Modern Management 现代管理, 2024, 14(4), 627-638
Published Online April 2024 in Hans. https://www.hanspub.org/journal/mm
https://doi.org/10.12677/mm.2024.144076

Hans汉斯

随机影响评估存在的问题及应对策略：一个研究综述

邓　雨

重庆大学公共管理学院，重庆市公共经济与公共政策研究中心，重庆

收稿日期：2024年3月6日；录用日期：2024年3月21日；发布日期：2024年4月24日

摘　要

随机影响评估的“反事实”设计可以获得最可信的因果推断，是公共政策或项目评估工具箱中的标准工具。然而，现实中开展随机影响评估面临着诸多问题，其中一些问题会导致较小的障碍，造成估计结果存在偏倚，而另一些问题会导致实验失败，使随机影响评估无法对感兴趣的假设提供有效的检验。缺乏对这些问题的深入认识及其应对策略的全面把握，阻碍了随机影响评估在实际政策或项目评估中的运用。本文通过对重要文献进行系统综述，围绕“存在问题－问题来源－潜在后果”，全面、深入地揭示了随机影响评估中存在的损耗、不依从、溢出效应、驱动效应、伦理、随机偏倚、外部有效性等七种问题，并对每一种问题的应对策略进行了系统梳理和概括。本研究有助于对评估者运用随机对照实验开展政策评估提供一定的指导，推进随机影响评估的开展。

关键词

随机影响评估，政策和项目评估，存在问题，应对策略

Problems and Countermeasures for Randomized Impact Evaluations: A Research Review

Yu Deng

Chongqing Research Center for Public Economy and Public Policy, School of Public Administration, Chongqing University, Chongqing

Received: Mar. 6th, 2024; accepted: Mar. 21st, 2024; published: Apr. 24th, 2024

文章引用: 邓雨. 随机影响评估存在的问题及应对策略：一个研究综述[J]. 现代管理, 2024, 14(4): 627-638.
DOI: 10.12677/mm.2024.144076

Abstract

The "counterfactual" design of randomized impact evaluations can obtain the most credible causal inference, which is a standard tool in the toolbox of public policy or project evaluation. However, in reality, there are a number of problems with conducting random impact assessments, some of which can lead to minor barriers that can bias the estimation results, while others can lead to experimental failures that prevent random impact assessments from providing a valid test for hypotheses of interest. The lack of an in-depth understanding of these issues and their countermeasures hinders the use of randomized impact evaluations in the evaluation of actual policies or projects. This paper systematically reviews the important literature, and comprehensively and deeply reveals seven problems in randomized impact evaluations, including attrition, non-compliance, spillover effect, driving effect, ethics, random bias, and external effectiveness, focusing on "existing problems, problem sources, and potential consequences", and systematically sorts out and summarizes the countermeasures for each problem. This study will help to provide some guidance for evaluators to use randomized controlled trials to conduct policy evaluations and promote the development of randomized impact evaluations.

Keywords

Randomized Impact Evaluations, Policy and Project Evaluation, Problems, Countermeasures

1. 引言

随机影响评估(Randomized Impact Evaluations)是指运用随机对照实验(Randomized Controlled Trials, RCTs)对政策或者项目的效果进行评估的方法[1]。在随机对照实验中，政策或项目被视为自变量或实验刺激，使用随机方法将实验对象分配到干预组和对照组，以创造一个有效的反事实，干预组和对照组之间结果的平均差异可以提供干预措施净效果的估计，因此能够准确识别政策或项目的因果效应。自 Fisher 的开创性工作以来，随机对照实验已经成为社会科学研究中研究因果问题的重要方法[2]。

从发展过程来看，随机影响评估经历了三个阶段。第一个阶段被视为随机影响评估的"曙光"。指 19 世纪 20 年代至 60 年代，Fisher 等人将随机化概念作为识别因果效应的关键因素，这一时期很少进行关于人的实验，主要回答关于农业生产力的重要经济问题，奠定了随机影响评估的理论基础。第二个阶段是随机影响评估拓展阶段。自十九世纪 60 年代至 90 年代，随机影响评估逐步转向人的研究，这一时期政府部门进行了一系列大规模的社会实验，包括就业计划、电价和住房津贴。该阶段初期专注于测试新项目的可扩展性，后期倾向于对现有项目进行审查以推动渐进式改革，对政策评估研究产生了重要影响，同时引发了实验研究和观察性研究之间权衡的争论。第三个阶段是随机评估快速发展阶段。自十九世纪 90 年代至今，许多学者在随机影响评估的理论、实验设计、数据分析方法上进行了完善和创新，使得可以采取随机对照实验探索问题的数量和类型大大增加，并广泛运用于公共政策和项目评估领域[3]。20 世纪末，随着"基于证据的政策制定"在公共管理领域兴起，在一些学者倡导的评估方法层次结构中，随机影响评估被置于顶端，俨然成为了识别因果关系的"黄金标准"[4]，越来越多的学者主张运用随机

影响评估方法为循证决策提供高质量的科学证据。特别是 2003 年以来，美国麻省理工学院成立的贫困行动实验室(J-PAL)，在全球 83 个国家开展了 1200 多项随机影响评估，在消除贫困和应对其他发展挑战方面作出了重大贡献。随机影响评估在现在的政策对话中占据了优势地位，自 2000 年以来，大约 60%的影响评估使用了随机化[5]。

随机影响评估也面临着现实争论，许多学者对随机影响评估作为因果识别“黄金标准”提出质疑，认为“来自随机实验的证据没有特别的优先权”[6]。这些质疑和争论核心是随机影响评估的内部有效性和外部有效性问题。在现实中开展随机影响评估是极其复杂的，面临着许多实施、操作以及统计分析的实际问题，制约了随机影响评估的方法优势。Heckman 和 Smith 分析了随机影响评估的局限性和面临的困难，认为随机对于许多感兴趣的问题提供的证据太少(如家庭背景、市场条件等对是否决定退出项目和完成项目所需时间的影响，各种替代干预的成本是多少等)、随机实验证据存在内在的变异性(实验未能为干预对结果分布的许多有趣特征的影响提供清晰和令人信服的证据)，实施过程中还面临随机偏倚、社会实验的制度限制以及替代偏倚，这些问题如果处理不好可能影响评估结果的准确性，甚至导致评估的失败[7]。Deaton 指出，在理想情况下，对项目进行随机影响评估有助于获得对政策或项目平均效果的令人信服的估计。但随机实验经常受到现实问题的限制，两个关键问题是对外部性的误解和对异质性的处理，这削弱了随机评估声称在统计或知识上有优势的说法[6]。Eble 等认为，随机影响评估中容易出现随机偏倚、测量误差、损耗、小样本和发表偏倚，并分析了这些问题如何导致有偏差的效应估计[8]。Burtless 和 Orr 讨论了随机影响评估的限期偏倚、排队偏倚、霍桑效应等问题[9]。Glennerster 对不依从、损耗、伦理、发表偏倚等问题进行了细致分析[10]。Greenberg 和 Barnow 总结了效果估计不精确、效果估计中的偏差、未能衡量干预项目的效果、原计划实验设计的削弱四类问题[11]。另外，还有学者关注随机影响评估的溢出效应、一般均衡效应等问题[12] [13]。

众多学者在分析随机影响评估存在问题的同时，也对如何应对这些问题进行了大量的探索和实践。如，为了解决损耗问题，Greenberg 和 Barnow 专门开发了一种模型方法[11]，Macours 和 Millan 利用追踪信息结合统计工具以纠正损耗偏倚；针对干预效果异质性问题[14]，Young 主张运用分层实验设计探索多种干预方案的效果和干预在亚群体中的不同效果[15]。为了指导初学者更好地开展一项随机评估，Choi 和 Kim 撰写了涵盖规划和实施随机评估一系列重要步骤的简单指南，以应对实施随机影响评估过程中面临的潜在问题[16]。这些研究对实施随机影响评估存在的问题进行了一定解释，并且对处理方法和应对策略进行了一些探索和介绍。然而，这些研究比较分散，没有全面、深入地揭示随机影响评估存在的若干困难，也没有系统地厘清应对这些困难的策略和方法，一定程度上限制了评估者对随机影响评估存在问题的准确认识和相关应对策略的灵活运用。

基于此，本文试图通过对重要文献进行系统梳理和综合分析，围绕“存在问题－问题来源－潜在后果”的逻辑，系统全面地揭示随机影响评估存在的问题，并厘清和概括相关问题的应对策略，旨在对评估者运用随机影响评估开展政策或项目评估提供有益指导，推动随机影响评估的实践运用。本文主要分析了随机影响评估存在的损耗、不依从、溢出效应、驱动效应、伦理、不依从、外部有效性等七个问题及其应对策略。

2. 随机影响评估存在的问题

2.1. 损耗问题

随机影响评估的社会实验设计包括对个体的实验前观察和相关数据收集，以及观察同一个体接受一段较长时间的实验处理。但是在实验中加入时间因素会引发一个经典实验中不存在的问题：损耗(Attrition)。损耗是指未能从原始样本中的某些人那里收集结果数据导致的样本流失[17]。通过调查而不是通过行政记

录获得后续数据时，最有可能发生这种情况。出现损耗的原因很多，最主要的是实验对象由于移民、搬家、转学、参军、死亡等退出实验，或者在最终的调查中调查员找不到实验对象；也有些实验对象认为保存实验所需的详细记录是不值得的，导致相关数据丢失；或者由于涉及个体隐私和一些其他问题，实验对象可能拒绝参与或者拒绝回答一些问题；也可能在一些实验中，实验对象没有得到实验干预的好处，因此退出实验。

损耗并不是实验研究的特有问题，其他非实验研究同样会面临类似情况。但是，损耗问题对于随机影响评估而言，可能会带来更大的挑战。一方面，损耗会降低研究的统计功效，另一方面，损耗会威胁研究的内部效度和外部效度。如果样本流失是随机的，仅减小了样本量的大小，这种情况下只会降低统计功效。但样本流失导致的关键问题是它不太可能是随机的，这种损耗可能会否定最初实验设计中的随机化，干预组和控制组之间由于参与者的不同损失而产生系统性差异，破坏了通过随机化获得的干预组和对照组的可比性。使用传统的统计技术将导致对实验效果的有偏差和不一致的估计，例如，如果那些从项目中受益最少的人倾向于退出实验，忽略这一事实将导致高估项目的效果，这会降低实验的内部效度。在干预效果存在异质性的情况下，它也可能对评估的外部有效性产生重要影响[14] [18]。因此，评估者需要谨慎处理损耗问题。

2.2. 不依从问题

干预在人群中随机分配并且依从性是完美的，是随机评估能够识别干预措施的因果效应的前提假设。随机化只影响个体接受干预的概率，而不是干预本身[18]。即使随机实验设计得很好，实施过程中也可能会出现实验对象未完全遵守随机分配的情况，即不依从问题(Noncompliance)。不依从是指干预组中一部分实验对象没有接受或没有完成干预，对照组的一些实验对象可能会接受干预。如果仅存在违反第一类干预分配的情况，称为片面不依从。当分配到对照组的个人可以有效地被禁止接受干预时，可能会出现这种情况。如果分配到控制组的某些个体确实设法接受了干预，被称为两方面不依从[19]。常见的不依从可能原因主要有：首先，不依从性的一个常见原因是研究人员很少能完全控制对照组的选择。例如，当随机化是在学校层面时，一些学生可能决定从对照组转移到干预组，以便从提供给干预组的项目中受益。其次，一些情况使研究人员无法强制干预组依从性。例如，在营养干预计划中，只有周一至周五老师才能现场监督学生服用多维元素片，而周末在家中跟踪每一个学生服用的成本很高且困难[20]。因此，并非接受干预的学校的每个学生都能得到完整干预。再次，随机化水平也可能是不依从性的驱动因素，在纸面上看起来可分离的随机化单元在实际情况中存在混淆。例如，卫生部门在地图上划定的清晰的诊所服务区可能与谁实际就诊于哪个诊所没有关系。政府规定的政治边界，如城镇、村庄等，也并不总是与可能推动项目实施和溢出效应的日常互动模式相对应。另外，研究人员或项目执行人员干预分配程序以及干预实施执行不到位也是出现不依从问题的可能原因。不依从问题对随机评估结果的影响与不依从水平有关，如果不依从水平较低，则意向干预估计(ITT) (按照分配方案计算)仍然有效，但通过减少干预和比较之间的对比，不依从性会显著降低统计功效。如果不依从水平较高，并且不依从行为不是随机的或偶然的，而是样本之间行为或特征的系统性差异的结果，将对随机影响评估的内部有效性构成威胁，影响项目净效果的估计[10]。

2.3. 溢出效应

溢出效应(Spillover Effects)是指未经实验干预的个体受到干预的影响。一些溢出发生在实验内部，即控制组的一些成员接收到他们应该被拒绝的干预，一些溢出发生在实验之外，使实验对象之外的个体受到干预的影响。溢出通常以两种形式出现，一些溢出与技术或物理有关，未经实验处理的个体间接获得

实验干预的好处，例如，肠道蠕虫具有传染性，如果一个孩子被驱虫，这会降低其邻居受感染的可能性。针对学校中的某些儿童的干预措施，通过同伴效应或者调整学校内的教学也可能使学校中处于对照组的其他儿童受益。一些溢出渠道是信息性的，主要以学习和模仿的形式出现，使得未经干预的个体得到类似的干预处理[21]。例如，Dupas 评估了肯尼亚免费发放长效杀虫剂处理蚊帐的影响，当随机选择的家庭在初始分配中收到高度补贴的蚊帐时，他们的邻居一年后购买蚊帐的意愿更高，这表明他们正在学习相关技术[22]。溢出效应可能发生在实验的内部或外部，也可能以多种形式出现，会导致对治疗效果的估计有偏差。如果对未干预个体的溢出效应是正的，那么意向干预估计通常会小于没有溢出效应的情况，导致低估干预效果。如果溢出为负，则估计值将向上偏[18]。随机影响评估的因果推断设计在于随机分配的干预组和对照组构造的反事实，溢出效应破坏了这种设计，干预组和对照组之间结果的平均差异不再提供干预对被干预者的平均影响，评估者必须采取措施识别和处理溢出效应。

2.4. 驱动效应

驱动效应是指参与者改变他们的行为或对问题的反应，因为他们意识到正在实验中[8]。两种情况下容易出现驱动效应，第一种情况是结果是主观的(例如，疼痛的自我报告或个人意见)，真实结果报告依赖于参与者主观意识和态度，第二种情况是参与者知道正在研究的干预以及对干预或控制的分配，他们可能会改变他们的行为。驱动效应主要有霍桑效应(Hawthorne Effects)、约翰·亨利效应(John Henry Effects)、调查效应(Survey Effects)、怨恨和士气低落效应(Resentment and Demoralization Effects)等。霍桑效应是指当实验干预组可能会感激接受干预并意识到被观察，这可能会促使他们改变自己的行为，例如，更加努力地使其成功。约翰·亨利效应是指参与者意识到自己是控制组的一部分时可能会感到被冒犯，并通过改变他们的行为来做出反应，例如，比较组中的教师可能会与干预教师“竞争”，或者决定懈怠[18]。调查效应与数据收集者过度影响参与者的行为或以在实验组之间产生人为差异的方式收集的数据有关。怨恨和士气低落效应是指群体内的随机化导致怨恨和士气低落情绪，当随机化涉及收入和利益的分配时，对照组中的人看到同一组中的其他人接受干预好处而他们没有接受好处时，他们可能不太可能配合计划实施或调查参与，也可能倾向于寻求可替代的干预方法[16]。当驱动效应的影响在干预组和控制组之间存在系统差异时，会使得实验的净效果的测度变得困难，此时实验净效果不仅包括干预组和对照组个体接受实验刺激和非实验刺激的差异，还包括了实验组和对照组个体对评估本身的意识导致的行为和态度的差异，降低了干预组和对照组的可比性，这可能会向上或向下扭曲干预效果估计。

2.5. 伦理问题

随机影响评估涉及人类受试者的项目或政策，很容易引发伦理问题(Ethical Issues)。Deaton 在讨论随机研究的相关问题时认为：“最令人不安的问题与伦理有关，尤其是在对非常贫穷的人进行实验时”[13]。Burtless 和 Orr 指出，对实验伦理的讨论往往集中在三个中心问题上。第一，如果实验处理本身具有对参与者造成重大伤害的风险，就会产生严重的伦理问题。第二，伦理问题涉及保护隐私和机密性，充分披露实验程序，以及拒绝参与或退出实验的权利。第三，对控制组拒绝提供潜在的有益服务也会引起伦理问题[9]。具体来看，伦理问题来自三个方面：首先，来自干预措施本身的伤害风险。除了对实验参与者的自身伤害风险，干预也不能为了促进社会的整体改善，反而必须不可避免地帮助一个群体而牺牲另一个群体。例如，开展一项帮助一些女性建立小企业的研究是否不道德，这可能对现有的当地企业产生潜在的负面外部影响。其次，来自随机评估过程中可能对参与者自由选择权的干扰和隐私权的侵害。最后，来自不同形式的随机化的潜在危害。反对意见认为随机分配资源是一种错误定位，理由是随机化者愿意为了研究而牺牲研究参与者的福祉。有些人需要干预的人没有得到干预，而其他人则接受了他们不需要

的干预。随机评估中不认真对待伦理问题会对评估造成不利影响。首先，伦理问题可能引起部分样本群体的流失，造成损耗，影响评估效果。其次，伦理问题也可能导致不依从情况，如某些对照组的参与者为了获得干预可能带来的好处而违背分配方案。另外，评估也可能会因为政治风险和公众舆论太大而被阻止，造成评估失败。因此，进行随机评估时，评估者必须审慎地考虑研究伦理以及围绕干预措施的伦理和法规。

2.6. 随机化偏差

随机化偏差又称随机偏倚(Randomization Bias)，是指由于随机化，实验样本可能与其他人群不同[7]。这是随机评估受到许多学者质疑的一点。理论上，随机评估是从所有目标群体中通过随机化选择部分对象样本，使在实验期间参与的人与在没有实验的情况下参与的人没有区别，并确保干预组和对照组之间没有系统的差异，以创造一个有效的反事实，因此干预组和对照组之间结果的平均差异可以提供干预措施净效果的估计。随机化的逻辑本身没有问题，但随机评估的社会实验很难真正实现随机化。随机偏倚的主要的来源是实验需要受试者和执行该项目的组织的同意，这些组织和个人可能出于多种原因拒绝参加实验：因为厌恶风险，因此不愿意接受随机分配；因为干预会使他们的情况变得更糟；或者因为他们怀疑干预的合法性等[9]。特别是一些申请参与制的实验，往往对照组的拒绝率较低，干预组的拒绝率相对较高。例如，Hotz 开展的一项职业培训伙伴项目(JTPA)是通过地理上分散的培训地点组织的，这些培训地点参与实验并不是强制性的，初始时超过 90%的培训中心拒绝参与该实验，评估者只能改变招募和吸纳程序，这正是产生随机偏倚的行为[23]。同时，一些资源限制导致项目管理者将参与者限制在满足某些标准的人，也会一定程度地导致随机偏倚[3]。另外，随机化偏差的一种来源可能是由于实验地点的选择是非随机的。本质上，随机化偏倚是样本代表性问题。随机化偏差意味着实验样本将代表相关目标人群的非随机子集，会降低实验样本的代表性，难以准确估计干预的效果，同时对干预效果的外部有效性造成威胁。

2.7. 外部有效性

对随机评估最重要的担忧是外部有效性(External Validity)，随机评估的外部有效性问题比其内部有效性问题更受争议。学界普遍认为，从内部有效性的角度来看，随机对照实验在建立因果关系方面比较可信。通过随机化可以消除干预组和对照组之间比较中的选择偏差，在执行良好并采取预防措施来规避任何可能的偏倚的情况下，随机实验结果具有内部有效性，即能够获得该项目效果的无偏估计[24]。Athey 和 Imbens 认为，“外部效度是指推广针对特定人群和环境得出的因果推论，其中这些替代环境可能涉及不同的人群、不同的结果或不同的环境”[19]。换言之，结果是否具有普遍性和可复制性。随机评估能够很好地回答政策或项目“是否有效”，但是无法为决策者提供更多他们更关注的信息，找出“什么在什么地方和什么情况下对什么人起作用”对提供科学准确的决策证据至关重要[9]。从根本上说，大多数对随机评估外部有效性的担忧都与干预效果的异质性有关。随机评估是在特定区域的特定人群和特定实验条件下进行，其他环境的普遍性永远无法保证。外部有效性还与干预效果的一般均衡效应有关，用分散样本进行的实验旨在估计个体层面的行为反应，仅提供政策变化的部分均衡效应的估计。因此，国家政策的一般均衡结果可能不同于实验中估计的部分均衡效应[25]。更有学者质疑，由非政府组织开展的小型“概念验证”研究的结果是否可以或应该直接转化为政府大规模实施的政策建议[6]，因为实验项目通常在特别谨慎和高度监督的情况下运行，也可能分配给它的资源比在更现实的情况下分配给它的资源要多，而大规模推广时难以保障这种条件。Banerjee、Chassang 和 Snowberg 认为，外部效度涉及三个焦点问题：其一，干预的可扩展性有多大？核心问题是干预措施如何扩展，即如果干预措施在一个省、国家或地区

推出，实验中衡量的干预效果会如何变化？其二，干预对不同群体有什么作用？如果一个项目在一个地区或一个国家有效，那么它对另一个国家也有效吗？如果一个项目对特定的社会群体有效，那么它对同一国家的不同群体是否有效？其三，在不同的情况下，干预对同一人群的作用是什么？相同人群在不同情况下对干预的反应可能不同。例如，如果一项干预措施帮助人们储蓄更多，但随着人们的储蓄不断积累，它是否会继续有效[26]。外部有效性不是随机评估特有的问题，其本身并不构成对随机评估结果准确性的威胁，但是无法确保外部有效性会影响随机评估结果的使用，难以对决策者决定是否推广或终止某项政策或项目提供科学准确的证据和指导。因此，随机评估者不应该回避外部有效性问题，而应该积极主动探索随机评估的外部有效性，推动循证决策的发展。

3. 应对策略

3.1. 损耗问题应对策略

关于如何解决损耗问题，目前学界没有标准化的方法。研究者主要通过在实验实施和数据收集过程中尽可能限制样本流失的数量来事前纠正样本选择，或者使用参数和非参数计量方法事后纠正损耗偏差。

事前规避损耗问题的各种措施都是尽可能获得更多的样本数据，以减少损耗偏差，从而确保更可靠的估计。首先，在干预设计中要能够考虑到将来损耗的发生，并且努力设计被试者持续参与项目的机制，保证项目在任何时间都可以进行。其次，通过改变随机化水平减少损耗，对于那些参与者可能从中受益的项目，个人一级的随机化可能会引起对实施组织的不满，在村庄、社区或学校一级进行随机化被认为是可取的计划，这样可以保证同一组的参与者保持相同待遇，有助于减少耗[24]。再次，改进数据收集方案是一个有效降低损耗的做法。例如：通过改善数据收集工具、合理化数据收集程序和规范收集管理来保证数据收集的效率，继而来降低损耗；通过持续不断追踪参与者，来尽可能减少损耗，这需要在基线处收集跟踪数据；通过收集参与者的代理信息，即通过认识参与者的其他人的收集关于他们的数据，这将最大限度地降低跟踪成本并减少流失，但必须在基线调查时收集并取得许可。另外，通过一定补偿或激励措施降低流失率，当调查需要很长时间或需要参与者主动前往调查时，这可能取得不错效果。最后，也可以使用管理数据解决损耗问题，即使用与干预信息相关联的管理数据(实施组织收集的作为其正常运作的一部分的数据)，可以大大降低数据收集的成本并减少损耗，例如，Angrist、Bettinger 和 Kremer 通过将代金券数据与哥伦比亚学校结业和大学入学考试注册数据联系起来，检验了哥伦比亚代金券计划的中期影响，但是要确保干预组和对照组之间的数据具有可比性[27]。

即使通过事前各种措施尽可能获得更多的样本数据规避损耗问题，一些自然样本流失几乎总是存在，这可能是非随机的。事后纠正损耗偏倚，主要是使用参数和非参数统计方法来模拟损耗过程，以获得更加可靠的估计。首先，在基于可观察的选择假设下，可以使用加权最小二乘回归获得无偏估计。逆概率加权(IPW)使用预测的被调查概率来纠正对可观察对象的非随机选择，即对每个可观测的观察值的概率取倒数，作为被观测的观察值权重，修正由数据缺失和损耗造成的估计偏差，提供可以推广到目标人群的结果[28]。其次，在由不可观察因素驱动的非随机选择的情况下，则可以使用 Heckman 样本选择校正模型修正损耗偏倚。Heckman 样本选择校正模型的本质是选择方程模型使用一个 probit 二分类模型，得出一个逆米尔斯比率(IMR)纳入第二个结果方程模型中，进行 OLS 回归[29]。另外，非参数界限也是最常用的方法。学者根据关注的结果和不同假设，估计不同类型的界限。一是通过假设那些缺失的人代表“最坏情况”来构建界限，并且使用结果变量的最小和最大可能值来估算缺失信息，界限比较宽泛。二是使用观察到的干预和控制分布的均值和标准差来构建界限，该模型通过假设每个实验组中的流失者的行为与该组中观察到的个体有些相似，从而导致更紧密的界限。三是对那些在干预组之间的损耗不平衡时总是被观察到的人的干预估计进行限制，不是构建最坏的情况，而是通过从上方或下方修剪样本的一部分

来估计界限，获得了更严格的界限。

总的来说，可以用统计技术来处理损耗问题，但最有效的方法是尽量限制样本流失。评估者应该将几种事前规避措施混合使用，尽可能地减少损耗，并且一开始就确保样本足够大。在足够大的样本中，跟踪那些丢失的随机样本，可以提供初始目标人群的样本代表性和具有高内部有效性的估计。在当损耗导致样本变得不平衡时，通过统计方法进行调整和消除估计偏差才是潜在选择。

3.2. 不依从问题应对策略

随机评估中处理不依从问题具有多种可行策略。首先，在事前对不依从问题进行考量，通过实验设计进行规避。主要有两种方法，其一是鼓励设计，随机选择的个体会被问到他们是否想参加该计划，他们可以选择是否参加。然后，评估会将接受者与不接受者进行比较。其二是超额认购设计，要求参与者进行申请，该计划将在申请者中随机分配[18]。其三是通过合适的随机化水平来限制不依从问题，例如在村庄、学校等群体层面而不是个体层面随机分配，以减少混淆干预组和对照组带来的不依从问题。其四是在项目实施过程中通过对项目执行人员的管理来减少不依从。选择正确的合作伙伴是依从性的关键，通常最好的策略是确保任何项目执行者完全与干预或控制人员一起工作；同时，有必要对项目执行者进行一定的培训和指导，或提供一些激励措施，以增强执行人员的工作热情；还必须在整个实施阶段监控依从性和接受度，及时向执行人员提供反馈，以便他们解决项目运行中出现的任何问题。其五是在事后通过数据分析策略或者计量方法对不依从问题进行处理，以减轻不依从问题导致的估计偏差。第一种是忽略干预的实际接受，进行意向干预分析(ITT)，即按照随机分配方案来分析，而不管是不是真正接受了干预。第二种是使用工具变量(IV)的方法来估计局部平均干预效果，即接受干预的因果效应。第三种是使用部分识别或边界分析来获得对整个人群接受干预的平均因果效应的值范围。评估者需要综合考量评估特定阶段和实际情况，合理搭配使用应对措施规避不依从问题。

3.3. 溢出效应应对策略

怎样有效处理溢出效应？常用的方法是调整随机化水平，也就是调整实验的随机设计来限制或者减轻溢出效应对平均效应的影响。应用该方法的基本假设是群体可以划分为组或集群，而相互作用仅限于同一集群内的个体，则组级随机化足以确定整体干预效果。例如，史耀疆等进行的陕西贫困农村学生的营养干预实验，选择在学校层面进行随机化，以限制干预措施的溢出效应[20]。如果对溢出效应很关注，可以专门设计实验来估计它们的范围和幅度。第一种技术是有目的地改变组内治疗的暴露水平。第二种技术是利用随机化自然产生的跨组暴露变化。第三种技术是将个人随机分配到不同的对等组[18]。计量方法和模型也被用来尝试考虑可能产生的偏差。例如，Orr 提出了一个简单的修正模型，当评估者知道干预组和对照组接受实验干预的比例时，可以使用该修正模型来调整溢出效应的估计偏差。Banerjee、Cole 和 Duflo 等则通过工具变量模型(IV)识别溢出效应[12]。另外，也有部分学者认为，如果溢出不明显，不对评估结果造成严重偏差的情况下，可以选择不进行处理。当然，如果溢出效应是全球性的(例如，世界价格的变化)，那么任何方法对项目效果的识别都是有问题的。可以说，如何处理溢出效应需要评估者对其有正确的认识，能够事前识别此类干预措施，评估者才能更好地通过实验设计去主动规避和溢出效应，而不是被动使用计量方法去调整和修正。

总之，溢出效应的规避要综合考虑特定评估项目评估进展的具体情况，评估者需要对溢出效应有全面的认识，并结合自己的认识考虑到底要不要去规避、选择何种方式规避。

3.4. 驱动效应应对策略

怎样克服随机评估的驱动效应问题？首先，理想的方法是使用经典实验的盲法设计，使干预组和控

制组都不知道哪个参与者属于哪个组，包括随机对照单盲设计和随机对照双盲设计。但通常情况下，盲法设计在随机评估的社会实验中比较困难[6]。其次，通过适当的随机化水平克服驱动效应，例如，在学校或社区等群体层面而非个人层面进行随机分组，这样会避免干预组和对照组之间的互动和干扰产生驱动效应，因为同一群体中的人受到类似对待。需要注意的是，更高级别的随机化水平和可能会降低统计功效，需要评估者进行权衡，同时也要考虑到干预组和对照组的结果差异也有可能受到群体特征的影响[24]。另外，还有学者尝试收集更长期的数据将驱动效应与项目的长期影响区分开来。Duflo 和 Hanna 在正式实验结束后的一年多时间里继续监测摄像项目的影响，当项目不再被正式评估时，结果与评估期是相似的，这表明存在的最初结果不是由于霍桑效应[30]。评估者需要主动识别驱动效应的潜在来源，以采取恰当的方法和手段去克服驱动效应对随机评估结果的影响。

3.5. 伦理问题应对策略

怎样有效规避随机影响评估的伦理问题。首先，评估者需要认真了解和遵守研究伦理，并在实验之前自觉接受伦理机构的审查。贝尔蒙报告中阐明了三个关键原则：其一，尊重人(Respect for Persons)。人应该被视为自主的代理人。他们有自己的目标，并且有权利和能力决定追求目标的最佳方式。这一原则要求研究人员清楚地向潜在参与者说明研究的风险和收益，并让他们决定是否要参与。其二，益处(Beneficence)。研究人员应避免故意造成伤害，并寻求最大限度地提高研究对象的利益并尽量减少研究对象的风险。然而，避免所有的伤害风险是不现实的，并且会阻碍来自研究的社会收益。因此，需要权衡伤害风险与可能从研究中获得的社会利益。其三，公正原则(Principle of Justice)。力求避免一个群体(例如穷人或囚犯)承担与研究有关的风险，而另一个群体则获得好处的情况[10]。其次，可以通过实验设计解决道德问题。Ravallion 讨论了可能规避伦理问题的几种实验设计[5]，第一种选择是使用“等效实验”(Equivalence Trial)，对照组获得被认为是下一个最佳的治疗。第二种选择是自适应随机化。即根据实验中陆续得到的结果及时调整分组概率同时不破坏随机性，从而使参与者最大程度地受益或最小程度地受害，这适用于分阶段的随机实验。第三种使用鼓励设计，鼓励设计在道德上的争议较小。另一种选择是使用条件随机化(也称为“封闭”或“分层”随机化)的方法缓解伦理问题。首先根据有关可能收益的先验知识选择符合条件的参与者类型，然后再随机分配干预措施，这适合并非所有参与者都可以覆盖的干预措施。再次，可以通过补偿性或激励性的支付来克服对随机实验的强烈道德反对。作为奖励金向所有同意参与的申请人支付(即接受随机分配的结果)或作为补偿金仅支付给那些实际拒绝服务的申请人(即分配到对照组的个体)。另外，研究人员需向相关人员解释与参与研究相关的任何伤害风险，在进行随机评估前征得同意[31]，并在数据收集和分析过程中采取相应措施以保护参与者信息的机密性。

3.6. 随机化偏差应对策略

如何回应随机化偏差带来的挑战？为了解决随机化偏倚的可能性，首先需要根据测量的基线特征分析参与是否是随机的，即干预组与对照组之间的个体(也可以是实验地点)特征是否存在显著差异，选择什么特征需要研究者以先验知识对感兴趣的人群进行定义，可以结合现有的非实验研究结果。如果存在随机化偏倚，研究者需采取有效方法进行处理。首先，可以通过优化实验方案设计解决随机偏倚问题。Banerjee 和 Duflo 等人认为，反复设计－再设计－再实验，并且不断更换实验时间、地点、实验被试者等可以有效解决随机化偏倚对样本代表性的质疑。Banerjee 等认为，可以扩大项目规模来获得足够大的样本量，进一步保证样本的代表性。并提出如果要扩大规模，实施第二阶段实验的组织必须是最终将大规模实施的组织，它必须由正式员工来实施，而不是外部专家来实施，并逐步转向更加放任自流的方法，但要继续在少数代表性的地点样本中仔细监控流程[21]。这样可以消解实验和项目实际运行的差异，更好

地评估干预措施的真实效果。其次，为了处理拒绝参与可能带来的随机化偏差，可以通过补偿性或激励性的支付来克服对随机实验的强烈道德反对。作为奖励金向所有同意参与的申请人支付(即接受随机分配的结果)或作为补偿金仅支付给那些实际拒绝服务的申请人(即分配到对照组的个体)。而对于项目的合法性持怀疑态度，以及对承诺的利益或服务是否会真正实现的不确定性导致的拒接参与。确定实验合法性和可靠性的最可靠方法是通过管理相同类型常规项目的同一项目机构来管理它[9]。另外，可以强化宣传，让人们更多地意识到干预措施的好处，以减轻对实验的抵制。总而言之，在设计和执行得当的实验中，随机化偏倚不一定会对样本的代表性产生严重的不利影响。

3.7. 外部有效性应对策略

如何解决或回答随机评估的外部有效性？目前随机评估者对这个问题进行了大量的探索。对于干预的可扩展性问题。首先，可以通过更大区域内、更大规模的实验研究均衡效应。但是一般均衡效应可能在国家甚至世界层面起作用时，实施随机评估可能是个不可能的挑战。其次，可以通过在相关市场水平上随机化来估计均衡效应，例如，代金券对学校之间的竞争、分类和留在公立学校的孩子的影响，可以通过在社区层面随机分配代金券来分析(假设一个社区足够大，可以容纳几所学校，一些公立学校和一些私人学校)。但同样，当整个国家都是相关市场时，均衡效应无法通过实验估计。再次，可以将非实验研究与实验研究结合起来，利用大规模经济政策变化的微观经济数据进行研究，可以分析区域层面的准实验研究结果是否与更多局部随机实验的结果一致。另外，可以将实验数据与微观均衡模型相结合，并利用对背景的理解来对可能的市场均衡做出一些预测[21]。

对于干预效果的异质性问题。首先，为了分析项目是否可以推广到其他环境，可以事前将实验设计为多点实验设计，在许多不同的背景下复制实验。多点实验设计可以根据地点特征分布有所不同，也可能根据干预的具体性质或干预率有所不同，以便评估推广到其他情况的可信度[32]。例如，Banerjee、Karlan和 Zinman 在六个独立的国家进行了六次小额信贷项目的实地实验，所有国家的基本干预措施都是一样的[33]。这样可以通过荟萃分析综合不同背景实验的结果数据，以量化异质性的水平，也可以分析该项目在不同背景中效果差异的来源。其次，为了分析同一项目对不同群体的影响是否具有异质性，评估者可以事前进行分层实验设计，提前确定可以推广的群体的特征，并对这些特征进行分层随机化。也可以事后通过统计方法探究干预效果的异质性，最常用的是亚组分析，即根据年龄、种族、性别、收入等群体特征，估计不同人群的平均干预效果，以考虑效果分布的差异。另外，通过参数和非参数统计方法分析干预效果的异质也是可行策略。参数方法，是指设定一个处理干预效果异质性的参数模型并报告估计值。具体来说，就是在回归模型中加入评估者感兴趣的个体特征协变量和处理状态变量的交互项。若存在的协变量过多，还需要使用正则化回归处理多重共线性。非参数方法，包括 K-最近邻匹配、核估计、回归树、随机森林、LASSO、支持向量机等。使用非参数方法来估计干预效果异质性，可以构建置信区间，帮助研究人员深入了解哪些类型的个体具有最高和最低的干预效果，得出个性化的政策建议[10]。

处理随机评估外部有效性问题的方法有很多，各自有优势和局限，评估者需要根据评估的性质和目的、误差的容忍度、数据统计分析上的考虑等条件来确定。但毫无疑问，实验设计和统计方法的搭配使用，往往能够对随机评估的外部性做出更加科学、准确的推断。

4. 总结与启示

本文通过重要文献进行系统梳理和综合分析，围绕“存在问题－问题来源－潜在后果”，全面、深入地揭示了随机影响评估存在的损耗、不依从、溢出效应、驱动效应、伦理、随机偏倚、外部有效性等七种问题，并对每一种问题的应对策略进行了系统梳理和概括。综合来看，上述随机影响评估问题的应

对策略大致分为优化实验方案设计、调整随机化水平和方法、改进实施管理程序、对受试者提供激励或补偿、对项目执行者进行培训和指导，以及开发各种统计工具和分析策略。但并不是每一种应对策略仅仅对应一种问题，也不是一种策略单独发挥效果。评估者需要对随机影响评估可能存在的问题和应对策略进行全面和深入的了解，在评估实践中面临这些问题和挑战时，需综合评估的性质、评估的阶段和具体情况，选择合适策略来进行处理。

本文也存在以下局限：首先，开展随机影响评估面临的问题远不止上述几种，由于时间精力等限制，本文选择其中最为核心和紧要的、会对评估结果产生重要影响的几个问题进行了分析。其次，将某些问题划分为特定类别是不够精确的，因为它们可能以不止一种方式影响估计，比如伦理问题可能会导致损耗，也有可能导致不依从问题，驱动效应可能会导致损耗和不依从问题，随机偏倚也有可能是由于伦理和损耗等问题引入的。最后，仅对相关理论讨论文献进行梳理，没有对现有的大量实践评估报告进行分析，缺乏对一些问题和应对策略的深入认识和细节把握。

本质上，许多问题本身不是随机影响评估的缺陷，也不是随机评估特有的，是由于被评估的项目是如何设计或实施的。随机影响评估在实施或操作中面临的这些问题，其中一些问题会导致较小的障碍，造成估计结果存在偏倚，而另一些问题会导致实验失败，使随机影响评估无法对感兴趣的假设提供有效的检验，甚至得出错误的结论。随机影响评估存在问题和挑战是客观事实，评估者要做的不是回避、忽视，或者放弃使用随机影响评估方法。相反，评估者要在整个随机影响评估中发挥更加关键的作用，在评估问题选择、实验方案设计、项目执行、数据收集和结果分析的各个环节准确识别和把握潜在问题，并采取有效的应对策略，增强评估结果的准确性和可信度。另外，评估者也要积极主动地去探索随机影响评估问题的创新性应对策略，在实验方案设计、随机化方法、数据分析方法等方面进行突破，推动随机影响评估在理论和实践上的纵深发展。

基金项目

重庆市教委人文社科重点研究基地项目“基于混合方法的‘双减’政策绩效评估与作用机制研究”(22SKJD003)；国家社科后期资助项目“西方政策评估理论与方法研究”(20FGLB043)。

参考文献

[1] Rachel, G. and Kudzai, T. (2013) Running Randomized Evaluations—A Practical Guide. Princeton University Press, Princeton.

[2] Fisher, R.A. (1925) Statistical Methods for Research Workers. Genesis Publishing Pvt Ltd, New Delhi.

[3] Levitt, S.D. and List, J.A. (2009) Field Experiments in Economics: The Past, the Present, and the Future. *European Economic Review*, **53**, 1-18. https://doi.org/10.1016/j.euroecorev.2008.12.001

[4] Malina, D., Bothwell, L.E., Greene, J.A., *et al.* (2016) Assessing the Gold Standard—Lessons from the History of RCTs. *The New England Journal of Medicine*, **374**, 2175-2181. https://doi.org/10.1056/NEJMms1604593

[5] Ravallion, M. (2020) Should the Randomistas (Continue to) Rule? NBER Working Papers No. 27554. https://doi.org/10.3386/w27554

[6] Deaton, A. (2010) Instruments, Randomization, and Learning about Development. *Journal of Economic Literature*, **48**, 424-455. https://doi.org/10.1257/jel.48.2.424

[7] Heckman, J. and Smith, J. (1995) Assessing the Case of Social Experiments. *Journal of Economic Perspectives*, **9**, 85-110. https://doi.org/10.1257/jep.9.2.85

[8] Eble, A., Boone, P. and Elbourne, D. (2016) On Minimizing the Risk of Bias in Randomized Controlled Trials in Economics. *The World Bank Economic Review*, **31**, 687-707. https://doi.org/10.1093/wber/lhw034

[9] Burtless, G. and Orr, L. (1986) Are Classical Experiments Needed for Manpower Policy. *The Journal of Human Resources*, **21**, 606-639. https://doi.org/10.2307/145769

[10] Glennerster, R. (2017) The Practicalities of Running Randomized Evaluations: Partnerships, Measurement, Ethics, and

Transparency. *Handbook of Economic Field Experiments*, **1**, 175-243. https://doi.org/10.1016/bs.hefe.2016.10.002

[11] Greenberg, D. and Barnow, B.S. (2014) Flaws in Evaluations of Social Programs: Illustrations from Randomized Controlled Trials. *Evaluation Review*, **38**, 359-387. https://doi.org/10.1177/0193841X14545782

[12] Banerjee, A., Cole, S., Duflo, E., *et al.* (2005) Remedying Education: Evidence from Two Randomized Experiments in India. NBER Working Paper No. 11904. https://doi.org/10.3386/w11904

[13] Deaton, A. (2020) Randomization in the Tropics Revisited: A Theme and Eleven Variations. NBER Working Papers No. 27600. https://doi.org/10.3386/w27600

[14] Macours, K. and Millan, T.M. (2017) Attrition in Randomized Control Trials: Using Tracking Information to Correct Bias. IZA Discussion Papers No. 10711.

[15] Young, A. (2019) Channeling Fisher: Randomization Tests and the Statistical Insignificance of Seemingly Significant Experimental Results. *The Quarterly Journal of Economics*, **134**, 557-598. https://doi.org/10.1093/qje/qjy029

[16] Choi, E.S. and Kim, B. (2016) A Beginner's Guide to Randomized Evaluations in Development Economics. *Seoul Journal of Economics*, **29**, 529-552.

[17] Hausman, J.A. and Wise, D.A. (1979) Attrition Bias in Experimental and Panel Data: The Gary Income Maintenance Experiment. *Econometrica*, **47**, 455-473. https://doi.org/10.2307/1914193

[18] Duflo, E., Glennerster, R. and Kremer, M. (2006) Using Randomization in Development Economics Research: A Toolkit. NBER Technical Working Paper No. 333. https://doi.org/10.3386/t0333

[19] Athey, S. and Imbens, G. (2016) The Econometrics of Randomized Experiments. *Handbook of Economic Field Experiments*, **1**, 73-140. https://doi.org/10.1016/bs.hefe.2016.10.003

[20] 史耀疆, 王欢, 罗仁福, 等. 营养干预对陕西贫困农村学生身心健康的影响研究[J]. 中国软科学, 2013(10): 48-58.

[21] Banerjee, A., Banerji, R., Berry, J., *et al.* (2017) From Proof of Concept to Scalable Policies: Challenges and Solutions, with an Application. NBER Working Paper No. 22931. https://doi.org/10.3386/w22931

[22] Dupas, P. (2014) Short-Run Subsidies and Long-Run Adoption of New Health Products: Evidence from a Field Experiment. *Econometrica*, **82**, 197-228. https://doi.org/10.3982/ECTA9508

[23] Hotz, V.J. (1992) Designing an Evaluation of JTPA. In: Manski, C. and Irwin, G., Eds., *Evaluating Welfare and Training Programs*, Harvard University Press, Cambridge, 76-114.

[24] Duflo, E. (2003) Scaling up and Evaluation 1. ABCDE Working Paper.

[25] Banerjee, A.V. and Duflo, E. (2009) The Experimental Approach to Development Economics. *Annual Review of Economics*, **1**, 151-178. https://doi.org/10.1146/annurev.economics.050708.143235

[26] Banerjee, A.V., Chassang, S. and Snowberg, E. (2016) Decision Theoretic Approaches to Experiment Design and External Validity. NBER Working Paper No. 22167. https://doi.org/10.3386/w22167

[27] Angrist, J., Bettinger, E. and Kremer, M. (2016) Long-Term Educational Consequences of Secondary School Vouchers: Evidence from Administrative Records in Colombia. *The American Economic Review*, **96**, 847-862.

[28] Wooldridge, J.M. (2002) Econometric Analysis of Cross Section and Panel Data. The MIT Press, Cambridge.

[29] Heckman, J.A. (1979) Sample Selection Bias as a Specification Error. *Econometrica*, **47**, 153-161. https://doi.org/10.2307/1912352

[30] Duflo, E. and Hanna, R. (2006) Monitoring Works: Getting Teachers to Come to School. NBER Working Paper No. 11880. https://doi.org/10.3386/w11880

[31] List, J. (2008) Informed Consent in Social Science. *Science*, **322**, 672. https://doi.org/10.1126/science.322.5902.672a

[32] Banerjee, A.V., Duflo, E. and Kremer, M. (2016) The Influence of Randomized Controlled Trials on Development Economics Research and on Development Policy. In: Basu, K., Rosenblatt, D. and Sepúlveda, C., Eds., *The State of Economics, the State of the World*, The MIT Press, Cambridge, 439-487. https://doi.org/10.7551/mitpress/11130.003.0015

[33] Banerjee, A., Karlan, D. and Jonathan, Z. (2015) Six Randomized Evaluations of Microcredit: Introduction and Further Steps. *American Economic Journal*: *Applied Economics*, **7**, 1-21. https://doi.org/10.1257/app.20140287