如何科学评估公共政策？

——政策评估中的反事实框架及匹配方法的应用

刘玮辰　郭俊华　史冬波*

摘要：公共决策科学化依赖于科学的公共政策评估。随着国外公共政策评估研究越来越依赖于旨在评估政策因果效应的因果推断方法，国内公共政策学者也开始进行初步尝试，但是如何有效选择和运用因果推断方法开展政策评估研究仍有待系统梳理。本文介绍了公共政策评估方法的反事实框架，精确定义了公共政策的因果效应。在此基础上，本文将匹配方法总结为距离测算与配对两个步骤，并详细阐述了协变量匹配、粗糙完全匹配、马氏距离匹配、倾向值匹配和熵平衡匹配的原理，比较了不同方法的优势与劣势。结合公共政策评估的前沿实证研究，本文介绍了如何具体运用匹配方法进行政策评估研究。本文还探讨了应用匹配方法的注意事项，包括匹配方法的适用性、匹配与回归的关系、匹配方法对样本数量的要求及是否允许放回等。

【关键词】公共政策评估　因果机制　反事实　匹配

【中图分类号】D63　　**【文献标识码】**A

【文章编号】1674－2486（2021）01－0046－28

一、引言

对于从事公共政策评估研究的学者来说，这是最好的时代。当前，我国政府绩效管理越来越依赖于第三方政策评估，这为公共政策评估研究提供了“用武之地”，相关研究数量迅速增长。在 CNKI 数据库，以“政策评估”为关键词的研究论文从 2010 年的 144 篇增加到 2017 年的 238 篇。对于从事公共政策评

* 刘玮辰，清华大学公共管理学院博士后；郭俊华，上海交通大学国际与公共事务学院教授；通讯作者：史冬波（shidongbo@ sjtu. edu. cn），上海交通大学国际与公共事务学院特别副研究员。本文曾于《公共行政评论》第四届青年学者论坛（2020 年 1 月，中山大学）宣读并获优秀论文奖。感谢与会专家的点评意见，感谢新加坡国立大学博士生葛野嫣然为本文做出的贡献，感谢匿名评审人的意见。

基金项目：上海市哲学社会科学规划课题《上海市需求侧创新政策评估及优化路径研究》（2020BGL006）。

估研究的学者来说，这也是个混乱的时代。研究领域快速发展的背后是普遍存在的研究方法不规范、不科学的问题，这导致了相关领域在研究方法上缺少共识，在研究结论上缺少稳健性，低水平评估的学术成果屡见不鲜（陈光、方新，2014）。

国际上主流的公共政策评估理论一直存在着强调技术中立的实证主义取向与强调多元价值的规范主义取向的方法论之争（高雪莲，2009）。前者将公共政策评估视作一个技术问题，旨在分析公共政策是否达成了其预定目标；而后者则认为公共政策评估是一个政治问题，需要考虑不同参与主体的利益诉求与政治影响。两种方法论的争论实际上只是强调了公共政策评估的不同阶段，而我国方法论的迷思在于还未完成技术性阶段便直接跳到了“价值多元”的争论当中。因此，我国政策评估研究应该坚定不移地走向实证主义（和经纬，2008）。

实证主义取向的政策评估研究的核心在于厘清政策干预的因果效应。政策评估的目的是通过科学、严谨的研究设计，判断某项政策实施先后的效果差异，只有正确分离出政策干预的因果效应，才能科学地指导后续政策调整。遗憾的是，对于“因果效应”的关切，在我们走向实证主义政策评估研究的过程中却没有得到足够的重视。诚然，我国的政策评估研究已经从规范性的辩论逐渐发展为更加依赖客观数据与实证经验的定量评估。然而，现有的政策评估研究中，仍然有相当比例的研究仅仅满足于通过设计指标体系来评估政策实施的效果，在评估伊始便难言其具有科学性。即使部分研究开始使用回归分析等方法来试图分析政策效应的因果效应，但是其方法使用中依然存在诸多不规范之处。如果方法使用不当，公共政策评估的结果不仅会错误估计政策效应的强度，甚至会错误估计政策效应的方向，从而误导公共政策制定。

造成这种现象的原因有三个：其一，公共政策因果效应的定义不够清晰，已有的研究比较欠缺。缺乏对因果效应的统一认识，自然在研究实践上存在诸多不当。其二，国内公共政策学界对于如何估计因果效应，以及各种方法之间的联系与区别缺乏深入了解。其三，尽管部分学者已经在研究实践中开始尝试使用一些估计方法，但是对于该方法的原理以及适用范围存在误解，造成了使用不当的问题。

正是在这样的背景下，我们希望通过本文来回应上述三个问题，以探讨并构建我国的公共政策因果效应评估方法论体系（如图 1 所示）。本文将主要探讨上文提及的前两个问题，并以匹配方法为例回应第三个问题。本文的安排如下：第二部分梳理国内外公共管理领域方法论的变迁过程；第三部分将通过“反事实”理论来定义因果关系；第四部分将指出因果估计的核心问题是消除选择性偏误，并将因果估计的方法做了分类比较；第五部分综述了最常用的匹配方法的原理和应用；第六部分讨论匹配方法的常见问题。

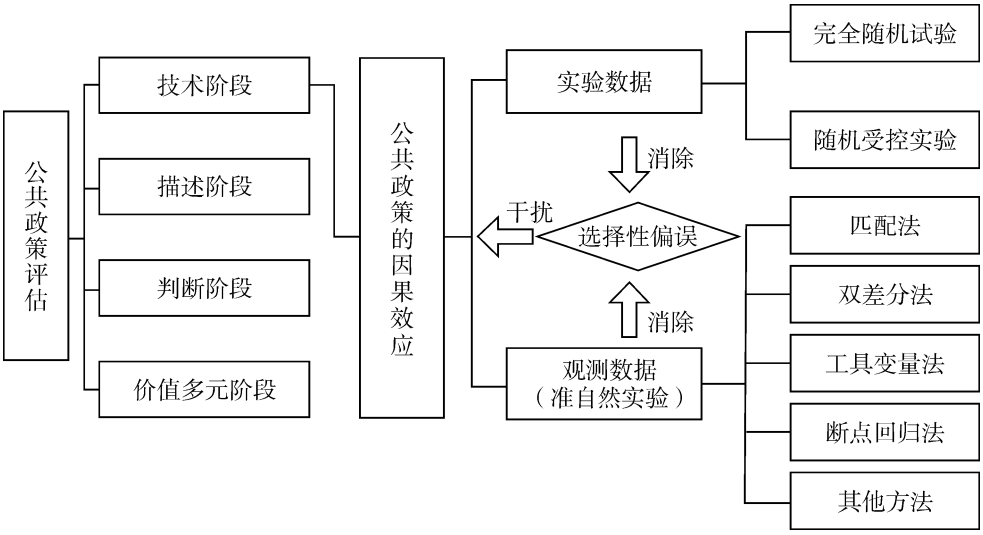

图 1　公共政策评估理论框架

资料来源：作者自制。

二、公共管理研究方法论变迁

本文对 2002－2017 年发表在国际和国内公共管理领域高水平期刊①的论文的研究方法进行了总结。我们将搜集到的公共管理论文按照图 2 进行分类，发现英文论文 3130 篇，中文论文 7314 篇。在英文论文中，共有 1170 篇规范研究，1960 篇实证研究，其中定性研究论文 595 篇，定量研究论文 1365 篇。在中文论文中，共 5934 篇规范研究，1380 篇实证研究，其中定性研究论文 666 篇，定量研究论文 714 篇。我们进一步进行了年份的趋势分析，发现从 2002 年到 2017 年，中英文实证论文的数量和比重均逐年增加。英文实证论文的比重从 50% 增加到 75% 以上，中文实证论文的比重从 5% 上升到 30% 左右。由此可见，实证研究已经成为国际公共管理占据绝对主导地位的研究范式。而国内公共管理学术研究的实证化程度仍然远远低于国际水平，存在明显差距。

① 中文期刊为《公共行政评论》《公共管理评论》《公共管理学报》《管理科学学报》《管理世界》《政治学研究》《中国管理科学》和《中国行政管理》，其中《管理科学学报》《管理世界》《政治学研究》三本杂志仅保留了其中公共管理领域的论文。英文期刊为 *Governance*，*Journal of Policy Analysis and Management*，*Journal of Public Administration Research and Theory*，*Public Administration* 和 *Public Administration Review*。在数据处理过程中，剔除了非学术论文的笔谈、书评、会议综述与合集、翻译论文、博士论文摘要以及专题讨论会的论文（symposium）。

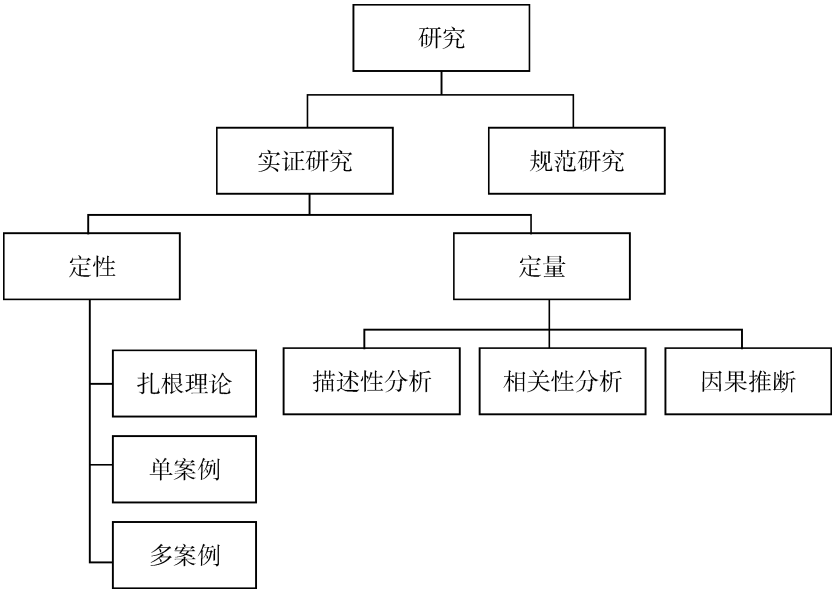

图 2　公共管理论文分类框架

资料来源：作者自制。

我们对实证研究论文的研究方法进一步细分，发现英文论文中相关性分析所占比重最高，从 2002 年到 2017 年占所有实证论文的比重从 30% 上升至 50%，其次是单案例研究，在多年份的使用占比仅次于相关性分析。从数据可以看到一个明显的趋势，那就是学者对于因果机制的关切。从 2002 年到 2017 年，因果研究的比重从个位数逐渐上升到 20% 以上（如图 3 所示）。在中文论文中，单案例研究和相关性研究论文数量最多，但单案例研究的比重在逐渐下降，从 2002 年近 70% 下降到 2017 年 30% 左右。而相关性研究的比重在逐渐增加，在 2017 年已成为实证研究中占比最多的方法。同样可以看到的是，学者逐步开始关心因果研究，但相比英文论文，因果分析的研究增加缓慢，增量较小（如图 4 所示）。

为了进一步分析因果关系方法的变化，本文对因果分析方法进行细分，从表 1 可以看到中文论文和英文论文的差距。共有 221 篇论文使用了因果分析的相关方法，其中，中文论文 24 篇，英文论文 197 篇。从中文论文来看，在公共管理领域，每年使用因果方法的数量仅为个位数，大部分研究都使用了工具变量和匹配的方法，也有少量的实验和双重差分方法，方法逐步多样化并且每种方法有增长趋势。从英文论文来看，因果分析的使用已有一定基础，各种方法都稳步发展并有所涉及，其中，实验和双重差分数量增长较快，工具变量和匹配相对稳定。

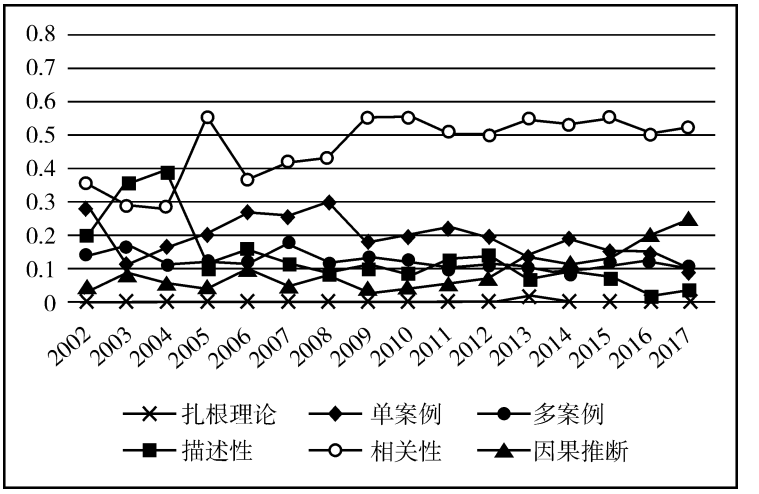

图 3　公共管理领域主要英文期刊中因果推断方法的衍变

资料来源：作者自制。

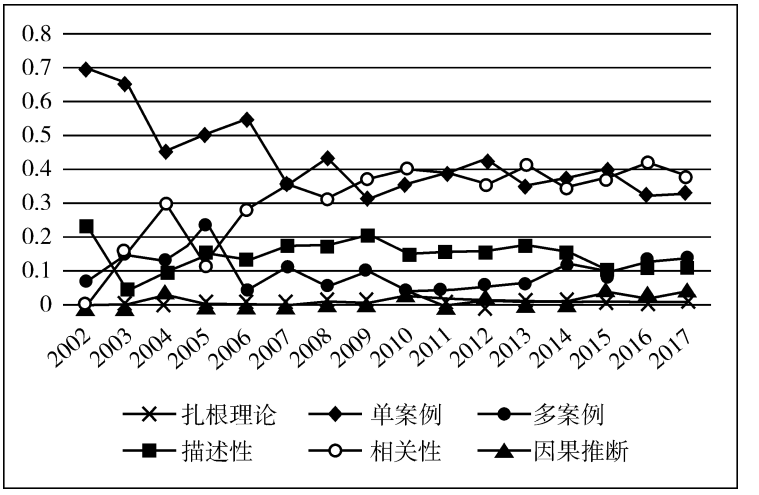

图 4　公共管理领域主要中文期刊中因果推断方法的衍变

资料来源：作者自制。

表 1　因果推断中具体方法的运用频次

	中文论文	英文论文
匹配	7	22
双重差分	1	36
工具变量	9	27
断点回归	0	9
实验	7	103

资料来源：作者自制。

综上所述，我们发现国际论文使用因果分析的方法已经有一定基础。而国内公共管理论文的研究方法与和国际论文存在较大差异，国内论文对因果机制的关切严重不足。特别是在公共政策评估研究中，因果推断方法的应用尤其欠缺。因此，急需完善公共政策评估理论，尤其需要对因果机制框架和相关方法进行比较分析。

三、反事实理论与公共政策因果效应

完善公共政策评估理论首先要精准定义公共政策的因果效应（Morgan & Winship，2015；Rosenbaum & Rubin，1983；Rubin，1997）。公共政策可以看成是对其作用对象的一种干预，政策评估便是研究这种干预对于其作用对象的影响。实际上，自 Fisher（1951）提出实验设计方法以来，统计学与计量经济学界对于因果推断的研究已经日臻成熟，Rubin（1973）、Angrist（1998）等学者在反事实理论的基础上构建了严密的因果推断方法论体系。起初，作为对自然科学的回应，经济学的实证研究中运用实验法逐渐受到认可（Kahneman & Smith，2002）。随着实验方法在时间成本、道德考量等方面的局限，实证研究中基于实验数据的自然实验得到经济学家的推崇（Angrist & Krueger，2001；Meyer，1995）。自然实验可以说是因果推断模型中的黄金原则，也被称为随机试验，具体指实验参与者被随机分配到处理组或者对照组，分配过程与实验参与者的特征无关。但是，由于其在实施过程中难以实现真正的随机分配，研究者往往会寻找自然发生的实验或者基于现有的观察数据来构造实验条件。那么，如何利用非随机的观测数据进行统计推断？此时，因果推断就成了一个难题，基于观测数据进行一系列准实验，包括工具变量、双重差分、匹配法、断点回归等也由此应运而生。

Rubin（1973）首次在理论上给出了正式阐述，匹配法的应用理论发展也始于此。其中 Rubin（1974）以及 Cochran 和 Rubin（1973）对此做出了开创性的贡献，他们关注了单个协变量的情形下平均处理效应的估计问题。随后，Rosenbaum 和 Rubin（1983）开创性地提出了倾向值匹配的方法，通过倾向值这一维度变量进行匹配以减轻协变量匹配对数据以及计算上的要求。近年来很多文献将匹配法与其他方法相结合，提出了一些新的评估思路，并通过实证研究表明这种方法上的结合有助于得到更为精确的匹配估计结果（Heckman et al.，1998）。

双重差分（DID）在公共管理领域已有小规模的运用，其中重要的应用在于研究政策实施前后的效果评估，通过自然形成的对照组来代替其本身存在的时间趋势，进而进行因果关系的估计。Heckman 等学者（Heckman & Robb，1985；Heckman & Robb，1986）最早提出使用 DID 方法对社会公共政策的实施效应进行评估，此后对 DID 方法的研究和应用成果层出不穷（Card，1990；Card & Krueger，2000；Chen et al.，2008；Donohue & Wolfers，2005）。值得注意的是，双重差分相较于其他的自然实验设计，能够通过差分解决不随时间变化的遗漏变量问题。

利用工具变量处理内生性的方法也得到实证研究者的重视和运用（Angrist & Keueger，1991）。然而，有学者针对工具变量的选取准则、弱工具变量问题、工具变量的效率问题等提出了质疑（Bound et al.，1995）。此后，有关工具变量的研究大多集中在如何寻找最优的工具变量上（Canay，2010；Okui，2009）。

断点回归作为一种准实验设计，其主要思想是如果政策在一个关于个人背景的连续变量（例如考试成绩、家庭人均收入等）上设定一个临界值，使得临界值一侧的个体接受政策干预，而在临界值另一侧的个体不接受干预，则在临界值附近就构成了一个准实验。Lemieux 和 Milligan（2008）在关于社会救助与失业率的研究中提供了断点回归的一个经典研究。Hahn 等人（2001）为断点回归的模型识别和模型估计进行了严格意义上的理论证明，并提出了相应的估计方法，自此断点回归在经济学上的应用才逐渐盛行。

从已有的文献来看，迄今为止，基于因果推断的政策评估研究成果还主要集中在劳动经济学、发展经济学等领域，涉及得多的是就业培训、社会保障等社会政策。国外的政策效应评估从理论研究到实证研究近年来取得了从无到有，再到逐渐成熟的巨大发展。然而受传统政治学取向以及政策科学本身特性的影响，我国政策评估研究仍以定性方法居多，忽视了定量方法的运用，对于因果推断的重视仍处于起步阶段，这与日益增长的政策效应评估需求显得极不平衡。并且，已有的研究大多局限于某个领域或具体的分析方法，对于整个公共政策的方法论体系的构建仍付之阙如。因此，通过反事实框架对政策的效应进行评估还是一个较新的领域，无论是从方法论上，还是从实证角度都有待发展。在表 2 中，我们罗列了在因果推断理论发展历程中的重要工作。接下来，我们将利用反事实理论定义公共政策的因果效应。

表 2　因果推断理论发展代表文献

代表学者	主要贡献
Fisher（1951）	最早提出用实验设计的方法来评估因果效应
Cochran & Rubin（1973）；Rubin（1973；1976）	首次从理论上阐述了如何利用观测数据进行统计推断以达到和随机试验相近的效果。早期研究匹配法的代表文献
Rosenbaum & Rubin（1983；1985）	提出条件独立假设与倾向值匹配法，并证明倾向得分定理，为后续倾向值匹配的理论与应用做出了开创性贡献
Theil（1953）	首次将工具变量用于处理选择性偏误
Angrist & Krueger（2001）	梳理了工具变量方法的发展历史
Thistlethwaite & Campbell（1960）；Campbell & Stanley（1963）	首次提出使用断点回归设计研究处理效应
Cook（2008）	梳理了断点回归的发展历史
Heckman（1974；1976；1979）	提出选择偏差并将选择偏差引入模型建立估计，从而修正偏差
Lalonde（1986）	用实验数据作为基准评价了非实验数据估计方法的优劣，其数据被广泛用于后续不同方法效果的比较
Goldberger（1972）	断点回归设计首次在经济学领域被提出
Moffitt（1991）；Card & Krueger（1994）；Eissa（1996）	发展双重差分方法，并使用 DID 方法对社会公共政策的实施效应进行评估

资料来源：作者自制。

反事实理论（counterfactual）是定义公共政策因果效应的基本框架。不失一般性，我们从最简单的政策情形出发，即一项不存在个体差异的政策干预，其干预方式对于所有被干预对象都是同质化的。当政策执行时，其作用对象总体就被该政策分为两组：其中受到政策干预的被称为处理组（treated group），未受政策干预的被称为对照组（control group）。假设我们关心政策干预对其作用对象某种特征的影响，这里的特征可以用 Y 来观测。此时，公共政策的因果效应便是比较处理组的特征和假定其未受到政策干预的特征的差异（后面将其定义为基于处理组的政策效应），或者比较假定对照组受到了政策干预的特征与其实际特征的差异（如图 5 所示）。

数学模型能帮助我们精确阐述上文的思想。假定用 D 来表示政策干预，对于任何个体 i，当 $D_i=1$ 时，表示个体 i 受政策的干预；反之，当 $D_i=0$ 时则表示个体不受政策干预。若用 Y_i 表示个体 i 特征（即政策的潜在结果），那么，Y_i 有两种可能，分别表示受到政策干预与未受政策干预的特征，即

$$Y_i=\begin{cases}Y_{(1,i)}, & \text{当 } D_i=1 \text{ 时}\\ Y_{(0,i)}, & \text{当 } D_i=0 \text{ 时}\end{cases}$$

上式可简化为：

$$Y_i=(1-D_i)\ Y_{(0,i)}+D_i\ Y_{(1,i)}$$

此时，对于个体 i，政策 D 的因果效应（Treatment Effects）被定义为：

$$\tau=Y_{(1,i)}-Y_{(0,i)}$$

但是，对于任何个体 i 而言，$Y_{(0,i)}$ 与 $Y_{(1,i)}$ 是不可能被同时观测到的。对于处理组 $Y_{(0,i)}$ 无法观测，对于对照组则无法观测到 $Y_{(1,i)}$，也就是说，存在着两组“反事实”，这便是反事实理论名称的来源。

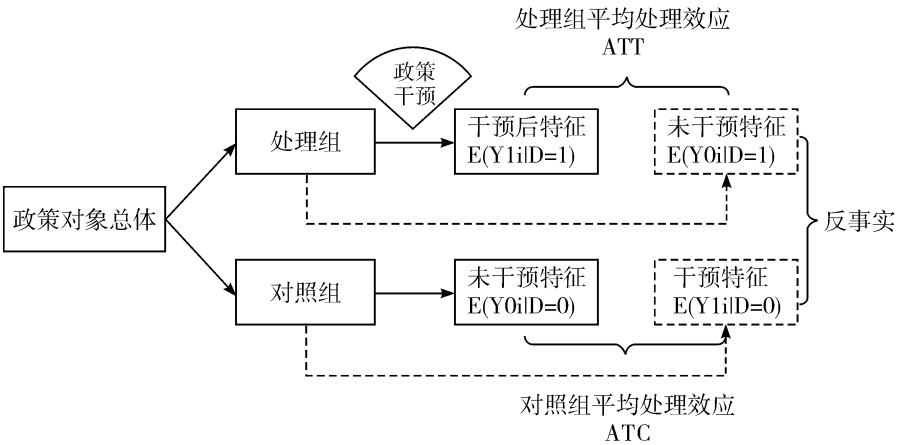

图 5　公共政策因果效应的定义

资料来源：作者自制。

实际上，反事实理论已经被普遍应用在社会科学的研究当中。社会科学的核心任务就是解释（刘骥等，2011），因果机制方法可以在较大的范围内对社会现象进行准确地描述，也可以解释不同现象之间的关系，其程序规范、严格，可信度高，在当代社会科学研究中得到了广泛的应用。表 3 列举了近年来发表在国际期刊的反事实理论在公共管理领域代表性的应用。

表 3　近年来国际公共管理顶级期刊因果推断的代表文献

文献	方法
Herbst & Tekin（2016）；Hong（2016）；Buckles & Guldi（2017）；Jimenez（2017）；Amuedo-Dorantes & Arenas-Arroyo（2019）；Bettinger & Evans（2019）	工具变量
Scott（2015）；Autio & Rannikko（2016）；Holt（2019）；Waddington & Berends（2018）；Rosholm et al.（2019）	匹配法
Dee & Wyckoff（2015）；Robinson-Cimpian & Thompson（2016）；Ellen et al.（2016）；Myerson et al.（2020）	断点回归
Bertelli et al.（2015）；Kaestner et al.（2017）；Cárdenas & Ramirez de la Cruz（2017）；Shinohara（2018）；Jo & Nabatchi（2019）	DID
Olsen（2015）；Andersen & Hjortskov（2016）；Hjortskov（2017）；Chiang et al.（2017）；Bellé et al.（2018）；Keiser & Miller（2020）	实验法

资料来源：作者自制。

反事实理论也清晰地反映了公共政策评估研究方法论争论的焦点。公共政策因果效应评估关注的是利用科学严谨的方法测算公共政策 D 对于结果 Y 的影响，而政策评估研究关注政策对象的哪些特征（或者说哪些价值）。如何评估公共政策的因果效应，则超出了因果效应评估的讨论范围，成为一个高度政治性的过程，这样的评估便从技术阶段转入判断阶段与多元价值阶段。这恰恰体现了因果效应评估在政策评估研究中不可替代的价值，只有基于科学客观的技术评估结果，政策价值判断与多元价值的讨论才有意义。否则，政治博弈的过程不但缺乏了事实基础，其结果更有可能南辕北辙。

四、公共政策因果效应的估计方法

在反事实框架下，政策因果效应评估就是对处理组的平均处理效应 ATT 的估计，其中如何构建反事实是政策效应评估中的关键，即克服 $E[Y_{0,i}|D_i=1]$ 的不可观测性。本节将介绍公共政策因果效应的主要估计方法。在此之前，我们需要引入一个重要的概念，即选择性偏误（Selection Bias）。回忆一下，在没有反事实框架之前，我们是如何估计政策因果效应的。最常见的做法是直接用 $E[Y_{0,i}|D_i=0]$ 替代 $E[Y_{0,i}|D_i=1]$，从而用观测效应（Observed Effccts）$E[Y_{1,i}|D_i=1]-E[Y_{0,i}|D_i=0]$ 来估计 ATT。这样的做法便会带来选择性偏

误。实际上，观测效应可以分解为处理效应与选择性偏误两部分，即

$$\underbrace{E[Y_{1,i}|D_i=1]-E[Y_{0,i}|D_i=0]}_{\text{观测效应}}=\underbrace{E[Y_{1,i}|D_i=1]-E[Y_{0,i}|D_i=1]}_{ATT}$$
$$\underbrace{+E[Y_{0,i}|D_i=1]-E[Y_{0,i}|D_i=0]}_{\text{选择性偏误}}$$

选择性偏误是处理组在不接受政策干预时的潜在结果与对照组结果的差异。通过上式可以清楚地看到，当选择性偏误不为零的时候，传统的政策评估研究伴随着高估或者低估政策因果效应的风险。

那么如何消除选择性偏误呢？以数据获得的方式可以分为实验方法与准实验方法。其中，最理想的情形是随机实验。当政策选择变量 D_i 与潜在结果变量 Y_i 相互独立时，$E[Y_{0,i}|D=1]-E[Y_{0,i}|D=0=E[Y_{0,i}]-E[Y_{0,i}]=0$。换句话说，当个体是随机地接受政策干预时（或者称为完全随机政策实验），研究人员可以直接用观测效应来估计政策的处理效应。退一步，如果存在一个协变量 X 使得 D 与 Y 相对于协变量 X 条件独立，可以局部地消除选择性偏误，Angrist 和 Pischke（2008）将其称为条件独立性假设（Conditional Independent Assumption，CIA）。CIA 成立时，即 $E[Y_{0,i}|D_i=1,X_i]-E[Y_{0,i}|D_i=0,X_i]=0$，根据期望迭代定律，政策的处理组平均处理效应可以通过局部政策处理效应的期望来得到，即 $E[E[Y_{1,i}|D_i=1,X_i]-E[Y_{0,i}|D_i=0,X_i]$。CIA 最常见的情形是随机受控实验（Randomized Controlled Experiment）。在随机受控实验中，决策者首先按照一定条件（即协变量 X）对个体进行分组，然后在每组内随机决定哪些个体进入实验。注意，此时不同个体进入实验（或者接受政策干预）的概率不一定相同（区别于完全随机实验）。但是一旦两个个体属于同一分组（即协变量 X 相等），其进入实验的概率便相同，因此 CIA 成立。

无论是完全随机政策实验还是随机受控实验，研究人员均可以直接通过观测效应来评估政策的因果效应。因此，政策实验被视为可信度最高的政策评估研究方法①。但是，实验方法也有明显局限。首先，政策实验实施起来成本高，时间长。政策实验几乎是可遇不可求的事情。其次，技术上讲，对照组需要完全不受政策作用，但实际上仍会存在政策的溢出效应、替代效应等问题。最后

① 如 King 等人对墨西哥的全民医保政策组织了非常严谨的政策评估，从 2006 年开始使用社会实验方法进行了历时数年的评估，是有史以来世界范围内覆盖最大的社会实验之一。其以随机分配为核心的政策评估可信度非常高，为全国层次的医疗卫生政策提供了宝贵的经验（和经纬，2009）。

也是最重要的，政府部门不能轻易将其用作政策实施前的决策依据。此外，实验本身伴随着风险，必然承受着社会压力，因此政府对待政策实验的态度是慎之又慎的。

绝大部分政策评估研究中，我们需要寻找其他方式来消除选择性偏误。政策实践中，个体（无论是个人还是组织）并非通过随机委派的方式接受政策干预。例如，多数情况下政策的实施是区域性的，区域中的个体会预测自己接受政策干预时可能获得的净收益，进而通过迁移等反应来决定是否接受政策影响。此时只能使用观测数据（observed data）来估计政策的因果效应，使用观测数据进行政策评估研究需要特别注意选择性偏误问题。研究人员为了解决这一问题，开发出了一系列研究工具，这些工具可以统称为准实验（Quasi-Experiment）方法。

准实验方法将政策实施视为一项“实验”，或通过为接受政策干预的处理组寻找一个对照组，或运用其他数学方法来消除选择性偏误。简单来讲，可以将准实验方法分为两类，第一类是在 CIA 的启发下，寻找对照组，包括匹配（Matching）和双重差分（Difference in Differences，DID）。第二类是工具变量方法（Instrumental Variables，IV）及其启发下的断点回归方法（Regression Discontinuity，RD）。不同的准实验方法各有利弊，例如工具变量方法虽然在数学性质上清楚，但寻找工具变量确实是一个非常需要“运气”的事情。研究人员需要根据实际的观测数据类型和政策情景选择合适的方法来进行政策评估研究。在众多准实验方法中，匹配是最直观也是应用范围最广的方法，因此，本文中我们将特别介绍当前主要使用的匹配方法的原理以及应用。

五、匹配方法的原理与应用

匹配方法是最直观且应用范围最广的准实验方法，已成为公共政策评估研究的主要工具（Autio & Rannikko，2016；Holt，2019；Rosholm et al.，2019；李绍平等，2018；刁伟涛、任占尚，2019）。本节中我们先介绍匹配的一般思路，然后分别介绍五种常见的匹配方法。所有匹配方法的理论假设都是 CIA，即 $[Y_{0,i}\,|\,D_i=1,\ X_i]-E\ [Y_{0,i}\,|\,D_i=0,\ X_i]\ =0$，此时 $ATT=E\ [E\ [Y_{(1,i)}\,|\,D_i=1,\ X_i]-E\ [Y_{0,i}\,|\,D_i=0,\ X_i]]$，换句话说，当我们已知 X_i 与 D_i 的联合分布时，公共政策的因果效应可以分两步进行估计。第一步，计算给定 X_i 时，处理组与对照组的潜在结果差异；第二步，以 X_i 的概率为权重对上一步中得到的差异求平均，这种方法称为基于协变量的匹配。这其中，第一步实际上就是选择可以互相比较的处理组与对照组样本。在一般意义上，匹配的基本原理在于找到与处理组

尽最大程度“相似”（协变量意义上）的个体作为对照组。当处理组和对照组分别具有足够的可观测变量时，其结果的差异就取决于是否接受政策干预。单一协变量匹配如图 6 所示。

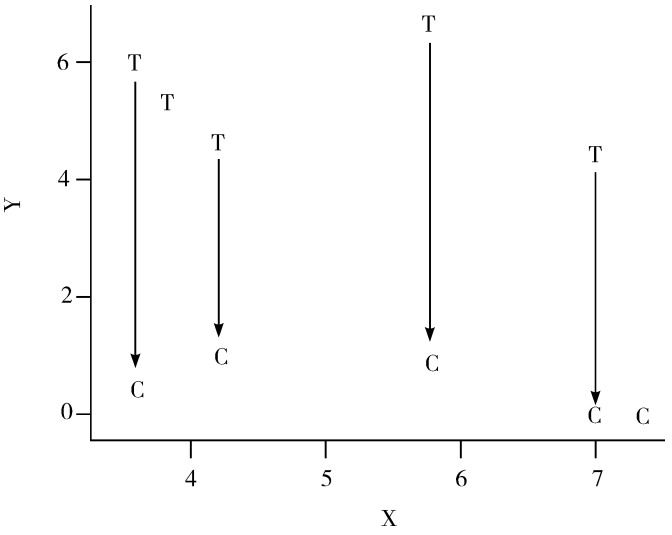

图 6　单一协变量的匹配示意

资料来源：作者自制。

将协变量匹配的思想稍作推广便可以得到匹配方法的一般思路。匹配方法主要部分均包含两个步骤，第一步是计算个体之间（也可以是群体之间）的距离（或者相似性），第二步是根据前面计算得到的距离，采取一定的配对方法进行匹配。当然在完成匹配后需要检查处理组与对照组之间的差异性（样本平衡性），如果满足预先设定的目标，即完成匹配；若没有完成，需要重新回到第一步开始匹配。目前，应用较多的计算距离的方法有四种：协变量、马氏距离、倾向值距离和信息熵增量，配对方法则可以分为完全配对、区间配对、最邻近配对、卡钳配对、局部平均回归和整体权重配对六种。根据这两个维度，可以将主要匹配方法归类如下（见表 4）。

表 4　主要匹配方法的比较

匹配方法	距离算法	配对方案	优势	劣势
协变量匹配	协变量	完全配对	确保 CIA 成立	局限于离散变量、依赖样本数量
粗糙完全匹配	协变量	区间配对	适用连续变量	依赖样本数量

（续上表）

匹配方法	距离算法	配对方案	优势	劣势
马氏距离匹配	马氏距离	区间配对、最邻近配对/卡钳配对	匹配率高	无法确保 CIA 成立
倾向值匹配	倾向值距离	区间配对、最邻近配对/卡钳配对、局部平均回归整体权重、配对	确保 CIA 成立、匹配率高	模型依赖性强
熵平衡匹配	信息熵增量	最邻近配对	确保样本平衡	理论不清晰

资料来源：作者自制。

（一）协变量匹配（Covariant Matching，CVM）

如前文所述，从 CIA 框架出发，协变量匹配是最平凡的匹配方法。协变量匹配使用协变量的联合分布作为距离计算的基准，个体之间的距离只有两个取值 0 或者 $+\infty$，即个体 i 和 j 的距离定义为：

$$d(i,j)=\begin{cases}0, & \text{当 } X_i=X_j \text{ 时}\\ +\infty, & \text{当 } X_i\neq X_j \text{ 时}\end{cases}$$

式中 X_i 为样本 i 的协变量。当 $d(i,j)$ 为 0 的时候，i 与 j 被匹配在一组。其实，协变量匹配法会将所有协变量相等的个体归为一组，最后根据 $E\{E[Y_{1,i}|D_i=1,X_i]-E[Y_{0,i}|D_i=0,X_i]\}$ 来计算平均处理效应。Angrist（1998）在研究服兵役对收入的影响时曾使用服役申请人的出生年份、受教育水平以及服役考试的分数作为协变量进行匹配，解决了本来存在的正向选择性偏误问题。国内研究中，胡吉祥等（2011）基于 1998—2007 年中国制造业企业层面的面板数据，从盈利、效率、投资三个方面，考察了国有企业公开上市对企业绩效的影响。作者发现公开上市的企业绩效，在上市之前表现就相对较好。直接对比公开上市和未公开上市的企业绩效，会导致对结果估计的偏差。并且，政府选择上市企业时具有倾向性，导致国有企业上市的选择并非随机。因此，作者使用了企业在上市前的盈利、杠杆率、规模、行业和区域特征等协变量为上市企业匹配了在上市前特征相近的对照组，来减少选择性偏误。研究发现，公开上市提高了企业的盈利水平、运营效率和投资能力。

基于协变量的匹配方法，优点在于思想简单直接。但是其适用场景受限，存在两个局限性。一是仅适用于离散变量的情形，二是随着协变量的增多，匹配成功的样本比例会迅速降低，设想存在 n 个二值协变量，那么观测值会被分

成在 2^n 组，当 n 非常大时，很难成功匹配处理组与对照组。

（二）粗糙完全匹配（Coarsened Exact Matching，CEM）

Iacus 等（2012）提出的粗糙完全匹配是协变量匹配的一种变体。CEM 的基本思路是将连续变量根据其分布切割成为若干区间，从而与离散变量一起将协变量空间分成若干栅格（Strata）。每个个体唯一落在其中一个栅格内，落在同一个栅格内的个体间距离定义为 0，否则为 $+\infty$。最后，采用完全匹配的策略，仅保留同时包含了处理组与对照组的栅格①。

史冬波（2016）在对国家杰出青年科学基金的评估中应用了 CEM 方法，研究使用科学家的最高学历、最高学历国家、年龄、性别、从国家自然科学基金委获得课题经费数量、所在机构的学科经费占比、发表论文数量与被引数量作为协变量，从获得国家自然科学基金面上项目的科学家中选取获得杰青基金科学家的对照组，有效解决了选择性偏误问题。刘玮辰等（2020）使用 CEM 方法研究了青年科学家的跨国流动对知识扩散的影响。CEM 的过程分为两步。第一步，作者为回国的青年科学家匹配了来自同一博士院校、同一专业、毕业年份相近但没有回国的科学家作为对照组。第二步，作者在回国前的发文数量、回国前加权期刊影响因子的发文数量、回国前第一作者发文数量和最后作者发文数量四个协变量上进行匹配，将连续变量切成不同区间，保留了在同一格子内的处理组和对照组作为最后的样本进入实验。研究发现，中国青年科学家的跨国流动对其知识扩散有正向作用。

CEM 通过连续变量离散化克服了协变量匹配的第一个缺陷，同时有效降低了干预效果对于模型的依赖性（Blackwell et al.，2009），这使得 CEM 的适用范围更广。CEM 中，栅格划分越细，匹配越精确，但是栅格数量也越多，从而对样本量依赖越大。CEM 通过灵活的栅格宽度选取在一定程度上减少了对样本量的依赖，但是其匹配成功率依然很低，对样本量的依赖依然很大。

（三）马氏距离匹配（Mahalanobis Metric Matching，MMM）

马氏距离匹配是利用马氏距离的匹配方法（Rubin，1980）。马氏距离是印度学者马哈拉诺比斯（P. C. Mahalanobis）提出的计算 n 维空间中两个点的协方差距离。首先计算处理组个体 i 与对照组个体 j 之间的马氏距离 d（i，j）可用下

① 每一个处理单元的权重是 1，每一个控制单元的权重等于将其层级中的处理单元的数量除以同一层级控制单元的数量，然后标准化。这样权重加总等于全部匹配的样本大小。

面公式计算：

$$d_{(i,j)} = (X_i - X_j)^T C^{-1} (X_i - X_j)$$

式中，X_i 和 X_j 分别表示 i 和 j 的协变量，C 是总体的协方差矩阵。两组个体间的马氏距离越小，则表示两组个体越相似，协变量之间的分布越均衡，匹配的效果越好，从而控制混杂因素的影响。马氏距离将高维空间的两个样本点之间的距离降维成一个单维距离。然后，可以采用最邻近匹配（Nearest Neighbour Matching）和卡钳匹配（Caliper Matching）的方法来寻找对照组。最邻近匹配指先将处理组的研究对象随机排序，然后从处理组的第一个研究对象开始，为其在对照组寻找一个马氏距离最小的个体作为匹配对象，直到所有处理组均在对照组找到匹配对象。卡钳匹配是指在最邻近匹配的基础上，设定一个卡钳值，只有当不同组间个体的倾向性评分值之差小于或等于卡钳值时才允许匹配。卡钳匹配保证了匹配的差异不会太大，同时也会减小匹配成功的比例，损失更多的样本信息。

马氏距离匹配成功解决了协变量匹配的两个问题。杨德斌（2016）在探析中国企业跨国并购的真实绩效问题时利用中国工业企业数据库，运用马氏距离匹配方法为具有跨国并购行为的企业寻找对照组，作者使用的协变量是包括企业全要素生产率、企业规模、资本密度和企业所属行业，这些都会显著影响企业是否进行跨国并购。马氏距离匹配后，实验组与对照组在协变量维度上不存在显著差异，从而减少了选择性偏误。研究发现，跨国并购显著提升了工业企业生产率，并且随着时间的推移，提升作用愈发明显。王孝松等（2020）从厂商层面考察了企业异质性对其遭遇反倾销贸易壁垒的影响。对于每一个遭遇反倾销诉讼的出口企业，作者使用马氏距离匹配方法筛选出了同年在该行业中未遭遇反倾销诉讼的出口企业作为对照组。作者选取了企业规模、融资能力、盈利水平、出口交货值和中间投入五个指标作为匹配变量。匹配后，减少了选择性偏误。研究发现，企业生产率是中国厂商面临反倾销诉讼的主要影响因素，低生产率的企业更容易在国际市场遭遇反倾销诉讼。

马氏距离匹配的优势在于匹配率高，应用范围广泛。相比于倾向值匹配，它的运算更简便。但是从理论上来看，是否两个马氏距离相近的样本点一定可以匹配却没有定论。换句话说，无法证明当 X 满足 CIA 时，$M(X)$ 也满足 CIA，这使得马氏距离匹配虽然在实际操作中具有更强的便利性，但却是以损失理论完备性为代价的。

（四）倾向值匹配（Propensity Score Matching，PSM）

倾向值匹配使用倾向值作为距离计算标准，是另一种将高维协变量转化为

一维距离的匹配方法。Rosenbaum 和 Rubin（1983）在 1983 提出 PSM 方法，并证明了著名的倾向值定理，即如果协变量 X 能使得条件独立假设成立，那么协变量倾向值函数 $p(X_i) = p(D_i = 1 \mid X_i)$ 也能保证条件独立假设成立。倾向值定理保证了 PSM 与 CVM 在数学上完全等价，并且成功克服了 CVM 的两个局限，这使得 PSM 成为应用最广泛的匹配方法。实际操作中，PSM 有两种情况，一种是当确定研究人员了解处理组的选择机制时，可以直接得到真实的倾向值计算个体间距离；另一种情况是处理组的选择机制并不确定，那么就需要研究人员首先通过 Logit 回归估计实验选择机制，并以估计的倾向值（Estimated Propensity Score）代替真实的倾向值计算个体间距离。在计算出距离后，可以利用最邻近匹配和卡钳匹配进行配对，也可以利用局部线性回归进行配对。后者是利用倾向值的核密度函数（Kernel Density）作为倾向值概率分布密度函数的估计，选取某个宽度阈值来定义处理组个体的局部区间，然后将区间内的处理组与对照组进行加权线性回归来计算局部的处理效应，即 $\hat{E}[Y(0)] \mid D = 1] = \frac{\sum_{\{i \mid D=0\}} Y_i d_i}{\sum_{\{i \mid D=0\}} d_i}$。其中，对照组中的每一个个体则被赋予权重为 $d_i = \frac{\hat{p}(x_i)}{1 - \hat{p}(x_i)}$，并通过 probit 或 logit 回归计算出 $\hat{p}(x_i)$ 为个体 i 的倾向值①。

PSM 方法作为评估政策效应的有效工具，被应用于多个政策评估领域。国外有大量研究广泛应用 PSM 进行政策和项目评估，如 Heckman 等人（1997）运用 PSM 方法对一项就业培训项目的效应进行了实证分析，结果表明来自于不可观测因素的选择偏差（Selection Bias）只占误差的很小部分。Gilligan 和 Hoddinott（2007）对埃塞俄比亚农村实施的应急食品救援政策进行了效应评估。在国内的公共政策评估研究中，倾向值匹配是常用方法。李彰（2017）评估了 863 计划对企业创新的影响。在公共资源的配置中，资助对象的选择并非随机。政府很可能倾向于选择那些即使未获得政府资助也有较大概率成功的项目或企业，从而导致高估资助效果。直接比较受资助企业和未受资助企业在创新产出上的差异，会带来选择性偏误。作者采用了 PSM-DID 来估计政策的平均处理效应。具体的操作步骤，首先选取企业所有制、企业规模、企业年龄、资本密集度、营业利润率等变量作为匹配的协变量，通过 logit 回归估计了企业获得 863 计划资助的倾向值；然后采用二次核匹配把倾向值相近的企业进行匹配来控制企业资质的差异，匹配范围仅限于满足共同取值假定的样本企业；随后对样本

① 实际操作中 logit 与 probit 模型的结果一般不会有显著差异。

进行筛选，剔除倾向值过高的样本，保证匹配后各协变量组间无差异，确定对照组样本；最后计算对照组样本在干预前后的变化，并计算出干预效应 ATT。研究发现，863 计划的资助显著提升了企业的创新绩效。

倾向值匹配是目前政策评估研究中的主要匹配方法。当我们对那些影响干预变量的其他混淆变量具有清楚的理论支持时，或对需要纳入模型中的混淆变量有清楚的认识时，倾向值匹配比较适用。但是，PSM 在应用中仍存在一些挑战，其中最大的挑战是模型依赖性。在实际评估中，倾向值往往是不清楚的，研究人员需要首先估计倾向值，这就造成 PSM 的效果对其估计模型的选择极其敏感（King et al.，2011），在实际应用中难以形成客观的标准。

（五）熵平衡匹配（Entropy Balancing Matching，EBM）

熵平衡匹配是一种另辟蹊径，以信息熵增量为距离进行整体匹配的思路（Hainmueller et al.，2012）。熵平衡采用了与其他匹配方法相反的思路。CVM、CEM、MMM、PSM 等方法是首先匹配个体，再检验处理组与对照组的平衡性，最后根据平衡表调整匹配，例如调整 CEM 的区间长度或者修正倾向值估计中的模型。EBM 则在控制处理组与对照组矩相等的情况下，优化对照组个体的权重。不同权重赋值的对照组可以看成是一组新的对照组，以信息熵增量作为处理组与对照组的距离，最后选择熵增量最小的一组权重作为匹配结果，来估算反事实项，即 $\hat{E}[Y(0)\,|\,D=1]=\dfrac{\sum_{\{i|D=0\}}Y_i w_i}{\sum_{\{i|D=0\}}w_i}$，其中 w_i 为对照组样本权重。

近几年熵平衡匹配法得到了越来越多的应用，其基本步骤是：首先对可能导致偏误的特征变量设定一系列矩条件，对处理组和对照组数据进行平衡，获得相应的权重，之后再利用该权重进行加权后的回归分析。Truex（2014）在关于企业高管成为人大代表为企业带来附加收益的研究中，利用熵平衡方法构建了一个中国企业的加权投资组合，使之与全国人大代表为主管的企业在财务特征方面相匹配。通过加权固定效应分析，全国人大的一个席位会带来相应的收益回报，证实了威权政治理论的假设。马恩和王有强（2019）采用熵平衡匹配方法，结合双重查分模型对开发区政策对企业创新的影响进行了分析，即企业进入开发区后创新绩效的变化。然而，企业是否进入开发区，并非是随机的，存在选择性偏误。一方面，地方政府会为了特定的政策目标，对一些企业或产业重点扶持；另一方面，企业会权衡开发区的条件，从而作出是否进入开发区的决策。作者选取了不在开发区的企业作为对照组，选取了企业年龄、企业规模、出口、盈利能力、资产负债率等协变量进行匹配。用熵平衡计算出的权重

为对照组进行加权。熵平衡后，处理组和对照组在实验年份前的企业平均创新绩效的变化趋势保持一致，确保得到了两组平衡的数据。研究发现，开发区政策对企业创新有显著的正向促进作用。

相比于其他匹配方式，熵平衡匹配在样本特征的平衡、样本信息的保留以及提高计算速度方面都有其优势。但也存在一定的缺陷。例如理论上并不确定熵平衡的样本是否满足 CIA 条件，另外，当处理组与对照组总体的差异过大，或者对照组中的极端情况过多时，熵平衡则难以得到有效解。

总的来讲，匹配方法是当前基于反事实思想用于公共政策评估研究的主要方法，在使数据平衡的过程中，不同的匹配方法展现出了各自的优势和适用场景。归纳来讲，这类问题的一个共性特点是我们要估计一个二分型变量对于另外一个变量的因果效应（胡安宁，2012）。因此，匹配方法从性质和目标上来讲具有一致性，即为了控制和消除选择性偏误，将各个协变量在处理组和对照组之间达到数据平衡的状态。

六、针对匹配方法的讨论

匹配方法是若干种消除选择性偏误的方法中思想最朴素的，通过寻找与处理组最接近的对照组来进行因果推断。因此，匹配方法在社会科学中得到了大规模的应用。但需要注意的是，在实际研究中仍有一些问题是容易被研究人员忽略的，例如匹配方法的适应性、条件独立假设的稳健性以及样本的依赖性等。下面我们将逐一讨论匹配方法在实际应用中需要注意的问题。

（一）匹配方法的适用性

匹配方法在多大程度上可以消除选择性偏误？这是关乎匹配方法适用性的关键问题。在社会科学中，实验是因果识别的理想选择（Fisher，1951），也是检验准实验方法适用性的黄金准则。自 Lalonde（1986）起，使用实验数据来评估非实验估计量成为基点，Dehejia 和 Wahba（2002）将倾向值匹配方法的估计结果与 Lalonde（1986）使用的实验数据进行了比较后发现两者之间不存在显著差异，从而使研究人员对匹配方法的信心大幅提升。然而好景不长，Smith 和 Todd（2005）使用同样的数据证实 Dehejia 和 Wahba（2002）的结论仅仅是一个巧合，PSM 方法并不能估计出准确的因果效应。理论上，匹配方法的适用性依赖于条件独立性假设，匹配方法仅仅可以消除可观测变量带来的偏误，而不能消除不可观测变量带来的偏误。当存在遗漏变量时，匹配法是无能为力的。

（二）关于匹配和回归的关系

线性回归是最常用的定量方法，那么回归与匹配是什么关系呢？可以证明，当条件独立性假设（CIA）成立，线性回归的系数也是对政策因果效应的无偏估计。此时可以证明，线性回归也可以看成是针对局部因果效应的一个加权平均。在这个意义上，回归与匹配是等价的。但是，回归最主要的缺陷是对于处理组和对照组之间的不平衡没有很好的检测，常常对观测数据外推。而匹配恰好解决了这个问题。

（三）匹配方法对样本数量的要求

这是匹配方法最大的局限之一，在数据量较少或者协变量较多的情况下，会存在所谓的“维度诅咒”问题。即基于协变量的匹配会造成样本数量骤减，无法满足匹配条件，给因果推断的准确性带来极大挑战。同时，基于协变量的匹配面临计算上的困难。

（四）是否允许放回

这是匹配法实际操作中涉及的问题。允许放回是指对照组的个体成功匹配后仍可以放回参与下一次的匹配。如果允许放回，可提高匹配成功率。但允许放回会导致匹配后的数据集中包含重复的研究对象，重复出现的个体会对因果推断有更大影响，降低结果的有效性。因此，在实际应用中，如果不存在个体重复率过高的问题时，放回与否对结果的影响并不显著。当然如果在样本数量足够的情况下，一般采取不放回匹配。

七、总结

本文从因果推断的视角出发，构建了公共政策评估研究的反事实理论框架，比较了不同估计方法的差异，特别是深入讨论匹配方法的理论与应用，为公共政策评估研究方法论体系化做了初步尝试。

匹配方法与工具变量、双重差分、断点回归一起构成了公共政策评估研究的准实验方法。准实验方法在公共政策评估研究中具有显著优势（Grant & Wall，2009）。首先，当无法进行随机分组或违背伦理时，准实验可以有效地增强因果推论。其次，准实验可以尽可能降低伤害、不公平、欺骗等引发的道德困境。最后，准实验有利于促进研究人员与实践者的合作等。但是，准

实验方法在其应用的过程中，也存在各种困境和挑战（王思琦，2018）。例如，本文提到的倾向值匹配方法虽然解决了传统协变量匹配的困境，但对于样本量大小要求较高，匹配过程仍是基于可观察的协变量进行，难以控制不可观察因素的影响。

与准实验方法相对应的是实验法。实验法是揭示因果关系的新方法。实验法的控制、随机性、操控等特征可以最大程度地有助于得到“纯粹”的因果关系。从数据收集方式来讲，实验法具有三项显著优势（耿曙等，2016）。其一，实验方法能够有效控制干扰因素的影响，得到纯粹的因果关系。并且，也便于将实验过程在不同情境、不同实验对象中进行重复研究。其二，从数据收集的成本来讲，实验方法更容易获得质量较高的微观数据用于观察个体行为，比大规模调查访谈更节省人力时间成本。其三，小规模的实验方法相较于大规模调查研究更有助于减少测量误差，强化测量的信度与效度。实验法的科学性和严谨性推动了研究者们在公共管理和政策领域不断尝试并初有成效，对于方法论的探讨和理解也在不断加深并推广，实验法在公共政策和管理领域的应用也已逐渐得到学术界的认同。将实验方法应用于公共政策研究，将是未来公共政策研究发展的重要趋势。但是，实验方法在运用中也存在一定的局限性。一方面，实验研究的结果是否具有可推广性在公共政策和公共管理领域备受争议；另一方面，实验方法在个体层面更易于操作，但上升到团队、组织以及政策、制度甚至包括国家等宏观层面的研究则限制了该方法的应用和推广。

在中国场景中，大量的政策创新、变革和实验，都为公共管理和公共政策研究者提供了绝佳的实验研究场域和话题（马亮，2017），基于因果推断的政策评估研究将大有作为。近年来，在卫生与教育政策（Glewwe et al.，2016；Lu & Anderson，2015；Luo et al.，2015；杨钋、徐颖，2017）、科技政策（李彰，2017；刘玮辰等，2020；吴翌琳、黄筝 2018）、社会政策（罗仁福等，2013；齐良书、赵俊超，2010；王增文、邓大松，2012）与劳工政策（郝明松，2020；周茂等，2019）等领域，我国学者都开始使用因果推断的方法来进行政策评估研究，提高了国内政策评估研究的可信度，以更具融合性与创新性的视野推动政策评估研究范式与方法的本土化建构，以更加开放多元的话语来讲述中国故事，向世界传播中国政策经验。

但是，在中国场景中开展基于因果推断的政策评估研究仍然面临不少制约。首先，我国政府相关数据的开放程度不足，数据可获得性较差，直接制约了政策评估研究的开展。其次，我国的公共政策繁多，在评估单一政策的效应时难以剥离其他政策的影响。再次，条块分割的治理结构导致政策在地方落实上呈

现很大的差异性，政策执行过程中的多方行动者带有的主观因素及行政干预因素致使评估研究具有极大的主观性和建构性（张晓丰、赵峰，2012）。最后，也是最重要的，从研究主体来看，没有经过严谨的研究方法训练的研究者仍然占据多数，国内学界对基于反事实框架的政策评估研究方法了解还不够深入，具体使用中存在误用。要克服上述制约，需要公共管理学者对政策研究的方法论体系化做出更多研究。虽然相关的方法论讨论仍然有限，但毋庸置疑，随着反事实框架和因果推断在公共政策评估研究中逐渐得到认可与运用，有关其方法论的研究与讨论，势必发展成为公共政策研究的重要议题。

参考文献

陈光、方新（2014）. 关于科技政策学方法论研究. 科学学研究，32(03)：321－326.
Chen, G. & Fang, X.（2014）. Methodology Study on the Science of Science and Technology Policy. *Studies in Science of Science*, 32(03)：321－326.（in Chinese）

刁伟涛、任占尚（2019）. 公众参与能否促进地方债务信息的主动公开——一项准实验的实证研究. 公共行政评论，12(05)：100－121.
Diao, W. T. & Ren, Z. S.（2019）. Can Initiative Disclosure of Local Government Debt Be Promoted by Public Participation? —An Empirical Study Based on a Quasi-Experiment. *Journal of Public Administration*, 12(05)：100－121.（in Chinese）

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，5(03)：45－52.
Geng, S., Yu, S. & Kong, Y.（2016）. Studying Public Policy with Experimental Methods：Lessons from a Tax Compliance Experiment. *Public Administration and Policy Review*, 5(03)：45－52.（in Chinese）

高雪莲（2009）. 政策评价方法论的研究进展及其争论. 理论探讨，5：139－142.
Gao, X. L.（2009）. The Research Progress and Controversy of Policy Evaluation Methodology. *Theoretical Investigation*, 5：139－142.（in Chinese）

郝明松（2020）. “找关系”的作用及其边界——基于人职匹配过程的交叉检验研究. 社会科学战线，2：224－236.
Hao, M. S.（2020）. The Function and Boundary of “Relationship Seeking”—A Cross-check Study Based on the Process of Job Matching. *Social Science Front*, 2：224－236.（in Chinese）

和经纬（2008）. 中国公共政策评估研究的方法论取向：走向实证主义. 中国行政管理，9：118－124.
He, J. W.（2008）. Methodological Orientation for China's Public Policy Evaluation：Towards Positivism. *Chinese Public Administration*, 9：118－124.（in Chinese）

和经纬（2009）. 全国性医疗卫生政策评估的方法论策略——墨西哥全民医保政策评估的经验. 公共管理评论，8：161－170.
He, J. W.（2009）. The Methodological Strategy for National Health Policy Evaluation：The Case of Policy Evaluation of Mexico's Seguro Popular de Salud, *China Public Administration Review*, 8：161－170.（in Chinese）

胡安宁（2012）. 倾向值匹配与因果推论：方法论述评. 社会学研究，1：221－242.
Hu, A. N.（2012）. Propensity Score Matching and Causal Inference：A Methodological Review.

Sociological Studies, 1: 221 – 242. (in Chinese)

胡吉祥、童英、陈玉宇(2011). 国有企业上市对绩效的影响：一种处理效应方法. 经济学(季刊), 10(03): 965 – 988.

Hu, J. X., Tong, Y. & Chen, Y. Y. (2011). Evaluating the Effects of Public Listing on SOEs' Performance in China: A Treatment Effect Approach. *China Economic (Quarterly)*, 10(03): 965 – 988. (in Chinese)

李绍平、李帆、董永庆(2018). 集中连片特困地区减贫政策效应评估：基于 PSM – DID 方法的检验. 改革, 298(12): 142 – 155.

Li, S. P., L. F. & Dong, Y. Q. (2018). The Impact Evaluation of the Policy of Poverty Alleviation in Concentrated Poverty-stricken Areas: An Investigation Based on PSM – DID Method. *Reform*, 298(12): 142 – 155. (in Chinese)

李彰(2017). 科技计划对企业创新的作用评估：基于 863 计划的实证研究. 公共管理评论, 1: 39 – 54.

Li, Z. (2017). Evaluation of an R & D Program on Firm Innovation: Evidence from the 863 Program. *China Public Administration Review*, 1: 39 – 54. (in Chinese)

刘骥、张玲、陈子恪(2011). 社会科学为什么要找因果机制——一种打开黑箱、强调能动的方法论尝试. 公共行政评论, 4: 50 – 84.

Liu, J., Zhang, L. & Chen, Z. K. (2011). Why Social Science Needs to Pursue Causal Mechanisms. *Journal of Public Administration*, 4: 50 – 84. (in Chinese)

刘玮辰、郭俊华、史冬波(2020). 科学家跨国流动促进了知识扩散吗？——基于青年千人的实证分析. 图书情报知识, 2: 32 – 41.

Liu, W. C., Guo, J. H. & Shi, D. B. (2020). International Mobility and Knowledge Diffusion: An Empirical Study of the Thousand Youth Talents Plan. *Documentation, Information & Knowledge*, 2: 32 – 41. (in Chinese)

罗仁福、张林秀、王晓兵、易红梅、史耀疆(2013). 家长健康信息干预对贫困农村学生贫血的影响. 中国学校卫生, 2: 225 – 227.

Luo, R. F., Zhang, L. X., Wang, X. B., Yi, H. M. & Shi, Y. J. (2013). The Impact of Parental Health Information Intervention on Anaemia in Poor Rural Students. *Chinese Journal of School Health*, 34(2): 225 – 227. (in Chinese)

马恩、王有强(2019). 区位导向性政策是否促进了企业创新？——以我国开发区政策为例. 科技管理研究, 11: 35 – 42.

Ma, E. & Wang, Y. Q. (2019). Does Place-based Policy Promote Firms Innovation: Evidence from China's Development Zones? *Science and Technology Management Research*, 11: 35 – 42. (in Chinese)

马亮(2017). 实证公共管理研究日趋量化：因应与调适. 学海, 5: 194 – 201.

Ma, L. (2017). Empirical Public Management Research is Becoming More and More Quantitative: Response and Adaptation. *Academia Bimestrie*, 5: 194 – 201. (in Chinese)

齐良书、赵俊超(2012). 营养干预与贫困地区寄宿生人力资本发展——基于对照实验项目的研究. 管理世界, 2: 52 – 61.

Qi, L. S. & Zhao, J. C. (2012). Nutritional Interventions and Human Capital Development for Boarders in Poor Areas: A Study Based on a Controlled Trial Project. *Management World*, 2: 52 – 61. (in Chinese)

史冬波(2016). 科研资助与科研产出：以国家自然科学基金为例(学位论文). 北京：清华大学.

Shi, D. B. (2016). *Funding and Research Output: National Science Foundation for*

Distinguished. Young Scholars (*Dissertation*). Peking: Tsinghua University. (in Chinese)

王思琦（2018）. 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. 公共行政评论，1：87－107.

Wang, S. Q. (2018). Field Experiments in Research of Public Administration and Public Policy: Causal Inference and Impact Evaluation. *Journal of Public Administration*, 1: 87－107. (in Chinese)

王孝松、林发勤、李玏（2020）. 企业生产率与贸易壁垒——来自中国企业遭遇反倾销的微观证据. 管理世界，36(09)：54－67.

Wang, X. S., Lin, F. Q. & Li, L. (2020). Productivity and Trade Barrier: Evidence from Chinese Firms Encountering Antidumping Petitions. *Management World*, 36(09): 54－67. (in Chinese)

王增文、邓大松（2012）. 倾向度匹配、救助依赖与瞄准机制——基于社会救助制度实施效应的经验分析. 公共管理学报，9(2)：83－88.

Wang, Z. W. & Deng, D. S. (2012). Propensity Score Matching, Aid Dependence and Aiming Mechanism: Based on the Experience Analysis to the Implementation of Social Assistance System. *Journal of Public Management*, 9(2): 83－88. (in Chinese)

吴翌琳、黄箐（2018）. 基于倾向得分匹配法的创业政策实证研究——以财税政策评估为例. 宏观经济研究，9：123－138.

Wu, Y. L. & H, Z. (2018). Empirical Research on Entrepreneurship Policy Based on Propensity Score Matching Method—Taking Fiscal and Tax Policy Evaluation as an Example, *Macroeconomics*, 9: 123－138. (in Chinese)

杨德彬（2016）. 跨国并购提高了中国企业生产率吗？——基于工业企业数据的经验分析. 国际贸易问题，4：166－176.

Yang, D. B. (2016). Do Cross-border M & As Boost the Productivity of Chinese Enterprises? An Empirical Analysis. *Journal of International Trade*, 4: 166－176. (in Chinese)

杨钋、徐颖（2017）. 数字鸿沟与家庭教育投资不平等. 北京大学教育评论，15(4)：126－154.

Yang, P. & Xu, Y. (2017). Digital Divide and Inequality in Household Education Investment. *Peking University Education Review*, 15(4): 126－154. (in Chinese)

张晓丰、赵峰（2012）. 科技政策评估的方法论体系及相关性研究. 北京化工大学学报(社会科学版)，2：1－6.

Zhang, X. F. & Zhao, F. (2012). Methodology and Correlation Studies on Science and Technology Policy Evaluation. *Journal Beijing University of Chemical Technology* (*Social Science Edition*), 2: 1－6. (in Chinese)

周茂、李雨浓、姚星、陆毅（2019）. 人力资本扩张与中国城市制造业出口升级：来自高校扩招的证据. 管理世界，35(05)：64－77.

Zhou, M., Li, Y. N., Yao, X. & Lu, Y. (2019). Human Capital Accumulation and Urban Manufacturing Export Upgrading in China: Evidence from Higher Education Expansion, *Management World*, 35(05): 64－77. (in Chinese)

Amuedo-Dorantes, C. & Arenas-Arroyo, E. (2019). Immigration Enforcement and Children's Living Arrangements. *Journal of Policy Analysis and Management*, 38(1): 11－40.

Andersen, S. C. & Hjortskov, M. (2016). Cognitive Biases in Performance Evaluations. *Journal of Public Administration Research and Theory*, 26(4): 647－662.

Angrist, J. D. & Keueger, A. B. (1991). Does Compulsory School Attendance Affect Schooling and Earnings? *Quarterly Journal of Economics*, 106(4): 979－1014.

Angrist, J. D. (1998). Estimating the Labor Market Impact of Voluntary Military Service Using Social Security Data on Military Applicants. *Econometrica*, 66(2): 249－288.

Angrist, J. D. & Keueger, A. B. (2001). Instrumental Variables and the Search for Identification: From Supply and Demand to Natural Experiments. *Journal of Economic perspectives*, 15(4): 69 – 85.

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

Autio, E. & Rannikko, H. (2016). Retaining Winners: Can Policy Boost High-growth Entrepreneurship? *Research Policy*, 45(1): 42 – 55.

Bellé, N., Cantarelli, P. & Belardinelli, P. (2018). Prospect Theory Goes Public: Experimental Evidence on Cognitive Biases in Public Policy and Management Decisions. *Public Administration Review*, 78(6): 828 – 840.

Bertelli, A. M., Sinclair, J. A. & Lee, H. (2015). Media Attention and the Demise of Agency Independence: Evidence from a Mass Administrative Reorganization in Britain. *Public Administration*, 93(4): 1168 – 1183.

Bettinger, E. P. & Evans, B. J. (2019). College Guidance for All: A Randomized Experiment in Pre-college Advising. *Journal of Policy Analysis and Management*, 38(3): 579 – 599.

Blackwell, M., Iacus, S., King, G. & Porro, G. (2009). CEM: Coarsened Exact Matching in Stata. *The Stata Journal*, 9(4): 524 – 546.

Bound, J., Jaeger, D. A. & Baker, R. M. (1995). Problems with Instrumental Variables Estimation When the Correlation Between the Instruments and the Endogenous Explanatory Variable Is Weak. *Journal of the American Statistical Association*, 90(430): 443 – 450.

Buckles, K. & Guldi, M. (2017). Worth the Wait? The Effect of Early Term Birth on Maternal and Infant Health. *Journal of Policy Analysis and Management*, 36(4): 748 – 772.

Campbell, D. T. & Stanley, J. C. (1963). *Experimental and Quasi-experimental Designs for Research. Handbook of Research on Teaching*. Chicago, IL: Rand McNally.

Canay, I. A. (2010). Simultaneous Selection and Weighting of Moments in GMM Using a Trapezoidal Kernel. *Journal of Econometrics*, 156(2): 284 – 303.

Card, D. (1990). The Impact of the Mariel Boatlift on the Miami Labor Market. *ILR Review*, 43(2): 245 – 257.

Card, D., Katz, L. F. & Krueger, A. B. (1994). Comment on David Neumark and William Wascher, "Employment Effects of Minimum and Subminimum Wages: Panel Data on State Minimum Wage Laws". *ILR Review*, 47(3):487 – 497.

Card, D. & Krueger, A. B. (2000). Minimum Wages and Employment: A Case Study of the Fast-food Industry in New Jersey and Pennsylvania: Reply. *American Economic Review*, 90(5): 1397 – 1420.

Cárdenas, S. & Ramirez, E. E. (2017). Controlling Administrative Discretion Promotes Social Equity? Evidence from a Natural Experiment. *Public Administration Review*, 77(1): 80 – 89.

Chen, S., Mu, R. & Ravallion, M. (2008). Are There Lasting Impacts of Aid to Poor Areas? Evidence from Rural China. *Policy Research Working Paper Series*, 93(3): 512 – 528.

Chiang, H. S., Clark, M. A. & McConnell, S. (2017). Supplying Disadvantaged Schools with Effective Teachers: Experimental Evidence on Secondary Math Teachers from Teach for America. *Journal of Policy Analysis and Management*, 36(1): 97 – 125.

Cochran, W. G. & Rubin, D. B. (1973). Controlling Bias in Observational Studies: A Review. *Sankhyā: The Indian Journal of Statistics, Series A*: 417 – 446.

Cook, T. D. (2008). "Waiting for Life to Arrive": A History of the Regression-discontinuity Design in Psychology, Statistics and Economics. *Journal of Econometrics*, 142(2): 636 – 654.

Dehejia, R. H. & Wahba, S. (2002). Propensity Score-matching Methods for Nonexperimental

Causal Studies. *Review of Economics and Statistics*, 84(1): 151 – 161.

Dee, T. S. & Wyckoff, J. (2015). Incentives, Selection, and Teacher Performance: Evidence from IMPACT. *Journal of Policy Analysis and Management*, 34(2): 267 – 297.

Donohue, J. & Wolfers, J. (2005). Uses and Abuses of Statistical Evidence in the Death Penalty Debate. *Stanford Law Review*, 58: 791 – 846.

Eissa, N. (1996). Tax Reforms and Labor Supply. *Tax Policy and the Economy*, 10: 119 – 151.

Ellen, I. G., Horn, K. M. & Schwartz, A. E. (2016). Why Don't Housing Choice Voucher Recipients Live Near Better Schools? Insights from Big Data. *Journal of Policy Analysis and Management*, 35(4): 884 – 905.

Fisher, R. A. (1951). *The Design of Experiments*. Edinburgh: Oliver & Boyd Publishing.

Gilligan, D. O. & Hoddinott, J. (2007). Is There Persistence in the Impact of Emergency Food Aid? Evidence on Consumption, Food Security, and Assets in Rural Ethiopia. *American journal of Agricultural Economics*, 89(2): 225 – 242.

Glewwe, P., Park, A. & Zhao, M. (2016). A Better Vision for Development: Eyeglasses and Academic Performance in Rural Primary Schools in China. *Journal of Development Economics*, 122: 170 – 182.

Goldberger, A. S. (1972). *Selection Bias in Evaluating Treatment Effects: Some Formal Illustrations*. Wisconsin: University of Wisconsin-Madison.

Grant, A. M. & Wall, T. D. (2009). The Neglected Science and Art of Quasi-experimentation: Why-to, When-to, and How-to Advice for Organizational Researchers. *Organizational Research Methods*, 12(4): 653 – 686.

Hahn, J., Todd, P. & Van der Klaauw, W. (2001). Identification and Estimation of Treatment Effects with a Regression-discontinuity Design. *Econometrica*, 69(1): 201 – 209.

Hainmueller, J. (2012). Entropy Balancing for Causal Effects: A Multivariate Reweighting Method to Produce Balanced Samples in Observational Studies. *Political Analysis*, 20(1): 25 – 46.

Heckman, J. (1974). Shadow Prices, Market Wages, and Labor Supply. *Econometrica*, 42(4): 679 – 694.

Heckman, J. J. (1976). The Common Structure of Statistical Models of Truncation, Sample Selection and Limited Dependent Variables and a Simple Estimator for Such Models. *In Annals of Economic and Social Measurement*, 5(4): 475 – 492.

Heckman, J. J. (1979). *Statistical Models for Discrete Panel Data*. Chicago, IL: Department of Economics and Graduate School of Business, University of Chicago.

Heckman, J. J. & Robb Jr, R. (1985). Alternative Methods for Evaluating the Impact of Interventions: An Overview. *Journal of Econometrics*, 30(1 – 2): 239 – 267.

Heckman, J. J. & Robb Jr, R. (1986). *Alternative Methods for Solving the Problem of Selection Bias in Evaluating the Impact of Treatments on Outcomes. In Drawing Inferences from Self-selected Samples*. Springer: New York, NY.

Heckman, J. J., Ichimura, H. & Todd, P. E. (1997). Matching as an Econometric Evaluation Estimator: Evidence from Evaluating a Job Training Programme. *The Review of Economic Studies*, 64(4): 605 – 654.

Heckman, J., Ichimura, H., Smith, J. & Todd, P. (1998). Characterizing Selection Bias Using Experimental Data. *Econometrica*, 66(5): 1017 – 1098.

Herbst, C. M. & Tekin, E. (2016). The Impact of Child-care Subsidies on Child Development: Evidence from Geographic Variation in the Distance to Social Service Agencies. *Journal of Policy*

Analysis and Management, 35(1): 94 – 116.

Hjortskov, M. (2017). Priming and Context Effects in Citizen Satisfaction Surveys. *Public Administration*, 95(4): 912 – 926.

Holt, S. B. (2019). The Influence of High Schools on Developing Public Service Motivation. *International Public Management Journal*, 22(1): 127 – 175.

Hong, S. (2016). Representative Bureaucracy, Organizational Integrity, and Citizen Coproduction: Does an Increase in Police Ethnic Representativeness Reduce Crime? *Journal of Policy Analysis and Management*, 35(1): 11 – 33.

Iacus, S. M., King, G. & Porro, G. (2012). Causal Inference without Balance Checking: Coarsened Exact Matching. *Political Analysis*, 20(1): 1 – 24.

Jimenez, B. S. (2017). When Ties Bind: Public Managers' Networking Behavior and Municipal Fiscal Health after the Great Recession. *Journal of Public Administration Research and Theory*, 27(3): 450 – 467.

Jo, S. & Nabatchi, T. (2019). Coproducing Healthcare: Individual-level Impacts of Engaging Citizens to Develop Recommendations for Reducing Diagnostic Error. *Public Management Review*, 21(3): 354 – 375.

Kaestner, R., Garrett, B., Chen, J., Gangopadhyaya, A. & Fleming, C. (2017). Effects of ACA Medicaid Expansions on Health Insurance Coverage and Labor Supply. *Journal of Policy Analysis and Management*, 36(3): 608 – 642.

Kahneman, D. & Smith, V. (2002). Foundations of Behavioral and Experimental Economics. *Nobel Prize in Economics Documents*, 1(7): 1 – 25.

Keiser, L. R. & Miller, S. M. (2020). Does Administrative Burden Influence Public Support for Government Programs? Evidence from a Survey Experiment. *Public Administration Review*, 80(1): 137 – 150.

King, G., Nielsen, R., Coberley, C., Pope, J. E. & Wells, A. (2011). Comparative Effectiveness of Matching Methods for Causal Inference (Unpublished manuscript).

LaLonde, R. J. (1986). Evaluating the Econometric Evaluations of Training Programs with Experimental Data. *The American Economic Review*, 76(4): 604 – 620.

Lemieux, T. & Milligan, K. (2008). Incentive Effects of Social Assistance: A Regression Discontinuity Approach. *Journal of Econometrics*, 142(2): 807 – 828.

Lu, F. & Anderson, M. L. (2015). Peer Effects in Microenvironments: The Benefits of Homogeneous Classroom Groups. *Journal of Labor Economics*, 33(1): 91 – 122.

Luo, R., Miller, G., Rozelle, S., Sylvia, S. & Vera-Hernández, M. (2015). Can Bureaucrats Really Be Paid Like CEOs? School Administrator Incentives for Anemia Reduction in Rural China. NBER Working Paper, No. w21302.

Meyer, B. D. (1995). Natural and Quasi-experiments in Economics. *Journal of Business & Economic Statistics*, 13(2): 151 – 161.

Moffitt, R. (1991). Program Evaluation with Nonexperimental Data. *Evaluation Review*, 15(3): 291 – 314.

Morgan, S. L. & Winship, C. (2015). *Counterfactuals and Causal Inference*. Cambridge: Cambridge University Press.

Myerson, R. M., Tucker-Seeley, R. D., Goldman, D. P. & Lakdawalla, D. N. (2020). Does Medicare Coverage Improve Cancer Detection and Mortality Outcomes? *Journal of Policy Analysis*

and Management, 39(3): 577 – 604.

Okui, R. (2009). The Optimal Choice of Moments in Dynamic Panel Data Models. *Journal of Econometrics*, 151(1): 1 – 6.

Olsen, A. L. (2015). Citizen (dis) Satisfaction: An Experimental Equivalence Framing Study. *Public Administration Review*, 75(3): 469 – 478.

Robinson-Cimpian, J. P. & Thompson, K. D. (2016). The Effects of Changing Test-based Policies for Reclassifying English Learners. *Journal of Policy Analysis and Management*, 35(2): 279 – 305.

Rosenbaum, P. R. & Rubin, D. B. (1983). The Central Role of the Propensity Score in Observational Studies for Causal Effects. *Biometrika*, 70(1): 41 – 55.

Rosenbaum, P. R. & Rubin, D. B. (1985). Constructing a Control Group Using Multivariate Matched Sampling Methods that Incorporate the Propensity Score. *The American Statistician*, 39(1): 33 – 38.

Rosholm, M., Mikkelsen, M. B. & Svarer, M. (2019). Bridging the Gap from Welfare to Education: Propensity Score Matching Evaluation of a Bridging Intervention. *PloS One*, 14(5): e0216200.

Rubin, D. B. (1973). The Use of Matched Sampling and Regression Adjustment to Remove Bias in Observational Studies. *Biometrics*, 29(1): 185 – 203.

Rubin, D. B. (1974). Estimating Causal Effects of Treatments in Randomized and Nonrandomized Studies. *Journal of Educational Psychology*, 66(5): 688 – 701.

Rubin, D. B. (1976). Inference and Missing Data. *Biometrika*, 63(3): 581 – 592.

Rubin, D. B. (1997). Estimating Causal Effects from Large Data Sets Using Propensity Scores. *Annals of Internal Medicine*, 127(8Part2): 757 – 763.

Rubin, D. B. (1980). Randomization Analysis of Experimental Data: The Fisher Randomization Test Comment. *Journal of the American Statistical Association*, 75(371): 591 – 593.

Scott, T. (2015). Does Collaboration Make any Difference? Linking Collaborative Governance to Environmental Outcomes. *Journal of Policy Analysis and Management*, 34(3): 537 – 566.

Shinohara, S. (2018). Exit, Voice, and Loyalty under Municipal Decline: A Difference-in-differences Analysis in Japan. *Journal of Public Administration Research and Theory*, 28(1): 50 – 66.

Smith, J. A. & Todd, P. E. (2005). Does Matching Overcome LaLonde's Critique of Nonexperimental Estimators? *Journal of Econometrics*, 125(1 – 2): 305 – 353.

Theil, H. (1953). *Repeated Least Squares Applied to Complete Equation Systems*. The Hague: Central Planning Bureau.

Thistlethwaite, D. L. & Campbell, D. T. (1960). Regression-discontinuity Analysis: An Alternative to the Ex Post Facto Experiment. *Journal of Educational Psychology*, 51(6): 309 – 317.

Truex, R. (2014). The Returns to Office in a "Rubber Stamp" Parliament. *American Political Science Review*, 108(2): 235 – 251.

Waddington, R. J. & Berends, M. (2018). Impact of the Indiana Choice Scholarship Program: Achievement Effects for Students in Upper Elementary and Middle School. *Journal of Policy Analysis and Management*, 37(4): 783 – 808.

责任编辑：王秋石