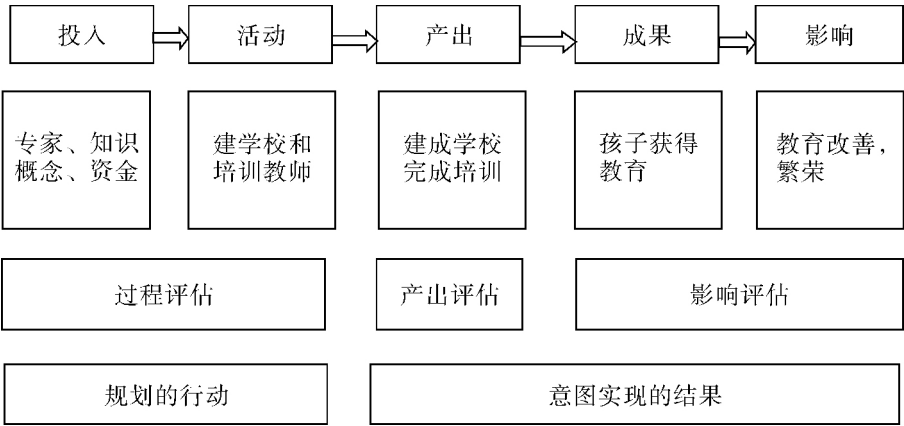


图1 儿童教育的结果链

资料来源：根据（Zewo Foundation，2011）翻译。

（二）结果导向以及结果导向的管理

结果导向（Results - Based）又翻译为基于结果，可以用来修饰绩效监测、绩效评估、管理、预算、审计、领导力等词语，其基本含义是以结果为基础（决策重点、焦点、引领、依据）来设计某一体系。由于结果本身是一个涵盖了产出、成果和影响的概念，在使用结果导向这一术语的时候，有时候需要严格地区分基于产出（Output - Based）和基于成果（Outcome - Based）。如果混淆使用，会导致预算、监测、管理中的理念和实践误区。

第一，就预算模式的演变而言，基于产出的预算与基于成果的预算具有根本性的差异。传统预算以投入为决策重点，按照科目体系对投入进行分解，基本职能取向是控制。绩效预算是以产出作为预算决策的焦点，基本职能取向是管理。新绩效预算以成果作为预算决策的焦点，基本职能是计划和管理（马骏、赵早早，2011）。关注产出仅仅说明政府完成了什么工作，并不能说明该项工作的目标本身是否合理。关注产出与关注投入是一个硬币的两面，形成了以过程控制为核心政府管理模式，传统的分项列支预算和绩效预算都是过程控制型政府管理的预算工具。关注成果则意味着更加关注产出与预期目标的关系。在新公共管理看来，为了实现特定的成果，可以引入竞争性、分权化、契约化的管理手段，公共服务提供的多元主体可以在实现结果的前提下，进行自主化和灵活化的管理。换言之，新绩效预算与新公共管理倡导的分权化、执行机构化、契约化、灵活化改革相适应，为其提供了预算工具。

第二，就监测与评估系统而言，基于结果并不等于无视产出。一般而言，监测分为实施监测与结果监测。对于追踪结果而言，两者都是很重要的。实施监测包括投入、活动和产出三个环节，结果监测则包括成效和目标（影响）两个环节。结果导向的监测由上述五个环节共同构成（库赛克、瑞斯特，2011）。结果导向的评估涉及相关性、效率、影响和持续性等方面，对效率的考察本身就同时涉及投入和产出，故结果导向的评估系统基于结果，以结果为焦点，但是也同时关注结果生成的过程。“在建立以结果为导向的监测与评价体系的每个环节都应制定指标，换句话说，在投入、活动、产出、成效和目标等所有环节都需要指标来监测实际的进展情况”（库赛克、瑞斯特，2011：72）。

第三，结果导向的公共管理区别于基于产出和过程的公共管理。结果导向作为新公共管理的基本原则，已经成为20世纪80年代以来西方国家政府改革的基本趋势和主要内容。与传统的产出和过程导向公共管理模式相对，结果导向的公共管理在结果设定、结果测量、结果实现、结果责任等方面塑造了政府

治理的全新形态。(1) 在结果设定上不仅仅关注产出，更关注由利益相关的参与形成目标的实现程度和产出对利益相关者的影响。(2) 在结果测量上，坚持经济、效率和效益的统一。(3) 在结果的实现上，通过契约化、分权化、市场化提供公共产品，公共产品的提供者在结果承诺的同时拥有自主性和灵活性。(4) 在结果责任方面，通过绩效测量与评估系统的建立，使得政府为结果负责，政府为结果负责就是履行公民的受托责任。如果将结果与产出混淆，新公共管理将会被抽离掉市场化、分权化、契约化的灵魂，退回到大政府和官僚控制的传统模式中。

需要采用恰当的方法厘清产出、成果和影响的关系，本文重点介绍将净影响从结果中分离出来的实验方法。

三、影响评估的实验方法

影响评估侧重对干预结果的评价，旨在通过系统的调查和数据收集，测量与评价项目的净效果，并证明干预与净效果之间存在因果关系（Bamberger et al., 2004）。干预的结果可以是预期的结果，也可能是执行过程中的阶段性结果，还可能是最终的产出和效果。因此，“影响评估可以在干预进行的任何阶段进行，从实施前的政策制定到计划、设计和实施”（罗希等，2002：203）。在结果链中，影响评估（Impact Assessment）重点关注如何将净影响（Net Effect）从总效应（Gross Effect）中分离出来，厘清干预与净影响的关系。一个社会干预引起的所有变化被称为总效应。总效应公式简单表述如下（李允杰、丘昌泰，2008：192）：

总效应 = 干预效应（净效应）+ 其他因素的效应（外部的无关干扰因素）+ 设计效应

影响评估的实验方法通过建立控制组，通过比较，探索干预与影响存在的因果关系，将净效应从总效应中分离出来。1963年唐纳德·坎贝尔（Donald Campbell）与朱利安·斯坦利（Julian Stanley）出版了一本探讨实验设计的著作《实验与准实验研究》（*Experimental and Quasi-Experimental Design for Research*），该书根据样本的控制程度和内外在效度的水平将实验分为前实验（The Pre-Experimental Design）、真实验（The True Experimental Design）和准实验（The Quasi-Experimental Design）三大类十六小类（Campbell & Stanley, 1966）。这些方法已经成为影响评估的常用方法。

（一）前实验设计

前实验设计又称作弱实验设计（Weak - Experimented Designs）或无设计实验（The Non - Design Experiment）。前实验设计对实验中涉及的干扰变量几乎没有任何控制，从而无法验证实验干预与实验结果之间的因果关系，实验的内外效度也非常低。前实验设计包含实验处理、测验或观察两个基本要素。前实验设计包括三种基本的类型，即单组个案研究（One - Shot Case Study）、单组前后测设计（The One - Group Pretest Posttest Design）和静态组比较设计（The Static - Group Comparison Design）。

单组个案研究。单组个案研究又称作单组后测设计（One - Group Posttest - Only Design），是先对一组受试者进行实验处理，然后再测量或观察因变量。

单组前后测设计。研究者选择一些受试者为对象，在进行实验处理（X）之前，先增加一项观察（前测 O1）。实验处理后观察或施以后测（O2），应用适当的统计检验因变量的变化是来自于自变量。该设计引入了前后比较，如果前后组相同则提供了控制的作用。

静态组比较设计。该设计使用实验组和控制组两组为实验对象，实验者仅对实验组进行实验处理（X），然后对两组皆观察或施以后测（O2），比较已接受处理的实验组与未接受处理控制组的区别。该设计另增控制组可供初步的比较。

（二）真实验研究

真实验设计是指具有实验组与控制组的对照（处理比较）、前测与后测的对照（基线比较）、以及实验组与对照组的随机分派三个要件的实验设计。在影响评估中，常用的真实验设计包括如下几种类型：

前后测控制组设计。前后测控制组设计（The Pretest - Posttest Control Group Design）又称作随机化实验控制组前后测设计（Randomized Control - Group Pretest - Posttest Design）。该设计随机分派受试者于实验组、控制组中，二组均进行前测，再将实验组进行实验处理，然后对实验组和控制组都进行后测，分析实验处理是否具有影响。

所罗门四组设计。所罗门四组设计（The Solomon Four - Group Design）将前测纳入实验设计中，先用随机方法选择受试者，并以随机分派方式将受试者分到不同的四组，接受不同的实验处理。前二组均接受前测，而对第一组和第三组分别给予实验处理，在实验处理之后，四组均接受后测。在所罗门四组设计

中，实验的总影响 = 前测的影响 + 实验刺激的影响 + 实验以外因素的影响 + 实验刺激与前测的交互影响，其中实验刺激与前测的交互影响可以用测量到的数据计算得出。

单后测控制组设计。单后测控制组设计（The Posttest – Only Control Group Design）又称为随机化控制组后测设计（Randomized Control – Group Posttest Design）。该设计随机化分派受试者于实验组、控制组中，实验组进行实验处理后，让实验组和控制组都接受后测，并比较两组后测的差异。哈里斯·迈凯（Harrison McKay）等人完成的论文《改善长期贫困儿童的认知能力》就是使用单后测控制组设计的例子（宾厄姆、菲尔宾格，2008）。

上述真实实验设计主要针对一个变量展开，社会科学中的很多事件往往是由许多变量交互作用形成的结果，很难完全孤立地分析某一个变量对结果的单独作用。基于此，在实验研究中，研究者同时操纵两个以上的变量，以探究每个自变量对因变量的影响，以及各个变量之间的交互作用，称为多因子实验设计。在统计学中，随机区组（Randomized Blocks）、分块（Split Plots）、希腊和拉丁方格（Greco – Latin Squares）、分数配置（Fractional Replication），混杂法（Confounding）等方法用于处理多因子之间的排列问题（Campbell，1966）。统计学已经发展出了多种针对多因子分析的技术和方法，可以将实验问题转换为统计问题，使用统计控制的方法达到变量控制和寻找因果关系的目的。

（三）准实验设计

准实验设计指的是具有实验研究的基本逻辑，但是缺乏严谨的实验控制，并且没有随机化分配实验组或控制组的实验研究。比如田野实验法（Field Experiment）无法对环境进行精密严谨的控制，对于被试者也无法进行随机选择与分派，就属于准实验设计的一种类型。准实验设计是影响评估中较为常见的类型，包括以下几种常见的形式：

时间序列设计。时间序列设计（Time – Series Design）指的是对某个测量指标进行周期性的追踪测量，以了解该指标所反应属性的变化情形。时间序列设计多用于小样本研究或者案例研究，特别是在政策干预、社会变迁和行为矫正等方面的应用较为广泛。过去事件、成熟、测量对象的适应或疲劳效应都会影响到研究的效度。

单组时间序列设计（Single – Group Time – Serious Design）是最简单的时间序列设计形式。单组时间序列设计与前实验设计中的“单组前后测设计”类似，不同的是，单组时间序列设计会对单组受试者在实验处理前后进行若干次周期

性的测量，分析多次测量结果的变化趋势，用以判断实验处理的成效。

如果在一个完整的时间序列中，插入一个实验处理，再观察整个时间趋势的变化情形，又称为间歇性时间序列设计（Interrupted Time - Series Design）。间歇时间序列设计对变量的变化采取多次测量，可以更为精确地刻画自变量引起的效应变化程度和时间跨度，可以在较长时间跨度上展示自变量与因变量之间的关系，做出的因果判断也更加有力。

如果在单组时间序列设计的基础上增设控制组，则称为控制组时间序列设计（Control Group Time - Series Design）。

相等时间取样设计。相等时间取样设计（The Equivalent Time - Samples Design）在只有实验组的情况下，让该组在一段时间内反复、系统地接受某种处理（ X_1 ），并交替使用另一种处理（ X_0 ）。该设计适用于一组可当控制及实验组，且缺乏相等的个人样本的情况。该设计可以看作是时间序列设计的变化形式，适合于处理效应比较短暂的社会干预。该设计的优点是：可以降低同时事件对研究效度的损害，满足内部效度的要求；可以控制历史的影响；可以控制外部因素的影响。

相等材料设计。相等材料设计（The Equivalent Materials Design）又称为相等材料前测后测设计（The Equivalent Materials, Pretest, Posttest Design）是以同组当作实验组与控制组，且进行两次以上的循环，特定组在第一循环可能当作控制组，至第二循环则当作实验组。

非对等控制组设计。非对等控制组设计（The Nonequivalent Control Group Design）与真实验设计中的“前测后测等组设计”类似，唯一的差别在于实验组和控制组不是随机抽取和随机分派产生的。

对抗平衡设计。对抗平衡设计（Counterbalanced Designs）又称为拉丁方格设计（Latin Square Design）、轮换设计。该设计让同一组被试者接受不同的实验处理时，因为只有一群被试者，因此不同的实验状况之间并不需要进行样本随机分派处理，但是被试者的反应却有可能受到实验顺序的影响，造成实验效果的混淆。对此，可以使用对抗平衡（Counterbalancing）原理来处理实验顺序的问题。拉丁方格实验设计中，首先根据自变量处理的水平数确定两个额外变量的水平数，然后利用两个额外变量的各个水平结合在一起构造一个拉丁方格，最后再将自变量的不同处理平衡地安排在这个方格中。这种研究方案要保证自变量的每一个水平在拉丁方格的每一行和每一列都出现且只出现一次。

分离样本前后测设计。分离样本前后测设计（Separate Sample Pretest - Posttest Design）以随机方法选择受试，并将其随机分派到两组，只有一组接受

◆ 专栏：因果推理

前测，两组均接受实验处理，但第一组接受的实验处理与研究结果无关，只有一组接受后测。此种方法的优点是能够测量大批被试者，如果设计包括多个组，能够彻底观察变量，便于排除历史的影响和客观自然因素等的影响。

分离样本前后测控制组设计。分离样本前后测控制组设计（The Separate - Sample Pretest - Posttest Control Group Design）与分离样本前后测设计类似，但增加了控制组。

连续拼凑式设计。连续拼凑式设计（The Recurrent Institutional Cycle Design: A “Patched - up”）对一个组进行实验刺激并进行后测，对另一组进行前测、实验刺激和后测。虽然该设计能较好地控制历史效应、试验效应、测量效应和选择效应，但是成熟效应和回归假象效应仍然可能存在。例如在 20 世纪 70 年代，美国的墨城（Model City）实施了一个对低收入人群进行免费医疗的项目试验，主要是为了评价该项目的实施对低收入人群健康水平的影响（Richardson & Others, 1976）。由于该项目试点无法随机选择对照组和试验组，又无法采用时间序列设计，故采用了该设计。

多组时间序列。如果在单组时间序列设计的基础上增设实验组，则成为多组时间序列（The Multiple Time - Series Design）。与单组时间序列设计的思想基本一致，不同的是它包含两组或两组以上的实验组。

回归 - 间断点设计。回归 - 间断点设计（Regression - Discontinuity Analysis）通过实验处理与事后测量回归线的间断点的特征，确定准实验处理的主效应。即如果实验处理前后的回归线出现了间断，说明实验处理是有效果的。

上述三类实验设计已经被广泛地运用于影响评估中，成为推动影响评估科学化的方法支撑。

四、影响评估实验方法的应用和成果

（一）实验方法在公共管理影响评估研究中运用的困境

公共管理领域影响评估的实验研究远远不如心理学、管理学、经济学、信息科学、社会学、政治学等领域的实验研究。按照 1992 年巴瑞·波兹曼（Barry Bozeman）和帕特里克·斯科特（Patrick Scott）在《公共行政研究和理论杂志》上发表的《实验室研究法在公共政策和管理领域》一文的阐述，公共管理领域实验研究之所以较少的原因主要有以下几点（Bozeman & Scott, 1992）：

公共管理研究往往忽视内部效度（PPM Researchers Bargain Away Internal

Validity)。实验室的实验研究通过较少的信息获得较多的知识，实地研究则通过较多的信息获得较少的知识。公共管理和政策研究是现实主义取向的，并且需要提供明确的政策导向和行动准则，因此在研究方法上，公共管理和政策研究往往更重视建构效度和外在效度，而较忽视内部效度；重视行动导向而不是理论导向，因此在方法论上较少采用实验方法。

对公共管理研究而言实地研究是首选的研究方法。实验室研究与实地研究相比较而言，实地研究往往成为首选，基于准实验研究的实地研究则更是首选。原因是实地研究和准实验研究能够较为严格地揭示因果关系，而且可以使政策制定和管理者与公民进行有效的沟通。基于准实验研究的实地研究的诸多发现，已经成为众多新公共政策制定的源泉（Greenberg & Robins, 1986）。

可信程度像真理一样重要（Credibility Is as Important as Truth）。公共管理和政策研究更倾向于获得特定情境下的发现，而不是获得一般性的结论；公共管理者和政策制定者更相信个人的判断和经验而不是依靠理论推导；公共管理和政策问题也很难分解为单一的变量进行线性的因果分析。总之，公共管理和政策研究更加重视可信程度，故多采用实地研究而不是实验室实验研究。

很少对博士生进行实验室方法的训练。正是因为公共管理和政策研究更多关注特殊问题而不是一般问题，学校很少给博士生提供实验室方法的训练。这与博士论文缺少严格性、对知识的积累的贡献很少、往往写成后就束之高阁不能出版的现状是一致的（White, 1986）。

公共管理和政策研究很少进行个体层面的分析。大多数实验研究都是进行个体层面的分析，尽管研究个体行为在公共管理学中具有重要的地位，但是公共管理中的个体行为往往更多地依赖集团和组织特征。聚焦于集团和组织的公共管理研究是难以进行实验变量的统计控制和调整的。

（二）公共管理领域影响评估的成果

尽管如此，实验思想和方法还是被广泛地运用于影响评估领域。可以认为凡是涉及社会干预和社会变迁的地方就可能开展影响评估。影响评估是一个对象、内容和范围都非常广泛的领域。一般人们往往使用社会干预影响评估、项目影响评估、政策影响评估等术语表述影响评估。从影响评估开展的时机来看，影响评估可以是事前，也可能在事中和事后。事中影响评估可以理解为阶段性的事后评估与预期性的事前评估的统一。

事前影响评估。事前的影响评估一般会假想出来一个对照组，对干预前后的效果进行预期性的对照，前实验设计和准实验设计的思路运用较为广泛，单

组个案研究、单组前后测设计、时间序列设计被广泛地运用到事前影响评估中。在公共管理领域，事前的管制影响评估、事前的环境影响评估和事前的社会影响评估在总体框架上就是一种前实验设计和准实验设计的思路。

事前的规制影响评估在美国和英国伴随着放松管制和新公共管理运动兴起，20 世纪 80 年代以来，澳洲（联邦层级）、奥地利、加拿大、丹麦、芬兰、法国、德国、冰岛、意大利、日本、南韩、墨西哥、荷兰、新西兰、挪威、葡萄牙、瑞典、英国、美国、欧洲执委会等国家或地区都建立了规制影响评估制度（Kirkpatrick，2006）。事前的规制影响评估在单组个案研究、单组前后测设计、时间序列设计的总框架中引入成本收益分析进行预期性的比较和方案之间的选择。事前环境影响评估和社会影响评估也基本上遵循这一思路。

干预后的影响评估。干预后的影响评估在公共管理领域涉及非常广泛，包括：决策与政策制定；生产力和绩效；评估研究；公共服务、公共事务和社会管理司法；犯罪和警察；组织行为；公共财政；管理信息系统；等等（曹堂哲，2014）。这些领域的众多议题都属于干预后的影响评估的范畴。

在生产力和绩效领域，学者们重点关注组织凝聚力、监督行为、管理行为对组织生产力和绩效的影响，使用的方法主要涉及真实验方法、受控现场实验方法、准实验方法等。比如：三隅二不二（Jyuji Misumi）和白樫三四郎（Sanshiro Shirakashi）通过实验方法研究了科层组织中监督行为对组织生产力的影响（Misumi & Shirakashi，1966）。尼古拉斯·小洛夫里奇（Nicholas P. Lovrich，Jr）等以四个州政府机关为对象进行准实验研究，研究了参与性绩效评估对工作满意度、组织氛围和工作价值的影响（Lovrich et al.，1981）。

在评估研究领域，对计划、规划和社会政策的影响进行评估是公共管理实验研究的重要议题。评估的对象包括规划、社会政策、教育培训、医疗、公共卫生等多方面的内容，常用的方法包括现场实验方法、真实验等。比如：拉普雷勒·约翰（LaPrelle John）等选取 10 个社区，对青少年吸烟进行了现场实验研究，提炼了影响吸烟的共同变量（John et al.，1992）。迈克尔·佛伦奇（Michael T. French）等使用现场实验的方法对门诊吸食大麻的青少年进行治疗进行经济评估（French et al.，2003）。理查德·桑迪（Richard T. Santee）和芭芭拉·戴维斯（Barbara G. Davis）使用实验研究方法探讨了预备教育规划（The Summer Threshold Program at the University of California at Berkeley）实施的影响因素（Santee & Davis，1980）。

在公共服务、公共卫生、公共事务和社会管理领域，影响评估的议题主要涉及人口控制、公共卫生和医疗、教育、社区等方面的内容。比如：库尔特·

巴克 (Kurt W. Back) 等对波多黎各人口控制问题进行了现场实验 (Back, 1957)。贝基·佩蒂特 (Becky Pettit) 和萨拉·麦克拉纳汉 (Sara McLanahan) 使用实验研究方法探讨了居住地流动和儿童社会能力之间的关系 (McLanahan, 2003)。苏珊娜·罗卜 (Susanna Loeb) 等在康涅狄格州进行随机实验研究了福利改革对儿童的影响 (Loeb et al., 2003)。哈里斯·迈凯 (Harrison McKay) 等完成的论文《改善长期贫困儿童的认知能力》就是使用单后测控制组设计的例子。该项实验的全部时间是3年,即从1971年2月至1974年8月。每9个月为一阶段,分为四个治疗阶段,加上阶段间的休整期。该项目以尽可能接近3岁的儿童开始研究,这主要是考虑到在2年的实验研究中所设计的治疗和衡量系统是为这个年龄的儿童准备的。预计的可能参加治疗的时间是180~200天,从而使每个计划阶段与哥伦比亚的一学年长度相似,第四阶段的末期安排得与适龄儿童进入一年级的时间一致。整个实验设计如下表所示:

表1 改善长期贫困儿童的认知能力单后测控制组设计

组	数量		特征
	1971年	1975年	
T ₁ (a)	57	49	SES低,体重和身高低于正常。一个治疗阶段,1973年11月至1974年8月(75到84个月的年龄)
T ₁ (b)	56	47	SES低,体重和身高低于正常。一个治疗阶段,1973年11月至1974年8月(75到84个月的年龄),预先进行营养补充及健康护理
T ₂	64	51	ES低,体重和身高低于正常。两个治疗阶段,1972年11月至1974年8月(63到84个月的年龄)
T ₃	62	50	SES低,体重和身高低于正常。三个治疗阶段,1971年12月至1974年8月(52到84个月的年龄)
T ₄	62	51	SES低,体重和身高低于正常。四个治疗阶段,1971年2月至1974年8月(42到84个月的年龄)
HS	38	30	SES高,未受治疗,但与T ₁ —T ₄ 组同时衡量
T ₀	116	72	SES低,体重和身高正常。未受治疗

注:SES指家庭社会经济地位。

资料来源:宾厄姆、菲尔宾格,2008:83。

在研究过程中有5个测量时点:(1)第一治疗阶段开始;(2)第一治疗阶

段结束；(3) 第二治疗阶段结束至第三阶段开始；(4) 第三治疗阶段结束至第四治疗阶段开始；(5) 第四治疗阶段结束之后。为了衡量治疗对认知发展单方面的作用，在每一测量点使用的是多个小测试而非单一的智力测试。为了避免检测者的偏见，儿童被随机安排给检测者，不提供治疗组或营养或社会经济水平的任何相关信息。通过观察多元矢量，形成了上述各组在五个点的测量数据矩阵。数据分析研究显示，在3岁至7岁之间进行营养、健康及教育组合治疗可以防止潜在认知能力的大量损失，而且，治疗开始得越早，效果越好。即使仅在小学前进行9个月的治疗，能力上似乎也有明显提高，尽管与持续了两次、三次和四次治疗的儿童的成果相比很小。治疗效果持久性的确定有待进一步研究，但目前数据显示能够延续到8岁（宾厄姆、菲尔宾格，2008）。

在司法、犯罪和警察领域影响评估涉及的议题主要包括犯罪学、监视系统、诉讼、假释、巡逻和治安等方面。比如：约耳书·加纳（Joel H. Garner）和克里斯蒂·维舍（Christy A. Visser）对犯罪的产生进行了实验研究（Garner & Visser, 2003）。大卫·法林顿（David P. Farrington）回顾了犯罪学随机化实验的发展历史（Farrington, 2003）。琼·彼得斯利尔（Joan Petersilia）采用随机化实验方法对密集监视项目（BJA's Intensive Supervision Project）进行了研究（Petersilia, 1989）。

五、结语

影响评估重点关注干预与净影响的关系，目的是将净影响从总影响中分离，并证明干预与净效果之间存在因果关系。20世纪70年代以来公共管理领域中的影响评估成为了重要的理论和实践领域。世界银行、国际影响评估协会、麻省理工大学等机构在影响评估领域取得了有益的成果，国内学者在环境、大气治理、教育、医疗等领域也推进了影响评估的研究。

就方法论而言，影响评估的方法包括定性和定量两类，实验方法是定量影响评估研究的基本方法。由于公共管理研究往往忽视内部效度、对公共管理研究而言实地研究是首选的研究方法、可信程度像真理一样重要、很少对博士生进行实验室方法的训练、公共管理和政策研究很少进行个体层面的分析等原因，公共管理领域影响评估的实验研究远远不如心理学、管理学、经济学、信息科学、社会学、政治学等领域的实验研究。

国内一些文献中存在结果（Results）与成果（Outcome）翻译混用的情况。表面上是名词术语的混淆，实质上反映了人们对影响评估属性和方法论认知的

混乱，进而影响到人们对结果导向管理模式的正确认知。有必要在结果链准确定位产出、成果和影响的意义。一般而言，成果评估是描述性的，侧重产出与对利益相关者目标的满足程度，并不重点关注如何将干预的净效应分离出来。影响评估则是因果分析式的，目的是将干预的净效应从总效应中分离出来，发现干预与净效应之间的因果关系。因此从方法论上来讲，影响评估通常需要使用独立的鉴别器、建立控制组、测量一个长期的改变。

本文尝试以 1963 年唐纳德·坎贝尔（Donald Campbell）与朱利安·斯坦利（Julian Stanley）出版的著作《实验与准实验研究》（*Experimental and Quasi-Experimental Design for Research*）为框架，阐明公共管理影响评估的实验方法体系，该体系分为三大类十六小类。实验方法已经广泛地运用于公共管理事前和干预后的影响评估。事前的影响评估一般会假想出来一个对照组，对干预前后的效果进行预期性的对照，前实验设计和准实验设计的思路运用较为广泛，单组个案研究、单组前后测设计、时间序列设计被广泛地运用到事前影响评估中。干预后的影响评估在公共管理领域涉及非常广泛，包括：决策与政策制定；生产力和绩效；评估研究；公共服务、公共事务和社会管理司法；犯罪和警察；组织行为；公共财政；管理信息系统；等等。公共管理领域影响评估的实验方法运用前景广阔，是一片学术沃土，亟待开垦耕耘。

参考文献

- 彼得·罗希、马克·李普希、霍华德·弗里曼(2002). 项目评估：方法与技术(第6版). 邱泽奇等译. 北京：华夏出版社.
- 曹堂哲(2014). 公共管理研究方法：基于公共管理问题类型学的新体系. 北京：北京大学出版社.
- 理查德·D. 宾厄姆、克莱尔·L. 菲尔宾格(2008). 项目与政策评估：方法与应用. 杨国庆、朱春奎译. 上海：复旦大学出版社.
- 李允杰、丘昌泰(2008). 政策执行与评估. 北京：北京大学出版社.
- 马骏、赵早早(2011). 公共预算：比较研究. 北京：中央编译出版社.
- 乔迪·扎尔·库赛克、雷·C. 瑞斯特(2011). 十步法：绩效监测与评价手册. 梁素萍、韦兵项译. 北京：中国财政经济出版社.
- 邱需恩(2004). 美国《1993 政府绩效与结果法案》译文. 中国行政管理, 5: 28-28.
- 亚洲开发银行(2011). 影响评估最新进展评述. 菲律宾曼达卢永市：亚洲开发银行.
- 詹姆斯·C. 麦克戴维、劳拉·R. L. 霍索恩(2011). 项目评价与绩效测量：实践入门. 李凌艳、张丹慧、黄琳等译. 北京：高等教育出版社.
- Allio, L. (2007). Better Regulation and Impact Assessment in the European Commission. In Kirkpatrick, C. & Parker, D. Eds. *Regulatory Impact Assessment: Towards Better Regulation*.

◆ 专栏：因果推理

Cheltenham : Edward Elgar.

- Back ,K. W. ,Hill ,R. & Stycos ,J. M. (1957). The Puerto Rican Field Experiment in Population Control. *Human Relations* ,10 :315 – 334.
- Bamberger ,M. ,Rugh ,J. ,Church ,M. & Fort ,L. (2004). Shoestring Evaluation: Designing Impact Evaluations under Budget ,Time and Data Constraints. *The American Journal of Evaluation* ,25 (1) : 5 – 37.
- Bozeman ,B. & Patrick ,S. (1992). Laboratory Experiments in Public Policy and Management. *Journal of Public Administration Research and Theory* ,2 (3) :293 – 313.
- Campbell ,D. T. & Stanley ,J. C. (1996). *Experimental and Quasi – Experimental Designs for Research*. Chicago ,IL :Rand McNally.
- Farrington ,D. P. (2003). A Short History of Randomized Experiments in Criminology : A Meager Feast. *Evaluation Review* ,27 :218 – 227.
- French ,M. T. ,Roebuck ,M. C. ,Dennis ,M. L. ,Godley ,S. H. ,Liddle ,H. A. & Tims ,F. M. (2003). Outpatient Marijuana Treatment for Adolescents : Economic Evaluation of A Multisite Field Experiment. *Evaluation Review* ,27 (4) :421 – 59
- Garner ,J. H. & Visser ,C. A. (2003). The Production of Criminological Experiments. *Evaluation Review* ,27 :316 – 335.
- Greenberg ,D. H. & Robins ,P. K. (1986). The Changing Role of Social. Experiments in Policy Analysis. *Journal of Policy Analysis & Management* ,5 (2) :340 – 362
- Hatry ,H. (2007). *Performance Measurement : Getting Results*. London : Rowman & Littlefield Publishers.
- John ,L. ,Bauman ,K. E. & Koch ,G. G. (1992). High Intercommunity Variation in Adolescent Cigarette Smoking in a 10 – Community Field Experiment. *Evaluation Review* ,16 :115 – 130.
- Kirkpatrick ,C. (2006). Regulatory Impact Assessment. In Crew ,M. A. & Parker ,D. Eds. *International Handbook on Economic Regulation*. Northampton ,MA :Edward Elgar.
- Khandker ,S. ,Koolwal ,G. B. & Samad ,H. (2010). *Handbook on Impact Evaluation*. Washington ,D. C. :The World Bank.
- Loeb ,S. ,Fuller ,B. ,Kagan ,S. L. & Carrol ,B. (2003). How Welfare Reform Affects Young Children : Experimental Findings from Connecticut—A Research Note. *Journal of Policy Analysis and Management* ,22 (4) :537 – 550.
- Lovrich ,N. P. ,Hopkins ,R. H. ,Shaffer ,P. L. & Yale ,D. A. (1981). Participative Performance Appraisal Effects Upon Job Satisfaction ,Agency Climate ,and Work Values : Results of a Quasi – Experimental Study in Six State Agencies. *Review of Public Personnel Administration* ,1 (3) : 51 – 73.
- Misumi ,J. & Shirakashi ,S. (1966). An Experimental Study of the Effects of Supervisory Behavior on Productivity and Morale in a Hierarchical Organization. *Human Relations* ,19 :297 – 307.

- OECD. (2011). Better Regulation in the United Kingdom. Available at (Jan. 26, 2018) : <http://www.oecd.org/dataoecd/54/33/07706.pdf>.
- Patton, M. Q. (2011). *Developmental Evaluation: Applying Complexity Concepts to Enhance Innovation and Use*. New York: The Guilford Press.
- Paul, J. G., M., S., Premand, P., Laura, B. R. & Vermeersch, C. M. J. (2011). *Impact Evaluation in Practice*. Washington D. C. : The World Bank.
- Petersilia, J. (1989). Implementing Randomized Experiments: Lessons from BJA's Intensive Supervision Project. *Evaluation Review*, 13: 435 – 458.
- Pettit, B. & McLanahan, S. (2003). Residential Mobility and Children's Social Capital: Evidence from An Experiment. *Social Science Quarterly*, 84(3): 632 – 649.
- Richardson, W. C. & Others. (1976). *The Seattle Prepaid Health Care Project: A Comparative Analysis of Health Services*. Springfield, VA: National Technical Information Service.
- Santee, R. T. & Davis, B. G. (1980). The Summer Threshold Program: An Experiment in Preparatory Education. *Evaluation Review*, 4: 215 – 224.
- White, J. D. (1986). Dissertations and Publications in Public Administration. *Public Administration Review*, 46, 3: 227 – 34.
- Zewo Foundation. (2011). Outcome and Impact Assessment in International Development. Available at (Jan. 26, 2018) : https://impact.zewo.ch/english/docsZewo_Wirkungsmessung_E_web.pdf.

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