第 8 章

田野实验

吴珏彧　戴恒琛

学习目标

1. 了解田野实验的定义和优势
2. 熟悉田野实验常见的随机化方式，能够分析一个田野实验所采用的设计方法
3. 学习田野实验中常见的统计问题，能够对田野实验研究的案例进行解析和评判
4. 了解田野实验的实操建议，提升使用田野实验的信心

近些年，田野实验（field experiments）在管理学中获得了越来越多的关注。相较于其他管理学常用的实证方法，田野实验既保存了实验的内部效度，又提升了因果推论的外部效度。本章我们首先定义田野实验并阐释其相对于其他研究方法的优势。其次，我们介绍田野实验设计中常见的随机化方式，并剖析田野实验在设计和分析上容易出现的问题。最后，我们分享田野实验的实操建议。本章吸取了多个社会科学学科（包括管理学、经济学、政治学、心理学）在田野实验上的研究成果，并结合作者在实操中汲取的经验，希望给年轻学者提供对于田野实验的直观认识和实际建议，鼓励更多的管理学学者尝试田野实验。

8.1　田野实验简介

8.1.1　田野实验的定义

从广义上说，我们把田野实验定义为在实际环境中进行的高度趋近于现实的实验研究（research with a high degree of naturalism）。田野实验也被叫作实地实验，是随机对照试验的一种，通常与实验室实验和准实验相比较（详见本书第 6 章和第 7 章）。值得注意的是，学者们对“田野”这个词会有不同的理解。一些学者将“田野”定义为研究对象和他人互动的物理环境（Wood，2009）；另一些学者则不强调物理环境而强调研究本身的现实属性（Paluck & Cialdini，2014），毕竟我们可以在实验室中设计出非常贴近自然行为的研究，也可以在非实验室中进行不贴近自然行为的人为干预。一般来说，与实验室实验相比，田野实验专门用来指代那些高度贴近现实环境的研究，而不是单单在实验室或网上进行的相对简单快速的研究。田野实验定义中的一个关键在于，田野实验研究者是在现实场景中收集采证的。这些现实场

景体现了人们在现实生活中真真切切的决策、工作状态、社会互动和群体动态等情况。

如何区分一个实验研究是否高度接近现实呢？我们可以从以下四个维度来评估（Gerber & Green，2012；Wu & Littman，2022）：

① 参与实验的被试是否是和研究现象相关的现实人群？

② 实验所处的环境是否是研究者感兴趣的现实环境？

③ 被试所接受的干预是否和干预在现实中的形式相似？

④ 实验测量的结果变量是否能体现人们在现实中的反应？

如果对于一个实验来说，以上四个问题的答案都是肯定的（换句话说，四个维度都非常接近现实），那么这个实验一定是田野实验。如果某些问题的答案是肯定的而其他是否定的，那这个实验未必是田野实验。所以，我们可以说田野实验虽然有多种形式，但它必须具备“田野性”或者“现实度”的特征。

我们来举个例子。请考虑这个研究问题：团队目标设定对员工绩效有何影响？表 8.1 显示了研究者如何在每个维度上使用不同现实度的设计来研究这个问题。“低现实度”一栏的组合将构成一个拥有低现实度的研究，这类研究不会被视为田野实验。“高现实度”一栏的组合很明显地构成了田野实验。在某些维度上具有高现实度但在其他维度上具有低现实度的实验仍可能被视为田野实验。比如 Harrison & List（2004）把只拥有高现实度被试的实验称为人造田野实验（artefactual field experiments），通常也被称为田野中的实验室实验（lab-in-the-field）；把只拥有高现实度被试以及高现实度干预或结果变量的实验称为框架田野实验（framed field experiments）；而把四个维度的现实度都较高的实验称为自然田野实验（natural field experiments）。本章重点讨论的就是现实度高的自然田野实验。

表8.1　田野实验的评估方式举例：团队目标设定对员工绩效有何影响？

	低现实度	高现实度
被试	在Amazon Mechanical Turk或问卷星上招募的便利样本	与一个企业合作获得的远程工作的员工样本
环境	被试在家中的电脑上填写在线调查，实验被嵌入调查问卷里	被试在办公电脑上通过远程工作软件与同事互动
干预	问卷文字说明让被试花五分钟的时间想象自己与同事讨论工作目标	员工持续两个月通过每周晨会和同事在线讨论工作目标
结果变量	被试的预期绩效	被试的实际绩效

田野实验可以将现实环境中的个人、团体或机构随机分配到不同实验条件下，即不同的干预组和控制组中，并在现实环境中测量有关作用机制和变化过程的变量。田野实验对不同被试、环境、干预和结果测量的包容性极大。回顾文献，我们几乎可以看到在任何现实条件下做的田

野实验，包括办公室（Eden & Moriah，1996；Greenberg，1988）、工厂（Hossain & List，2012；Wu & Paluck，2020，2022）、政府单位（Bandiera et al.，2021）、医院（Dai et al.，2021；Reiff et al.，2022）、托儿所（Gneezy & Rustichini，2000）、学校（Grant，2008；Paluck et al.，2016）、战后社区（Paluck，2009）和媒体节目（Blair et al.，2019）等任何你能想到的地点。这些田野实验针对现实环境中不同的议题测量不同的行为或心理结果，例如企业员工绩效、群体归属感、对社会权威的态度等。田野实验可研究的问题是无穷的。

8.1.2　田野实验的优势

根据现代心理学鼻祖威廉·詹姆斯（William James）的说法，研究人类心理和行为的最终目的是解释和预测人们在现实世界中的行为表现（Gantman et al.，2018；James，1907）。在管理学中，我们的研究主要用来解释和预测人们在组织中的行为、动机、决策、人际关系等。要做到这一点，管理学理论就需要与员工在企业和现实生活中的行为表现密切相关。如果不关注现实场景，理论研究可能会得到没有意义甚至误导大众的结论。从这点来说，田野实验具有明显的优势。田野实验在现实环境中进行，直接研究与所感兴趣的现象相关的现实人群，提供更实际和更有影响力的干预方式，并直接测量在该现实环境中发生的结果。相对于实验室实验、档案数据分析与定性分析来说，田野实验有如下几个明显的优势。

高生态效度（ecological validity）。相对于实验室实验来说（参照本书第 6 章），田野实验最明显的优势是其高度的现实性，这对应着研究的高生态效度。生态效度是指一项研究的变量和结论在多大程度上可以推广到现实环境中。在实验室实验里，被试的行为表现是基于对研究者呈现的实验材料（stimuli）做出的反应，这样的反应不一定代表现实环境中被试的真实反应，实验结果不一定能很好地运用到真实社会和企业中去。相比之下，田野实验在现实环境中进行，更有可能保证被试行为和心理的真实度，更容易逼近个人在现实生活中的决策、人际互动以及情感和动机表达中的心理过程，因此结果通常可以更好地预测目标人群在现实环境中的行为。

研究议题的广泛性（broadness of research constructs）。相较于标准化的实验室研究范式，田野实验扩大了管理学研究议题的范围。比如说，在实验室中很难研究企业中长期的群体动态行为，也很难在不违背伦理的情况下研究和再现现实社会中暴力和冲突的心理过程。相反，对于此类无法在实验室重现的议题，田野实验研究者可以在现实社会中选取合适的场景来做研究和干预。再比如说，研究者虽然可以在实验室中研究短期贫困给人们造成的影响，但无法研究长期贫困给人们造成的影响。一般来说，田野实验比实验室实验更适合帮助研究者调查涉及持续和动态的行为模式和行为变化的研究议题，更利于研究者捕捉在现实世界中多样的行为表现形式，从而扩大了研究议题的广泛性。

另外，田野实验能够突破传统实验室实验在样本与环境上的局限，进而扩大研究本身的应用价值。田野实验可以在尚未建立实验室的社区和不发达地区展开，如直接招募那些离高

校较远的乡镇企业和农民工群体作为被试。因此，田野实验在获得非常规样本和研究环境方面更具有包容性，这也间接扩大了研究议题的广泛性。

因果关系检验（causal testing）。相对于实地问卷调查和档案数据分析来说（详见本书第 9 章和第 10 章），田野实验在保留外部效度的同时也提供了因果测试。正如传统实验室实验将各种元素从现实环境中抽象出来导入实验室中，田野实验将实验中的干预组和控制组从实验室输出到现实环境中（Paluck & Cialdini，2014）。有了随机的干预组和控制组，田野实验让我们能够在多种社会环境因素的影响下，得出自变量与因变量之间的因果关系（causation），而不单单是变量之间的相关性（correlation）。

因果关系对于理论的建立是非常重要的。比方说，你假设“团队参与度高的员工绩效更高”，并用问卷搜集了每个员工在团队中的参与度和绩效信息。这个设计仅仅能验证团队参与度和绩效的相关性，但不能说明两者是否存在因果关系。比如是因为员工积极地参与了团队建设使得绩效更高，还是因为员工绩效更高所以更有动力来参与团队建设呢？一个相关性的研究是无法回答因果问题的。实验则可以验证因果关系，来更好地建立理论。回到上面的例子，如果能确定是团队参与度影响了绩效，那么我们接下来就可以更深入地研究为什么这个因素会影响绩效。同时，弄清因果关系也可以帮助我们对如何提高员工绩效提出实质性的建议。

企业项目评估（program evaluation）。田野实验的一种形式是评估某企业或政策项目在特定人群中的有效性，旨在衡量资源的有效部署程度（Gerber & Green，2012）。比如 Bloom et al.（2013）用田野实验在印度 17 个纺织企业的 28 个工厂中对不同管理方式的作用进行评估。他们把工厂随机分配到干预组和控制组中，给干预组的工厂提供免费的咨询服务。他们发现，与没有收到免费咨询服务的控制组工厂相比，干预组工厂的生产力在第一年提高了 17%，并在 3 年内开设了更多家工厂。这种与企业项目评估相关的长期干预在实验室中是无法进行的，而且田野实验的因果推论的有效性也更高，所以田野实验更加适合。

我们再以消息传递干预举一个例子。为了测试疫苗接种广告是否可以提高社区成员的疫苗接种率，田野实验会在不同的社区随机分配疫苗接种广告的传播情况，并测量干预组和控制组社区之间疫苗接种率的差异。从项目评估的角度来看，田野实验方法优于实验室实验。在传统实验室实验中，被试会看到不同的广告，然后研究者会询问他们是否有意愿接种疫苗（Cui et al.，2022）。虽然实验室实验可能能够检测到影响方向（effect direction），即某些信息是否比其他信息更有效，但它们不太可能捕捉到干预在现实中能产生的影响大小（effect size）。例如，实验室实验可能捕捉不到目标社区的一些居民会错过广告、不专心看广告或在生活的其他干扰中忘记信息的可能性。只有田野实验才能帮助研究者衡量在现实情况下一项干预的影响方向和大小，并衡量这个干预所需资源的成本大小（Saccardo et al.，2023）。

社会影响力（social impact）。从应用角度来说，田野实验的结论在学术界之外可能更有影响力。因为田野实验真真切切发生在企业的日常运作和人们的生活中，既有现实相关

性，又有因果效应，所以更容易让企业和政府的政策制定者信服（Dolan & Galizzi，2014；Hansen & Tummers，2020）。

8.2　田野实验设计中的随机化方式

实验一个的必要特征是有不同的实验条件（conditions）。研究者根据研究议题来设计一个或多个干预组和控制组。干预组包含了研究者感兴趣的实验干预内容。控制组可以完全不包含干预内容，作为单纯控制组（pure control）；控制组也可以是安慰剂控制组（placebo control），即包含与研究者关心的干预所不相关的实验内容。与单纯控制组相比，安慰剂控制组的被试在体验和经历上一般更接近于干预组，但是他们的体验和经历与研究者最感兴趣的干预不相关。

实验的另一个必要特征被试是被随机分配到不同的实验条件下。随机分配这个概念在实验设计中非常重要，直接影响研究变量之间的因果关系测试的有效性。Kenny（1979）甚至强调随机化是实验设计中最重要的环节。当被试被随机分配到不同的实验条件下后，除了实验干预是否存在，不同的实验条件理论上不存在任何系统差异。换句话说，当没有实验干预时，随机分配到干预组的被试表现不能显著优于或劣于控制组的被试。因此，在样本量足够大并且随机分配得当的情况下，我们可以说任何观察到的不同实验条件之间的差异都可以归因于特定的实验干预本身。在本节中，我们简要阐述田野实验常用的随机化方式，并讨论一些在田野实验中常见的设计和因果估计问题。

8.2.1　简单随机化（simple randomization）

最常用的随机分配化是简单随机化。随机单位为个体。每个被试通过一个随机程序（如掷硬币、抓阄、电脑随机算法）按照一定的概率被分配到研究者预先设定的不同实验条件下。简单随机化不一定是等分（更多关于分配比例的讨论见 8.3 节），但研究者一定要事先设计好被试进入不同组别的概率。在随机分配前，每一个被试和研究者都不知道他将会被分配到哪个组别中。

每个被试的随机分配互不影响。比如，如果用简单随机化的方式把同一个工厂的所有员工随机等分分配到一个干预组和一个控制组中，那同一个车间的员工的随机分配是独立的：通过掷硬币，员工甲有 50% 的概率进入干预组，同一车间的员工乙也有 50% 的概率进入干预组；如果员工甲被分配到干预组，这并不影响同一车间的员工乙被分配到干预组（仍是 50% 的概率）。

简单随机化的优势是方便易行，实验室实验基本上都用这样的随机化方式。但是在田野实验中，简单随机化不是在所有条件下都适用的。第一，简单随机化不能研究群体层面的因果关系。第二，简单随机化在田野实验的实际操作中未必现实。第三，简单随机可能面临溢出效应（spillover），详见本章 8.3 节。在这三种情况下，我们就需要用集群随机化来探索因果关系。

8.2.2　集群随机化（cluster randomization）

集群随机化是指以集群（cluster）为单位来进行随机分配。为了更好地解释个人和集群层面随机分配的差异，我们看 Wu & Paluck（2022）中参与式小组结构对工人绩效和企业归属感影响的实验。Wu & Paluck（2022）在一家大型纺织企业中做田野实验，样本是缝纫工厂的 65 个小组（工人人数为 1 752 人）。每个组有 20—30 人一起工作，每组都由自己的组长监督。一旦工人被工厂雇用并分配到特定的小组，他们就不会轮换。换句话说，这些小组是固定的。研究者的议题是如果员工们有机会在小组会议上自由发言，这会如何改变工人绩效和企业归属感。怎么设计实验来研究这个议题呢？研究者可以把每个小组（集群）随机分配到干预组和控制组中，即同一小组内的所有工人作为一个整体被同时分配到不同的实验条件中。所以在这个田野实验中，研究者是对 65 个小组进行随机分配，而不是对 1 752 个工人进行随机分配。

相较于简单随机化，集群随机化有三点好处：

第一，从理论上说，当管理学学者研究个体（如单个员工）相对独立的行为、态度和心理状态时，随机实验通常发生在个体层面，但很多研究所需的实验单位超越了个体。比如上文提到的以小组为单位的研究议题。由于工人们只在他们现有的小组中举行会议，将单个工人分配到干预组和控制组的简单随机化方案根本不可行，否则就把一个小组的工人们分开了。所以研究者需要进行以小组为单位的集群随机分配。

总的来说，当管理学者感兴趣的是集群层面的研究问题时，就需要用集群作为实验单元。比如，特定的工作环境对员工绩效有什么影响？夫妻心理咨询会如何影响夫妻关系？教师的教学风格会如何影响学生成绩？老板的领导方式是否影响团队矛盾？在这些例子中，我们需要得到自变量（特定的工作环境、夫妻心理咨询、教师的教学风格、老板的领导方式）与因变量（员工绩效、夫妻关系、学生成绩、团队矛盾）之间的因果关系。这四个研究议题的一个共同特点是：我们在现实环境中进行实验干预时，通常不止一个人会受到影响。干预工作环境时受影响的是在这个环境中工作的所有员工，干预夫妻心理咨询时受影响的是夫妻双方而非个人，干预教师的教学风格时受影响的往往是整个班级的学生（除非研究的是私人教师），干预老师的领导方式时受影响的是整个团队而不仅仅是某一个员工。这些同时受到干预影响的单位称为实验集群。不同于将每个被试单独分配的简单随机化，集群随机化将每个集群作为一个实验单元整体分配到不同的实验条件中。集群随机化使研究者能够在超越个体的层面研究群体和社区现象的因果关系。

第二，集群随机化可以解决某些干预只能在集群层面上实施的实操问题。比如通过有线电视或广播来进行干预时，我们不太可能将干预随机分配给特定的观众，因为有线电视和广播节目通常是播放给某个地理区域的全部观众的。同样，在大多数学校里我们无法在一个教室内对每个学生实施不一样的教学大纲。在这两个例子里，研究者只能在地理区域维度和教

室维度进行随机化。

第三，集群随机化是解决溢出问题的方法之一。比如我们担心同一个工作环境下的不同实验条件之间可能存在互相干扰的情况（例如同一车间的员工会观察到其他员工所处的实验条件，从而调整自己的行为），那么可以把关联的个体以集群的形式来随机分配。

8.2.3 区组随机化（block randomization or stratified randomization）

区组随机化是指将被试分成区组（blocks / strata）并在每个区组内进行完全随机分配。我们来看一个区组随机化的示例。假设某公司想知道团建是否会影响新老员工的工作积极性，并计划做一个田野实验，于是需要把自己的员工随机分配到干预组（进行团建）和控制组（不进行团建）。由于这个公司成立时间较短，他们的新员工比例远远大于在公司工作三年及以上的老员工的比例。公司希望干预组和控制组中的新老员工比例相等。如果使用简单随机分配的话，那么在新老员工比例失衡的情况下，尤其是样本量不够大的时候，不同实验条件之间的新老员工比例很可能会有所不同。此时，区组随机化可以确保不同实验条件下的新老员工比例相同。

如何进行区组随机化呢？在上面这个示例中，你首先需要先将这个公司的员工分成两个区组：进入公司三年以下的新员工和进入公司三年及以上的老员工。然后，在新员工区组中，你需要将他们随机分配到干预组和控制组中。同样，在老员工区组中，也随机把他们分配到干预组和控制组中。在实际操作中，你其实在同一个公司样本中根据区组进行了两次简单随机分配。这样的好处是不用担心干预组和控制组中新老员工比例严重失衡。总的来说，区组随机化实际上是创建了一系列的简单随机化：在每个区组中分别进行一次简单随机化，来最终确保不同的实验条件中区组内不同被试的比例大致相等。

区组随机化通常被用于解决两种类型的问题。首先，区组随机化可以解决实际操作或实验伦理的问题。例如，一个教育项目可能对年龄和性别等人口指标有要求，希望不同实验条件拥有比例相等的人口指标。区组随机化可以帮助研究者围绕这些约束条件来进行随机实验。比方说，可以先将年龄接近、性别相同的被试放在一个区组里，再在每一个区组里（比如20—30岁女性区组，40—50岁男性区组）进行简单随机化。最终的效果是你的干预组和控制组中被试的年龄和性别比例相当。

其次，区组随机化解决了重要的统计问题，包括减少抽样变异性和确保某些区组可以用于单独分析（Gerber & Green，2012）。区组随机化可以提高因果推论的精准度（precision），尤其是在样本量相对较小的情况下。一个变量和因变量相关性越大，基于这个变量做区组随机化就越能提高因果推论的精确度。我们可以通过查看文献和进行实地观察（field observations）来预测哪些区组可能和我们需要测量的因变量有关联。比方说，如果干预旨在提高员工绩效，研究者可以查看哪些基线因素（如工作经验、性别、年龄等）和员工绩效相关，并把这些基线因素作为区组随机化的根据，以便更好地在实验设计中控制这些变量。

然而，区组随机化并非总是可行的。有时田野实验有严格的时间限制或财务限制，研究者往往没有时间精力、财力来进行区组随机化。又或者，研究者在实验设计阶段无法获得区组随机化所要求的背景信息。但由于区组随机化仍具备许多优势，我们还是应该秉承 Gerber & Green（2012）对于田野实验的经典建议：在条件允许的情况下，尽量进行区组随机化。

8.2.4　被试内设计（within-subjects design）

上面说到的随机方式都属于被试间设计（between-subjects design），即在同一时间对随机分配到不同实验条件中的实验单元（如单个被试或集群）进行比较。而被试内设计指的是每个实验单元在不同的时间经历不同的实验条件。如果说被试间设计是把若干实验单元随机分配到不同实验条件中来比较不同实验条件下的结果，被试内设计则是把每个实验单元在不同时间点分配到不同实验条件中，来比较不同时间点下每个实验单元在不同实验条件下的结果。如果一个实验使用的是纯粹的被试内设计，那所有被试在同一时间点接受的是同一个实验条件，尽管不同实验条件开始的时间点是研究者随机选的（比如为期四周的实验，用掷硬币的方式决定哪两周是干预、哪两周是控制）。被试内设计不只适用于被试，也适用于单个集群。比如，Brownell et al.（1980）为了鼓励更多出行人员走楼梯而不是坐电梯，在多个地铁站入口放置一个彩色的标识来鼓励进站人员走楼梯。这些标识有些星期会摆放，有些星期会拿掉。摆放的星期是随机挑选的，对于每个地铁站入口是相同的。因变量是每星期使用楼梯的人数——由研究助理在每星期的固定时间在地铁口蹲点记录。在这个实验里，实验单元是因变量的测量地点（楼梯口），属于集群，而每个集群在每个测量的时间点都会产生数据。

这种纯粹的被试内设计的优势在于，对每个实验单元我们都能观察到干预组和控制组比较的一个精确数值估计。这个数值的精确性就在于被试内设计中的单个实验单元都同自己做比较，相当于控制了被试内的其他变量。但是，如果我们想让被试内设计产生无偏估计，我们需要做出无干扰（non-interference）假设，即在不同时间点下的被试反应是独立的。

无干扰假设存在两个子假设：无期盼（no-anticipation）假设和无持续（no-persistence）假设。举个例子，假设我们做一个为期四周的实验，通过掷硬币的方式来决定哪两周被试会经历一个自我肯定（self-affirmation）的干预项目（干预组），哪两周不经历这个干预项目（控制组）。因变量是被试下班后的主观心情。如果使用被试内设计的话，比较的就是每个被试在经历自我肯定的干预期间每天下班后的心情和不在干预期间的下班后的心情。试想一下，在这个例子中使用被试内设计会存在哪些问题？如果被试签了实验知情书，那么他们可能预期自己在某个时刻会经历某种干预。如果被试首先进入的是没有任何的控制组，那么被试也可能会对干预产生期盼，这种心态可能会影响他们下班后的心情，从而违反无期盼假设。我们可以使用安慰剂控制解决这类问题。抑或被试首先进入的是干预组，两周内每天都接受自我肯定干预，两周后即使进入了无干预的控制组，但干预的效果很可能是持续性的，进而影响被试在接下来控制组中下班后的心情，这就违反了无干扰假设下的无持续假设。如果无干

扰假设不成立，被试内设计下比较每个被试不同时间点的结果就存在偏差，也就无法得出可靠的因果推论。

所以，当研究者们在田野实验中使用被试内设计时，必须认真考虑无干扰假设在理论设计层面和实际操作层面是否成立。在考虑这个问题的时候，可以思考过去文献中有没有用被试间设计研究过你感兴趣的这个现象，有无发现这个现象有持续性？如果有持续性的话，你有没有可能在实验设计中设立一个洗脱期（wash-out period）？比如某干预的效果持续期为一周，你有没有可能将干预后的一周作为洗脱期、不用于数据分析，而是再过一周才开始控制组的实验？有没有可能被试会对接下来的干预内容有期待？这种期待是否影响你的结果测量？这些问题需要研究者了解田野实验的环境和干预内容后再做出判断。被试内设计如果要得出可靠的因果推论，对实验设计本身的要求是很高的，比如研究者要保证不同时间点内外部条件是可控制的、不互相干扰的，干预后设置洗脱期。因为无干扰假设在社会科学研究的问题中很难完全成立，纯粹的被试内设计得出的因果推论往往不是最有说服力的。如果因为种种原因，一个实验必须涉及被试内设计的，我们建议研究者考虑下节介绍的候补名单设计，将被试内设计和被试间设计相结合。

8.2.5　候补名单设计（waitlist design）

候补名单设计，也被称为阶梯楔形设计（stepped-wedge design）或随机推出设计（randomized rollout design）。候补名单设计随着时间的推移跟踪被试从他们未曾经历实验干预的情况转换到接受实验干预的情况，或者从实验干预转换到控制观察的情况。候补名单设计的核心在于，不同被试进入干预组的时间是不等的，并且每个被试接受实验干预的时间点是随机分配的。如果我们画一张图（见图 8–1），被试进入干预组的时间看起来就像一级级的阶梯一样，因此候补名单设计也被称为阶梯楔形设计（Hussey & Hughes，2007）。

被试间设计

被试	时间 1
1	1
2	1
3	0
4	0

被试内设计

被试	时间 1	2
1	0	1
2	0	1
3	0	1
4	0	1

候补名单设计

被试	时间 1	2	3	4	5
1	0	1	1	1	1
2	0	0	1	1	1
3	0	0	0	1	1
4	0	0	0	0	1

图8–1　干预组和控制组的被试间设计、被试内设计、候补名单设计分配方式

注：0 表示控制组，1 表示干预组。被试间设计通常在一个时间段内随机分配，被试内设计和候补名单设计通常包含多个时间段。

候补名单设计有两个主要优势。第一，这种设计结合了被试间设计和被试内设计的元素。与被试内设计一样，候补名单设计会生成时间序列数据，单个被试在不同时间点经历不同的实验条件。但与传统的被试内随机化不同的是，候补名单设计在同一时间点，一些被试接受

干预，而另一些被试则在无干预的控制组，这又与被试间设计类似。因此，候补名单设计产生的数据是丰富的，我们既能做同一时间点上的被试间比较，也能做同一被试在不同时间点的比较。它使研究者能够从相对较少的受试中提取更多具有统计意义的因果估计值，因此能增加统计功效（Aronow & Samii，2012；Gerber & Green，2012；Rubin，2001）。

第二，候补名单设计能解决田野实验在实际操作或实验伦理要求上的一些限制。有的时候，单纯的被试间设计是现实条件不允许的。比如我们想研究某电视广告对产品销售的影响。假设我们的实验对象是 10 个独立媒体市场，且广告公司需要 10 个市场近期都播出测试广告。但由于财务或日程安排的限制，广告同一时段只能在 2 个或 3 个媒体市场播出，这种情况就不适合只有一个时间区间的被试间设计了。这时，我们可以向广告公司推荐随机抽取不同的媒体市场在不同的时间点播放新的电视广告。比如我们可以随机抽取 3 个媒体市场在第一周内进入干预组（即播放新电视广告），另外 3 个随机抽取的媒体市场在第二周进入干预组，等等。Wu & Paluck（2021）有关工厂车间的实验呈现了一个更复杂的候补名单设计的例子。

有时，为形成一个纯粹的控制组而拒绝对一部分被试进行干预，在实践或道德层面是不可行的。比如，当实验干预可能对被试产生很大的正向影响时（如技术培训或新的教学方法可能帮助员工 / 学生提高工作效率或学习成绩），田野实验合作机构对于设置一个完全不受干预的控制组可能是有顾虑的。在这种情况下，研究者可以使用候补名单设计来保证所有被试最终都会接受实验干预，从而削弱合作机构的顾虑。

8.3　田野实验设计的注意事项

8.3.1　样本量（sample size）和功效分析（power analysis）

和实验室实验一样，田野实验需要事先做功效分析。如果研究者的目的是测试干预组和控制组是否具有显著的差异，那么就需要了解检验功效（statistical power），它是指在干预效果确实存在的情况下，研究者有多大概率能够拒绝干预效果为 0 的这个原假设。我们接下来着重介绍的是在田野实验中功效分析的目的以及注意事项。关于功效分析的具体计算公式，感兴趣的读者可以阅读 Cohen（2013）。[①] 一般来说，在一定的统计显著水平下，样本量越大、干预效果越强、结果变量的标准差越小，实验的检验功效越大。

对于田野实验而言，做功效分析是为了实现以下三个目的中的一个或者多个（Duflo et al.，2007）：第一，判断需要多大的样本量才能达到一定的功效，进而判断田野实验的合作机构或者研究者自己需要投入多少资源来获取足够的样本。第二，考虑到田野实验的样本量很大程度上取决于经费、干预的实施难度和合作机构能提供的被试数量（例如某个公司可能只有一

① 除了套公式，研究者也可以通过模拟（simulation）来做功效分析，具体如何操作我们这里不做讨论。

个 100 人的工厂能做实验），研究者可以通过功效分析来判断在给定的样本量和预期的干预效果下，一个实验能有多大的功效，从而决定做这个实验是否值得。第三，功效分析可以帮助研究者思考在给定的限制条件下，如何设计实验来增大检验功效。接下来我们逐一讨论这三个目的。

对于第一个目的，即计算需要多大的样本量才能达到特定的检验功效，我们需要知道的因素包括希望达到的检验功效（一般设定在 80% 以上），预定的统计显著水平（通常是 5% 或 10%），结果变量的标准差，以及希望能检测到的干预效果大小。最后两点信息可以通过合作公司的历史数据（historical data）、研究者的前期调研（baseline survey）、前导实验（pilot test），以及在类似人群上研究类似干预的文献来获得。关于希望能检测到的干预效果大小，如果研究者关注的是一个干预的应用价值（例如允许在家办公能否提升员工的工作效率），那需要考虑干预效果至少得达到什么水平（例如员工的工作效率至少要增加多少）才对公司有足够的实际意义，或者才能使得这个干预是经济有效的（cost effective），而不仅仅是考虑统计意义上的显著性。如果研究者只是想从理论角度知道一个干预有没有可能引起行为改变，那么即便很小的实验效果可能也是研究者感兴趣的。

如果绝对的干预效果大小（例如干预组的员工能比控制组的员工多完成多少订单）或者结果变量的标准差难以判断，研究者可以在功效分析时输入他们希望检测的标准化效应量（例如干预组和控制组员工的工作效率差异是多少个标准差）。当结果变量是连续变量时，标准化效应量对应的就是常说的 Cohen’s d。根据 Cohen（2013）的分类方法，0.2、0.5、0.8 的标准化效应量分别对应的是很小、中等、很大的实验效果。这个捷径能够帮研究者大体判断他们需要多大的样本才能有 80% 的功效来检测一个很小、中等、很大的效果。如果研究者没法获得合适的参考信息，我们建议采取保守的策略，预设一个偏小的标准化检验效应量（Richard et al.，2003）。

值得注意的是，如果研究者采用的是集群随机化，那么每个集群内的个体之间的结果变量很可能具有相关性，功效分析就需要考虑集群内个体之间的结果变量相关性大小。更准确地说，应该是考虑集群内相关系数（intraclass correlation coefficient），即所有被试的结果变量的变异在多大程度上是来自集群之间结果变量的变异（而非集群内部被试之间的变异）。集群内相关系数越大，说明集群内部的被试行为更加同质，那么依照普通的功效分析计算出来的功效就需要打更大的折扣；在一定的样本量和一定的集群内相关系数下，集群内的被试数目越大（意味着集群数目越小），那么功效需要打折扣的幅度也就越大（Killip et al.，2004）。如果没有被试的历史数据，直接估算可靠的集群内相关系数会比较困难。研究者可以假设不同程度的相关系数，分别计算需要多大的样本量，从而得到所需样本量的区间。[①] 切记，在设计实验时，增加集群的数目对于增加实验功效的作用会大于增加集群内被试数目的作用。直

① 在涉及区组或集群随机的非简单随机的情况下，通过模拟进行功效分析尤其有用。我们推荐使用Declare Design，这是由政治学学者专门为田野实验的功效分析而编写的R语言的统计方法包。

观来说，我们可以这么理解：当集群内的被试高度相关时，如果研究者想要增加一个被试，那么找一个新的集群里的个体作为新增被试所能提供的信息量要大于从已有的一个集群里再找一个新个体作为被试所能提供的信息量。

对于第二个目的，即计算在某个样本量下能实现多大的功效，计算方法和第一个目的本质上是一样的，只不过从限定功效来求样本量变成了从限定样本量来求功效。第二个目的是田野实验相较于实验室实验比较特殊的一点，因为田野实验的样本量会受到合作机构所能提供的资源（例如可以用于实验的员工或者用户数目、可以用于推动实验和实施干预的人力和资金）的限制。那么在决定开展合作之前，研究者可能需要知道在给定的样本量下，自己能有多大的概率检测到一定的实验效果，从而判断是否值得花时间来做这个实验。概率越大（即功效越大），那么研究者自然是越有信心。但如果功效低，这个实验还是否值得做呢？这个问题比较难回答。我们赞成 Duflo et al.（2007）的观点：从研究者个人利益的角度，最好避免在一个功效很低的实验上花费精力；但是站在整个学科的角度，功效低的实验不是没有意义的，因为一个重要的学术议题一般会有多个团队用不同的实验来推进迭代，那么最初的一些小样本、低功效的实验可以为这个议题的后续发展提供重要的数据，而且多个功效低的实验也可以合并起来做元分析（meta-analysis）。

对于第三个目的（借助功效分析，思考如何设计实验来增加功效），一个显而易见的思路是，如果我们发现自己样本量所能达到的统计功效较低，那我们需要考虑如何增强干预的效果（例如提升干预的频率和强度）。除此以外，我们可以通过在和结果变量高度相关的基线变量上做区组随机化或者控制这些变量，来提高实验的功效。另外，研究者也可以从样本分配的角度动脑筋，这里举几个例子，请参见本章线上资源。

8.3.2　溢出效应（spillover）

随机实验方法的一个重要假设是一个被试的潜在结果（potential outcomes）不受到其他被试所处实验条件的影响。这个假设是随机分配实验条件也未必能够保证的，需要研究者在设计实验和分析数据时格外注意。溢出效应指的就是这个假设不成立的情况——被试的潜在结果会根据其他被试所处的实验条件而变化。假设说有一个以计件工资方式支付薪酬的工厂，设计了一个实验来评估员工技术培训对于生产效率的影响。随机分配到干预组的员工接受了技术培训，随机分配到控制组的员工则不接受任何的培训。如果同一个车间里的控制组员工观察和模仿干预组员工的操作流程，因此也提高了工作效率，那么这个干预（技术培训）就产生了正向的溢出效应。在这种情况下，控制组和干预组员工在培训结束之后的工作效率差异要小于培训对于生产效率产生的真实影响。如果同一个车间里的控制组员工发现干预组的员工接受了培训，感到不公平，因此消极怠工降低了工作效率，那么技术培训这个干预就产生了负向的溢出效应。在这种情况下，控制组和干预组员工在培训结束之后的工作效率差异要大于培训对于生产效率产生的真实影响。一般来说，在正向溢出的场景里，比较控制组和

干预组在结果变量上的差异会低估干预的真实效果；而在负向溢出的场景里，比较控制组和干预组在结果变量上的差异会高估干预的真实效果。

怎么解决溢出效应对于评估干预效果的影响呢？如果溢出效应只存在于每个集群内部而不存在集群之间，那么集群维度的随机分配可以得到对于干预效果的无偏估计。比如说在前面提到的例子里，如果工厂有多个独立的车间，可以以车间为单位，一部分的车间随机分配到干预组接受培训，另一部分的车间随机分配到控制组不接受培训。如果不同车间的员工不会相互交流，那么比较干预组车间员工和控制组车间员工的生产效率就可以得到技术培训对于员工生产效率的影响；否则这种实验分配方式依然不能解决溢出效应的问题。

如果研究者希望能够测量出干预的溢出效应，那么只做集群随机化是不够的。一种解决办法是在每个集群内部，让不同的被试在不同程度上接触干预。比如说在上述的工厂例子里（假设溢出效应只可能存在车间内部），可以先将一部分的车间随机分到干预组，然后在干预组的车间里再随机挑选一部分员工来真正接受培训。通过将被随机挑选真正接受培训的员工和控制组车间的所有员工做比较，可以得出培训对生产效率的直接影响；而将干预组车间里没有被随机选中做培训的员工和控制组车间的员工做比较，可以得出培训对于生产效率的溢出效应。使用这个方法的案例参见 Banerjee et al.（2015）和 Duflo & Saez（2003），背后的统计原理参见 Hudgens & Halloran（2008）。

做集群随机化需要较多的集群，因为如前文所说，实验的功效受到集群数量的影响，且这个影响一般要大于每个集群内被试数量所产生的影响。而且这些集群需要相对独立来避免集群间的溢出效应，因此这种方式并不适用于所有情况。在这种情况下，我们可以尝试从数据分析上检验溢出效应的可能影响。读者可以通过本章线上资源获取 Bai et al.（2022）和 Bloom et al.（2015）的案例，也可以参考 Aronow et al.（2021），Aronow & Samii（2017），Gerber & Green（2012）学习更多的消除和估算溢出效应的统计方法。

除了在实验随机化方式和数据分析上做文章，研究者在实验实施阶段也可以注意降低溢出效应。比如说，干预的实施尽量不打乱员工正常的工作流程，尽量不要让控制组员工知道干预的存在，这样控制组员工不会因为自己没有受到干预而改变对于公司、领导或者工作的态度，而且控制组员工也不会因为预期自己之后会受到干预而调整现在的工作状态。类似的，尽量不要让干预组员工意识到自己处于干预条件下，最好也不要让管理者知道员工（或团队）分别被分配到了哪个实验条件下，以防止管理者区别对待不同实验条件下的员工（或团队）。总之是在条件允许的情况下，尽量使用双盲的实验设计。另外，也尽量避免不同实验条件下的员工交流和实验相关的内容，比如若干预是在一个会议里实施的，那么公司可以明确规定员工不可以和他人讨论会议上的内容。

8.3.3　样本流失（attrition）

在实验情境下，样本流失指的是研究者不能从所有最初被放到实验里的被试处获得结果变量。这个问题在实验室实验中也存在（见本书第 6 章）。但田野实验里有更多的因素可能导

致样本流失。比如说，研究者需要被试在下班后一个小时内完成一份问卷，但并非所有被试都愿意或者有空在下班后一小时内填写问卷；研究一个干预的长期效果时，如果被试后期更改了姓名、地址或电话号码，那么研究者可能会和被试失联，无法再获得被试后期的结果变量；如果被试在研究的观察期内离开了公司，那么这个被试的最终结果变量（比如在观察期结束时的职位是否高于实验开始时的职位）也是缺失的。另外，有的时候，实验设计的因素可能会导致某些被试缺失结果变量。比如，研究者想看的是政府提供的技能培训对于只有高中学历的市民一年后工资的影响，那么对于一年后没有在工作的人而言他的结果变量就是缺失的。

如果样本流失是完全随机发生的，那么它只会降低实验的有效样本量，进而减少功效；但如果样本流失和被试接受的实验条件有关，那么它就可能带来对干预效果估计的偏差。假设一个公司想知道消耗资源进行企业文化培训是否值得。他们设计了一个实验，随机分配到干预组的员工参加一整天的企业文化培训，控制组的员工不参加且不知道有这样的培训。一周之后，人力资源部给干预组和控制组的员工都发了邮件，请他们填写一份问卷，其中包括了这个实验的一些结果变量（如企业忠诚度和工作投入度）。如果说干预组里那些对企业文化最不感兴趣的、受到培训的正向影响最小的那部分员工也是最懒得填写问卷的（即这部分员工最容易产生样本流失），那么在填写问卷的员工当中比较干预组和控制组的企业忠诚度将会高估这个培训的作用。总体来说，随机分配可以保证潜在结果独立于被试最初被分配到的实验条件。但是，一旦发生样本流失不随机的情况，在有完整结果变量的被试中，潜在结果与实验条件的独立关系就无法得到保证了，就不能通过简单地比较没有流失的被试在干预组和控制组的差异来准确获得对于干预效果的因果推断。

值得注意的是，即便样本流失率在不同实验条件之间看起来是很接近的（或者至少是没有统计意义上的显著差异），样本流失也依然可能是带有选择性的，导致没有流失的样本在不同实验条件之间依然有质的差异。比如说，一方面，企业文化培训可能导致一些对培训有抵触情绪的员工更不愿意填写人力资源部发的问卷，这是可能导致干预组问卷回收率低于控制组的一个原因；另一方面，企业文化培训可能让另一些员工觉得填写问卷本身就体现了他们的集体精神，因而填得更积极了，这是可能导致干预组问卷回收率高于控制组的一个原因。最终导致干预组和控制组的问卷回收率差不多，但是是否填问卷这个决策是带有选择性的，和员工所在的实验条件有关，最终填写了问卷的干预组员工和填写了问卷的控制组员工其实并不可比。

样本流失问题最好能从实验设计和数据收集的角度来避免或者尽量削弱。通过和合作机构的深入沟通，在条件允许的情况下与被试做焦点小组或者进行小规模的前导实验，研究者可以了解在实验过程中哪些环节与因素可能导致样本流失，哪些类型的被试更可能受到这些因素的影响，哪些变量更可能面临样本流失问题，从而寻找解决方案。假如一个研究团队想知道给没有高等学历的市民提供技术培训能否长期提高他们的就业率和生活质量，租房生活

的市民可能在实验观察期会搬家，那么利用实验开始时登记的住址就没法找到这些市民来追踪就业和生活情况。如果预料到这种情况，研究团队应在实验开始前尽量获得被试多个联系方式（比如电话号码、微信号），用其他联系方式找到被试，尽可能去他们的新住址记录就业和生活情况。如果说追踪每一个流失样本的成本太高，也可以从流失的样本中随机选择一部分进行紧密追踪（案例见 DiNardo et al.，2021）。

在选择结果变量时，考虑什么结果变量不太会受到样本流失的影响。比如说最好能使用合作机构本身就有记录的行政信息（administrative data），而不仅仅依赖需要员工自愿填写的问卷。在上述研究企业文化培训的例子里，可以看到公司本身就会定期记录员工工作绩效数据，比如销售业绩、客户满意度、360 度绩效评估等。这样的结果变量对于所有在职的员工都是完整的，可以和通过问卷收集的、可能存在缺失的员工工作态度数据一起使用，案例参见 Chang et al.（2019）。 有的结果变量即便被试“失联了”，也依然能够得到有效的数据。这样的变量可以考虑加入到数据分析中，案例参见 Linos et al.（2021）。以上两个案例请参见本章线上资源。

总体而言，在数据收集过程中，研究者应该尽力获得完整的行政记录，与合作机构中负责收集数据的工作人员保持密切的沟通和良好的关系，向被试说明填写问卷的重要性，并及时跟进被试来尽量减少因为被试不积极参与而带来的结果变量缺失。最好能够从多种渠道收集多个结果变量，这样不同变量有着不同的缺失程度、缺失原因和缺失对象，那么一个变量的缺失数据也许能够通过其他变量估算出来，或者至少可以通过多个变量的实验结果来相互佐证。

有的时候结果变量的数据缺失是无法避免的，那么数据分析的第一步应该是检查不同实验条件下的结果变量流失率，并且看流失的样本和未流失的样本是否在基线数据上有着系统性的差异，从而判断样本流失是否存在一定选择性。如果样本流失率在不同实验条件下有差异或者存在选择性（比如和基线的行为与态度数据相关联），那么可以使用一些统计方法来调整样本流失对于实验效果评估可能会带来的偏差。一个简单的方法是计算平均处理效应的上下限。假设干预组缺失的数据等于样本里观测到的最大值，控制组缺失的数据等于样本里观察到的最小值，那么填充数据之后再比较干预组和控制组的被试得到的是干预效果的上限；反之，得到的会是下限。这种方式在数据缺失量小且样本里观察到的最大值和最小值差距不大时，比较有用。否则，还需要额外的假设和调整（Gerber & Green，2012）。感兴趣的读者还可以参考其他的方法，如 Graham（2009）、Hausman & Wise（1979）、Horowitz & Manski（2000）、Lee（2009）。

8.3.4　不依从行为（noncompliance）

到目前为止，我们都是假设干预组和控制组被试的经历和实验设计是吻合的：干预组的被试都接受了干预，而控制组的被试都没有接受干预。但在田野实验中，现实往往并非如此。

单边不依从（one-sided noncompliance）指的是一些被分配到干预组的人实际上并没有受到干预，因为研究者并不能强迫干预组的被试接受干预。比如说 Zhang et al.（2020）和阿里巴巴合作研究折扣券对于一百万名用户短期和长期消费行为的影响。和在商店里直接给用户发折扣券不同，在天猫和淘宝里发折扣券时，干预组用户在实验期间未必会打开消息盒子，因此未必会看到优惠券，也就未必会真正地接受到这个干预。控制组用户没有被发送优惠券，也不可能接收优惠券。这个实验就存在单边不依从的情况。

双边不依从（two-sided noncompliance）指的是不仅干预组有一些被试没有接受干预，而且控制组有一些被试其实接受了干预。这是因为研究者不仅不能强迫干预组被试真正接受干预，而且他们也不能完全保证控制组被试不得到干预。比如说 Zhang et al.（2019）和阿里巴巴合作研究线下体验店对线上消费的影响。阿里巴巴于 2017 年秋季在杭州一个购物中心设立了一个关于牛仔品牌的线下体验店。他们定位了约 80 万名在购物中心周边居住的用户。研究者和阿里巴巴真正感兴趣的、希望用户能接受到的“干预”是让用户去线下体验店，但他们既不可能强迫一些用户去体验店，也不可能剥夺一些用户去体验店的权利。他们只能改变是否引导某些用户去体验店，这种设计叫作随机鼓励设计（encouragement design）。在 Zhang et al.（2019）的实验里，被随机分在了短信引导组的用户，收到了一条短信，短信中告知他们线下体验店的时间和地点，鼓励他们去线下店。被随机分在控制组的用户没有收到这条短信。在随机鼓励设计里，干预组的实验条件和研究者真正希望被试接受的干预内容是不一样的。比如，Zhang et al.（2019）里干预组的实验条件是收到一条引导短信，而研究者真正希望用户接受的干预是让用户去体验店。考虑到分在了干预组的用户未必会去体验店，控制组的用户也未必就不会去体验店，因此这里存在双边不依从的情况。

不依从行为是否会对研究造成问题取决于研究目标。如果研究者感兴趣的就是给被试提供一个干预（相较于不提供这个干预）对于结果变量的影响，而不在乎被试是否真正地接受了干预，那么不依从行为就不太重要。比如说，田野实验的目的是评估一个给员工提供部分在职 MBA 学费的项目，研究者从相对独立的部门中选了一部分作为干预组，告知员工如果他们未来 3 年内读在职 MBA，公司将提供 20% 的学费，而控制组部门的员工不知道也不享有这个学费减免待遇。公司关心的就是这个项目（即提供学费支持）本身的价值，而不考虑员工是否真的读了在职 MBA。举个学术的例子，Milkman et al.（2011）给一家公司的 5 000 多名员工发信件，鼓励员工打流感疫苗。发给干预组员工的信件额外地鼓励员工制订打疫苗的计划，并鼓励他们写下准备什么时间去打疫苗。Milkman et al.（2011）关心的就是鼓励员工做计划对于员工疫苗接种率的影响，收到鼓励的员工未必都做计划了，而没收到鼓励的员工也可能自发地做了计划。在这种情况下，正确的分析方式应该是不管这些被试是否做了计划，比较所有被分到干预组的被试以及所有被分到控制组的被试最终的疫苗接种率。这种分析方式叫作意向处理分析（intent-to-treat analysis）。

有的时候，研究者感兴趣的并不是（或者不仅仅是）他们通过随机分配来改变的实验条

件如何影响了被试的行为；他们想了解的是如果被试真正接受了干预，他们的行为会如何改变。比如说在前文提到的线下体验店的例子中，研究者和阿里巴巴真正想测量的并不是给用户发一个关于体验店的短信对于用户线上消费有什么影响，而是用户去线下体验店对他们的线上消费有什么影响。在这种情况下，意向处理分析（即比较所有分到干预组的用户和被分到控制组的用户）不能完全回答后面的这个问题。[①] 那我们能通过比较干预组收了短信之后去体验店的用户和所有控制组的用户来回答这个问题吗？这是错误的，因为是否去体验店本身是带有选择性的，而非随机的。假设说更喜欢购物的用户更可能在收到短信之后去体验店感受一下，那么这种错误的分析方式将会高估去体验店对于后续线上消费的拉动作用。那么有什么更好的办法来估算被试真正接受干预后的行为改变呢？

在满足一定的假设情况下，研究者可以使用工具变量法（instrumental variable analysis）计算出实际接受干预对于结果变量在依从者（compiler）中的影响，这种影响也被称作局部平均处理效应（local average treatment effect，LATE）（Imbens & Angrist，1994）。依从者指的是符合如下特点的被试：当他们被分到了干预组时，他们就会接受干预；当他们被分到了控制组时，他们就不会接受干预。工具变量法的核心思想是，被试是否被随机分到了干预组可以作为被试是否实际接受了干预的工具变量（instrument）。回到上文提到的线下体验店的例子。如果满足一定的假设，那么被试是否被随机分到了短信组（即是否有短信引导被试去线下体验店）可以作为被试是否去了线下体验店的工具变量，从而计算出去线下体验店对于依从者后续线上消费的影响。这里依从者指的是那些如果收到了关于线下体验店的短信就会去体验店，而如果没有收到短信提醒就不会去体验店的用户。使用工具变量法的假设请参见本章线上资源。

想要使用工具变量法计算依从者的处理效应，研究者应首先明确自己感兴趣的干预到底是什么，自己能够随机分配的实验条件又是什么，以及如何定义和测量“实际接受了干预”。如果没法测量被试是否实际接受了干预，那么局部处理效应也就无从谈起了。具体计算方法可以使用 Wald Estimator，等同于随机分到干预组和控制组的被试在结果变量上的差异除以这两组被试实际接受干预的比例的差异。也可以使用两阶段最小二乘估计法（2SLS）。需要注意的是局部平均处理效应未必能够代表整个被试群体的平均处理效应，因为依从者和其他的被试可能不同。比如那些因为收到短信而愿意去体验店的用户可能不同于那些不管是否收到短信都去体验店的用户，不同于那些不管是否收到短信都不愿意去体验店的用户，也不同于那些不受短信影响但会因为其他引导措施（比如发放体验店优惠券）而更愿意去体验店的用户。

如何降低不依从现象对实验结果的影响呢？一方面，研究者可以通过调整提供干预的方

① 虽然存在不依从问题时，意向处理分析不能得出“真正接受了干预对于被试行为的影响”，我们还是建议研究者要做这个分析。不同于我们后述的工具变量法，意向处理分析是不需要任何假设的。在随机分配的情况下，意向处理分析能让我们可靠地估算“提供干预或者引导被试接受干预对于被试行为的影响”。

式或者引导被试接受干预的方式来减弱不依从的现象。在实验设计阶段提高依从者占比的好处有三个：第一，如果依从者占比很高，干预组和控制组的被试实际接受干预的比例差异很大，此时即使排除性限制有微小的违反，局部平均处理效应的估算也不容易带来偏差。第二，依从者占比越高，局部处理效应的标准误会更小（假设其他条件不变），更容易得到显著的统计结果。第三，依从者占比越高，依从者对于整体样本的代表性也越高，那么得到的局部平均处理效应就越能代表实际接受干预在整个被试群体中的平均效应。因此在实验设计阶段，最好能够通过前导实验和利用合作机构对于被试群体的深入了解，来判断不依从问题是否严重及其可能出现的原因，并试图提高依从者占比。

另一方面，在给定的依从率（尤其是当依从率较低的时候）下，研究者可以使用安慰剂设计来降低统计不确定性。安慰剂设计包括两步：先找到更可能依从的被试，然后再将这部分被试随机分到干预组（实际接受干预）和控制组（不接受研究者感兴趣的干预，但会接受其他的、理论上不会影响结果变量的安慰剂处理）。假设我们研究上门拉票（face-to-face canvassing）对于居民参与选举投票的影响，一个思路是先把一个区域内符合条件的居民按照家庭地址分到干预组（上门拉票）和控制组（不上门拉票），然后只拜访分到干预组的家庭，如果居民开门了，就进行拉票游说。这样做的问题是，有的干预组居民不在家或者不给陌生人开门，那么这些居民将不会实际接受到干预（被游说）。所以这种设计存在单边依从问题。Nickerson（2008）采用了安慰剂设计：研究者首先将家庭分到了干预组（上门拉票）和安慰剂组（上门聊环保），然后训练有素的访问员拜访了所有符合研究条件的家庭；当某家开门的时候，访问员根据这个家庭的实验条件，要么进行拉票游说，要么聊环保。最终，研究者首先可以只分析开门的家庭，比较两个实验条件下人们的投票行为。使用这种安慰剂设计的注意事项，参见 Gerber & Green（2012）。参见本章线上资源了解 Brody et al.（2022）是如何减弱不依从现象的。

8.4　田野实验分析的注意事项

8.4.1　样本平衡检验（balance checks）

汇报田野实验的结果时，应该首先汇报样本平衡检验结果。这需要研究者比较被分到不同实验条件下的被试的基线数据（baseline data），尤其是和潜在结果相关的变量数据（比如说结果变量的历史值）。基线数据指的是在实验开始前就有的数据，可以是实验开始前研究者测量的或者公司档案里记录的，也可以包括年龄、性别、种族等基本不会因为实验条件而变化的特征（不过最好这些特征也是来自实验前的数据）。如果基线数据在不同实验条件之间是可比的，那么这可以作为实验条件被成功随机分配的证据。如果某些基线数据在某些实验条件之间出现了显著的统计差异，那么就需要判断这是随机机会（random chance）导致的，还是因为随机分配或者数据收集出了问题。我们推荐读者阅读 Gerber & Green（2012），了解如

何处理这种情况。另外，如果核心结果变量存在样本流失问题，那么我们建议样本平衡检验不仅应该在最初被随机分组的被试中进行，而且应该在最后用于分析的样本中进行。

8.4.2　协同控制变量（covariates）

分析田野实验常见的问题之一是回归分析的时候，要不要添加控制变量。答案取决于添加的是什么样的变量。如果我们说的是一个在实验前测量的变量（基线数据），那么是否控制这个变量是不影响实验效果估计值的期望值的。这是因为实验条件的分配是随机产生的，不同实验条件之间观察到的被试在基线数据上的差异来自随机性误差（而非不同实验条件之间的系统性差别）。如果一个实验前测量的变量能够较好地预测或者影响结果变量，那么加入这个变量一般可以降低实验效果估计值的标准误（standard error），从而提高实验的功效。这种变量值得添加。但如果一个实验前测量的变量不能预测或者解释结果变量，那么将其作为控制变量反而会因为降低统计上的自由度（degrees of freedom），提高实验效果估计值的标准误，进而降低实验的功效。这种变量就不值得添加了。更糟糕的是，如果我们说的是在实验开始后测量的变量，那么这个变量是有可能被干预所影响的。控制这样的变量可能会给实验效果的估计带来偏差，因为这个变量在不同实验条件间的差异也体现了干预的作用。切记，在田野实验的分析中，控制变量应该来自基线数据。

如果结果变量的基线值和其在实验观察期内的值是高度相关的话，那么它就是一种有价值的、特殊的控制变量。但如果一个结果变量的基线值和它在实验期间的值关联性不大或者基线值的测量存在误差，那么添加结果变量的基线值反而可能会降低实验的功效。在有必要加入基线值的情况下，除了在回归分析中加入结果变量的基线值作为控制变量，研究者也可以将结果变量实验期间的观察值和基线值相减来作为调整后的结果变量。如果研究者有纵向数据，从实验开始前的一段时间到实验开始后都有测量结果变量，那么也可以使用双重差分法来分析数据。案例参见 Milkman et al.（2022）。

如果控制变量的数值有缺失怎么办？比如说，控制变量是员工实验前一年的工作满意度，但是对于实验前刚加入公司的新员工这个变量是缺失的。第一步，创建一个新的虚拟变量（dummy variable），它对于基线工作满意度有缺失的员工等于 1，对于其他员工等于 0。第二步，对于基线工作满意度有缺失的员工，给他们的基线工作满意度补上一个值（什么值并不重要，只要这些员工用的是同一个值）。第三步，在回归分析的时候，既要加上补了值的基线工作满意度（现在每个员工都有值），也要加上新创建的虚拟变量。只要虚拟变量同时被控制，不管补的值是多少，都不会影响估算出来的实验效果（参见 Groenwold et al.，2012）。[①]

如果我们知道一个基线变量和潜在结果相关性很高，那么我们是在实验分析的时候控制这个变量，还是在实验设计的时候基于这个变量来做区组随机化呢？一般来说，区组随机化会比事后分析的时候再控制变量更有助于提高因果推测的精准度，因为前者可以保证在每个

① 注意这种情况下基线工作满意度自身的回归系数不好理解，且会受到添补值的影响。

区组里指定比例的员工进入了不同的实验条件，从而降低抽样变异性。有的读者可能会问：已经用于做区组随机化的变量是否还需要在回归分析的时候再作为控制变量处理呢？是否再作为控制变量处理都是可行的，理论上不影响实验效果的估计值，但是控制的话，可以进一步降低残差变异性（residual variance）（Duflo et al.，2007）。

我们建议研究者在实验开始前就弄清楚如何使用控制变量和使用哪些控制变量，防止在看到结果之后再根据如何能让结果更显著这一问题来挑选变量。如果将有控制变量的分析作为主要分析，我们也建议研究者汇报没有控制变量的稳健性检验分析结果。

8.4.3　集群随机化时的标准误处理

对于集群随机化的实验，如果最终的数据分析是以每个员工作为观察单位的话，那么需要考虑到每个集群中员工之间误差项（error term）的独立性。假设没有异方差性（no heteroskedasticity）并且集群之间有着相同的协方差结构（covariance structure），那么可以通过莫尔顿因子（Moulton factor）来调整回归计算出来的实验效果估计值的标准误（Moulton，1990），或者使用带有集群随机效应（random effect）的最小二乘法（generalized least squares）来分析。如果研究者不希望假设集群之间有着相同的协方差结构，那么可以使用 cluster-correlated Huber-White 协方差结构估计值来计算聚类调整标准误（cluster robust standard error）。这种方法需要集群数目相对比较大。[①]

如果集群数目相对较小，可以对聚类调整标准误进行调整（Cameron et al.，2008），或者使用随机推断（randomization inference）（Rosenbaum，2002）。在这个场景里使用随机推断的话，第一步是用集群分配方式重新随机分配集群进入不同实验条件。第二步是用重新分配会得到“假的”干预组和“假的”控制组，用和原始分析同样的回归方程来估计“假的”干预效果。第三步是多次重复前两步，每次都得到一个“假的”干预效果。最后随机推断的 p 值就等于“假的”干预效果小于原始分析估计出来的干预效果的比例。Bloom et al.（2006）和 Wu & Paluck（2022）的实验就使用了聚类调整标准误和随机推断来分析集群随机化实验。

8.4.4　检查实验效果是否来自控制组的变化

研究者在实验设计和分析阶段应该注意排除一种情况，即干预组和控制组之间存在的差异其实是由于控制组被试的行为变化了，而不是因为干预组被试受到了干预而产生行为变化。前面我们讨论的溢出效应是可能导致控制组被试行为变化的原因之一（Bloom et al.，2015）。即便没有溢出效应，控制组被试也可能由于自己的实验条件产生行为变化，造成研究者观察到控制组和干预组之间存在差异。我们鼓励研究者不仅在实验设计阶段想办法尽量减少控制组被试因为自己所在的实验条件而产生的认知或行为变化，并且在实验分析阶段也想办法用数据分析方法进一步说明这种情况不太可能解释他们看到的结果。案例参见 Zeng et al.（2022）。

① Duflo et al.（2004）的模拟发现，当集群数目小于50时，cluster-correlated Huber-White估计值表现比较差，会导致研究者过多地在实验没有效果的时候拒绝原假设。

Zeng et al.（2022）和一个中国短视频 App 合作，将田野实验和社交网络模型结合，研究短视频观看者给视频制作者发催更信息将如何影响 App 里的视频上传量。在田野实验开始的时候，催更在这个 App 里还是一个新功能。研究者随机选取了一群制作者可以收到来自观众的催更，另一些制作者则不能收到观众催更（即便观众给他们发了催更消息，消息也不会进入制作者的消息中心）。最基本的实验发现是，视频上传量在可以收到催更的干预组要高于控制组。研究者希望的研究结果是催更给干预组制作者带来了鼓励，激励他们上传作品；但是有可能控制组的制作者在和观众通过其他渠道交流的时候，发现有观众给自己发了催更信息，自己却从来没有收到，因此对 App 产生了不满的情绪，进而降低了创作热情。于是 Zeng et al.（2022）通过一系列的分析来排除他们的实验效果被控制组制作者所驱动的这个可能性。

8.5　田野实验的实操建议

8.5.1　寻找合作者

田野实验有时需要有田野合作者（field partner），他们可以是企业、各级政府、非营利机构或当地社区等。研究者和田野合作者的合作方式是多样的。有时候研究者用实验的方式帮助合作方评估一个已有项目的有效性，研究者的任务主要是提供实验方法论的支持；有时候研究者和合作方共同设计实验干预内容，共同测试干预的有效性；也有的时候，研究者从理论角度出发独立设计了感兴趣的干预内容，然后再寻找一个愿意提供田野实验被试的合作方。

与田野合作者建立合作关系的途径也是多样的。如果你的社交网络很广，或者正好有认识的合适的合作方，他们也许会主动来找你。我们也可以通过校友会等渠道接触合适的企业合作方。如果没有能直接接触合作方的渠道也不用担心，我们的建议是：你首先要想好自己感兴趣的合作方，然后你可以通过邮件或社交媒体来介绍自己，看对方是否对你的研究议题感兴趣。你也可以去拜访自己感兴趣的合作机构，参加合作机构的社交活动、信息交流会等，这样至少可以让你踏进合作方的大门，增加与其建立沟通渠道的机会。千万不要小瞧了这些最原始的手段，很多大规模的田野实验都是从这些尝试开始，通过与合作方一次次的试探与交涉逐步确立合作关系，最终扩大规模的。

值得注意的是，在寻找田野合作者的过程中，我们不仅要想自己感兴趣的研究议题和干预内容是什么，也要站在合作方的角度思考我们的干预内容有什么价值、能否契合合作方的动机（比如干预有可能产生提高员工绩效、减少离职率等一系列对合作方有利的正效应），还要思考如何消除合作方的顾虑（比如干预在时间和财务上有没有限制、干预会不会产生不好的影响、公司信息是否会泄露），从而达成互利共赢。结合 Eden（2017）与我们自身经验，我们对于如何消除合作方顾虑给出如下更细致的建议：

• 与企业合作方交涉时，尽量避免使用专业术语，特别是避免使用“实验”这个词。有些管理者对“实验”有误解，这个词会让他们联想到自然科学领域中做的实验，以为你要把员

工当成“小白鼠”而对你产生怀疑。在与合作方不熟悉的情况下，可以把你的实验宽泛地说成是一个研究、一个项目或一个课题。

• 一定要把随机化的过程与合作方说清楚。也许有的合作方声称他们对随机化这个概念很了解，但根据我们的经验，实际上很少有合作方对随机化很熟悉并且有自己做随机化的经验。为了使合作方了解随机化的重要性，我们建议你提前准备一个简短的对随机化及其重要性进行介绍的报告，与合作方说清楚随机化是必要的而且并没有想象中的困难。如果真的开始合作，我们建议研究者尽量自己来对被试进行随机化，至少能够亲自监督随机化的过程，而不是完全交由合作方处理。因为随机化这个过程在实验设计中至关重要，而且非常容易出错，并不是所有人都有随机化的训练与经验。

• 向合作方表示自己非常愿意接受他们的专业建议，让合作方觉得你是真心想建立合作关系、他们是有机会参与实验设计的，而不仅仅是你想利用他们的资源来实现你个人的想法。可以邀请合作方参与实验干预和问卷的设计。这样做有几个好处：让合作方更深入了解你的干预和测量方式，打消他们的疑虑；提高他们参与合作的积极性；以及帮助研究者预判被试在实验情境下的构建。

• 尽量利用合作方已有的活动、流程、政策和结构（如晨会、客户反馈表、团建活动等）来实施干预。这样在与合作方谈判的过程中，你可以强调你是帮助他们来改善已有的活动和政策，而不是重新创建一个新的项目。对于企业来说，调整一件已经在做的事情比创建一个新项目的风险要小得多，操作也相对容易得多。

• 在跟合作方沟通实验方案前，先在实验室中进行测试。已有支持性正向结果的干预内容更能让合作者信服。

• 从小规模的实验开始谈合作，如果实验成功，再逐步扩大规模。除非你与合作方已有充分的信任关系，或者你的干预内容极其简单易行，否则一般企业很难让你一开始就做大规模的干预。所以，我们建议从小规模实验开始逐步扩大，而非一开始就对企业提出全面、复杂的要求。

• 当合作方对实验干预内容犹豫不决时，可以先站在合作方的角度帮助他们解决一些他们最想解决的问题，让他们看到实验的好处和研究者专业背景带来的优势，建立双方的信任和互利互惠的合作关系。比如，本章作者之一为了说服企业做一个耗资相对较大的干预，首先无偿帮助企业分析他们现阶段最想解决的问题，为他们提供建议，并利用实验的方式评估解决方案的有效性，让合作方看到了研究者的诚意与专业水平，最后企业才同意实施研究者最初感兴趣的田野实验。

如果有田野实验合作方表现出与你合作的意愿，这当然是值得高兴的事，但不是所有的合作机会都对你的研究有用，都值得你全力以赴。有些企业也许不愿意投入大量的人力、物力来配合你来完成一个高质量的田野实验；有些企业也许规模太小，不具备足够数目的被试或集群来做在统计上有功效的随机分配；也有的企业虽然愿意配合田野实验，但是却不赞成

把数据在期刊上进行公开。与田野实验合作方建立正式合作关系前，研究者需要与合作方进行谈话，确保双方在合作预期上达成一致：研究者要确定合作方具备实施田野实验的条件，并且自己能完成合作方的要求。如果确立了合作关系，我们建议签署一个书面文件来规定双方的研究需求和责任（例如签署谅解备忘录和数据保密协议等）。

当然，田野实验也未必一定要和企业、政府等合作方有关系才能做。田野实验是极富创造性的，在没有合作机构的情况下也能做。一个经典的例子是 Bertrand & Mullainathan（2004）向美国波士顿和芝加哥地区招聘广告上的地址投送了上千份除名字外其他内容均相同的求职简历。他们发现以黑人名字投出去的简历相对于以白人名字投出去的简历更难收到面试通知。这类审计实验（audit experiment）可以通过网络和现实中的招聘广告来随机投送不同实验条件下的简历，并不需要寻求田野合作者。这个实验范式近些年也被管理学者运用（Kang et al.，2016；Milkman et al.，2012）。类似的，研究者可以化身为雇主，通过改变招工广告的信息来随机分配不同的实验条件，以便研究应聘者的行为（Dai et al.，2021；He et al.，2022；Leibbrandt & List，2015）；或者研究者化身消费者，通过随机分配不同的关于消费者的信息来研究商家的歧视行为（Cui et al.，2020）。

再举两个在日常生活情境里的例子。Epley & Schroeder（2014）让研究助理在地铁站和公交站蹲点，随机让一部分出行者尝试在地铁或公交上与陌生人交谈（干预组），另一部分则让他们在地铁和公交上自由表现（控制组）。结果他们发现，干预组比控制组的出行体验更好。Cohn et al.（2019）在全世界 355 个不同的城市中的特定地点丢下钱包，钱包内的金额有多有少，钱包里有主人的联系方式。研究者追踪世界各地有多少人会以发邮件的方式来寻找钱包“失主”。他们发现，相对于小金额的钱包，那些较大金额的钱包更容易被寻回。Cohn et al.（2019）这个大规模田野实验虽然有很多高校合作者，但没有特定的田野合作者。上述的例子只是“冰山一角”，有很多的学者非常有创造性地在没有田野合作者的情况下设计和实施了田野实验。我们鼓励读者们打开想象，在日常生活情境、自由职业者平台、线上或线下的消费场景里寻找可以实验的空间。

8.5.2　分析被试在实验情境下的构建

在现实环境中设计实验和收集数据时，一个重要的环节是分析被试在实验情境下的构建（construal），即被试对实验材料（如干预内容、调查问卷的内容等）和实验实施条件（如与被试接触的研究者、干预进行的地点、企业的大环境等）的主观理解（Paluck & Shafir，2016；Ross & Nisbett，1991）。首先，我们需要预判什么因素会影响被试在研究中的参与体验，并确定其行为是否体现了他们的真实反应。常见的影响因素包括被试对于实验目的的揣测、社会期望、自我展示的欲望、对研究者的信任、教育水平，以及简单的功利主义动机（如为获得奖金激励）。一个被试对干预内容的体验可能会受到非干预内容的影响，比如他们会猜测：谁在实施这个干预？实施这个干预的真实目的是什么？这些都是被试在实验情境下的构

建，都可能影响他们在实验中的反应。例如，研究者在企业中做问卷调查，如果没有解释，不少员工可能会把一个关于工作场所满意度的调查项目当成老板在变相考察其忠诚度。其次，我们需要注意田野实验实施的情境（context）也可能会影响被试对于干预的反应。情境包括实验所发生的地点，也包括被试所属行业的社会规范和所属国家的文化。例如，地点可以直接影响行为，当美国选民为一项增加教育支出的政策投票时，随机分配到学校投票的选民比分配到教堂投票的选民对教育支出的投票支持率增加了 0.5 个百分点（Berger et al.，2008）。

如果研究者没有充分了解被试在实验情境下的构建，田野实验很可能产生与预期相悖的结果。比如，研究者可能在设计中忽略了某些影响被试的环境因素，最终看起来干预没有产生显著的结果，其实只是因为被试误解了问卷题目或者误解了干预传递的信息。我们要了解到，被试不是干预的被动接受者，他们是有主观认知的。干预的过程是被试与干预内容交互的过程。用认知心理学家杰罗姆·布鲁纳（Jerome Bruner）的话来说，人们对干预的构建通常都超越了干预本身所提供的信息（Bruner，1957）。

如何分析被试的构建从而更好地设计田野实验呢？ Paluck & Shafir（2016）认为，田野研究者需要与被试实现“共享构建”（shared construal），从而了解被试对实验材料和实验环境的心理认知，以此来设计一个能高度还原感兴趣的干预内容的实验条件。了解被试在实验情境中的构建，即在研究者和被试之间实现共享构建，这并非易事。在这里，我们首先提供三个实操建议，最后分享一个田野实验案例。

第一，研究者可以实施前导研究（piloting）来了解被试和情境，以及最大程度地减少研究设计带来的意外后果。前导研究通常是指在实验开始之前调查实验情境、小规模测试研究范式。在田野实验中，前导研究也指在设计干预之前花时间调查和理解利益相关者和被试所属群体在现场环境中的心理构建——他们将如何理解干预涉及的相关行为以及实验情境对被试的影响。

第二，研究者可以在实验设计前用认知访谈（cognitive interviews）（Shafer & Lohse，2005；Willis，2004）来了解被试所在群体对相关信息的看法，以改进实验干预的设计。在认知访谈中，研究者在自然的环境下采访与被试类似的群体（而非实际参与实验的被试），鼓励他们思考并解释他们对干预内容的预期和反应，剖析这些反应，并询问他们反应背后的动机。目的是判断被试会如何理解你设计的干预内容，即在这个情境下，被试知道什么、想要什么、感知到什么、关注和记住了什么等。以此最大程度地保证研究者与被试达成了共享构建——研究者对感兴趣的干预内容的构建和被试对干预内容的构建是基本一致的。

第三，研究者可以寻找一个了解被试的构建和实验情境的合作者来加入田野实验设计。这样就保证了研究团队中至少有一个研究者充分了解被试所在群体，并且可以指导其他研究者共同设计出符合被试构建的干预内容。第三点建议在跨国田野实验中尤为重要，可以避免文化或语言在实验设计中的干扰。

我们举一个田野实验的案例来说明分析被试构建的重要性和过程。分析被试构建帮助 Wu

& Paluck（2021）想到可以运用铜钱在中国文化环境下的特殊意义来减少工人在车间乱扔垃圾的行为。

Wu & Paluck（2021）曾在一家大型纺织厂研究如何减少工人在车间乱扔垃圾的行为。工厂和研究者合作之前已经尝试发布“不准乱扔纺织废料”的指令，并且实行了罚款措施——如果工人周围的垃圾太多，每个月就会被扣钱。但这些指令和罚款措施的效果并不好。如何从行为科学的角度来进行干预呢？通过总结关于减少乱扔垃圾行为的文献，作者发现在英国和丹麦，在人行道靠近垃圾桶的地方贴上绿色的脚印能非常有效地鼓励行人将垃圾扔到垃圾桶（Hansen & Jesperson，2013；Keep Britain Tidy，2015）；但在有些国家，行人从未见过在路上画标记这样的事情，像贴脚印这样的干预就失效了（Sheely，2013）。放到工厂的环境中，作者觉得贴脚印这个干预也不会生效。为什么呢？

这就需要我们剖析工厂车间这个情境下，工人乱扔垃圾背后的行为动机和对实验干预的构建。Wu & Paluck（2021）通过实地观察和采访了解到，作为计件工人，时间就是金钱，工人们不想花费几秒钟的时间把垃圾扔进垃圾桶。然而，工人没有想到的是，虽然乱扔垃圾在短期内帮助他们加快了生产速度，但是如果员工周围垃圾太多，清洁工经过时就会产生较长的停顿，反倒降低了生产速度，而且粉尘堆积还会影响织布质量。因此作者推测前述绿色脚印的做法更可能驱动那些不排斥把垃圾丢进垃圾桶的工人，但对于那些只考虑当下绩效而乱扔垃圾的工人未必有效。如何设计一个能与工人的构建相辅相成的干预呢？最后，作者想到在地板上贴印有金币的贴纸，因为在中国的文化背景下，工人们认为金币是财富和好运的象征，不应该被垃圾玷污。工人本来就重视自己的绩效，而工人对金币的心理构建和自己想赚钱的动机是相辅相成的，所以工人不希望象征着财富和好运的金币被自己丢的垃圾覆盖。也就是说，保护金币所象征的财富与好运的动机能抑制工人乱扔垃圾的动机。通过 5 个月的观察，作者发现贴印有金币的贴纸使车间地面垃圾覆盖率减少了 20% 以上。在这个研究中，如果没有实地观察和采访，作者就不会了解工人乱扔垃圾这个行为背后的动机；如果不了解这个情境下的文化背景，也不会设计这个金币实验。

在田野实验中，了解被试在实验情境下的构建不仅有助于实验的设计，也有助于回答“为什么有些干预在一个情境下有效，换了另一个情境效果就不同了”这一问题。干预在不同场景下的效果很大程度上取决于在不同情境下被试的构建。感兴趣的读者可以阅读 Reiff et al.（2022），其中的研究者和加州大学洛杉矶分校医院合作测试一个过去文献里认为可以影响医生绩效的干预——告诉医生他们的绩效相较于同事是处于较好还是较差的水平。结果不同于过去文献的发现，在 Reiff et al.（2022）中，干预组的医生们觉得绩效比较（peer comparison）在他们的工作环境里不合适，因此他们觉得这个干预体现了医院领导对他们的支持不够，从而不仅没提升工作绩效，反而还降低了工作满意度。可见，充分了解被试在实验情境中的构建有助于研究者发展理论，总结不同情境下田野实验结果的可靠性和普遍性。

8.5.3　提升数据质量

在条件允许的情况下，建议研究者在实验开始前做基线调查或者从公司行政记录中获得基线数据。这不仅可以帮助研究者了解被试，并且收集的信息可以用于功效分析、区组随机化、平衡检验、作为控制变量来提高功效，或者用于做处理效应异质性分析（heterogeneous treatment effect）。另外，基线数据的收集过程可以帮助研究者预判实验开始后的样本流失问题，思考之后的数据收集有何改进之处。一般来说，在实验开始之后再通过问卷去补测实验开始前的信息（比如说让被试回忆实验开始前他们的企业忠诚度）是不合适的，因为干预可能会改变被试的记忆和回答（Wu & Coman，2023）。

如果数据主要是从合作机构获取的，我们有三个实操建议：第一，在决定合作之前一定要跟合作机构沟通清楚，确保其能提供研究者所需要的数据。研究者可以制作一个数据结构模板，将必须要有的数据和最好能有但并不一定要有的数据标识出来，由合作机构明确表态哪些信息是他们可以提供的。第二，最好能让合作机构提供一些历史数据。除了起到上述收集基线数据的作用，也给了研究者一个检查数据质量的机会，不仅可以确保合作机构能够有效地提供相关信息，而且知晓对方合作的诚意。同时，从心理学的角度，如果合作机构花费精力提供了数据，那么他们对于后续合作的兴趣也更浓厚。第三，如果研究者需要合作机构额外收集被试数据（而不是用他们本来就会自然追踪的数据），那么要确保合作机构在不同实验条件之间的数据收集方式和时间是一样的。合作机构有可能想尽快实施干预，于是先在干预组收集基线数据，然后等到实验开始了，再去控制组收集基线数据。这个做法显然是错误的，因为干预组和控制组的数据可能会随着时间变化而不可比。

如果数据的收集比较耗费人力（如需要专人去现场记录被试的行为或进行一对一的问卷填写），我们也有三个实操建议：第一，研究者可以雇用第三方数据收集机构或高校的研究助理来协助收集数据。值得注意的是，即使第三方有研究资质（如专业的访问员），研究者也一定要在数据采集之前对第三方做系统的培训，确保他们充分理解采样需求和数据收集的步骤。数据收集的培训资料可以记录下来，放在预注册（pre-register）[①] 中，以便其他学者能复制收集过程。第二，在数据收集的过程中，尤其是数据收集初期，研究者最好能陪同第三方共同收集数据，确保其完全按照研究者的步骤来收集数据。如果研究者无法直接参与数据收集，那么一定要进行严格的数据抽检来保证其质量，确保其中没有作假行为。第三，数据收集完成后，研究者需要查看数据分布，可以用统计方法或模拟来评估数据质量。感兴趣的读者可参考 Gomila et al.（2017）用智能手机音频监测来提高田野实验数据质量的方法。

8.5.4　增强可复制性（replicability）和普遍适用性（generalizability）

说到如何减少实验结果不可复制的情况，不少学者推崇预注册的方法（Nosek et al.，2018）。田野实验做预注册有不同于实验室实验的难点。有的时候公司不允许研究者预注册，

① 常用的预注册网站有Open Science Framework，As Pedicted AEA RCT Registry，clinicaltrial.gov。

尤其是做针对于用户的大规模实验（如测试 App 新功能）时。一个顾虑是即便预注册不透露公司名字，关于干预的信息也可能涉及商业机密（如公司不希望 App 新功能在公司正式推出之前就被竞争对手知道）。再者，田野实验往往存在很多不确定性，在干预完成和数据收集完之前，都不能保证预注册的信息不会因为研究者不可控的因素而发生变化。

我们的观点是，首先尝试和合作机构商量，把预注册设为隐藏模式，等论文发表的时候再公开，这样可以保留预注册的底案。如果公司允许预注册，实验中的不确定性不应该阻碍研究者预注册。即便有的信息不能准确地在预注册阶段提供，预注册依然可以起到对于研究者的约束作用。另外，即便情况在实验过程中发生了变化，只要研究者还没有分析数据、没有根据他们看到的结果来决定或者引导变化，他们可以修改预注册[①]或者再提交一个预注册作为对之前预注册的补充说明。参见 Dai et al.（2021）和 Reiff（2022）的例子。

Dai et al.（2021）和加州大学洛杉矶分校的医院合作，在 2021 年 2 月开始研究不同的助推文案对新冠疫苗接种率的影响。因为这家医院所能拿到的疫苗量、接种疫苗条件（即什么时候、针对什么样的人群会放开）都是不确定的，研究者并不能事先知道自己的样本量以及实验干预的持续时间。虽然预注册不能事先给出样本量的具体值，但研究者仔细解释了被试选择的标准、被试在符合接种疫苗条件后会如何被分配到不同的实验条件，以及样本量的最大可能值（基于医院所有的可能最终会符合疫苗接种条件的病人）。这些背景信息，加上研究者严格按照预注册执行的实验设计、结果变量和分析计划，依然增加了研究的透明度。前文提到过的 Reiff et al.（2022）研究的是医生绩效比较（peer comparison）对于医生工作表现、工作满意度和工作倦怠感的影响。研究者本来是计划和医院合作进行 9 个月的干预。但因为新冠肺炎疫情爆发，医院工作环境和就医情况发生明显的变化，他们停止了才进行了 5 个月的实验。研究团队决定停止实验的时候并没有看任何实验数据，完全是对外界条件的反应，并且研究团队立刻在网站（clinicaltrial.gov）中更新了他们的预注册。

在前面我们建议研究者和合作机构签署书面文件来明确双方的需求与责任，其中有两点需求可能是合作机构比较敏感、未必会答应的，但这两点对于研究的透明度和可复制性又比较重要。第一，研究者最好能让合作机构书面确认：不论实验的结果如何（是否在统计上显著、是否是合作机构期望的结果），研究者都有权发表实验结果，而不是由合作机构来决定什么样的实验结果可以发表。这样有助于减少由于合作机构的喜好带来的发表偏见（publication bias），当然也是对研究者自身利益的保护，避免研究者花费了很多时间完成实验却不能发表。为了减少顾虑，研究者可以告诉合作机构，文章可以进行脱敏处理。第二，现在越来越多的期刊要求研究者分享数据和代码（如 *Econometrica*，*American Economic Review*，*Management Science*，*Marketing Science*）。有很多原因会导致合作机构不允许研究者分享原始数据，甚至是已经被标准化处理的数据。有的期刊（如 *Management Science*，*Marketing*

① 不少预注册网站会记录下最初预注册提交的时间、修改提交的时间和修改的核心内容（比如说实验周期发生了变化）。

Science）允许研究者采用其他方法来提供可以让别的学者复制主要实验结果的信息。建议研究者能事先和合作机构沟通好数据分享权限。

相较于实验室实验，复制一个田野实验所需要的资源一般大很多，尤其是大型的田野实验。因此，我们很少见到研究者在一个文章里汇报多个田野实验或者复制前人做过的田野实验，这让判断一个田野实验结果的可复制性和普遍适用性尤为困难。我们简单介绍一个近些年逐渐开始得到关注和运用的有助于解决这个问题的方法——协同实验（coordinated studies）。协同实验指的是多个研究团队在多个场景和被试群体里同时或者先后用同样的实验设计回答同一个问题、测试同一个干预（Blair & McClendon，2021；Ferraro & Agrawal，2021）。比如 Banerjee et al.（2015）在 6 个发展中国家做田野实验（涉及 10495 名居民），来测试一个旨在提高最贫穷人群生活质量的项目。他们测试的项目在每个国家的具体内容会根据该国的具体情况和文化做一些调整，但是保持了同样的机理和原则。2020 年，EGAP（Evidence in Governance and Politics）提出了一套具体如何做协同实验的规范——Metaketa Initiative。根据这套规范，由专家组成的指导委员会先确定研究问题和干预内容，然后通过竞标的方式邀请其他研究者参与项目。研究者的选择很重要，团队必须足够多元化，整个团队才有能力和经验在不同的地点和被试群体中做田野实验。在组建了团队之后，所有研究者商讨如何在不同地点和被试群体之间确保干预和结果变量的可比性。最终由指导委员会负责分析从各个地点和被试群体收集来的数据，做元分析，并发表实验结果。遵循该规范来做协同实验的案例参见 Blair et al.（2021）。如果在不同的实验地点和被试群体中发现相同的干预有着类似的结果，我们会对这个结论的可复制性和普遍适用性更有信心。

最后我们简单介绍另一个田野实验范式——megastudy。这是由凯瑟琳·米尔科曼（Katherine Milkman）和安吉拉·达克沃思（Angela Duckworth）教授提出的一种范式，最初的案例来自由他们领导的一项名为“良好行为改变计划”（Behavior Change for Good Initiative）的项目（Milkman et al.，2021，2022）。megastudy 指的是在一个特大型的田野实验中同时测试多个干预对于同样的结果变量的影响，这些干预可以组合成多个规模小一些的子实验，可以由多个研究者独立设计。这个实验范式增加了不同干预的可比性，产生规模经济（因为是由一个核心团队来负责和合作机构沟通、执行多个干预），并且降低了不显著的实验结果被隐藏起来的可能性（因为当一个 megastudy 的结果被发表时，不管单个干预是否产生统计上显著的结果，所有的干预效果都会被呈现出来）。

8.6　总结

田野实验对于管理学的研究很有价值，但田野实验的长周期、高风险和高成本也让很多学者望而却步。基于班杜拉的自我效能理论（self-efficacy theory）（Bandura，1997），Eden（2017）指出，一个学者做田野实验的自我效能会影响到他们是否愿意在这个研究方法上投入

资源，而提升自我效能的最好方法可能就是让他们成功地完成一个田野实验。我们希望这个章节能够鼓励更多的管理学学者迈出田野实验的第一步，意识到田野实验的形式可以是多种多样的，帮助他们预判和应对在设计、执行和分析田野实验时常见的问题，从而更顺利地完成一个漂亮的、适合自己研究问题的田野实验。

思考题

1. 阐述田野实验相对于实验室实验和非实验研究的好处。

2. 常见的田野实验中的随机化方式有哪些？这些随机化方式之间有什么区别，分别适用于何种情况？

3. 在本章对于集群随机化的介绍中，提到了四个研究问题：特定的工作环境对员工绩效有什么影响？夫妻心理咨询会怎样影响夫妻关系？教师的教学风格会如何影响学生的成绩？老板的领导方式是否会影响团队矛盾？如果用田野实验来研究上述这些问题，可进行随机分配的集群分别是哪些？

4. 在给定的被试样本和预期干预效果下，有什么提高实验功效的办法呢？

5. 样本流失在什么情况下不会威胁实验结果的因果推断？什么情况下会威胁？如何减少样本流失的影响？

6. 假设一个实验存在不依从的情况，怎么做意向处理分析？该分析得到的结果代表的是什么？使用工具变量法的假设是什么？该分析得到的结果又代表了什么？

7. “既然田野实验采用了随机化，那么添加控制变量是没有好处也没有坏处的。”判断这句话是否正确并说明原因。

延伸阅读

Duflo，E. & Banerjee，A.（2017）. *Handbook of Field Experiments*. Amsterdam：Elsevier.

Duflo，E.，Glennerster，R. & Kremer，M.（2007）. Using randomization in development economics research：A toolkit. *Handbook of Development Economics*，*4*，3895–3962.

Eden，D.（2017）. Field experiments in organizations. *Annual Review of Organizational Psychology and Organizational Behavior*，*4*，91–122.

Gerber，A. S. & Green，D. P.（2012）. *Field experiments：Design，analysis，and interpretation.* New York：W. W. Norton & Company.