第1章 绪 论

1.1 历史回顾

1.1.1 随机对照实验与自然实验

17世纪以来，实验方法的系统性应用改变了自然科学的面貌，形成了数理科学之外的实验科学传统（库恩，2004）[37]。广泛开展的实验活动不断地创造新的现象和物质，极大地增强了科学技术改造自然的力量。随着实验方法的不断发展，实验的开展和应用不再局限于实验室之内，而是扩展延伸至真实的自然和社会场景之中。在21世纪以来蓬勃发展的多元化实验潮流中，随机对照实验与自然实验均是应用最为广泛、取得了最为突出成果的方法。

随机对照实验指的是，通过随机分配过程将实验对象分为实验组和对照组，对比两组样本在接受实验干预后的结果差异。历史上有记载的第一个对照实验来自圣经故事。尼布甲尼撒国王手下的犹太奴仆们不想食用违背自己宗教信仰的食物。为了说服国王允许他们食用素食，他们提出以分组的方式、在一段时间内采取不同饮食方案进行对照“实验”，来说明素食并不会让他们变得消瘦（珀尔，麦肯齐，2019）[113]。无论故事真假如何，这一例子与今天科学实验设计的基本逻辑完全一致。在早期有关疗法效果的研究中，最为著名的对照实验是18世纪中叶英国海军舰艇上对柑橘类水果治疗坏血病的小型对照（Matthews，2006）[2]。12名得了坏血病的船员被平均分成6组，每组人员分别服用苹果酒、硫酸丹剂（Elixir Vitriol，主要含硫酸和酒精）、肉豆蔻、醋、海水、柑橘和柠檬。食用柑橘和柠檬的2人出现了“最迅速和显著的好转”。

而最早有意识地采取系统性随机化分组的实验设计可能要等到19世纪80年代首次出现于心理学领域（Hacking，1988）。1883年10月至1884年7月，皮尔士（Charles S. Peirce）和他的学生加斯特罗（Joseph Jastrow）

在对感觉辨别(sensory discrimination)问题进行实验研究时,为了反驳当时学界被广泛接受的结论,采取了精确可靠的单盲随机化设计。统计学家斯蒂格勒(Stephen M. Stigler,1978)称赞他们的实验设计与今日的心理学实验无异。

当代随机对照实验的统计学基础在20世纪初得以初步建立。1923年,统计学家内曼(Jerzy Neyman)在其博士学位论文中提出了实验因果推理中的潜在结果模型,并论证了实验的观测结果是平均潜在结果的适当估计[①](珀尔,麦肯齐,2019)[236]。1925年,英国统计学家菲舍尔(Ronald A Fisher)发表的《给研究者的统计学方法》(*Statistical Methods for Research Workers*,Fisher,1925)一书从农业实验设计出发,首次将随机化过程确立为实验分组的标准方式,此后该方案迅速得以推广至各个需要探清相互关联的复杂因素之间影响的领域(Salsburg,2001; Hall,2007)。在医学领域建立现代随机对照实验范式的先驱者是英国流行病学家、统计学家希尔(Sir Austin Bradford Hill),他通过严格设计的随机对照实验证明了链霉素对结核病的疗效(Jadad & Enkin,2007)[xi]。很快,1962年美国食品及药物管理局(Food and Drug Administration)通过了一项修正案,要求制药厂商使用随机对照实验来证明疗法的效果和安全性(Bothwell & Podolsky,2016)。时至今日,随机对照实验已经被大多数科学家冠以研究方法的"黄金标准"称号(Jones & Podolsky,2015; Hariton & Locascio,2018),并将其视为最可靠的因果推理实验设计。

自然实验指的是将自然界或社会中不在实验者操控之下发生的事件作为实验干预,寻找其中自然形成的、可比较的实验分组,并对后续结果进行比较分析。由于自然实验的发生和设计往往需要机缘巧合与学者的敏锐眼光,在早期科学研究中并不常见,但却带来了非常重要的研究成果。例如,1835年达尔文在加拉帕戈斯群岛对地雀种群的不同进化方向研究是自然选择学说的出发点。彼此隔绝但气候条件近似的群岛正如天然的实验室,提供了在没有其他干扰因素影响下研究进化过程和结果的绝佳机会(Grant,1998)。1854年英国伦敦爆发了霍乱,医生斯诺(John Snow)通过在地图上标注病例位置,对比了两家供水公司提供服务的家庭的感染情况,

① "潜在结果"(potential outcomes)也称为奈曼-鲁宾因果模型,这个概念最初是由波兰统计学家耶日·奈曼(后来成为伯克利大学的教授)在20世纪20年代提出的。但是,直到20世纪70年代中期,唐纳德·鲁宾发表了关于潜在结果的研究论文之后,这种因果分析方法才真正开始得到不断的推进和发展。

最终论证了霍乱病菌通过水源而非空气进行传播(Snow,1855,详见第 5 章)。

但是,自然实验正式作为一种现代独立研究方法的确立可能要推后至 20 世纪末。根据统计,自 1990 年至今,自然实验的相关实证研究发表数量出现了明显的增加趋势(见图 1.1)。虽然自然实验的实际应用已经变得越来越流行,对其基本特征与适用范围的方法论研究还较为初步。直到 2012 年,第一本系统性论述自然实验的设计和应用的书籍才发表(Dunning,2012)。科学哲学领域的讨论就更少见(Morgan,2013)[342]。

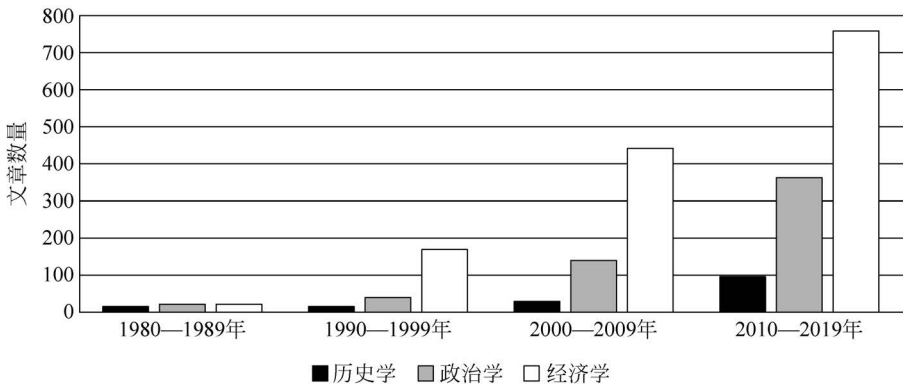

图 1.1　以“自然实验”为主题发表于主要历史学、政治学和经济学期刊的文章数量及其变化趋势

2022 年 3 月 6 日根据在 JSTOR 数据库的检索结果绘制

1.1.2　相关争议

然而,围绕这两类具有潜力、快速发展的实验方法却存在着不少争议。第一类论辩围绕着两种实验方法的地位和作用问题展开。随机对照实验的支持和使用者(尤其在经济学和医学领域)往往将自然实验归为观察研究(observational study)中的“准实验”或是“伪实验”(Meyer,1995; Shadish et al.,2001; Sims,2010; Clark et al.,2012),认为其并不是可靠的因果推断方法。自然实验者的回应则更多强调其具有处在随机对照实验与观察式研究之间的独立方法地位(Dunning,2012)[15]。

第二类批评对实验方法本身进行了批判性的剖析。一方面,从循证医

学领域流行的证据等级评价体系[①]可以看出，随机对照实验通常独占了金字塔的顶端位置（见表 1.1，图 1.2），研究者也往往给予其极高的评价，如："最佳的实验设计"（Norman & Streiner，2000）；"现代临床研究中最有力的工具"（Silverman，1981）；"随机对照实验是能够避免选择偏误和混杂偏误的唯一已知方法。该实验设计接近了基础科学中的受控实验。"（Schulz & Grimes，2019）[9] 纽卡斯尔大学的医用统计学教授马修斯（John N. S. Matthews）写道："在过去的 50 年中，随机对照实验已经成为最为基础的研究方法，在许多情况下甚至是衡量新疗法效果的唯一可靠证据来源"（Matthews，2006）[3]。然而，对这一看似完美科学方法的怀疑使得不少研究者致力于对其提出批评。如一部分对随机对照实验持保守立场的研究者强调，随机对照实验只是定量受控对照实验的类型之一，而不是对所有健康问题的万能灵药（Jadad et al.，2007）[8]，因而不该被授予黄金标准的地位。

表 1.1　美国预防服务工作组推荐证据等级

证据等级	研 究 方 法
Ⅰ	至少一项合理设计的随机对照试验（RCT）
Ⅱ-1	合理设计的非随机操控试验
Ⅱ-2	合理设计的队列研究或案例控制研究，最好是多中心或不同研究组进行的试验
Ⅱ-3	多个时间序列的观察研究，不一定进行实验干预。基于非控制试验的重要证据也可以认定为此类证据（如 20 世纪 40 年代引入的青霉素疗法）
Ⅲ	受尊敬的专家观点；基于临床经验、描述性研究或来自专家委员会的案例报告

Harris et al.，2001[9]

另一方面，对自然实验而言，使用者对该方法的定义与设计思路仍有分歧。部分学者将其视为对随机对照实验的模仿，并以此为基础来构建自然实验的设计标准（Dunning，2012）[3]；而另一部分学者则将自然实验视作"在自然的巨大实验室里不断流动运行着的实验"（Haavelmo，1944；Ozonoff et al.，1987），因此不必怀疑其作为一种实验的合法地位，也并不认为自然实验相比随机对照实验存在着本质性的缺陷（Diamond et al.，2010）。可见，无

① 医学中称之为证据等级，但有学者指出实际上该评价体系是针对实验方法而非实验结果（Bluhm，2005），即并不区分同一方法由于各种原因导致的结果质量的区别。因此称之为方法等级（hierarchy of methodology）也许更为贴切。本书仍采用证据等级的称谓。

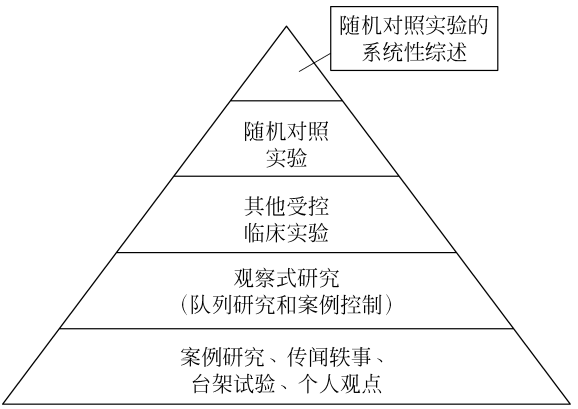

图 1.2　一个常见的证据等级示例

Greenhalgh，2014，著者译

论是这两类实验方法内部还是相比较而言，都存在着许多有待澄清的问题。

上述分歧不禁令人回想起社会科学研究中围绕着定量和定性方法的长久争议。在经典教材《社会科学中的研究设计》（*Designing Social Inquiry*：*Scientific Inferences in Qualitative Research*，1994）一书中，金（Gary King）、基欧汉（Robert O. Keohane）、维巴（Sidney Verba）三位作者指出：定量方法与定性方法之间存在共同的推理逻辑，其差异只是在于研究风格和具体实施细节上（King et al.，1994）[3]。而“大部分研究工作都不能被简单地划归到其中一类，好的研究方案总是试图将两种方法加以综合”（King et al.，1994）[5]。本书持有类似的立场，希望先澄清实验方法之间的争论，并试图找出进行综合性方法论说明的前提和基础，再以此来促进对随机对照实验与自然实验中的设计、选择、作用、缺陷等问题的理解和对话。

1.1.3　实验方法论的科学和哲学研究

在科学家积极参与方法论争论的今天，对特定类型的实验评议有多种研究进路（见 1.2.2 节）。在这里，笔者注意到研究工具的新进展为许多问题提供了新视角与答案。例如，对样本进行随机分配的方法从传统的物理随机系统（如抛硬币，抽取扑克牌）一度变为随机数表，近年来又被数据处理软件中内嵌的样本分组程序所取代。工具的不断发展革新是否仍然能够满足实验设计对随机化分配的内在要求？再比如，因本斯（Guido Imbens）等人对匹配方法的改进是否能使得恰当匹配的自然实验样本给出与随机对照实验同等质量的研究结论？元分析（meta-analysis）能否可靠地实现对现有

大量随机对照实验结果的评价和综合功能？对上述工具的关注有助于理解特殊实验方法中的实际问题与发展情况。

除了实验者内部的争论，在社会科学等传统上不涉及实验研究的领域中，以上述两类实验为代表、作为整体的实验方法的扩展和推广在积极推进的同时亦面临着巨大争议。20 世纪以来，实验方法开始较为普遍地进入社会科学。“二战”后，实验的兴盛和发展受到至少三个方面的历史性变化带来的影响，它们分别是新的研究对象（或现象）、新的理论、新的技术。就研究对象而言，在社会学和社会心理学中，围绕着人际影响、判断扭曲和从众过程等现象的研究议题受到了更多的关注。从理论发展来看，经济学开始将博弈论概念化，并对行为经济学产生了兴趣；政治学发展了投票选举的理性选择理论；社会学有了新的社会交往理论；心理学则进一步扩展了社会因素对个体影响的研究（Webster et al.，2014）[5]，等等。新技术的发展伴随着大学、政府、企业性质的各类研究机构中实验室的建立，为实验活动的开展提供了物质基础。自从 1879 年冯特（Wilhelm Wundt）在德国莱比锡大学建立了第一个心理学实验室后，各类实验室设备得以逐渐发明和成熟，如单向镜、录音录像设备、电视和电脑等，这使得对实验对象实施干预、控制、观察和记录得以可能。在数学工具方面，通过将结构功能概念转化为行为变量，以变量为核心的统计学方法使实验的语言得以规范化（罗斯，2007）[199]。

进入 21 世纪后，实验方法变得愈发受到关注，相关研究文献的数量急剧增长，其中产生的重要学术成果推动着相关方法论最终受到学界认可。以经济学为例，2002 年，基于实验室实验方法，卡尼曼（Daniel Kahneman）的决策行为研究和史密斯（Vernon Smith）的市场机制研究获得了诺贝尔经济学奖。颁奖词中写道：“实验室中的研究结果……能够对经济学理论的发展起到重要影响……正如物理学实验室对于微观现象的研究结果（如基本粒子和热力学）关键性地影响了理论物理学的发展那样”（Royal Swedish Academy of Sciences，2002）[3]。班纳吉（Abhijit Banerjee）、迪弗洛（Esther Duflo）与克雷默（Michael Kremer）利用随机对照田野实验（random controlled field experiment）进行的扶贫政策研究在 2019 年获得了诺贝尔经济学奖。2021 年，该奖项颁给了对自然实验方法以及实验的因果推断框架做出重大理论和实证贡献的卡德（David Card）、因本斯与安格里斯特（Joshua Angrist）。此外，在历史学、政治学等领域，也涌现出戴蒙德（Jared Diamond）、邓宁（Thad Dunning）等一批自然实验方法的主要推广者，以及一系列以实验方法论为主题的研究专著。

然而，社会科学共同体中否定或是反对使用实验方法的观点绝非少数。这类观点最早可以追溯到穆勒(John S. Mill)。他认为，从认识论上来说，人类行为和社会现象十分复杂，研究者无法逐一观察和记录实验过程中的事实和特征；即便用足够长的时间完成了实验结果的确认，现象本身也通常已经发生了变化。而从实践的角度，穆勒更加怀疑实施实验的可能性："当我们试图在研究社会现象中的规律时使用实验方法，第一个难题就是没有任何手段可以开展人工实验"(Mill，[1843]1965)[881]。反对者们普遍认为，社会科学中难以实现操纵和控制，如塞缪尔森(Paul A. Samuelson)和诺德豪斯(William D. Nordhouse)在早期经典教材《经济学》中所言："经济学中无法进行化学和生物那样的可控实验……就像天文学或气象学那样，经济学应该依赖观察"(Samuelson & Nordhaus，1985)[8]。劳森认为，经济学应该接受无法应用实验方法的现实，并应该以此为前提去讨论如何以受控观察和新方法来推动研究的发展(Lawson，1997)[199]。近年来，随着各类实验方法投入实际的使用，越来越多的学者对其进行了更有针对性的批评。如普林斯顿大学微观经济学家迪顿(Angus Deaton)认为，随机对照实验在设计上缺少理论和机制的指引，并不具有方法论上的优越地位(Deaton，2010)。宏观经济学家西姆斯(Christopher A. Sims)则批评自然实验等准实验方法不过是伪实验和"修辞设备"(rhetorical devices)，并且断言"经济学不是一种实验科学"(Sims，2010)。安格里斯特和皮施克(Jörn-Steffen Pischke)在同期期刊上与西姆斯展开了针锋相对的论辩(Angrist & Pischke，2010)。此外还有从实验设计、统计学工具、结论解读、应用前景等具体问题出发进行的技术性批评。迪顿和西姆斯分别是 2015 年和 2011 年的诺贝尔经济学奖获得者。由此可见，以随机对照实验和自然实验为代表的扩展性实验方法仍面临着巨大的争议，以及许多悬而未决的具体问题。本书希望通过说明社会科学中常用实验方法与自然科学实验方法的共同逻辑，来促进对科学的整体性理解。

上述实验的方法论问题不仅是社会科学家关心的重要话题，而且开始被科学哲学领域的前沿研究所关注。在近 20 年内出版的一般科学哲学和分支科学哲学的综合性著作中，常常能见到相关的议题。例如，2007 年出版的《一般科学哲学：焦点问题》(*General Philosophy of Science: Focal Issues*)中，第五章专门讨论了社会科学中实验的角色，并指出这是社会科学方法论中的核心问题(Kuiper，2007)[275]。2012 年出版的《牛津社会科学哲学手册》(*The Oxford Handbook of Philosophy of Social Science*)第十三章专门

讨论了在经济学中应用随机对照实验的证据质量和政策效果预测能力(Kincaid,2012)。《经济学哲学:当代导论》(*Philosophy of Economics: A Contemporary Introduction*)第十章中讨论了经济学实验的四种类型及其特点和作用(Reiss,2013)。《当代社会科学哲学导论》(*Philosophy of Social Science: A Contemporary Introduction*)中将实验与因果模型和案例研究作为社会科学中最重要的三种研究方法加以讨论(Risjord,2014)。

随着实验在社会科学领域迎来新的应用和发展,其形式和内容的不断扩充使其概念自身面临着挑战,其定义亟须寻求新的理解。保守立场坚持实验应受控地发生在实验室之中,因而认为实验概念在扩展至其他领域的过程中已经遭到了过度应用和延伸。"已经有太多的活动被当作是'实验'。例如在经济政策研究中,与实验室完全无关的'实验'声称自己得到了新的知识;甚至还有被认为能够产生知识的、基于严格的想象的'思想实验';又或者是通过使用统计学比较、理论和定量模型的计算机模拟来探索不同的场景。很明显,在经济学中,'实验'一词的真正含义一点也不清楚。"(Fontaine & Leonard,2005)[2] 这促使政治学家邓宁提出,为了防止实验概念的过度泛化和滥用,这些非传统实验应该以医学中作为方法论黄金标准的随机对照实验为设计标准(Dunning,2012)。而对于实验持开放性立场的研究者则更加强调以求异法、比较法为核心设计原则的方法论融合与创新。如在动物学、生理学和历史文化研究各领域均有建树的戴蒙德就将自然实验视作受控的实验室对照实验的可靠替代(戴蒙德,2017)[7]。可见,实验方法的定义一方面需要充分考虑和涵括其实践的多元特征,另一方面也需要说明其内在多样性在融合过程中存在的张力。

在上述历史背景下,本书关注于随机对照实验与自然实验两种特殊类型的实验中存在的方法论问题。通过对实验方法的设计与选择、实验结果的分析以及实验方法的评价等方面的评述,首先,梳理两类方法的构成要素,考察二者是否具有对话的共同基础。其次,分析各方法自身的特性,比较其间的差异,并尝试对相应的科学争论问题给出回答。

1.2 研究问题及研究现状

1.2.1 研究问题的界定

本书以两种特殊类型的实验为研究对象。谈到实验的科学哲学研究,首先需要对实验的概念进行适当界定。早期新实验主义和一般科学哲学中

对于实验的表述大多是以实验室为原型进行的抽象刻画，强调了实验作为人工可控环境中运行的物质性活动的特征。例如，《当代自然辩证法教程》中将科学实验定义为“人们根据一定的科学研究目的，运用一定的物质手段（科学仪器和设备），在人为控制或变革客观事物的条件下获得科学事实的基本方法……与自然观察不同，它是对于客体进行积极干预下进行的观察。”（曾国屏 等，2005）[159] 刘大椿在《科学哲学》中将实验定义为“人们根据一定的研究目的，利用科学仪器、设备，人为地控制或模拟自然现象，使自然过程或生产过程以纯粹的、典型的形式表现出来，以便在有利的条件下进行观察和研究的一种方法。”（刘大椿，2006）[99] 新实验主义为了强调实验与命题知识的区别，更是着重强调实验活动的物质性，如哈勒（Rom Harré）认为：“实验是对仪器的操控，即按照不同方式将特定排列的某种物质原材料整合到物质世界中。”（Harré，2003）[19] 类似地，拉德（Hans Radder）认为：“实验是人对物质世界的主动干预，涉及实验过程的物质实现（包括研究对象、仪器以及它们之间的互动）。”（Radder，2003）[4]

以上定义均强调了实验需要借助科学仪器进行人工控制和干预，但这些要素对于随机对照实验和自然实验来说似乎并非必要，因此本书需要对实验进行重新界定。在常用这些方法的社会科学领域的哲学讨论中，实验的定义仍然处于争论之中。问题的核心部分在于：是否应当以及该如何沿用和协调传统上的自然科学实验概念来理解和评价社会科学实验？冈萨雷斯整理了实验概念具备的七个特征（Gonzalez，2007），这为我们寻找一个适当的实验定义提供了很好的参考框架。

（1）语义学上，实验与“观察”有着不同的语义和指向。

（2）逻辑上，实验是科学的结构性组成部分，并且原则上与“理论”和“模型”区分开。

（3）认识论上，实验是一种通过非直接（non-immediate）过程获得的可靠知识。

（4）方法论上，实验应该与一个可重复的过程相联系，因此，它通常与可再现性（reproducibility）和可重复性（replicability）相联系[①]。

---

① 可重复性在英文中对应许多不同的词汇，典型的如“repeatability”“replicability”和“reproducibility”。进行相关讨论的学者经常混用以上词汇。少数研究者和学科领域会区分不同词语对应含义，如区分实验流程和结果的重复，重做完整实验或是重新进行数据处理，等等（肖显静，2018c；Fidler & Wilcox，2021）。拉德认为“repeatability”侧重于观察这一行为的再现，“replicability”侧重于物质性实验结果的复制，“reproducibility”则意味着对实验过程和结果看作一个流程式的整体（Radder，1992）。下文中不对相关用词进行特别的区分，主要参考诺顿（Norton，2015）并以“replicability”对应于中文的“可重复性”。

(5) 本体论上,实验与他物(otherness)的概念有关(即需要通过测试来确认真或假的事物)。

(6) 价值论上,实验可以根据不同的目标设立不同的价值取向(如基础科学实验可以与应用科学实验不同)。

(7) 伦理学上,对于与某些人类和社会事务有关的实验有着特别的关注。

虽然本书中的后续讨论与该框架的部分要求有着不同的观点(如针对第(1)点和第(4)点),但冈萨雷斯给出的七个特征在类目上无疑是实验的哲学研究必须关注的。冈萨雷斯还强调,这些新的实验方法是扩展的、多元化的,不同类型的实验在不同程度上展现了以上特征。以往的哲学家和社会科学家质疑实验,其实是基于传统上将实验视为一种受理论约束的、在可控的物质环境中稳定实现的人为干预的立场,因而难以设想其在社会科学研究情景中的可行性。而今天涌现出的多元化、成果颇丰的实验实践所带来的扩展版本的实验概念为打破这些质疑提供了新的观点和证据。

经济史家摩根(Mary S. Morgan)在对自然实验的研究中批评了新实验主义表现出的对实验物质性的强调:"奇怪的是,一方面新实验主义建议我们将实验视为有自己的生命,另一方面则仍然保留了实验本质是在物质性层面上进行可控的操控或干预的这种旧观念。"(Morgan,2003)[217] 受此启发,考虑到传统实验定义在此处的不适用,本书给出进行的实验定义只保留"干预"和"对照"作为核心设计要素,而不再限定其发生的环境和依赖的仪器。即,实验指:通过考察其他条件相似的干预组和对照组在受到干预后的表现差异,从而探究因果关系的研究。其中,干预指对实验目标系统施加的外部扰动,其效果必须能够使因变量的取值发生改变。干预本身不依赖于研究者的意向和能力范围:干预可以是自然发生的(如地震、火灾),也可以是实验目的之外的人为事件(如战争、政策变化)。对照,即目标研究对象在未受到干预和受到干预两种情形下状态的比较,从而获得因变量发生改变后自变量的变化值。这个定义可能稍显宽泛,但是它的确能够容纳选题所关注的两种实验的构造逻辑,同时并不违背自然科学实验的一般设计思路。例如,在新型冠状病毒感染流行前后对某类人群心理健康状态的访谈是否可以称为实验呢?[①] 本书认为这并不违背实验的定义:它包含了一个

① 该研究设计来自(Prati & Mancini,2021),标题为 The psychological impact of COVID-19 pandemic lockdowns: a review and meta-analysis of longitudinal studies and natural experiments。