**如何做 4.3**

描述后测仅控制组设计

作为Campbell和Stanley真实实验类型中使用最广泛和推荐的设计，后测仅控制组设计可能是你经常使用的设计之一。与使用Solomon四组设计的实验不同，大多数采用这种普遍设计的研究并不正式宣布自己的设计类型。相反，你可以通过寻找对控制组和随机分配的提及，以及没有提及预测的内容来确定它们使用了这种设计类型。

下面是两个例子，第一个例子明确指定了它是后测仅控制组设计，接着是第二个例子，没有明确指定。

Coleman, Renita, Paul Lieber, Andrew Mendelson, and David Kurpius. 2008. “Public Life and the Internet: If You Build a Better Website, Will Citizens Become Engaged?” New Media & Society 10 (2): 179–201.

“该研究采用了后测、对照组实验设计。该研究的实验刺激是一家关于州财政预算的网站，由大众传播学生在网站开发课程中创建...”（第188页）。“对照组网站是关于州财政预算的官方州政府网站。它在没有可用性测试或对上述任何问题的了解的情况下创建，这些问题指导了实验性网站的创建。重要的是，对照网站与实验网站在主题上相同，以排除由于主题而非内容、外观或导航产生的影响...”（第189页）。“这60名参与者被随机分配到查看对照网站或实验网站”（第190页）。

Bennion, Elizabeth A., and David W. Nickerson. 2011. “The Cost of Convenience: An Experiment Showing E-Mail Outreach Decreases Voter Registration.” Political Research Quarterly 64 (4): 858–869. “电子邮件实验的设计本身很简单。学生被随机分配到三个条件之一：（1）未接收电子邮件的对照组，（2）接收来自行政人员（如大学校长或学生事务主管）的三封电子邮件的处理组，或（3）接收来自学生领导者（通常是学生会主席）的三封电子邮件的处理组。电子邮件内容简短，解释了为什么注册很重要，并提供了Rock the Vote在线注册工具的链接”（第862页）。

独立性，也就是说，有些人可能属于多个组别，最后准实验无法像真实实验那样控制所有的外部因素。随机分配、独立性和控制这些特征并非总是可行的。当情况如此时，准实验被认为是一个可行的选择，甚至在某些情况下更可取。有些事情无法“分配”。例如，研究人员无法在道德上要求某人吸烟，而要求另一个人不吸烟，以实现随机分配的处理组和对照组。特别是在商业和教育环境中进行的实验，往往难以完全控制研究过程。28参与研究的企业可能希望自行挑选参与者。



当使用完整的班级中的学生进行实验时，这被称为准实验，因为这些受试者没有被随机分配。.

iStock.com/skynesher

出于某种原因，或者希望每个人都被平等对待，不显得偏袒。在其他情况下，参与者可能在不知情的情况下选择了他们所在的组别。例如，教育研究经常与完整的班级合作，而不是对学生进行随机分配的班级。教育工作者可能会根据他们想要多样性或男女混合，不想让双胞胎在同一个班级，或者想让两个孩子在一起而将学生分配到班级中。在大学阶段，很明显报名参加实验设计课程的人与报名参加民族志研究方法课程的人之间存在个体差异。学生可能会选择一门课程而不是另一门，因为他们听说某位老师教得好。或者这可能是一个时间问题，因为有些学生不喜欢早上早起，周三下午有兼职工作，或者想要在周五离开城市。造成这些个体差异的原因可能会影响实验的有效性。

仅靠准实验无法对个体之间的差异产生影响结果进行归因（2023）。

准实验仍然可以揭示重要的知识，但它们带来的有效性威胁与真实实验不同，必须加以解决和记录。在最终的论文中，重要的是指出受试者无法进行随机分配，并解释原因。论文还应该解释为了最小化任何系统性差异而采取了哪些措施，例如是否测量了关键变量并将其用作协变量？是否进行了统计检验以确定各组在重要变量上的等价性？重要的是对能够推断因果关系或推广到研究的一个案例之外进行可信的主张。最后，准实验无法像真实实验那样控制所有可能导致结果的外部因素。没有随机分配，准实验无法消除未受控制的变量混淆结果并影响进行因果推断的能力的可能性。然而，研究人员应始终尝试测量可能的混淆变量，然后通过使用协变量在统计上进行控制。每项研究都有其局限性，包括真实实验（有关这些的更多信息请参阅第5章）。局限性并不一定否定了研究结果的重要性，尤其是如果这是研究特定现象的唯一方法。准实验可能感觉比真实实验更真实，并且通常比受控实验更容易进行纵向研究-这是所有的好处。了解研究发现总比一无所知要好，但读者应始终被告知其局限性。

三角化，即通过使用另一种方法来确认已经找到的结果，也是很重要的。如果不同的方法证实了同样的发现，例如，在人工环境中进行的真实实验所得出的结论与准实验在真实世界中得出的结论一致，那么我们可以更加自信地认为这些研究结果代表着真实的情况。

社会工作教育中准实验的最佳实践之一是使用两个相似的班级研究在线或面对面教学哪种方式能够获得更大的学习效果。在这个准实验中，研究人员使用了完整的教室，因此无法将学生随机分配到不同教学风格的班级中。为了尽可能控制个体学生差异，他们使用了学生的年龄和经过成绩膨胀调整的平均绩点作为协变量。他们表示，一些限制因素是未考虑的外部变量和缺乏推广性。

在一项商业研究中，研究人员研究了正式导师制度对员工建立人际网络能力的影响。接受导师指导的人组成了处理组，而没有接受导师指导的人则组成了对照组。在这项研究中，参与者并非随机分配；相反，公司的管理层根据他们在公司内部晋升潜力的认识选择了将要接受导师指导的员工。为了减少不确定性，研究人员采用了“配对设计”，即处理组中的每个员工与控制组中尽可能相似的员工进行配对，例如工资水平、前一年的绩效评级、在公司工作的时间以及所在办公室等重要特征相同。该研究的作者提供了统计证据表明配对是成功的，即处理组和对照组在其他特征上没有显著差异，包括年龄、教育程度以及在接受导师指导之前建立的人际网络联系的数量。更多的人际网络联系是该项目的一个目标结果，因此这确保了处理组的人员在一开始并不更有可能建立人际网络。然而，处理组和对照组的员工在其他未知方面可能存在差异，这可能导致与导师指导相关的预期结果的差异，因此使用了其他各种策略来帮助克服缺乏随机分配的问题。然而，因为我们无法考虑到所有其他混淆变量，我们无法确保这些策略与简单的随机分配一样成功；总是存在更多的不确定性。

另一个准实验的例子是对虐待和忽视儿童进行干预研究，研究人员采访了这些儿童的心理健康和生活质量。

这类研究通常只对成年人进行访谈，因为人们认为儿童不可靠，并且不需要因为这类问题而感到不安。在这项研究中，只对成年人进行访谈的案例代表了对照组；对儿童和成年人进行访谈的案例代表了干预组。由于荷兰医学伦理委员会担心仅对成年人进行访谈的群体可能会得到较差的照顾，因此无法进行随机分配到治疗组或对照组。研究人员能够有治疗组和对照组，但参与者无法被随机分配到这些组。文章提到了选择偏差的可能性。

除了Campbell和Stanley的预实验、真实实验和准实验设计外，研究人员还使用自然实验或实地实验。下面简要介绍了这些实验类型，有许多好的书籍深入讨论了这些主题（请参阅“建议阅读”部分）。这只是作为一种引导，让您思考您的主题是否最适合进行真实实验、准实验、自然实验还是实地实验。

**自然实验**

自然实验是准实验的一个子集，因为研究人员不会随机将受试对象分配到条件或进行干预。相反，自然实验利用一些自然发生的现象来创建治疗组和对照组。在自然环境中，一组人暴露于某种因素，而另一组人则没有。这被视为尽可能接近随机分配的近似，一些研究人员称之为近似随机或仿真随机。32自然或社会创造了治疗或暴露，研究人员在其已经发生之后发现它，然后将其构想为一个实验。Crasnow称其实际上更像是一项观察性研究。33由于研究人员没有设计治疗或干预措施，他们无法控制可能引起观察结果的所有其他因素。这可能导致可信的替代解释，需要尽可能解释和说明。

经常将用于改善社会中某种状况的计划的评估作为自然实验进行。一个例子是对马里兰反对家庭暴力网络开展的计划进行评估的研究，该计划为受害者提供社会服务的支持者。34在这项研究中，警方和社会服务组织设计了他们自己的干预措施，为妇女提供支持、安全规划和转介服务。研究人员在事后进行研究以评估其效果。

在其他情况下，研究人员只是注意到自然发生的“干预”并对其进行研究。例如，研究人员比较了犯罪嫌疑人搬迁的人和未搬迁的人，以查看他们犯罪行为发生的地点是否更接近他们的新居住地。35研究人员没有设计干预措施——分配人们进行搬迁——它只是自然而然地发生了。研究人员使用了许多协变量来尽量减少可行的替代解释，包括犯罪类型、犯罪之间的时间间隔、先前犯罪发生的地点以及先前的住所地点。尽管采取了这些预防措施，但还是讨论了其他一些可行的解释，因为这是应该做的。

这是一个很好的地方来指出，研究人员有时将准实验和自然实验等术语互换使用；为了语言的准确性，本书将准实验用于研究人员设计干预措施的研究，将自然实验用于利用某些自然发生的干预措施的研究。这两种方法都不使用随机分配。报告真实实验或具有随机分配的实验室实验的文章通常不使用真实或实验室这些术语，只称之为实验，并指的是任何在其自然环境中进行的、使用随机分配的实验。

自然实验的另一个例子是传播学领域的经典研究，研究了三个加拿大小镇：一个没有电视，一个有一个电视频道，另一个有四个电视频道。这是在电视信号广播到所有小镇之前的时代。研究人员在不用电视的小镇开始电视广播后的三年重新访问了这三个小镇。这项研究增加了关于电视对儿童攻击行为、阅读能力、认知发展、休闲活动、其他媒体使用、性别角色态度和其他个性特征和态度的短期和长期影响的纵向知识。36在这些实验中，人们被分配到条件中，而不是由研究人员决定，即由他们居住在那些小镇的原因决定。存在某种干预或暴露；在这个例子中，是接触不同水平的电视——没有电视、一个电视频道、四个电视频道——导致能够说电视暴露导致了结果差异。由于这不是一个随机分配的实验室实验，还有许多其他因素可能导致差异或混淆结果，37但研究人员尽力控制了这些因素。

自然实验在许多不同学科中越来越受欢迎。38研究人员可以利用世界上的变化进行自然实验。例如，当商业学者研究秘鲁与美国签订自由贸易协定之前和之后的软饮料市场变化时，他们以玻利维亚作为对照组，因为该国没有这样的协定。39他们通过匹配人口统计学、收入增长等重要特征来选择对照国家。

人口和经济贸易指标。虽然比较的国家并不完全相同，但它们尽可能接近研究者所能获得的。作者讨论了限制，并提出了未来研究的建议，包括更多地控制工资水平、经济和政治稳定性，并使用多个国家作为比较对象。

伦敦的另一项研究观察了为年轻人提供免费公交车乘坐权之前和之后的公共健康情况。潜在混杂因素包括其他旨在改变人们交通选择的政策最近也在推出（例如，高峰时段驾车收费增加），整体文化态度的改变（例如，为了预防肥胖而增加步行），以及没有对照组（所有十八岁以下的人都获得了免费公交车通行证）。他们列出了自然实验的其他限制，包括相对于真实实验而言更弱的因果推断能力，以及难以推广到除了正在研究的单个案例之外。这篇论文很好地解释了现实环境和内部有效性之间的权衡，并提供了一些解决方案，包括混合设计和数据收集。他们表示这些结果是“足够好的”证据，是“尽可能健壮的评估”。

提前规划使研究人员能够从研究人员的研究伦理审查委员会获得研究人类受试者的批准。然而，并非所有适合进行自然实验的情况都有提前警告，例如，在飓风卡特里娜之后犯罪再投入率如何变化，迫使假释犯在整个州范围内分散，而不是集中在特定社区，或者在骚乱之前和之后对福利受助者的态度发生了变化。这些研究只有在研究人员使用事前收集的数据、事后获得批准，或者已经获得类似主题研究的伦理审查委员会批准，并能够快速获得修正批准的有利位置时才能进行。有些人甚至区分自然实验和“自然实验”，但本书不涉及此问题。随时留意此类机会从来都不是坏事。

**场外实验**

虽然自然实验通常无法进行随机分配，因为人们暴露于实验组或对照组的条件是基于研究人员无法控制的自然因素，但还有另一种实验类型，即场外实验，它充分利用真实世界的环境，同时采用随机分配来确定实验组和对照组45。

田野实验这个术语区分了在自然环境和实验室环境中进行的实验，尽管如今并没有多少实验是在实际的实验室中进行的。例如，使用调查软件进行的实验（在第7章中介绍）可以在受试者自己的家中进行。田野实验家认为，除了自然发生的环境之外的任何环境都可以视为实验室。除非实际背景是大学环境，例如关于考试作弊的研究，否则在大学校园进行的任何实验都属于实验室实验。另一个例子是让参与者在家中的电脑上观看电视广告并评估它们，这被认为是实验室实验。尽管受试者在自己的家中，但他们不是在电视上观看广告，并且他们知道自己在进行一项研究。这是田野实验家的另一个关键区别；受试者不应意识到自己正在接受研究。田野实验家“努力使实验尽可能真实和不显眼。”46对于实验对象在被观察时表现不同的担忧被称为霍桑效应，而参与者希望给出“正确答案”或实验者想要的答案则被称为反应性。目前尚不清楚这些是否是实验室实验的严重问题，因为很少有研究同时在田野和实验室条件下复制实验。田野实验家还指出，在实验室环境中有效的方法可能在现实世界中无效。48一些效应可能会立即产生强烈影响，在实验室实验中显现出来，但随着时间的推移逐渐减弱，在田野实验中可能显现较弱。

田野实验家对实验室实验的另一个反对意见是研究人员所创造的信息或刺激可能缺乏真实性。这些都是重要的关注点，实验室实验家应该注意确保他们的刺激尽可能地真实，让其领域内的从业人员创建或审查干预措施。这是第9章关于刺激的主题。

田野实验应该在四个方面具有真实性：参与者、处理、背景环境和结果指标。49参与者应该是真实的选民，而不是假装是学生。处理应该是真实的政治辩论，而不是由演员扮演的辩论。背景环境应该是选民在自己的起居室里看电视，而不是在大学教室里与一群陌生人一起观看。结果指标应该是他们的投票或对候选人的捐款，而不是他们打算投票或捐款的自我报告。

田野实验可能比传统的实验室实验更昂贵和困难，而且在伦理上也存在挑战，因为其中一个标志是受试者不知道他们正在参与实验（请参阅第11章关于伦理的内容）。Gerber和Green讨论了,

详细讨论三个最常见的挑战：50非服从性是指被分配到一种处理的受试者实际上获得了其他处理；流失是指无法获取每个受试者的结果指标；干扰是指受试者相互交流、比较笔记或记住处理的情况。这些问题在实验室实验的环境中被最小化或不存在。

然而，实验室实验通常用于测试理论命题，其中严密控制的条件非常重要，而田野实验更适合应用研究，例如对程序有效性的评估。51

一些田野实验的例子：

为了看看直接与政治家互动是否能够说服人们改变对问题的态度、对政治家品质的评估以及投票方式，研究人员使用在线城镇会议。他们指出，到那时为止，唯一的研究是在实验室环境中模拟个人互动的一些特征。先前的研究对于真实环境的结果是否与假设环境的结果相同存在矛盾。因此，研究人员招募了美国参议员和众议院议员，让他们与选民进行实时在线论坛互动。这个田野实验中的干预措施是由研究人员创建的，但比由演员扮演政治家的模拟城镇会议更接近真实生活。参与的公民被随机分配到与政治家进行在线讨论的处理组，或只接收有关问题背景的阅读材料的对照组。处理组的成员除了阅读材料外，还参与了在线会议。你可以看到在这里展现的努力与菲利普·辛巴多在招募帕洛奥尔执法人员“逮捕”他的监狱实验受试者时所做的努力相似。

其他研究人员在荷兰进行了一项关于扩展学日计划对小学生数学和语言学习效果的随机田野实验。他们指出，在教育中关于计划效果的研究很少使用随机实验，而没有适当对照组的准实验研究则成为常态。在这项研究中，研究人员随机选择学生，并给予他们参加该计划的机会，而不是允许他们或他们的教师选择参与者。与之前使用随机实验并发现微小或不存在效果的研究一样，这项研究也没有发现任何效果。了解不同方法产生相反结果的时机非常重要；例如，在这个问题上，没有随机分配的准实验可能会显示出效果，但更严格的随机真实实验则不会。在这种情况下，

• 当不是真正必要时使用前测

• 因为存在缺陷而将实验报告为“探索性研究”

• 未能将受试者随机分配到不同条件中。渴望成为真正实验的研究，但未能使用随机分配的研究很少被发表。

**常见错误**

现场实验的使用随机分配有助于教育工作者了解哪些结果更值得信任。

与准实验和自然实验的术语一样，研究人员对于字段实验也使用不同的含义。在本书中，我主张使用字段实验一词来指代在自然或现实生活环境中进行的实验，同时使用随机分配。在受试者没有随机分配到处理和对照条件的情况下，我使用准实验和自然实验。

本书主要致力于真实实验，也称为实验室实验，尽管这里介绍的概念也适用于其他类型的实验。进行准实验、自然实验或字段实验的人也应参考涉及这些设计的具体问题的文献。

总之，没有完美的实验。通常无法同时在真实世界环境中进行实验并进行随机分配。研究人员为了理解特定现象所能做的最好努力是进行多种不同设计的研究，这些研究至少具有其中一种特征，并使用不同的背景。在不同条件下进行复制实验可以更加确信所发现的结果是真实的。事实上，敢于理解机会何时出现，通过非传统方法进行研究是创造性研究人员的特点之一。

本章并不是对所有可用的众多不同实验设计的详尽总结。还有许多其他选择，富有创造力的实验者将会对发现重要因果关系的新方法保持开放的态度。下一章将专注于真实实验或实验室实验，并从内部和外部效度的问题开始。

**106** Designing Experiments for the Social Sciences

##### 测试你的知识

1. 根据坎贝尔和斯坦利的实验设计分类，哪种被认为是黄金标准？

a. 前测后测对照组设计

b. 索洛蒙四组设计

c. 仅后测对照组设计

d. 静态对照组比较设计

2. 根据坎贝尔和斯坦利的实验设计分类，哪种是目前最推荐的？

a. 前测后测对照组设计

b. 索洛蒙四组设计

c. 仅后测对照组设计

d. 静态对照组比较设计

3. 对于一个实验被视为真实实验，前测是必需的。

a. 真

b. 假

4. 使用随机分配的主要原因是什么？

a. 确保治疗组和对照组在重要特征上相等。

b. 确保治疗组和对照组的人数相同。

c. 确保同一人不会在多个组中。

d. 公平对待所有被试。

5. 哪种类型的实验在自然发生的环境中使用随机分配？

a. 实验室实验

b. 拟实验

c. 自然实验

d. 现场实验

6. 当前测不会威胁到被试的表现时，以下哪种设计被推荐？

a. 静态对照组比较设计

b. 前测后测对照组设计

Chapter 4 ■ Types of Experiments **107**

c. 单一组前测后测设计

d. 拟实验

7. 以下哪项是具有前测的设计的优点？

a. 它提供了一个基准度量，用于比较任何变化。

b. 被观察或测量两次可能会导致被试表现的变化。

c. 前测对于真实实验设计至关重要。

d. 它确保组之间在可能导致任何变化的重要特征上是相等的。

8. 一项研究使用了两个类似的完整班级，以确定在线教学或面对面教学是否会导致更好的学习。这是一个示例：

a. 自然实验

b. 现场实验

c. 实验室实验

d. 拟实验

9. 在自然或现实生活环境中进行的实验，同时使用随机分配的是：

a. 拟实验

b. 自然实验

c. 现场实验

d. 真实实验

10. 以下哪项允许研究人员比较治疗前后的差异，并判断是否存在前测效应？

a. 静态对照组比较设计

b. 索洛蒙四组设计

c. 前测后测对照组设计

d. 单一组前测后测设计

|  |  |  |  |
| --- | --- | --- | --- |
| d.  Answers |  |  | |
| 1. b | 4. a | 7. a | 9. c |
| 2. c | 5. d | 8. d | 10. b |
| 3. b | 6. b |  |  |

**108** Designing Experiments for the Social Sciences

1 使用Googlescholar.com或您学校图书馆的数据库，查找在您的学科中使用实验设计的研究。阅读您最感兴趣的三个实验，并确定实验的设计类型：是真实实验或实验室实验、准实验、自然实验还是场地实验？是预测-事后对照组设计、Solomon四组设计还是其他设计？有哪些局限性，作者如何解决这些局限性？

2. 检查有关您的主题的文献，以了解使用的方法学。网格型图表将帮助您看到使用最多的方法。在第3章的文献综述任务中阅读的25篇以上文章中，按照方法进行分类，包括批判性文章、焦点小组、访谈、民族志、内容分析、调查和实验等。哪种方法使用最多？现在进行旨在建立因果关系的实验是否合适？

**应用练习**

Crasnow, S. 2015. “政治科学中的自然实验和多元主义。”《社会科学哲学》45 (4/5): 424–441.

Gerber, Alan S., and Donald P. Green. 2012. 《实地实验：设计、分析与解释》。纽约：Norton出版社。

Levy, Y., T. J. Ellis, and T. Cohen. 2011. “信息系统研究中初学者的实验和准实验研究指南。”《跨学科信息、知识和管理杂志》6: 151–161。

Campbell, Donald T.和J.C. Stanley. 1963. 《研究的实验和准实验设计》。芝加哥：Rand McNally出版社。

**建议阅读**

Chapter 4 ■ Types of Experiments **109**

1. D. T. Campbell和J. C. Stanley，《研究的实验和准实验设计》。芝加哥：Rand McNally，1963年。

2. 同上，第6页。

3. David W. Catron和Claudia C. Thompson，《四个重测间隔后WAIS分数的重测增益》，《临床心理学杂志》35卷2期（1979年）：352–357页。

4. 同上。

5. Brendon R. Barnes，《发展中国家社区试验中的霍桑效应》，《社会研究方法国际期刊》13卷4期（2010年）：357–370页。

6. John G. Adair，《霍桑效应：对方法论人为因素的重新考虑》，《应用心理学杂志》69卷2期（1984年）：334–345页；

Ryan Olson等，《我们在I-O和OB教材的引言部分教给学生的霍桑研究内容回顾》，《工业组织心理学家》41期（2004年）：23–39页。

7. E. Mayo, 1933年; Chen-Bo Zhong和Julian House，《霍桑再访：物理工作环境的组织影响》，《组织行为研究》32卷（2012年）：3–22页。

8. Olson等人，《我们教给学生的东西》。

9. Henry A. Landsberger，《霍桑再访。管理与工人：批评以及工业中人际关系的发展》（纽约：康奈尔大学，1958年）。

10. J. G. Adair，D. Sharpe和C. Huynh，《教育实验中的霍桑控制程序：对其使用和效果的重新考虑》，《教育研究评论》59卷2期（1989年）：215–227页。

11. Barnes，《社区试验中的霍桑效应》；Zhong和House，《霍桑再访》。

12. Baptiste Leurent等，《通过医疗设施离开访谈监测患者护理：坦桑尼亚疟疾治疗指南遵循试验中的霍桑效应评估》，《BMC传染病》16卷（2016年）：1–9页；Jim McCambridge，John Witton和Diana R. Elbourne，《霍桑效应的系统性回顾：研究参与效应需要新的概念研究》，《临床流行病学杂志》67卷3期（2014年）：267–277页；Magnus Hansson和Rune Wigblad，《重新解读霍桑效应》，《斯堪的纳维亚管理期刊》22卷2期（2006年）：120–137页。

13. Olson等人，《我们教给学生的东西》。

14. Martin T. Orne，《需求特征和准控制概念》，收录于《行为研究中的人为》：Robert Rosenthal和Ralph L. Rosnow的经典著作，编者Robert Rosenthal和Ralph L. Rosnow（牛津：牛津大学出版社，2009年）：110–137页；D. Steele-Johnson等，《目标导向和任务需求对动机、情感和绩效的影响》，《应用心理学杂志》85卷5期（2000年）：724–738页。

15. Martin T. Orne，《心理实验的社会心理学：特别涉及需求特征及其影响》，《美国心理学家》17卷（1962年）：776–783页；《需求特征和准控制概念》，收录于《行为研究中的人为》，编者R. Rosenthal和R. Rosnow（纽约：学术出版社，1969年），143–179页。

16. Barnes，《社区试验中的霍桑效应》。

17. Adair，《霍桑效应》。

18. David Salsburg，《品茶女士：统计如何在二十世纪革命性地改变科学》（纽约：W.H. Freeman，2001年）。

19. Campbell和Stanley，《实验和准实验设计》。

20. Kaanan Butor-Bhavsar，John Witton和Diana Elbourne，《研究评估本身是否会导致行为变化试验中的偏倚？基于Solomon四组研究的证据的系统综述》，《PLoS ONE》6卷10期（2011年）：1–9页；Campbell和Stanley，《实验和准实验设计》；Shlomo Sawilowsky和D. Lynn Kelley，《荟萃分析和Solomon四组设计》，《实验教育学杂志》62卷4期（1994年夏季）：361页。

21. R. Barker Bausell，《有意义的实验：成为科学家的40个步骤》（千橡，加利福尼亚州：Sage出版社，1994年），90页。

22. Yair Levy，Timothy J. Ellis和Eli Cohen，《信息系统研究中实验和准实验研究初学者指南》，《跨学科信息、知识和管理杂志》6卷（2011年）：154页。

23. Stephen Gorard，《实验数据分析中的误差传播：预测-事后测试设计的比较》，《教育研究与方法国际杂志》36卷4期（2013年）：372–385页。

24. 同上，372页。

25. Campbell和Stanley，《实验和准实验设计》。

26. V.L. Willson和R.R. Putnam，《实验设计中预测-事前测试敏感化效应的荟萃分析》，《美国教育研究杂志》19卷（1982年）：249–258页。

27. 有关示例，请参见Charles Boy Kromann、Morten L. Jensen和Charlotte Ringsted，《测试对技能学习的影响》，《医学教育》43卷1期（2009年）：21–27页。

28. Levy, Ellis和Cohen；J.W. Creswell，《教育研究：定量和定性研究的规划、实施和评估》，第2版（新泽西州，Upper Saddle River：Pearson，2005年）。

**110** Designing Experiments for the Social Sciences

29. Ralph Woehle和Andrew Quinn，《比较HBSE社会工作研究生课程的实验：面对面和远程教学》，《社会工作教学杂志》29卷4期（2009年）：418–430页。

30. Sameer B. Srivastava，《网络干预：评估正式辅导对工作场所网络的影响》，《社会力量》94卷1期（2015年9月）：427–452页。

31. Froukje Snoeren等，《使用儿童访谈干预在对儿童虐待和/或忽视报告后的调查中的有效性和成本效益的准实验设计》，《BMC公共卫生》13卷1期（2013年）：1–16页。

32. Sharon Crasnow，《政治科学中的自然实验和多元主义》，《社会科学哲学》45卷4/5期（2015年）：429页。

33. 同上。

34. Jill Theresa Messing等，《俄克拉荷马致命性评估研究：对致命性评估计划的准实验评估》，《社会服务评论》89卷3期（2015年）：499–530页。

35. Andrew Wheeler，《搬迁对效果的影响：评估家庭位置对犯罪地点的影响的准实验》，《定量犯罪学杂志》28卷4期（2012年）：587–606页。

36. Tannis MacBeth Williams，《电视的影响：三个社区的自然实验》，编者Tannis MacBeth Williams（奥兰多，FL：学术出版社，1986年）。

37. Campbell和Stanley，《实验和准实验设计》。

38. Crasnow，《自然实验和多元主义》。

39. Phillip Baker等，《贸易和投资自由化、食品体系变革和高度加工食品消费：秘鲁和玻利维亚软饮料市场的自然实验对比》，《全球化与健康》12卷（2016年）：1–13页。

40. Judith Green等，《在评估中整合准实验和归纳设计：免费公交对公共卫生的影响的案例研究》，《评估》21卷4期（2015年）：391–406页。

41. 同上，395–396页。

42. David S. Kirk，《集中安置前囚犯在同一社区的后果的自然实验》，《美国国家科学院院刊》112卷22期（2015年）：6943–6948页。

43. Aaron Reeves和Robert de Vries，《媒体报道是否影响公众对福利领取者的态度？2011年英国骚乱的影响》，《英国社会学杂志》67卷2期（2016年）：281–306页。

44. Mary S. Morgan，《自然实验和社会科学中的自然实验》，《社会科学哲学》43卷3期（2013年）：341–357页。

45. Alan S. Gerber和Donald P. Green，《实地实验：设计、分析和解释》，（纽约：W.W. Norton，2012年）。

46. 同上，9页。

47. 同上。

48. 同上。

49. 同上。

50. 同上。

51. 同上。

52. William Minozzi等，《国会成员在在线城市厅会议中对实质、归因和行为的说服力：现场实验证据》，《美国国家科学院院刊》112卷13期（2015年）：3937–3942页。

53. Erik Meyer和Chris Van Klaveren，《扩展日计划的有效性：来自荷兰的随机实地实验的证据》，《教育经济学评论》36期（2013年）：1–11页。

5

# 内部有效性与外部有效性

#### *没有什么问题比关于社会科学实验结果的有效性和普遍性的争论更长久或更频繁。1*

#### *— 约翰·A·考特赖特*

学习目标

• 阐述内部和外部有效性以及权衡之间的关系。

• 确定实验中如何实现普遍性。

• 比较逻辑推断与统计推断。

• 讨论实验中复制的作用。

• 解释随机分配如何提供内部有效性。

• 确定七类威胁内部有效性的外部变量。

前一章讨论了不同类型的实验——实验室实验、准实验、自然实验和野外实验，并得出结论：没有完美的实验。就像生活中一样，实验涉及权衡，这在有效性的两个方面——外部有效性和内部有效性中尤为如此。就像跷跷板上的孩子一样，一个有效性类型的上升往往伴随着另一个类型的下降。

**111**

下降。在社会科学实验中，目标是同时最大化两者，但通常很难实现。就像开头的引用所示，这个争论可能会持续很长时间。"有效"意味着某物是有根据、准确或权威的。在科学的情况下，研究的结论至少在可接受的概率水平上应该是有效或真实的。有效性的话题对于实验尤为重要，实验常因内部有效性而受到赞扬，但同时也因缺乏外部有效性而受到批评。内部有效性指的是实验效果实际上是由处理引起的程度，而外部有效性涉及研究在超出特定研究中的被试、环境和处理的能力上进行推广——也就是说，实验效果也会在现实生活中得到证实。因为外部有效性对实验的威胁程度与内部有效性对观察性研究的威胁程度一样大，本章从外部有效性的讨论开始。

**生态学和外部有效性**

广义而言，外部有效性是实验结果能否推广到其他人、环境和处理上的程度。影响外部有效性的一个关注点是实验的真实性（或人为性），即生态有效性。当实验反映了真实生活环境或模拟了环境，为被试提供了与实际发生情况相似的体验时，可以说具有生态有效性。尽可能营造自然情境增加了被试正常反应的机会。准实验、自然实验和野外实验在生态有效性方面比真实实验或实验室实验更高。实验室实验者面临的挑战是在保持控制的同时最大限度地增加真实感。我们从第二章中可以看到，斯坦利·米尔格拉姆（Stanley Milgram）和菲利普·辛巴多（Philip Zimbardo）为此付出了努力。现今的实验者仍然可以通过额外的努力和想象力达到一定程度的真实感。例如，在实际的投票场所进行政治科学实验，使用真实的投票设备和注册选民，有助于增强生态有效性。与真实的教师互动比要求他们根据阅读内容想象互动更加真实。某些学科的研究人员发现近似真实生活比其他学科更容易。例如，研究加薪是否能提高绩效，通过将人们带入实验室，创造工作岗位，操纵奖励，然后观察或测量绩效似乎对于商业实验来说是不可思议的。但在现实生活环境中无法控制，人们显然不能随机分配加薪。让被试阅读关于一份工作的信息，想象自己处于那种情境中，并回答关于他们可能会做什么的问题，可能是目前可行的最佳选择。这是对外部有效性的一个例子。"

权衡。实验经济学长期以来一直关注实验的人为性，不仅因为实验环境和任务与现实世界不同，还因为反应性和需求效应，这些内容在第二章中已经讨论过。研究人员应该了解自己所在领域的思考方式。创造真实感通常需要研究人员更多的工作和创造力，但结果可能是值得的。我个人亲身体验到了这一点，当我还是研究生时，我帮助进行了一个实验，旨在确定不同框架的新闻报道是否会导致人们将问题的责任归因于社会或个人，并支持不同的解决方案。在第一次尝试中，我的导师和我使用了真实的新闻报道，但将它们以打印在8.5 x 11英寸纸上的形式呈现给被试。除了有标题和署名外，它们并不太像真实报纸中的报道。我们没有发现任何效果。在一番辩论之后，我们决定更加真实地呈现这些报道；我们不再将这些报道交给被试在一张打字纸上阅读，而是生成了带有被操纵的报道的模拟大报纸，这些报道嵌入在其他真实新闻报道、照片、报头和页码旁边。我们使用一个超大复印机将它们打印在22英寸长和13英寸宽的纸上，与大报纸的尺寸相同。之前没有产生差异的报道现在展现了理论所预测的效果。我们所做的只是改变了外观以更加真实。这样做的成本更高，时间更长，但模拟报纸页面的真实感似乎是关键所在。我们并没有将所有的事情都做得非常真实；例如，我们的被试仍然在大学教室中阅读模拟报纸，而不是在早餐桌前阅读，而且报纸的名称是虚构的，而不是真实存在的。但这一个改变似乎已经足够了。一个更便宜、更简单的妥协方法是13 x 22英寸的大报纸中的一个剪报，这是报纸记者所说的，“剪报”看起来像是从报纸上剪下来的一篇报道。我发现使用剪报也会产生效果，并且制作起来更容易。今天，研究人员更有可能需要创建网页或推文。（第9章将讨论创建真实刺激的方法。）

**推广性**

真实性是推广性的关键之一，即高度人为的实验效果如何在现实生活中转化。除了看起来有多真实之外，外部有效性还涉及到特定实验的结果是否适用于其他人、其他环境、不同时间和不同处理的情况。当结果超出研究的具体情况而具有普遍适用性时，我们称之为“稳健”。大多数实验不像调查和其他定量研究那样从人群中进行随机抽样。

这是一个报纸报道的例子，上面是低生态效度的版本，下面是具有高生态效度的“剪报”版本。

糖尿病班促进社会支持

社区帮助糖尿病患者过上幸福健康的生活，班级宣传

泰勒·丹尼尔斯报道

每日记录记者

安吉拉·米切尔没有晚期肾脏疾病或不可手术的脑肿瘤。她的疾病更容易管理。事实上，有数千万美国人与安吉拉一样。他们都是糖尿病患者。

米切尔参加了由西安德森社区医院提供的糖尿病教育班，并发现美国糖尿病或糖尿病前期的人数现在已经达到7500万。

糖尿病患者和糖尿病前期患者占美国人口的25%。预计到2050年，美国糖尿病患者的人数将达到3900万。

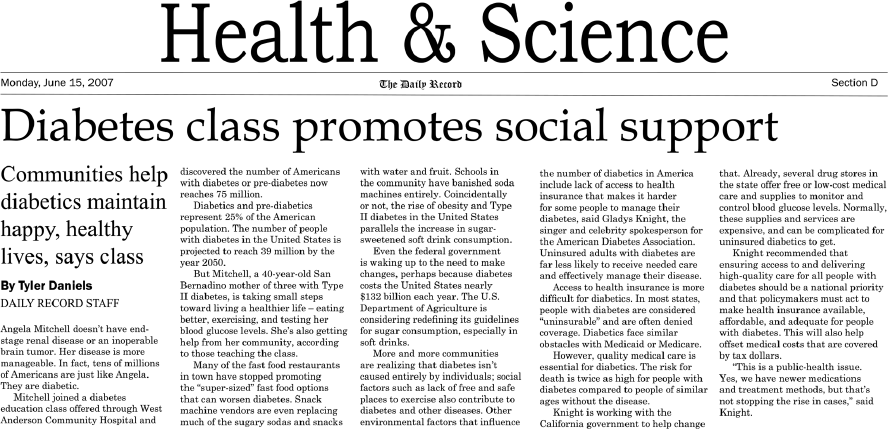
但是，40岁的圣贝纳迪诺三个孩子的母亲、患有2型糖尿病的米切尔正在朝着过上更健康的生活迈出小小的步伐——改善饮食、锻炼身体并检测血糖水平。根据教授这门课程的人所说，她还得到了社区的帮助。

镇上的许多快餐店已经停止推广那些可能加重糖尿病的“超大份快餐”选项。零食机供应商甚至用水和水果取代了大部分含糖的汽水和零食。社区的学校已经完全取消了苏打水机。美国肥胖和2型糖尿病的增加与摄入含糖软饮料的增加有着不可忽视的相关性。

甚至联邦政府也开始意识到需要进行变革，或许是因为糖尿病每年给美国造成近1320亿美元的费用。美国农业部正在考虑重新定义其对糖的摄入指南，特别是对于含糖饮料。

越来越多的社区意识到糖尿病不完全是个体所致。缺乏自由和安全的锻炼场所等社会因素也对糖尿病和其他疾病起到了一定的贡献。美国糖尿病协会的代言人、歌手格拉迪斯·奈特表示，影响美国糖尿病患者数量的其他环境因素还包括缺乏健康保险，这使得一些人更难控制他们的糖尿病。没有健康保险的糖尿病患者更不可能获得所需的医疗保健和有效地管理他们的疾病。

对于糖尿病患者来说，获得优质医疗保健至关重要。与没有糖尿病的同龄人相比，糖尿病患者的死亡风险是其两倍。



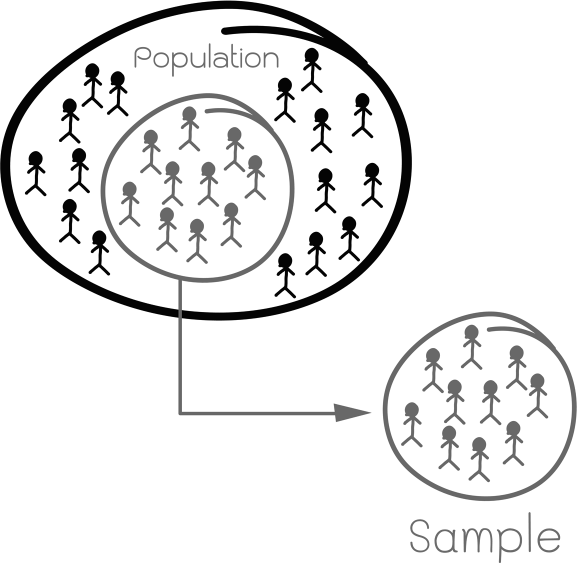
方法可以做到。随机抽样可以让人群中的每个人都有相等的被选择机会，从而得到的样本使得研究者能够以一定的概率说研究结果适用于更大的人群。大多数实验不会随机抽样被试者，但它们通过其他方式实现了泛化的能力，包括使用尽可能接近被泛化对象的人群。例如，以20岁大学生作为研究对象的实验可能无法推广到缓刑和假释的人群。关于志愿者的研究结果可能无法适用于不参与志愿活动的人群。13对于捐助健康相关慈善机构的原因，艺术爱好者与普通观众可能存在差异。14使用会计师的实验结果可能无法推广到工厂工人等其他职业群体。实验中治疗方法只对实验所选用的特定人群有效的可能性是外部效度的一个关注点。15实验的一个最严重的限制是无法推广到研究所涉及的特定人群之外的人群。16 Campbell和Stanley举了一个假设的例子，其中一个研究者试图邀请十所不同的学校参与实验，但被九所学校拒绝，他们说：“这第十所学校几乎肯定在很多具体方面与其他九所学校和我们希望推广到的学校群体不同，因此它是不具代表性的。”17那所学校同意参与实验的原因可能也导致治疗方法的有效性，例如，如果研究仅适用于实验中的一个人群，而未包括在研究中的更大群体，则不能将结果推广到未被研究涵盖的群体。18这导致了一个明显的结论，研究者应该使用与他们打算研究的人群或其他单位（如企业、学校或组织）尽可能相似的样本，例如使用注册选民而不是未注册选民，使用全国各地的公民而不仅仅是一个州的公民，或者使用会计师而不是学生，如果这些第一组是要推广的目标人群。这反对了现如今对学生进行实验的常规做法；如果实验的目的是研究军事领导者的决策或最能鼓励家长给孩子接种疫苗的公共服务公告的类型，那么在系部的学生实验池中使用二年级广告专业学生就不是一个好选择。当被试者不能代表研究结果所应用的人群时，外部效度就较低，并且研究结果的可信度会受到质疑。例如，如果研究面向的是广告学生，甚至是广告专业人士，那么实验对象的选择就更具合理性，因为广告学生很可能很快就会成为广告从业人员。19使用实验对象池中的学生可能更容易，但如果这些学生无法接近实验所要泛化结果的人群，那么为了寻找更具代表性的被试者而付出更多努力可能是有益的。第8章将详细讨论样本选择的机制，特别是学生样本的选择。

Campbell和Stanley描述了威胁效度的十二个因素。20本章不会详细讨论它们所有，而是专注于几种应对低外部效度的较流行方法。有时，实验者会发现他们的实验被论文审稿人评估，而这些审稿人是该领域的专家，但对实验方法并不熟悉。这有时会导致反馈意见，认为如果被试者没有进行随机抽样，研究就不应该被认为是有效的，也不应该发表。在这些情况下，准备一份知识渊博、恭敬有礼的回应是很有帮助的。本节旨在帮助设计实验，从一开始就避免这样的批评，并在需要时作出回应。

最近的研究解决了实验的泛化问题，并提出了解决问题或至少估计结果准确代表研究对象之外特定人群的可能性的方法。21其中一些方法包括选择被试者的不同方法，而其他方法则通过结果变量或协变量创建测量指标，以估计实验样本对人群的代表性。22其中一些技术可以在选择样本时使用，其他技术则在数据收集后使用。23这些都是相当高级的技术，本教材不会详细讨论；相反，本章涵盖了基本问题和解决方案。

**随机抽样**

正如随机分配是实验中实现等价组的黄金标准一样，实现泛化的黄金标准是随机抽样。24自从1936年《文学指南》杂志错误预测了美国大选的胜利者以来，我们就知道了这一点。

就像乔治·盖洛普的民意调查机构在1936年的总统选举中正确预测结果一样，随机抽样是从较小样本中高效、有效地估计整个人群某种特征的方法。但在实验中很少看到它的使用。首先，我们应该明确随机分配和随机抽样之间的区别。随机抽样是选择研究对象的过程，以使他们充分代表更大的人群。它确保每个人、地点或其他单位都有平等的被选中机会，而且比测量整个人群更便宜、更容易。

尤其是在不可能完成时，随机抽样确保每个人、地点或其他单位从人群中有平等的被选中机会，而且相比于测量整个人群而言更为便宜和容易。

相比之下，随机分配是决定哪些对象接受特定的实验处理或分配到特定组的行为。随机分配通常使用方便样本，即容易获取的对象，而不是随机抽样的对象。两者的目的有很大的区别。根据一种观点，"随机分配...通过使样本之间随机相似来促进因果推断，而随机抽样则使样本类似于人群"。在实验中，重要的不是研究对象是否与不参与研究的人相似，而是研究中不同组别的研究对象是否相似。

**两步随机化**

如果每种随机化方法都是黄金标准，那么将两种方法结合使用肯定是白金标准。事实上，一些实验确实使用了两步随机化模型，首先从人群中随机抽样，然后将被选中的对象随机分配到处理组中，但这种情况非常少见。最明显的原因之一是研究人员需要一个完整且最新的人群名单，以便进行随机抽样。人群越大，这种名单存在的可能性就越小。例如，美国没有完整的所有适龄学童的家长名单，包括那些就读于私立学校的家长（2023年数据）。虽然有登记选民的名单，但没有未登记选民的名单。将人群限制在具体范围内可以使随机抽样在某些情况下更可行；例如，可以获得某个学区所有学生的名单，但是样本范围的缩小会引发之前提到的同样问题：一个学区的效果是否适用于其他学区？如果研究人员可以从全美各地进行随机抽样，那么将很难、甚至不可能将实验者所使用的某些处理应用于这些遍布全国的人群。研究人员将不得不在全国各地进行旅行或将对象带到实验室，这两者都是昂贵的，或者依赖他人进行处理操作，这会危及内部有效性，因为对象对不同的实验者有不同的反应，甚至可能以不同的方式提供刺激。除了这些物流考虑因素外，还有预算限制。例如，商业上可以购买到一个包含所有美国工作记者的数据库，但需要花费数千美元。考虑到编制此类信息需要多少时间和精力，这是可以理解的。大多数研究人员没有这样的时间和资金。研究人员应该权衡收益和成本。

**非响应偏差**

在获取人群名单较为容易的情况下，实验研究者会遇到与调查研究人员在招募被试者时遇到的相同问题。为了估计

一个人群所需的参与者比发现效应所需的参与者要多得多；例如，为了进行推广，可能需要400名参与者，而只寻找效应的实验通常只需要40名参与者。因此，如果实验者还希望进行推广，就需要更多的参与者，这会增加成本。此外，并不是每个被随机选择的人都愿意参加研究；民意调查和调查研究人员在面对最大的外部有效性问题之一——非响应偏差时就会遇到这个问题，即当人们不回答问题或拒绝参与时。不回应的人可能在未知的方面与回应的人非常不同，就像在 Campbell 和 Stanley 的虚构示例中拒绝参与的学校一样。非响应偏差对推广性构成威胁，多年来，响应率一直在下降，甚至跌至个位数以下。还有一个关于人员流失的担忧，即在完成研究之前退出的受试者。对于实验研究人员来说，费时、费钱和麻烦从人群中获取一个随机抽样名单，然后发现不足够的人同意参与以确保推广性是没有意义的。当可能获取到研究中所有单位的名单时，例如 Malesky 等人在越南的一项关于贿赂的在线实验中所做的，然后在随机分配之前对它们进行随机抽样，必须计算和报告响应率，并提供对响应偏差的讨论（见研究亮点5.1，"Generalizability and External Validity Concerns"）。然而，即使有这样的操作，研究的推广性问题也并未消失。作者提供了一张表，比较了样本的特征与人群的特征。

这些问题使得随机抽样成为从实验中进行推广的"不充分的解决方案"。此外，一些企业、政府机构或其他组织不允许对其员工或成员进行随机抽样；他们可能认为给予一些人机会而其他人没有机会是不公平的，等等。例如，如果研究人员提供职业发展研讨会作为处理，雇主可能不愿意告诉一些员工他们不能参加，并面临偏袒和歧视的指责。由于随机抽样需要"高水平的资源和很少可行的后勤控制"，大多数实验研究人员使用其他在科学实验领域被接受的抽样方法。这些抽样方法将在第8章进行讨论。为了更加有信心地推广实验的结果，研究人员使用其他方法，接下来将对其进行描述。

**代表性**

从更大的人群中进行随机抽样的目标是确保较小的样本与较大的人群相似，也就是说，它是具有代表性的。在缺乏

## 研究重点 5.1

普遍适用性和代表性

下面是一段关于实验推广性的讨论的示例。作者包括了一张比较样本与总体代表性的表格。

Malesky, Edmund J.，Dimitar D. Gueorguiev 和 Nathan M. Jensen. 2015. “Monopoly Money: Foreign Investment and Bribery in Vietnam, a Survey Experiment.” 《美国政治科学杂志》第59卷（2期）：419–439。

附录4：2010年省级竞争力指数样本和普查数据特征

*A screenshot of a computer screen

Description automatically generated(未完待续)*

"我们的最终样本由19,363家国内企业和3,888家外资企业组成，它们在2000年之后注册，并分布在该国的63个省份。在这三年中，选择样本的框架是国家税务管理机构数据库中注册的国内企业和外资企业的名单。对于国内企业，调查的回应率约为30%，对于外资企业则为25%，这比国际商业文献中通常报告的回应率要高得多（White和Luo，2006年），但仍然足够小以引起可靠性的担忧（Dillman等，2002年）。因此，有理由问是否非响应会导致选择偏倚，从而可能影响我们的结论（Jensen，Li和Rahman，2010年）。在支持信息的附录4中，我们将PCI数据与2012年的国家统计局企业普查和税务管理机构数据库中的可用信息进行了比较，结果显示给定年份的PCI数据反映了国家人口的可观特征，因此可以高度准确地描述越南的外资和国内投资者。”31

能够进行随机抽样，并采取预防措施确保样本充分代表整个人群，是增加实验结果外部有效性信心的一步。35 Shadish、Cook和Campbell讨论了在实验中使用的各种增强代表性的方法，包括将样本作为人群的缩影，称之为“在目的意义上代表性的描述样本”。36 目标是了解人群的关键特征，并确保样本充分代表尽可能多的这些关键特征。许多实验会努力比较样本的特征与目标人群（例如全国普查）的特征，以此来验证样本的代表性。37 同时，也需要指出样本与人群的差异之处。38 在某些人口统计学上，样本不必与人群的比例完全一致是可以接受的，因为只需要在一些而不是所有关键特征上与人群相似即可。39 例如，Klaiman等人的样本在性别、年龄、收入和农村与城市地区与普查结果相似，但受教育程度稍高于人群。40 研究人员应特别关注影响结果变量的变量。

无论匹配有多好，这种广泛使用于实验中的推广方法从统计逻辑上都无法得到支持。41 Verschoor等人对经济风险承受能力的研究专门测试了实验如何反映现实行为的问题。在这项针对乌干达农民的研究中，研究人员使用了代表性样本，并将他们的研究称为“实验方法外部有效性的测试”，43 结果发现，如果在抽样方法和其他有效性方面采取了足够的注意，使用代表性人群的实验确实具有外部有效性，并且尽量减少混杂变量的影响。

## 更多关于...的内容...框5.2

## 随机抽样和随机分组的实验

使用随机抽样和随机分组的真实实验是罕见的。虽然它不被推荐作为推广到整个人群的唯一方式，但Shadish、Cook和Campbell确实表示，“当可行时，我们明确支持它”。34

“可行”通常意味着人口被限制在可管理的范围内，例如一个城市或一个州的居民，或者当存在整个人群的名单时，例如对于持牌职业的成员。当数据库是公开可用的、研究人员寻求具有访问人群信息的官员的帮助，或者可以购买数据库时，随机抽样也会更加便利；例如，公司Cision跟踪美国所有的印刷、广播和在线记者，还有顶级社交媒体影响者（http://www.cision.com/us/pr-software/media-database/）。

下面总结了两个使用两步随机化的实验的例子：

陆方文、张佳楠和Jeffrey M. Perloff。2016。“制止交通违规的一般和特定信息：来自随机实验的证据”。《经济行为与组织杂志》123：97-107。

这个实验在中国青岛进行，使用了该市的注册车辆数据库，该数据库也列出了车主的手机号码。研究者与青岛市公安局合作提供了数据。该研究还将注册车主随机分配到几个处理条件中，他们会收到包含不同内容的短信，内容涉及他们收到的交通违规、一般安全信息或者对照组不收到任何信息。然后他们追踪观察哪种安全信息能够导致驾驶员在未来遭受更少的违规行为。

史颖。2016。“跨领域信息与选民投票率：来自同性婚姻修正案的证据”。《政治沟通》33（3）：433-459。

这项研究使用了北卡罗来纳州公开可用的选民注册和选举历史数据库作为人群样本，然后随机抽取了注册的民主党和共和党选民的样本，给他们发送直邮明信片，内容要么支持他们所属政党在同性婚姻问题上的立场，要么反对，以观察跨领域观点是否刺激或抑制政治参与，参与程度以是否在收到明信片后进行投票来衡量。

有多种方法可以增加样本代表性的可能性，例如，对某些难以获取的人群进行过度抽样，使用多个地点进行研究，这应该能增加样本的多样性。44 使用在线调查方法进行实验，通常被称为调查实验，也增加了实验使用代表性样本的能力，45 此外，使用由抽样公司招募的调查受访者面板也能增加代表性样本的使用。46 这些将在第8章中详细讨论。

当实验样本不能准确代表总体时，论文的讨论部分应包括对这一局限性的承认，并解释为什么这并不成问题或不应该否定研究结果。47（请参阅“How To Do It”框5.3，了解一些实验研究人员处理这个问题的示例。）一个好的做法是，如果样本不代表总体，总是要谨慎地说出“结果应该谨慎解释”，并小心谈论“这些受试者”或“样本中的人”，“我们的受试者”或类似的语言，以避免暗示超越样本之外的情况。

**因果关系**

此外，在文章中指出实验并非旨在推广，而是旨在发现效应的事实也是一个好主意。在大多数实验中，

**如何操作5.3**

当样本不具有代表性时，应该说些什么

对于非代表性样本，实验研究者会坦率地指出这一点，但会以其他各种理由为结果辩护。以下是一些研究中关于这方面讨论的摘录：

Kim, Mirae, 和 Gregg G. Van Ryzin. 2014. “Impact of Government Funding on Donations to Arts Organizations: A Survey Experiment.” Nonprofit and Voluntary Sector Quarterly 43 (5): 910–925.

“与美国人口相比，面板中女性和白人的比例相对较高，这是Berrens、Bohara、Jenkins-Smith、Sliva和Weimer (2003)在其他美国志愿在线面板中发现的共同特点。尽管这限制了使用在线面板进行研究的外部有效性，但有一些证据表明，至少对于某些主题，这种志愿者面板可以近似基于美国人口概率抽样的调查结果。虽然不能代表一般人口，但我们研究中的参与者明显对艺术感兴趣并支持，因此可能更接近美国许多艺术非营利组织的潜在捐助者群体”（第916-917页）。

Dunlop, Claire A., 和 Claudio M. Radaelli. 2016. “Teaching Regulatory Humility: Experimenting with Student Practitioners.” Politics 36 (1): 79–94.

“无论是公共部门行政人员还是公共管理硕士学生，两个群体都不是代表性样本。然而，这并不成问题，因为我们的目的是说明在课堂上使用实验帮助学生从批判性角度进行反思，而不是对控制错觉文献做出任何实证贡献。所收集数据的实用性是教育导向的，而不是以研究为导向”（第82页）。

并不是旨在代表较大人群。48相反，实验的目的是研究理论上重要的关系。这是公众、媒体和其他研究方法使用者偶尔会产生的对实验研究的根本误解。49研究人员可能会陷入诸如普遍性等问题中，并忽视实验的目的是建立因果关系——即确定某个处理变量是否是某个结果或效应的原因。实验室实验的目的并不是将结果推广到特定人群。50在这方面，随机分配的问题更为关键；关于此问题将在第7章中进一步讨论。

调查、内容分析和其他方法与实验相比存在相反的问题；如果采样正确，它们可以推广到人群，但不能推断它们所发现的关系的原因。51这类研究在描述时应该谨慎使用“相关性”或“关系”而不是宣称或暗示“因果关系”。对于调查研究人员来说，广泛获得人们的态度和观点更为重要。询问这些态度和观点是否被他人广泛持有是更合理的。但对于试验来说，旨在确定因果关系，这个问题相对次要，重要的是要确定结果是否受到了其他因素的干扰。

对于审查——非商业使用（2023）

逻辑推断

对于试验试图进行的推断的种类也可能存在误解。实验研究并不是进行统计推断，也就是说可以通过数学计算来确定样本与较大人群之间的拟合程度。实验研究依赖于逻辑推断，即根据常识可以得出合理的结论。例如，如果一项实验测试某个与种族、性别或所在地等因素关系相对不大的功能，例如意见形成的认知过程，那么可以合理地推断到样本之外的其他人群。Basil举了一个从近距离拍摄和拍摄较远处的照片导致的生理反应的例子，他说对这样的图像的激动应该是相同的，无论在样本中有谁。如果没有证据证明相反，那么可以合理地假设在非样本中也会有类似的反应。例如，如果大学生比非大学生更害怕恐怖的近距离镜头，为什么我们不期望非大学生有相同的反应呢？同样，已经证明媒体的影响，如议程设置、启动和框架在全国乃至世界各地的各种人群中的作用方式相同。54然而，如果所研究的内容因不同条件而变化——例如，在具有

国营媒体——我们不会从不同条件下的结论中推广到它们，例如在自由新闻社会中。除非某个现象对不同的人群有不同的影响，否则研究结果应该适用于所有人。对于实验来说，真正关注的是样本是否适用于所研究的内容，而不是样本是如何获取的。55

最后，如果基于所使用样本的普遍性来否定实验结果，那将意味着拒绝医学、心理学、化学、物理学等学科在很大程度上基于非代表性样本对科学和日常生活做出的有价值的贡献。

**复制**

与其过分关注样本如何获取或者样本是否与更大的人群相似，实验主要依赖于复制来推广其结论。56复制是指在不同的主体、不同的条件、不同的时间和地点、不同的问题、变量等情况下重复实验结果的能力。如果在不同研究中发现了一致的效应，那么可以更有信心认为这些发现代表着某种真实和可推广性，而不仅仅是基于个别样本的情况。57复制是一种特定的方法重复使用的重复研究，用于特定评估先前研究结果的真实性。58一些依赖于循证实践的专业对于复制尤为关注，以确保在进行昂贵的实践变革之前，结果是可重复的。59政治学在20世纪90年代开始了复制运动，60社会心理学最近也在20世纪70年代开始了自己的复制运动。61不幸的是，复制并不像呼吁进行更多复制研究那样普遍存在；例如，在特殊教育领域，仅有0.5%的文章试图复制先前的发现，62而在教育领域总体上则仅有0.13%。63心理学期刊的比例最高，为1.07%，64但这仍然似乎很低。

几乎每个研究方法的社会科学家都经常听到单一研究不能证明任何事情；这同样适用于实验。研究人员还建议使用谨慎的语言——“证明”一词不应该即使在结果复制时也不应使用；相反，好的表述包括“这增加了我们推广的信心”或“通过复制，我们在进行推广或因果推断时处于更有力的位置”。

**自我复制**

研究人员可以通过自我复制来加强自己研究结果的推广能力，称为自我复制。在某些领域，这是为了证明结果不是由某些奇怪的因素造成的必备条件。通过使用新的样本、增加样本量、在不同的环境中重复研究，研究人员可以验证自己的结果的稳健性，并进一步验证其推广性。自我复制对于验证实验的可靠性和推广性具有重要意义，并对于进一步验证和发展研究领域的认识至关重要。

同一篇文章中的第二个研究是实现复制的一种方式。到1998年，一本社会心理学期刊上的文章中的实验数量已经多达三个。65这也符合学术部门鼓励的系统性研究计划。例如，一个关于新闻道德判断的研究计划始于一项实验，通过操纵新闻记者在做出伦理决策时是否看到照片，发现照片能提高他们的道德判断质量。66这个主要结果在同一篇文章中进行了第二个实验的复制。这些实验引发了更多关于记者道德决策过程的问题，因此研究继续探索该主题，进行了三项关于种族是否产生差异的研究，67还进行了一项使用观众而不是记者的研究。68所有这些实验都测试了照片是否能改善道德判断，并发现确实能够改善。通过使用学生、专业人士和不同种族的观众，在国家的不同地区和不同时间进行的五项研究得出一致的结果，使我们对照片提升用于做出伦理决策的理由质量的能力更有信心。

**精确复制与概念复制**

一个重要区别是这些复制研究并非对第一个研究的精确复制或直接复制，即它们并没有完全重复原始的方法和步骤。在社会科学领域，这种类型的复制研究并不像在医学等其他领域那样受到重视或推荐。69在社会科学中，精确复制甚至可能是不可能的，因为必须存在某种差异，比如研究对象可能会有所不同。70相反，复制研究还应该在某种理论意义上对研究进行扩展。这被称为概念复制，71或称为再现，72在关键特征上有意地变化，如环境、研究对象、变量或干预措施。概念复制测试研究的理论，而直接或精确复制测试研究的操作化。概念复制扩展了实验的推广性，73因此，本章重点关注这种类型的复制研究。在前面的例子中，研究了通过照片传达的种族如何影响新闻学学生的道德判断是扩展研究关于照片影响的发现的第一步。之后的研究通过调查黑人新闻学学生的内外群体过程来进行同样的研究，因为种族对于少数族裔来说可能有不同的作用，结果确实如此。下一项研究将其从新闻学生扩展到了三个不同种族的专业记者，并研究了同理心的中介作用。最后一项研究将焦点从记者转移到新闻观众，并研究了认知详述的工作方式。这些研究代表了一个系统性的复制进程，通过在一个类型的研究对象之外逻辑地扩展，并纳入不同的概念进行。

因果机制是涉及照片的道德判断过程中的变量或路径，它们是某种结果发生的途径，是将某种原因与效果联系起来的东西。在这种情况下，种族、同理心和详述是导致更好的道德判断（结果或效果）的因果机制，而看到照片（原因）则起到了连接作用。回想一下第1章关于为什么某种效果发生的解释的讨论。斯坦利·米尔格拉姆关于服从权威实验系列的书籍是将发现系统地扩展到不同种群、不同环境和不同时期，并纳入不同的理论概念作为因果机制的复制的一个很好的例子，该实验在第2章中有介绍。74

**多个实验**

许多社会科学实验报告的是单一研究，但有些学科更倾向于在一篇文章中报告多个实验。例如，社会心理学正朝着在一篇论文中报告多个研究作为可复制性证据的方向发展。75在传播学中，多个实验也变得越来越常见，连续的实验系统地排除了替代解释。76菲斯克建议研究人员在一篇论文中发表多个研究。77虽然将研究分割成多篇论文可能对个人职业发展更有帮助，但在一篇论文中报告多个实验有助于推广性的原因，因为一致的结果为结论提供了更强有力的支持。一个例子是施韦策尔的一篇文章，他在一篇论文中报告了两个实验的结果，这两个实验旨在评估教授学生一种促进创意生成和讨论的技巧的有效性。第二个实验复制了第一个实验的结果，并通过确定学生需要多少个教室会话来学习这种技巧来扩展了实验结果。

在一篇论文中报告多个实验也是有道理的，当一项研究引发问题或对结果提出了合理的替代解释时，需要进行第二个实验来回答这些问题或排除替代解释。单独一项研究可能无法通过期刊审稿人的审核，除非还有额外的实验来解释结果并回答问题。一篇论文中多个实验的一个例子是一篇关于伤害和厌恶对政治道德判断的影响的论文。第二个实验的设计是为了测试第一个实验中出现的替代解释，并且复制了那些结果。它还引入了新的政治议题，从而对三个不同的议题进行了结果的复制。有关复制的更多示例，请参阅5.4的“更多内容”框。

外部复制

自己复制自己的工作确实有助于提高研究的推广性，但是让独立的研究人员（即不包括你自己、你的合著者，甚至不包括你所在机构的其他人）来复制结果也是理想的。这被称为外部复制。如果实验可以被外部研究人员独立复制，与第一组实验无关，那么对结果的信心就会增加，因为这有助于确认之前的结果不是错误、质量不佳的工作，甚至不是彻头彻尾的欺诈行为。80一些研究发现，当一个或多个复制研究的作者也参与了原始研究时，他们复制出原始发现的可能性显著高于外部研究人员复制的可能性。81当然，这可能有很多原因，并不一定指向邪恶活动。而且，一个失败的复制并不一定意味着效应是错误的或研究是欺诈的。相反，结果是全面考虑的，当多个复制研究显示出类似的模式或接近原始研究的结果时，它们被认为是成功的。82当外部复制不能复制出原始结果时，可能是因为样本容量过小、测量不准确、实验者效应、执行问题或其他许多原因。这不是一种注定失败的迹象，它鼓励更多的研究。这可能表明发现了一些新的或有趣的东西；例如，如果某种政治策略一直有效，但突然在外部复制中未能复制，这可能表明实验存在缺陷。非统计显著的发现仍然有意义，并且应该被报告。虽然这里无法详述，但如果想看研究人员在外部研究人员未能复制其原始结果时作出的回应，请阅读Ebersole等人（2017）的文章。83

事实上，结果无法被复制或仅能复制部分结果是很常见的。一些研究发现复制率低于50%。84以照片对提高道德判断能力的研究为例，经过五次成功的概念复制，有一次未显示出相同的结果。85那篇论文提出了理论上的原因，并通过实证证据支持了这一点。有一点是不道德的，那就是不报告或不尝试发表复制失败的结果。这种缺乏透明度或试图掩盖不希望的结果的行为可能会导致有前途的研究生涯的终结，因为一些人已经承担起了监督专业的责任。86

科学知识是累积的，它随着相关研究的进行而积累。因此，概念复制不仅对推广性很重要，也对推进我们对重要的人类现象的理解很重要。社会科学研究是关于尽可能多地了解人类过程，而不是以任何代价支持自己的假设。

i我想感谢匿名评审人员建议添加这两个最后观点。.

## 更多内容...盒子5.4

###### 复制研究的例子

Krupnikov，Yanna和Spencer Piston。2015。“强调负面：候选人种族和竞选策略。”《政治传播》32（1）：152-173。

这项研究在实验中操纵了种族和语气，观察选民如何对黑人和白人候选人的负面竞选广告作出反应。它重复了先前研究的结果，并通过研究种族和负面性的互动来扩展了研究范围；在先前的研究中，这两者被独立考虑。作者进行了第二个实验，以测试第一项研究中出现的替代解释，并复制了自己的结果。这项实验使用了嵌入在调查中的具有全国代表性的样本。

Camerer, Colin F.，Anna Dreber，Eskil Forsell，Ho Teck-Hua，Jürgen Huber，Magnus Johannesson，Michael Kirchler，Johan Almenberg，Adam Altmejd，Taizan Chan，Emma Heikensten，Felix Holzmeister，Taisuke Imai，Siri Isaksson，Gideon Nave，Thomas Pfeiffer，Michael Razen和Hang Wu。2016。“经济实验的可复制性评估。”《科学》351（6280）：1433-1436。

这篇文章有很多作者的原因在于他们复制了经济学领域18个不同实验中最重要的发现。他们成功地复制了这些实验的结果（61％）。这篇文章中没有列出所有18个研究的详细信息，但可以在一个相关网站上找到。

Nazarian，Deborah和Joshua M. Smyth。2010。“背景调节表达性写作干预的效果：一项随机的两个研究的复制和扩展。”《社会与临床心理学杂志》29（8）：903-929。

这篇文章报道了两个实验，测试了表达性写作的效果，通过改变受试者的写作地点（家中或实验室）和实验者的权威性（通过穿着严肃程度）进行操作。该研究还涉及生态效度的问题，即实验室之外的普遍适用性，发现学生在实验室写作时表现更好，而成年人在家中写作时表现更好。一个实验中使用学生，另一个实验中使用成年人，也有助于推广到不同人群的普适性。

一个写得好的方法部分的主要目标之一是让其他人能够复制这项研究。方法部分就像实验的配方，提供足够的细节，以便其他人能够在不联系原作者的情况下重现一项研究。当方法和实验步骤清晰、完整且精确时，其他研究人员更有可能成功地复制结果，因为他们将获得所有必要的元素，恰好如原研究。

同样，目前正在进行的另一个运动是将数据和用于分析数据的编码免费提供，以便他人可以轻松检查研究工作。研究人员向其他请求者提供数据一直是伦理准则的一部分。随着互联网的发展，这种做法变得更加普遍，因为研究人员可以直接下载数据，而不需要向其他研究人员提出请求。越来越多的期刊鼓励作者在补充材料网站上不仅提供数据集和代码，还包括刺激材料和其他组成部分。一些研究人员甚至将它们放在自己的网站上或学术社交媒体网站上，如Researchgate.net和Academia.edu。虽然有少数期刊要求这样做，但大多数仍然是自愿的。同样，一些期刊还包括专门用于复制的常规部分。87

三角测量

建立某一发现的有效性的另一种方法是同时使用定性和定量方法。这对于泛化性也很重要。如果实验室实验的结果与访谈、人类学研究和其他方法的结论一致，那对于现实世界的相关性将具有重要意义。Fiske和其他人推荐将实地实验作为实验室实验的复制，因为它们具有较高的外部有效性。

最后一点，关于外部有效性的讨论将在第4章中重新讨论。除了前面概述的优缺点之外，出于外部有效性的考虑，实验设计时应避免预测测试。由于测试和处理之间的相互作用，将受试者暴露于两次测试可能会使他们对该主题变得敏感，或者产生相反的效果并减弱对它的敏感性，从而降低处理的效果。因此，多次测试的受试者的结果可能不能推广到没有进行多次测试的受试者。

这并不是对外部有效性所有威胁的完整列表。其他威胁可能包括信息变异和信息重复的缺乏，这涉及为每个处理水平的刺激提供超过一个版本给受试者。通过仅使用一个刺激，我们只能得出一个结论，即确切的刺激产生了效果，而不能得出所有该类型刺激都具有效果的结论。这将在第9章中详细讨论。

**内部有效性**

迄今为止讨论的关于泛化性和生态效度的所有问题，如果由于内部有效性问题导致实验结果不准确，那么这些问题都没有意义。Campbell和Stanley称之为“基本最低”要求。

对于任何实验，内部有效性都是真实实验的主要优势；没有其他方法能够给予研究人员如此高度的内部控制，即控制条件之间的因果机制。内部有效性涉及因果关系的核心问题，即实验所发现的原因实际上对结果负责，而不是其他因素。在外部有效性和内部有效性之间的权衡中，实验者通常会选择内部有效性。

**三个基本标准**

第一章中介绍的实验的三个基本标准直接涉及内部有效性，简要回顾如下：（1）原因在结果之前，（2）原因和结果之间存在关系，（3）对于结果没有其他合理的替代解释。前两个标准几乎不会在实验室实验中讨论；哪个先发生是田野实验和自然实验的关注点。例如，失业是否导致通货膨胀，还是通货膨胀导致失业，就像是鸡生蛋还是蛋生鸡的问题。如果你不能明确说哪个先发生，那就没有内部有效性。这主要是Campbell和Stanley的三个前实验设计（准实验或自然实验）的关注点。但是在实验室或真实实验中，研究人员应确保受试者在测量结果之前接受处理或操作。内部有效性的第二个标准是原因和结果之间存在关系，如第1章关于鸟儿带婴儿的例子所示。除了也许在1930年代生活在丹麦的人之外，鸟儿和婴儿之间的关系显然是不合逻辑的，也没有任何理论支持。这也是为什么实验者不会测试直觉，而必须首先进行文献综述，并建立任何提出的原因和效果之间的理论和实证联系。在第3章关于理论和文献的详细任务中执行这项任务，应该能够迅速消除研究人员对因果关系的错误观念。相反，第三个标准占据了实验者大部分的时间：确保没有混杂因素或合理的替代解释可以解释效果。在实验中实现内部有效性的关键在于对所有可能的混杂因素和合理的替代解释进行严格管理，减少偏差和误差，并建立强大的控制措施。这是实验者在内部有效性相关问题上面临的关键任务。

虽然在实验研究中，混杂因素和其他合理的解释可能有无限多种，但遗憾的是，没有一个全面的列表。然而，审稿人在指出它们方面是专家。关键是要提前识别它们，这可以通过彻底审查文献以及运用自己的专业知识和常识来实现。对于关于照片对道德判断影响的研究，可能存在的一个混杂因素的例子是照片中的人是近距离拍摄还是远距离拍摄。对文献进行全面审查的结果得到了

在非语言行为的亲密行为研究中，有证据显示，与他人保持亲密距离会使人们相信自己比与他人保持较远距离时更有说服力、可信度和社交能力。如果这些研究没有通过使用拍摄距离相同的照片来控制亲近距离，道德判断的提升可能是由于照片中的人与被操作的内容有关，而不是由于距离的远近。另一个例子是控制消息长度的实验；我们知道给予人们更多的信息会影响他们的知识水平、态度和其他结果变量，因此，将被操作的消息长度尽量保持相似以控制这种混杂因素是标准做法。

**随机分配**

随机分配的黄金标准也通过尽量使比较组相似来帮助内部有效性的实现。因此，一个组与另一个组在本质上并没有任何不同的因素导致了一个处理对一个组起作用而对另一个组不起作用。因此，我们可以认为是处理引起了结果。随机分配的好处不可言喻，这将在第7章中单独讨论。

**混杂因素的审查**

当可能的混杂因素在研究进行后才被提出时，这为扩展研究的复制提供了素材，正如前面在外部有效性部分所描述的那样。这种概念复制主要是为了外部有效性的目的，同时也通过排除混杂变量作为合理解释来帮助内部有效性的实现。当确定了可能的混杂因素时，有时可以通过对研究人员已经拥有的数据进行额外分析来排除合理的替代解释。例如，有一次，一位审稿人对宗教信仰的测量方式提出了质疑，因为它由三个问题组成。审稿人认为其中一个问题并不能真正衡量概念，并提出它可能对结果产生了混淆。为了确定这个担忧是否严重，作者重新分析了数据，删除了有问题的问题，只使用了一个包含两个项目的宗教信仰测量。结果是相同的。这使审稿人满意，并且文章被发表了。另一次，一位审稿人担心受试者只是试图支持他们现有的道德决定，而不是认真思考其他观点以作出慎重的判断，正如研究所提出的那样。通过回顾受试者给出的开放式回答，这些担忧得到了解决，证明他们确实考虑了与自己相反的观点，通过计算支持自身观点和反事实论证的数量，并举例说明。有关最终发表的解释，请参阅第5.5框“如何做”。

**如何做到这一点 5.5**

解释实验效度和混淆因素

在这项研究的审查过程中，审稿人指出结果可能被不同图像的暴力程度、它们在观众中引发的情绪或所描绘犯罪的严重性等变量混淆。在回应中，作者引用了实验的受控性质和先前的经验证据。以下是文章中的部分内容：

“因为这是一项受控实验，我们可以确定这种效应是由唯一的改变因素——视觉形式引起的。我们可以确定，情绪、暴力图像、犯罪的严重性或其他变量不会产生影响，因为这些因素对所有被试都是相同的——所有被试都观看了三个故事。视频是一样的，唯一的区别在于被试观看视频的次数，或者是否看到了直接从每个视频中截取的静止图像。”99

随后，文章再次进行了讨论：

“我们没有测量受众对这些故事的情绪反应，因为情绪在道德判断中反复被证明不是一个因果因素（Blanchette、Gavigan和Johnston，2014年；Hauser，2006年）。”

“责任归因、犯罪严重性和其他因素应该在未来关注与问题特殊性有关的研究中进行考察；然而，我们再次指出，这些或其他故事之间的差异不会影响我们的主要发现，即静止照片相比一次性视频能够提高道德判断水平。我们可以确信，静止照片导致了比一次性视频更高水平的道德判断，因为这是唯一发生变化的因素。所有条件下的被试都接收到了所有三个故事，因此故事本身引起的任何差异都得到了控制，因为不同条件下的所有被试都收到了完全相同的故事。”100

**Meader, Aimee, Lewis Knight, Renita Coleman, and Lee Wilkins. 2015. "数字时代的伦理：移动图像和照片对道德判断的影响比较。" 《媒体伦理学杂志》30（4）：234–251。**

关于内部效度的威胁最广泛的列表在坎贝尔和斯坦利的小册子中进行了详细概述，并包括与真实实验的内部效度相关的七个不同类别的外部变量。ii以下简要描述了它们，以及一些实际应用。

历史

这指的是在给予被试接受处理和测量结果之间发生的事件。例如，如果研究人员正在测试旨在减少对飞行恐惧的干预措施，并且在研究期间发生了飞机坠毁，参与者可能对飞行更加焦虑；干预措施可能是有效的，但飞机坠毁事件会产生影响。The most extensive list of threats to internal validity is outlined in Campbell and Stanley’s little book and includes seven different classes of extraneous variables relevant to internal validity in true experiments.ii They are briefly described here, along with some practical applications.

ii这里还有第八个威胁，与准实验设计相关。由于本书的重点是真实实验，本章将回顾与真实实验相关的七个威胁。

威胁实验效度。当考虑到长期效果并且需要多次测量结果时，历史也成为一个关注点。例如，新形成的态度往往随着时间的推移而减少，在某些情况下，时间间隔实际上会增加说服效果。对于对持久性说服和态度形成感兴趣的研究，重要的是在延迟后测量效果，而在这种情况下，研究人员无法控制的事件可能威胁到结果的有效性。时间间隔越长，威胁越大。

在其他情况下，治疗之前发生的事件可能会影响结果。例如，如果在实验之前，一个激进分子袭击了一个夜总会，那么研究操纵新闻报道中如何呈现穆斯林的实验可能会产生不同的结果。态度的改变可能是由于恐怖袭击而不是治疗措施。

通过前测-后测对照组设计可以控制由于历史引起的威胁，因为可能引起治疗组变化的任何事件也会在对照组中产生相同的变化。所有组应该同时运行，以确保对一个组的任何影响也同样影响其他组。此外，同时进行的实验有助于控制由于一天中的时间、一周中的日期、季节和其他变化引起的差异。在许多社会科学领域，研究人员可以通过暴露受试者于操纵因素后立即测量其反应来控制历史的威胁，例如向新闻记者展示照片并立即要求他们解决一个道德困境；这与新闻工作室在紧迫的截稿时间上的正常流程相符。没有延迟发生，因此历史不会对内部效度构成威胁。

**成熟**

这也与时间的推移有关，但更具体地关注受试者随时间变化的情况，例如年龄增长、智力增强、饥饿、疲劳或厌倦等。有效教学方法的实验可以在几周、几个月、几个学期甚至一年内展开。在此期间，受试者可能会发生生物和心理变化，这可能是结果产生的原因，而不是治疗措施。即使一项研究需要太长时间才能完成，也可能导致受试者变得厌倦或疲劳，从而改变他们的反应。Campbell和Stanley举例说明学习的累积效应和经验压力是改善结果的原因，而不是教育干预措施。

**测试**

这指的是之前讨论的前测的问题，前测不仅对外部效度构成威胁，也对内部效度构成威胁。当受试者经过多次测试时，前测可能会对结果产生影响。

测试的频率，无论是更好还是更差，都会因为各种原因而发生变化，例如他们学会了如何应对测试或感到厌倦等。人们在第二次进行人格、成就和智力测试时通常会表现更好，105但在刻板印象测试中显示出更多偏见。106所有这些都可以在没有任何干预的情况下发生。避免使用前测，除非必要，是规避这种威胁的一种方式。当需要进行前测时，使用相互替代但等价的测试形式可以缓解这种担忧。107减少测试威胁的另一种方式是使用非反应性或隐蔽性测量，即受试者不知道他们正在接受研究、对被测量的内容毫无察觉，或者无法改变他们的反应，例如通过测量汗水、心率或通过眼动跟踪设备或网页屏幕录制软件测量注视时间的方式。当受试者意识到自己正在被研究并能猜测假设时，需求特征和社会期望性会发挥作用。隐蔽性测量也可以设计成这样，使得虽然受试者知道自己正在接受研究，但他们不知道“正确”的答案或研究的具体假设。例如，在道德判断研究中，受试者知道他们正在进行一项“道德测试”，但除非他们还了解科尔伯格的道德发展理论，否则他们不知道“正确”的答案。受试者认为重要的问题不是他们会采取什么行动，而是他们选择了多少代表较高层次道德判断的陈述。在第10章中，将给出更多隐蔽性测量的例子。观察受试者可能是最常用的隐蔽性测量方法，例如记录受试者在接受某种治疗后在等候室读的杂志。 Festinger和Carlsmith的109认知失调度量指的是受试者是否试图说服他人实验有趣、有意义和值得，或者告诉真相说实验无聊和毫无价值。另一个观察测量可以是受试者在接受有关环境的某些信息后是否将塑料杯放入回收箱中。

**工具设备问题**

这种威胁涉及到实际仪器的问题，例如测量受试者皮肤上的汗水的导电皮肤反应仪，常用于指示心理或生理激活。当机器出现故障或需要重新校准时，这会威胁内部效度，因为它可能会导致记录的分数发生变化。另一种工具设备问题是人为因素，当人们用于观察行为、评分或编码时，他们的判断可能会有所不同，标准也可能会发生变化。他们还可能会随着实践而变得更好或随着时间的推移而感到疲劳。 Campbell和Stanley110建议随机分配观察者，并使用每个观察者在不同条件下对同一行为进行评分，以减少工具设备的威胁。

双盲实验和双重标记是一种应对这种威胁的方法，即不让受试者或观察者知道谁接受了治疗，以避免观察者的评分产生偏见。还有一些技术可以确定不同评分者之间的标准是否不同，或者评分标准是否随着评分时间的推移而发生了变化，这被称为评定者一致性，详细解释在第10章中。如果观察到的结果可以以隐蔽方式记录，例如使用摄像机或网络跟踪软件，那么可以评估两个评分者在测量方面的相似程度。

**统计回归**

统计回归，也称为回归至平均，表示当极端得分倾向于回归到群体平均值时，会对结果产生威胁，例如，最聪明的学生似乎变得更笨，而最迟钝的学生似乎变得更聪明。这种情况发生在最高和最低得分的情况下，因此建议研究人员不要选择因为在某些方面特别优秀或糟糕的受试者进行研究；在再次测试时，他们很可能更加平均。因此，例如，只使用考试分数最差的学生进行教学干预很可能会显示出干预的伪效应而不是真正的效应。当选择代表整个范围的受试者时，可以纠正回归至平均的效应。

**选择**

通过随机分配来控制导致偏差的受试者选择；任何差异都应平均分布在各组之间。当随机分配不可行时，例如当学生自主选择参加某个班级时，研究将变为准实验研究，读者应参考关于此主题的众多优秀书籍以获取减少这种威胁的建议。

**缺损**

也被称为流失，这是指受试者退出研究或未回答所有问题的情况。当缺损率在治疗组和比较组之间不同时，可能会损害效度。当存在系统性的缺损时，例如如果治疗组的受试者退出比控制组更多，那么任何效应可能是由于组别差异而不是治疗导致的。即使在研究之前两组是相等的，退出的受试者也可能使它们变得不相等。例如，受试者可能因为认为教育干预对他们无效而退出，或者因为认为锻炼计划太难而退出。对于缺损数据的后续分析可以评估缺损是否会对结果产生影响，有时也可以使用补偿方法来纠正缺损的影响。

• 设计的实验不够现实

• 使用的受试者与研究者打算进行逻辑推断的对象不够相似

• 仅仅复制他人的研究结果，而没有在理论上进行扩展

• 未能控制混淆变量

**常见错误**

组别平等将揭示是否存在这个问题。缩短研究的持续时间可以帮助缓解缺损的担忧。

总结一下，控制内部效度最有效的方法是使用随机化和带有对照组的设计。随机化不仅包括将受试者随机分配到条件中，还包括尽可能随机分配其他元素，例如随机安排实验会话的顺序或实验员在哪个会话中运行，如果所有组不能同时进行测试。这些将在第7章中更详细地讨论。对照组除了没有接受治疗外，其经历的一切与治疗组相同；它们面临着与历史、成熟、回归、测试、测量仪器和缺损相同的威胁。如果研究人员能够控制这些威胁对内部效度的影响，那么实验设计就被认为是严谨的，意味着是可以设计的最准确、精确或有效的设计。

读者也无疑会意识到本章中存在一些矛盾之处，例如，预测测试通过测试效应对内部效度构成威胁，但也是缓解历史威胁的一种方法。与所有关于实验设计的决策一样，研究人员必须权衡利弊，仔细考虑哪些问题对实验设计来说是重要的，一篇论文中应该对权衡进行明确的讨论。

还有其他更具体的内部效度类型，例如构念效度、面向效度、内部一致性和可靠性，这些将在第10章中进行讨论。接下来，本书将转向关于实验设计类型的更具体决策，具体来说，将操纵多少个自变量以及每个受试者将获得多少个自变量。

.

Chapter 5 ■ Internal and External Validity **137**

##### 测试你的知识

1. 实验室实验通常在 效度方面较弱于 效度。

a. 外部, 内部

b. 内部, 外部

c. 生态, 外部

d. 逻辑, 统计

2. 通常，实验室实验使用随机抽样以便具有概括性。

a. 真

b. 假

3. 在没有进行样本与总体拟合的数学计算的情况下，根据常识可以得出合理结论的观点被称为 。

a. 统计推断

b. 逻辑推断

c. 荒谬推断

d. 复制

4. 重复一项研究以测试理论，并在某种理论上有意义的方式进行扩展被称为 。

a. 自我复制

b. 直接复制

c. 外部复制

d. 概念复制

5. 内部有效性的基本标准是 。

a. 因果必须先于效果

b. 因果与效果必须相关

c. 没有其他合理解释可以解释效果

d. 以上皆是

6. 在七类影响内部有效性的外部变量中，以下哪一项描述的是记录观察结果的人在判断上存在差异或改变标准的情况？

a. 选择

b. 淘汰

**138** Designing Experiments for the Social Sciences

c. 平均回归

d. 工具测量

7. 随机分配如何提供内部有效性？

a. 通过使样本之间随机相似

b. 通过从要进行概括的总体中进行抽样

c. 通过确保来自总体的每个个体或单位有相等的被选择机会

d. 通过确保代表性

8. 在威胁内部有效性的七类外部变量中，哪一类描述了最聪明的学生看起来变得更迟钝，而最愚蠢的学生看起来变得更聪明的情况？

a. 选择

b. 退学

c. 平均回归

d. 工具测量

9. 在威胁内部有效性的七类外部变量中，哪一类描述了研究对象退出研究或未回答所有问题的情况？

a. 选择

b. 退学

c. 平均回归

d. 工具测量

10. 在威胁内部有效性的七类外部变量中，哪一类描述了学生选择参加某个班级的情况？

a. 选择

b. 退学

c. 平均回归

d. 工具测量

**答案**

1. a

2. b

3. b

4. d

5. d

6. d

7. a

8. c

9. b

10. a

Chapter 5 ■ Internal and External Validity **139**

1. Continue to develop your experiment by writing about how you will ensure external and internal validity in your study. Write one or two pages on how you will increase external validity by making your study more realistic and generalizable. Use the topics in this chapter as a guide. Write one or two more pages on how you will minimize threats to internal validity, specifically considering the seven threats described in the chapter. Also include a discussion of the trade-offs, as you can rarely have perfect internal and external validity. Which is more important, and what would you give up; why? Imagine reviewers have challenged you on these issues and respond to their concerns.
2. Find three published experiments in your field and write a one-page critique of each on how the authors addressed issues of internal and external validity. What trade-offs did they make? Was it convincing? Are there other arguments you would have used? Are there other threats they did not consider but should have?

**应用练习**

《用于广义因果推断的实验设计》。贝尔蒙特，加利福尼亚州：Wadsworth出版社。您不需要阅读构建效度部分，该部分将在下一章中涉及。

对于实验真实性的一个很好的例子，请阅读Iyengar, Shanto和Donald R. Kinder的《新闻的重要性：电视和美国舆论》。芝加哥：芝加哥大学出版社。这些研究尽可能模拟真实的国家新闻广播。

Stoker, Gerry. 2010. “探索政治科学中实验的潜力：微观基础洞见和政策相关性。” 政治研究 58: 300–319。本文讨论了政治科学中的有效性问题，并主张进行更多的实验，但这些论点适用于所有社会科学领域。

第2章，“统计结论效度和内部效度”，以及第3章的部分内容，“外部效度”。

**建议阅读**

|  |
| --- |
| **注释** |

**140** Designing Experiments for the Social Sciences

1. 约翰·A·科特赖特（John A. Courtright），《理性思考非概率抽样》，《广播与电子媒体杂志》40卷3期（1996年）：414页。

2. D. T.坎贝尔（D. T. Campbell）和J. C.斯坦利（J. C. Stanley），《实验与准实验设计研究》（芝加哥：兰德·麦克纳利出版社，1963年）。

3. E. A.斯图尔特（E. A. Stuart）等，〈当影响评估有意选择样本地点时的外部效度偏误估计〉，发表于《年度效果研究学会会议论文集》（2012年春季，华盛顿特区）。

4. Rebecca B.莫顿（Rebecca B. Morton）和肯尼斯·C.威廉姆斯（Kenneth C. Williams），《实验政治科学与因果关系研究：从自然到实验室》（纽约：剑桥大学出版社，2010年）。

5. Murray Webster和Jane Sell， 《社会科学中的实验室实验》（阿姆斯特丹：爱思唯尔，2007年）。

6. Morton和Williams，《实验政治科学》。

7. Maria Jimenez-Buedo和Francesco Guala，〈实验经济学中的人为性、反应性和需求效应〉，《社会科学哲学》46卷1期（2016年）：3-23页。

8. Renita Coleman和Esther Thorson，〈将犯罪和暴力置于环境背景中的新闻报道的影响：测试报道的公共卫生模型〉，《健康传播学杂志》7卷4期（2002年）：401-426页。

9. Campbell和Stanley，《实验与准实验设计》，17页。

10. Morton和Williams，《实验政治科学》；Campbell和Stanley，《实验与准实验设计》。

11. Morton和Williams，《实验政治科学》，253页。

12. Elizabeth Tipton，〈你的实验有多具有普适性？一种比较实验样本和总体的指数〉，《教育与行为统计学杂志》39卷6期（2014年）：478-501页。

13. Roger Kirk，《实验设计：行为科学的程序》，第4版（千橡市，加利福尼亚州：Sage出版社，2013年）。

14. Mirae Kim和Gregg G. Van Ryzin，〈政府资金对艺术机构捐款的影响：一项调查实验研究〉，《非营利与志愿部门季刊》43卷5期（2014年）：910-925页。

15. Campbell和Stanley，《实验与准实验设计》。

16. Morton和Williams，《实验政治科学》。

17. Campbell和Stanley，《实验与准实验设计》，19页。

18. 同上。

19. Annie Lang，〈在非概率样本的实验数据中使用推论统计的逻辑：受到Cooper、Dupagne、Potter和Sparks的启发〉，《广播与电子媒体杂志》40卷3期（1996年夏季）：422-430页；Courtright，《理性思考非概率》；Michael D. Basil，《在传播研究中使用学生样本》，《广播与电子媒体杂志》40卷3期（1996年夏季）：431-440页。

20. Campbell和Stanley，《实验与准实验设计》。

21. Tipton，《你的实验有多具有普适性？》

22. 例如，参见Elizabeth Tipton，《使用聚类分析的分层抽样：改进实验推广的平衡抽样策略》，《评估评论》37卷（2014年）：109-139页；E. A. Stuart，《因果推断的匹配方法：回顾与展望》，《统计科学》25卷（2010年）：1-21页；Stuart等，《外部效度偏误的估计》。

23. L. V. Hedges和C. A. O'Muircheartaigh，《改进设计实验的推广能力》，（伊凡斯顿，伊利诺伊州：西北大学，2011年）。引自Tipton，2014年。

24. Tipton，《你的实验有多具有普适性？》

25. David Freedman，Robert Pisani和Roger Purves，《统计学》，第4版（纽约：诺顿，2007年），335页。

26. William R. Shadish，Thomas D. Cook和Donald T. Campbell，《实验与准实验设计用于广义因果推断》，（贝尔蒙特，加利福尼亚州：沃兹沃思·森格学习出版社，2002年），342页。

27. 同上，248页。

28. 例如，参见Edmund J. Malesky，Dimitar D. Gueorguiev和Nathan M. Jensen，《垄断货币：越南的外国投资和贿赂，一项调查实验》，《美国政治科学杂志》59卷2期（2015年）：419-439页。

29. Lang，《使用推论统计的逻辑》。

30. Malesky，Gueorguiev和Jensen，《垄断货币》。

31. 同上，426-427页。

32. Shadish，Cook和Campbell，《实验与准实验设计》，373页。

33. 同上，23页。

34. 同上，348页。

35. P. Rossi，M. Lipsey和H. Freeman，《评估：系统方法》，第7版（千橡市，加利福尼亚州：Sage出版社，2004年）。

36. Shadish，Cook和Campbell，《实验与准实验设计》，355页。

37. Eric Kramon，《投票买卖在哪里有效？来自肯尼亚的一项列表实验的证据》，《选举研究》44卷（2016年）：397-408页。

38. 例如，参见Arnstein Øvrum和Elling Bere，《评估自由学校水果：来自挪威一项自然实验的代表性数据结果》，《公共健康营养学》17卷6期（2014年）：1224-1231页。

39. Shadish、Cook和Campbell，《实验与准实验设计》。

40. Kimberly Klaiman、David L. Ortega和Cloé Garnache，《美国消费者食品包装回收的认知障碍：来自选择实验的证据》，《食品控制》73卷（2017年）：291-299页。

41. Shadish、Cook和Campbell，《实验与准实验设计》。

42. Arjan Verschoor、Ben D'Exelle和Borja Perez-Viana，《实验室与现实：实验中观察到的风险选择行为是否反映现实世界中的行为？》，《经济行为与组织杂志》128卷（2016年）：134-148页。

43. 同上，136页。

44. Alese Wooditch、Lincoln B. Sloas和Faye S. Taxman，《关于涉毒缓刑者无缝关怀模式的多边随机区组实验》，《毒品问题杂志》47卷1期（2017年）：50-73页。

45. 例如，参见Malcolm Fairbrother，《地理工程、道德风险和对气候科学的信任：来自英国一项调查实验的证据》，《气候变化》139卷3/4期（2016年）：477-489页；Mogens Jin Pedersen和Justin M. Stritch，《内部管理和感知的管理值得信赖性》，《美国公共行政评论》（2016年）：1-20页。

46. 例如，参见Øvrum和Bere，《评估自由学校水果》；Rosie Campbell和Philip Cowley，《选民的期望：一项关于调查实验中候选人特征反应的研究》，《政治研究》62卷4期（2013年）：745-765页；Tianguang Meng、Jennifer Pan和Ping Yang，《对公民参与的有条件接受度》，《比较政治研究》（2014年）：1-35页；Klaiman、Ortega和Garnache，《感知食品包装的障碍》；M. L. Loureiro和W. J. Umberger，《关于牛肉的选择实验模型：关于食品安全、原产地标签和溯源的相对偏好》，《食品政策》32卷4期（2007年）：496-514页。

47. 例如，参见Claire A. Dunlop和Claudio M. Radaelli，《教授监管谦卑：与学生从业人员进行实验研究》，《政治学》36卷1期（2016年）：79-94页；Jin Huang等，《个体发展账户与低收入残疾成年人的住房权：来自一项随机实验的证据》，《应用社会科学杂志》10卷1期（2016年）：55-66页；Wooditch、Sloas和Taxman，《多侧随机区组实验》。

48. Lang，《使用推论统计的逻辑》。

49. 同上。

50. Courtright，《理性思考非概率性》。

51. Lang，《使用推论统计的逻辑》。

52. 同上。

53. Basil，《学生样本的使用》。

54. Joanne M. Miller，《审议议程设置的中介因素：一种新的实验范式揭示了情绪的作用》，《政治心理学》28卷6期（2007年）：689-717页。

55. Basil，《学生样本的使用》。

56. Shadish、Cook和Campbell，《实验与准实验设计》。

57. B. G. Cook、M. Tankersley和T. J. Landrum，《确定特殊教育的循证实践》，《特殊儿童》75卷（2009年）：365-383页。

58. S. Schmidt，《我们真的要再做一次吗？在社会科学中被忽视的强有力的复制概念》，《综合心理学评论》13卷（2009年）：7页。

59. Matthew C. Makel等，《特殊教育研究的复制：必要但非常罕见》，《补习和特殊教育》37卷3期（2016年）：1-8页；同上。

60. Gary King，《复制，复制》，《政治科学与政治学》28卷4期（1995年）：444-452页。

61. Susan T. Fiske，《如何在21世纪发表严谨的实验》，《实验社会心理学杂志》66卷（2016年）：145-147页。

62. Makel等，《特殊教育研究的复制》。

63. Matthew C. Makel和Jonathan A. Plucker，《事实比新奇更重要：教育科学中的复制》，《教育研究者》43卷（2014年）：304-316页。

64. Matthew C. Makel、Jonathan A. Plucker和B. Hegarty，《心理学研究中的复制：它们真正发生多频繁？》，《心理科学展望》7卷（2012年）：537-542页。

Chapter 5 ■ Internal and External Validity **141**

65. S. A. Haslam和C. McGarty，《100年的确定性？学校心理学、实验方法和科学不确定性的管理》，《英国社会心理学杂志》40卷（2001年）：1-21页。

66. Renita Coleman，《视觉对伦理推理的影响：对新闻从业人员进行道德决策的照片价值是多少？》，《新闻与传播季刊》83卷4期（2006年冬季）：835-850页。我在这里使用了我自己的工作示例，不是因为它们在实验设计方面是卓越的典范，而是因为我知道背后的故事和推理过程，并且可以从第一手的角度加以解释。

67. Renita Coleman，《种族和伦理推理：种族对新闻决策的重要性》，《新闻与传播季刊》80卷2期（2003年）：295-310页；Renita Coleman，《色盲：非洲裔美国人在新闻困境中的伦理推理》，《新闻与传播季刊》88卷2期（2011年夏季）：337-351页；Renita Coleman，《少数族裔记者的道德判断：亚裔美国人、黑人和西班牙裔专业记者的证据》，《大众传媒与社会》14卷5期（2011年9月-10月）：578-599页。

68. Aimee Meader等，《数字时代的伦理：移动图像和照片对道德判断的影响的比较》，《媒体伦理学杂志》30卷4期（2015年）：234-251页。

69. Shadish、Cook和Campbell，《广义因果推理的实验和准实验设计》，342页；Makel等，《特殊教育研究的复制》；Fiske，《如何发表严谨的实验》。

70. Christian S. Crandall和Jeffrey W. Sherman，《关于概念复制在科学进展中的科学优越性》，《实验社会心理学杂志》66卷（2016年）：93-99页。

71. Ibid.；Makel等，《特殊教育研究的复制》；Makel等，《事实比新奇更重要》；Makel等，《心理学研究中的复制》。

72. B. D. McCullough、Kerry Anne McGeary和Teresa D. Harrison，《从JMCB档案中得到的教训》，《货币、信贷和银行》38卷4期（2006年）：1093-1107页。73. S. A. Haslam和C. McGarty，《100年的确定性？学校心理学、实验方法和科学不确定性的管理》，《英国社会心理学杂志》40卷（2001年）：1-21页。

74. Stanley Milgram，《服从权威：一种实验观点》（纽约：Harper＆Row，1974年）。

75. Fiske，《如何发表严谨的实验》，《实验社会心理学杂志》66卷（2016年）：145-147页。

76. Esther Thorson、Robert H. Wicks和Glenn Leshner，《新闻学和大众传播研究中的实验方法论》，《新闻与传播季刊》89卷1期（2012年）：112-124页。

77. Fiske，《如何发表严谨的实验》。

78. Tim Schweizer，《将思维导图引入商业课程：两项实验结果》，《国际技术、知识和社会杂志》7卷1期（2011年）：190-200页。

79. Pazit Ben-Nun Bloom，《政治中的恶心、伤害和道德》，《政治心理学》35卷4期（2014年）：495-513页。

80. Makel等，《特殊教育研究的复制》；Makel等，《事实比新奇更重要》；Makel等，《心理学研究中的复制》。

81. Ibid.

82. Crandall和Sherman，《关于概念复制的科学优越性》。

83. Charles R. Ebersole等，《观察、假设、测试、重复：Luttrell、Petty和Xu（2017）展示出良好的科学》，《实验社会心理学杂志》69卷（2017年）：184-186页。

84. J. P. A. Ioannidis，《为什么大多数发表的研究结果是错误的》，《PLoS医学》2卷（2005年）：696-701页；Paul H. P. Hanel和Katia C. Vione，《学生样本能否准确估计一般公众？》，《PLoS ONE》11卷12期（2016年）：1-10页。

85. Rebecca S. McEntee、Renita Coleman和Carolyn Yaschur，《比较生动的写作和照片对公共关系中道德判断的影响》，《新闻与传播季刊》94卷4期（2017年）：1011-1030页。

86. Fiske，《如何发表严谨的实验》。

87. McCullough、McGeary和Harrison，《从JMCB档案中得到的教训》。

88. Eric M. Camburn等，《与校长进行随机实验的益处、限制和挑战》，《教育行政季刊》52卷2期（2016年）：187-220页；J. A. Maxwell，《教育中的因果解释、定性研究和科学探究》，《教育研究员》33卷2期（2004年）：3-11页。

89. J. K. Maner，《走进野外：野外研究可以增加复制性和实际影响》，《实验社会心理学杂志》66卷（2016年）：100-106页；Fiske，《如何发表严谨的实验》；Camburn等，《益处、限制和挑战》。

90. Campbell和Stanley，《广义和准实验设计》。

91. Kirk，《行为科学的实验设计程序》。

92. Thorson、Wicks和Leshner，《新闻学中的实验方法论》。

93. Campbell和Stanley，《广义和准实验设计》，5页。

94. Cengiz Erisen、Elif Erisen和Binnur Ozkececi-Taner，《政治心理学研究方法》，《土耳其研究》13卷1期（2013年）：13-33页。

95. Gustav Fischer，《鸟类学月报》，第44卷第2期（1936年）。

**142** Designing Experiments for the Social Sciences

96. Coleman：《种族和伦理推理》；《色盲：非洲裔美国人在新闻困境中的伦理推理》；《少数族裔记者的道德判断：亚裔美国人、黑人和西班牙裔专业记者的证据》。

97. Paul Ekman、Wallace V. Friesen和Phoebe Ellsworth，《人类面部表情的情感》（纽约：Pergamon，1972年）；A. Mehrabian：《从沟通者的姿势、方向和距离推断态度》，《咨询与临床心理学杂志》32卷3期（1968年6月）：296-308页；《一些非语言行为的指标和测量》，《行为研究方法与仪器》1卷（1969年）：213-217页；《沉默的信息》（Belmont, CA：Wadsworth，1971年）；《非语言交流》（芝加哥：Aldine/Atherton，1972年）；J. K. Burgoon、T. Birk和M. Pfau，《非语言行为、说服和可信度》，《人类沟通研究》17卷1期（1990年秋季）：140-169页。

98. Thorson, Wicks, and Leshner，《新闻学中的实验方法学》。

99. Meader等，《数字时代的伦理》，245页。

100. 同上，247页。

101. Alice H. Eagly和Shelly Chaiken，《态度心理学》（Fort Worth, TX：Harcourt Brace Jovanovich，1993年）。

102. Campbell和Stanley，《实验与准实验设计》。

103. Coleman，《视觉效果对伦理推理的影响》。

104. Campbell和Stanley，《实验与准实验设计》，8-9页。

105. Anne Anastasi，《差异心理学》，第3版（纽约：Macmillan，1958年）；V. R. Cane和A. W. Heim，《重复测试的效应：III. 进一步实验和总结》，《实验心理学季刊》2卷（1950年）：182-195页；C. Windle，《人格问卷的再测效应》，《教育心理测量学》14卷（1954年）：617-633页。

106. R. E. Rankin和D. T. Campbell，《对黑人和白人实验者的皮肤电反应》，《异常与社会心理学杂志》51卷（1955年）：30-33页。

107. Campbell和Stanley，《实验与准实验设计》。

108. Lawrence Kohlberg，《道德发展哲学：道德阶段和正义理念》（剑桥，MA：Harper & Row，1981年）；《道德发展心理学：道德阶段的本质和有效性》（旧金山：Harper & Row，1984年）。

109. L. Festinger和J. M. Carlsmith，《强制顺从的认知后果》，《异常与社会心理学杂志》58卷2期（1959年）：203-210页。

110. Campbell和Stanley，《实验与准实验设计》。

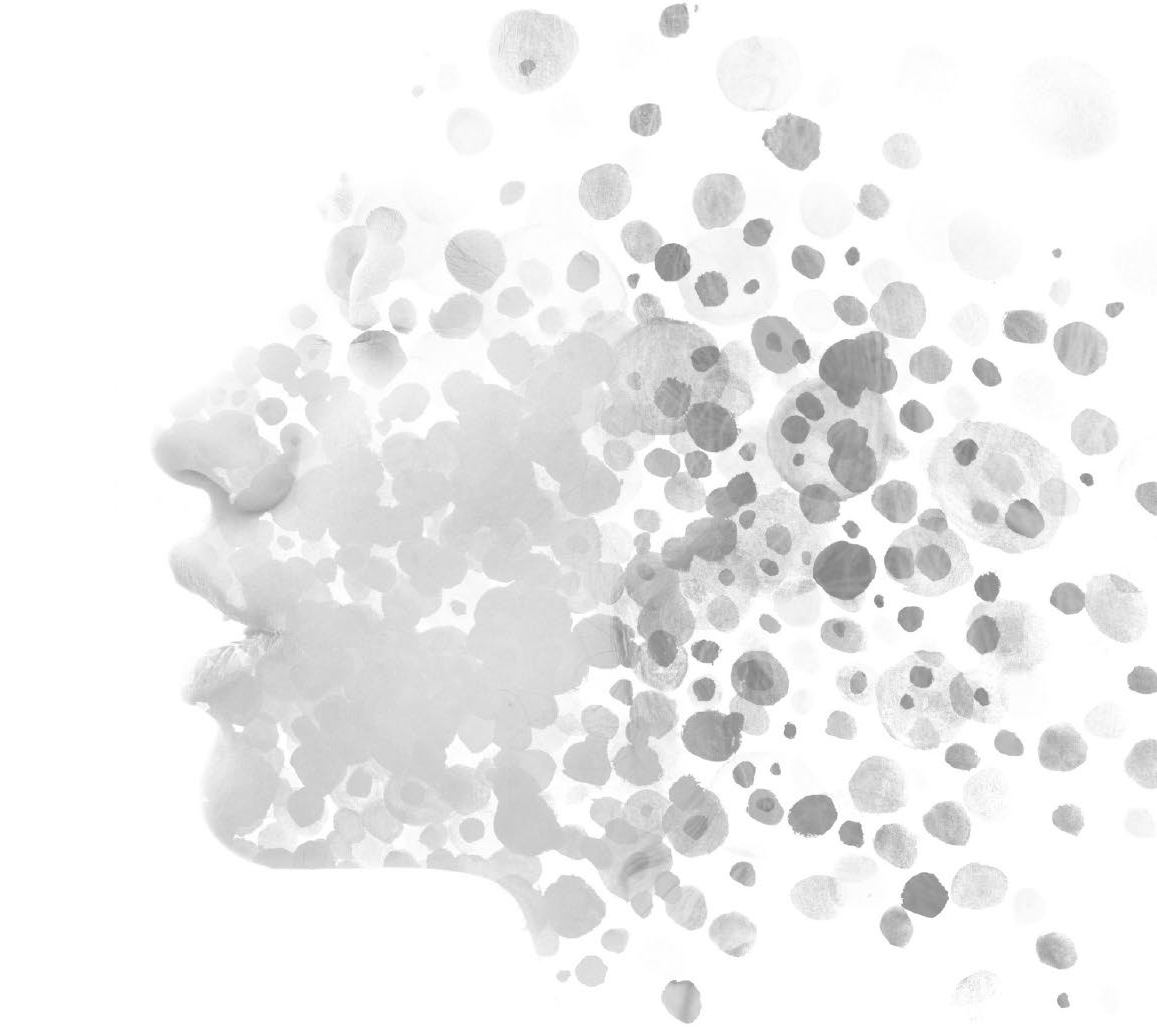
111. 同上。

112. 同上。

113. 同上。

114. 同上。

Chapter 5 ■ Internal and External Validity **143**



For Review-No Commercial Use(2023)

6

# 多因素设计

不必要的复杂性很少能够提升实验研究的质量。

—戴安娜·穆茨

• 确定单因素和多因素实验设计。

• 使用因子符号和设计表绘制实验图表。

• 解释纵向设计、横向设计、混合设计和不完全因素设计之间的差异

• 创建一个实验并解释因素的使用以及如何分配被试者。

• 规划如何实施对照组。

学习目标

在第4章中，本文以概述的方式讨论了不同类型的实验，包括自然实验和现场实验、真实实验和准实验，以及诸如前测-后测对照组设计和所罗门四组设计等分类。本章将聚焦于一种类型的实验——真实或实验室实验，并更详细地探讨其设计中的具体决策。具体而言，本章关注实验中的因素数量、不同设计中的被试者使用方式以及对照组。在阅读本章时，请记住开头引文中的建议，因为事情可能会很快变得复杂起来。

I

**145**

单因素设计

因素是指被操作的自变量，即感兴趣的处理变量，或是在真实实验中有意改变的东西。迄今为止，本书强调了操纵一个自变量的重要性，无论它被假设为会引起某种结果变化或对某个因变量产生影响。当实验中只有一个因素时，它被称为单因素设计或一因素设计。例如，在一项新闻报道的研究中，被操作的因素可能是报道的框架或所使用的角度。在一项人际交流研究中，感兴趣的因素可能是说谎。在一项心理学研究中，研究人员可能对人们的不安全感感兴趣。在一项政治科学实验中，因素可能是政治立场。框架、说谎、政治立场和不安全感是被视为因素的理论概念。

对于任何因素，必须存在两个或更多个水平，即每个因素的不同离散值。在一项新闻框架的研究中，一位研究人员研究了两个水平：客观性框架和倡导性框架。他将倡导性框架定义为只有一面之词的报道，给人一种共识的感觉；而将客观性框架定义为写作客观、两面性的新闻报道，不给出任何反驳的观点。在人际交流研究中，有研究人员使用了因素的两个水平：高或无。他们让被试者写下自己的财务未来（产生高度不安全感），或者写下听音乐这个中性的话题，预计不会引起任何不安全感。在一项政治科学研究中，使用了两个政治立场的水平，将一个立场标记为自由主义，将另一个立场标记为保守主义。

通常也会使用三个甚至四个水平的因素。例如，在一项有关说谎的研究中，研究人员使用了三个水平：遗漏式说谎、具有误导性的真实陈述和真相。超过这个水平的情况比较少见，因为太多水平会使研究变得复杂。如果要研究多个因素，并且每个因素有很多水平，通常会进行另一个实验。

因素设计

单因素实验是最简单的设计之一，也是学习进行实验的好方法。然而，在现实中，大多数实验同时操纵多个自变量，形成了因素设计。因素设计同时考察多个变量的效应，但这些变量之间仍然是独立的。这是一种更有效的假设测试方法，因为可以同时研究两个或多个事物，而不需要进行多个单独的实验。直到罗纳德·费舍尔开始主张多因素实验之前，大多数实验都是单因素设计。

在他的著作《设计实验》中，罗纳德·费希尔（Ronald Fisher）提出了因子实验设计这一术语，认为多因素实验更为经济。10 Campbell和Stanley认为一次只处理一个变量的方法效果不佳，是一种“粗糙的工具”。因子设计的价值在于不仅可以让研究者观察到每个因子独立变化时会发生什么，还能观察到这些因子相互组合时会发生什么。这种能力使我们更真实地了解复杂世界，因为许多结果是由多个因素共同作用产生的。11

例如，当人们在电视上看到公益广告（PSA）时，他们通常会受到广告的多个特征的影响。在一项关于HIV/AIDS的PSA研究中，研究人员考察了其中三个因素，分别是：（1）广告中论证的质量，水平分为强势和弱势论证；（2）HIV/AIDS主题对每个受试者的个人相关性（高/低）；以及（3）两种不同的广告形式，一种使用叙事故事情节，另一种依靠统计数据。结果显示，强势论证在改变人们对避孕套态度和意图方面比弱势论证明显更好，而广告形式并不重要；叙事和统计形式的效果大致相同。此外，那些认为HIV/AIDS与自己更相关的受试者更有意愿使用避孕套。这些发现是主效应，即一个因子的不同水平单独导致结果变化的能力。因此，在前面的例子中，强势论证在改变受试者的态度和意图方面产生了效应，而弱势论证没有产生效应；因此，论证力度是一个因素的主效应。个人相关性也有主效应，高相关性比低相关性更为有效。但广告形式没有主效应；无论广告是叙事还是统计形式，对结果的影响是相同的。

这个例子中的研究使用了三个因素，也被称为三因素设计。具有两个因素的实验被称为二因素设计，是因子实验设计的最简版本，每个因素有两个水平。根据因素的数量，实验的术语也很直观，例如四因素设计等。然而，四因素设计及以上的实验较少见，更复杂且难以分析和解释，因此不常见。

**主效应和交互作用**

因子设计有几种可能的结果：每个因素单独的主效应，前面已经描述过的；以及两个或多个因素的交互作用效应。交互作用发生在一个因素对结果变量的影响取决于另一个因素水平的不同。换句话说，一个因素的效应在不同的另一个因素水平下会发生变化。交互作用揭示了多个因素联合作用的效果与仅考虑每个因素独立效应所预期的效果之间的差异。

为了理解交互作用，使用因子符号和设计表进行可视化是有帮助的。因子符号表示每个因素的水平以及它们如何组合。例如，2x2的因子设计表示一个具有两个因素的二因素设计，每个因素有两个水平。第一个因素的水平标记为A1和A2，第二个因素的水平标记为B1和B2。不同水平的组合形成四个条件：A1B1、A1B2、A2B1和A2B2。

设计表用于组织和展示因子设计中收集的数据。每一行代表一个条件或因素水平的组合，列代表测量的结果或因变量。设计表允许研究人员比较每个条件的均值或效应，并确定任何主效应或交互作用。

在实验研究中，控制组是非常重要的，因为

当一个因素的效应因另一个因素而改变时，这被称为交互作用效应。换句话说，两个因素联合作用的效应不能从单独考虑每个因素的效应中预测出来。交互作用是因素之间的关系，而不是因素水平之间的关系，就像主效应一样。一个简单的例子是饮食和锻炼是影响体重的两个因素。如果仅改变饮食会影响体重，那就是一个主效应；如果仅改变锻炼计划会影响体重，那也是一个主效应。当饮食和锻炼同时改变时，可能会出现交互作用，例如，只有在同时采用低脂饮食和每周多锻炼三天时才能减肥。在前面描述的HIV/AIDS信息研究中，强势论证和个人相关性的低水平同时产生了效应。要报告交互作用，首先要说明哪些因素之间存在交互作用，然后继续描述产生效应的每个因素的水平。例如，HIV/AIDS研究可能会说：“论证力度和个人相关性这两个因素之间存在交互作用。对于那些对HIV/AIDS关注度较低的受试者，强势论证显著地更能鼓励他们使用避孕套。”这个研究展示了在一个实验中研究三个因素的好处；如果研究人员分别在一项研究中考察论证力度，而在另一项研究中考察相关性，他们就无法得知这种交互作用。

因子设计可能会产生不同的结果组合：

- 对于所有因素而言，主效应显著，但没有交互作用效应。

- 对于某些因素而言，主效应显著，但不是所有因素，同时可能存在显著或非显著的交互作用。

- 存在交互作用效应，但没有主效应。

Campbell和Stanley举了一个假设研究的例子，其中一个因素是三种不同的教师个性类型，另一个因素是三种不同的教学方法。如果所有三种个性和方法的效果相等，那么它们之间将没有主效应；也就是说，在每个因素内，不同水平的因素都没有明显优于其他水平。但是，教学方法和教师个性可能会发生交互作用，以至于对于 spontaneous教师来说，小组讨论比教学辅导更有效。

在许多研究中，交互作用效应通常是研究人员感兴趣的重点。例如，在健康传播研究中，研究人员指出，信息的表述方式

不同的表述方式的效果很小，19指出这可能是因为需要考虑人们对风险的评估以及表述方式的影响。该研究专门设计来研究表述方式和风险评估之间的交互作用，并对两个因素的主效应不太感兴趣，主要是因为主效应已经被了解。20

在进行因子设计时，应该对每个因素独立预测假设（即主效应），并对可能的交互作用提出假设或研究问题（RQ）。例如，一个具有两个因素的研究，如教学方法和教师个性研究，将有两个预测主效应的假设——一个是教师个性的主效应，一个是教学方法的主效应。还会有第三个假设或RQ，涉及两个因素之间的交互作用——教师个性和教学方法之间的交互作用。如果没有足够的证据来预测可能的交互作用效应，研究人员通常会提出一个RQ；以下是一个用叙述方式而不是正式RQ表述的例子。这是一项研究，研究不同类型的警示是否可以阻止人们购买与他们的电子设备（如手机）不兼容的配件，警示信息在该领域被称为“助推”：

该研究还测试了社会人口学变量（年龄和教育）与实验处理（警示信息...）之间的交互作用效应。

换句话说，如果年龄和教育对购买不兼容产品有影响，这种影响在所有实验条件下是否相同？或者在某种给定类型的助推存在时，这种影响是否增加（或减少）？由于很难预测这些交互作用效应的方向和幅度，因此该调查的这部分是探索性的，不受特定假设的引导。21

或者，可以用正式的RQ来表述：

RQ：社会人口学（年龄和教育）与警示信息风格（传统的、感情化的、无警示）之间是否存在交互作用？

在另一个例子中，利用关于表述健康信息和人们对患上人乳头瘤病毒（HPV）的风险水平的研究，研究人员主要关注风险和表述方式之间的关系，因此提出了两个只预测交互作用的假设；没有对主效应进行预测。22他们还有足够的证据来预测交互作用效应的方向。斜体字（我的）表示预测的每个因素的交互水平。

“H1：当父母认为他们的孩子患上HPV感染的风险较低时，使用损失框架（相对于获利框架）的信息将导致更强烈的意愿为他们的孩子免费接种疫苗和付费接种疫苗。”

“H2：当父母认为他们的孩子患上HPV感染的风险较高时，使用获利框架（相对于损失框架）的信息将导致更强烈的意愿为他们的孩子免费接种疫苗和付费接种疫苗。”23

换句话说，当低风险的人看到损失框架的信息时，预计他们会更有意愿为孩子接种疫苗；而高风险的人看到获利框架的信息时，预计他们会更有意愿为孩子接种疫苗。

我建议不要陷入太多因素或每个因素的太多水平之中，因为实验可能会变得非常复杂，难以在多个方面进行控制。例如，交互作用的数量会随着每个因素的增加而呈指数增长。在一个具有两个因素的实验中，存在一种可能的交互作用（因素1与因素2）。但是如果增加第三个因素，就可能出现四种交互作用（因素1与因素2；因素1与因素3；因素2与因素3；以及所有三个因素之间的交互作用，即因素1、2和3的交互作用）。具有十个因素和两个水平的实验会产生210 = 1,024种组合！而且，在存在两个以上因素的情况下，解释交互作用的意义可能会变得复杂。24最好通过控制其他所有可能影响结果和操作的变量来保持不同研究的一致性。

除了交互作用的数量之外，具有多个因素和每个因素多个水平的设计还需要创建更多刺激，这些刺激是处理方式传达的方式，例如通过信息、广告或教学技术。在一项政治传播研究中，研究人员不得不撰写256个新闻稿，因为该研究使用了五个因素，每个因素有两到四个水平。25幸运的是，新闻稿很简短。

不仅刺激物的数量是一个问题，所需受试者的数量也会增加，这将在后面的章节中介绍。由于交互作用比主效应更难检测，因此可能需要更大的样本量。26如果每个受试者都接受两个因素的影响，因子设计可能需要比单独的单因素研究更少的受试者，27但是随着需要进行多个统计检验，要找到显著性就更加困难，除非增加更多的受试者。28简而言之，最好设计几个连续的实验，而不是试图在一项研究中涉及过多内容。两个、三个或四个因素是最常用的因子设计。29

**阶乘符号**

当使用阶乘设计时，通常会使用阶乘符号来表示，例如2 x 2设计、3 x 3设计或2 x 2 x 3设计，其中的“x”读作“乘”，就像“two by two”一样。每个数字代表一个因素，因此数字的数量表示因素的数量。

A screenshot of a computer

Description automatically generated

数字表示因素的数量。每个数字的值指的是该因素的水平数。一个 2 x 2 表示有两个因素，每个因素有两个水平。一个 3 x 3 表示有两个因素，因为只有两个数字。每个因素都有三个水平。一个 2 x 2 x 3 表示有三个因素，因为有三个数字，其中第一和第二个因素各有两个水平，而第三个因素有三个水平。（参见更多相关内容6.1节，阶乘表示法。）

关于因素表示顺序的规则并没有固定的规定；一些研究者将具有更多水平的因素放在前面，而其他人则将最重要的因素放在前面。在实验的方法部分开始处，应使用这种表示法描述设计，通常将因素和水平的名称放在括号中，如下所示：

• “我们采用了一个 2（论据质量：强/弱）x 2（个人相关性：高/低）x 2（证据：叙事/统计）的混合因素设计。”30

• “第二项研究采用了一个 2（危机阶段：危机期/危机后）x 2（视觉非语言线索：无力/有力）的被试间因素设计。”31

单因素设计可以这样报告：

• “第一项研究采用了单因素设计，以研究声音音调的提高和降低对组织发言人的认知影响。”32 在这个例子中，单因素“声音音调”的两个水平是提高和降低。

有关如何报告因素设计的更多示例，请参见“How To Do It”框6.2。

**HOW TO DO IT 6.2**

Reporting Factorial Design

方法部分描述因素设计、每个因素的水平以及对照组的部分应该位于方法部分的开头，并且通常非常简洁，可以是一到两个句子或一个段落。应包括以下所有内容：

• 是单因素设计还是因素设计

• 使用阶乘表示法表示的因素数量和每个因素的水平。如果是不完全阶乘设计，应在此处进行说明。

• 紧接着表示法之后，用括号内简要描述每个水平

• 受试者分配方式的描述——被试间设计、被试内设计或混合设计

• 是否有对照组，如果有的话，是什么对照组

• 还要注意随机分配的使用，这是下一章的主题。

并不是所有的研究都按照这种精确的方法进行描述，有些研究以更加叙述性的方式描述这些特征。然而，按照上述方式描述是比较安全的做法。

接下来是描述设计的示例，为了确保包含了前面提到的所有要点，这些示例是根据真实研究进行了调整。

单因素设计1

这是一个单因素的被试间实验，反映了先前审查的内部管理的五个水平（目标、承诺、参与、反馈、奖励）。每个被试被随机分配到五个处理组或一个对照条件中的一个。处理组中的被试接收到与对照组完全相同的信息。此外，处理组的被试还收到了一行文字，描述了管理活动。

改编自：Pedersen, Mogens Jin, 和 Justin M. Stritch. 2016. “Internal Management and Perceived Managerial Trustworthiness: Evidence From a Survey Experiment.” The American Review of Public Administration 48 (1): 67–81. https://doi.org/10.1177%2F0275074016657179.

单因素设计2

我们使用了一个单因素的被试间实验，被试被随机分配到两个层次的公民参与建议来源处理组（正式机构/互联网）或一个对照组。对照组接收到与处理组完全相同的信息，只有一项关于正式机构或互联网居民建议的处理项例外。

改编自：Meng, Tianguang, Jennifer Pan, 和 Ping Yang. 2014. “Conditional Receptivity to Citizen Participation: Evidence From a Survey Experiment in China.” Comparative Political Studies 50 (4): 399–433. https://doi.org/10.1177%2F0010414014556212.

被试间因素设计1

这个实验采用了一个3（提醒频率：无提醒、每月提醒、每周提醒）x 2（捐赠方式：定期捐赠或一次性捐赠）的因素设计，共得到六个处理组。被试被随机分配到其中的一个处理组。

（注：该研究没有指定对照组，但无提醒条件代表了该因素的基线。）

改编自：Sonntag, Axel, 和 Daniel John Zizzo. 2015. “On Reminder Effects, Drop-Outs and Dominance: Evidence from an Online Experiment on Charitable Giving.” PLoS ONE 10, no. 8 (August 7): 1–17.

被试间因素设计2

这个研究是一个2 x 3的被试间设计，第一个因素是学业成就水平（高vs低），被试被随机分配到只接受有效实验设计指导（有效）或混合混淆和有效设计指导（无效），以及仅在后测之后进行教学的处理组（对照条件）。

改编自：Lorch, Robert F. Jr., Elizabeth P. Lorch, Benjamin Dunhan Freer, Emily E. Dunlap, Emily C. Hodell, 和 William J. Calderhead. 2014. “Using Valid and Invalid Experimental Designs to Teach the Control of Variables Strategy in Higher and Lower Achieving Classrooms.” Journal of Educational Psychology 106, no. 1 (February): 18–35.

（续）

被试内因素设计

有三个被试内因素，每个因素有两个水平。凝视线索因素操纵了凝视线索面孔的凝视方向；在提示条件下，凝视线索面孔朝向目标面孔，而在未提示条件下，凝视线索面孔朝向屏幕空白的一侧。情绪因素是凝视线索面孔的情绪表达（积极或消极）的操作。提示数量因素是单个或多个凝视线索面孔的操纵。单个凝视线索面孔条件下只有一个凝视线索面孔。在多个凝视线索面孔条件下，三个凝视线索面孔都被呈现。

（注：显然，当所有条件都是被试内因素时，被试被分配到所有组，指定他们是“随机分配”的是不必要的。此外，在这种情况下，没有指定对照组，因为面孔不能没有凝视方向。在这项研究中，组之间相互比较，而不是与没有接受处理的对照组进行比较。）

改编自：Landes, Todd Larson, Yoshihisa Kashima, 和 Piers D. L. Howe. 2016. “Investigating the Effect of Gaze Cues and Emotional Expressions on the Affective Evaluations of Unfamiliar Faces.” PLoS ONE 11, no. 9: 1–24.

混合因素设计

这是一个4 x 2 x 2的混合因素设计。第一个被试内因素是问题（卖淫、毒品、虐待老人、家庭暴力）。所有参与者都接收到这四个困境。第二个被试内因素是决策类型（刊登照片或使用匿名消息来源）；所有参与者接收到两个困境，询问是否在有人要求不刊登照片时刊登照片，以及是否在有很多消息来源但没有人愿意使用其姓名时刊登文章。被试间因素是照片条件（有照片/无照片）。被试被随机分配到处理组，在处理组中，他们在所有四个困境中看到了照片；或者被分配到对照组，在对照组中，他们没有看到照片。

改编自：Coleman, Renita. 2006. “The Effect of Visuals on Ethical Reasoning: What’s a Photograph Worth to Journalists Making Moral Decisions?” Journalism and Mass Communication Quarterly 83, no. 4 (Winter): 835–50.

设计表

另一种表示因素设计的方法是使用设计表。这些表格用于帮助实验计划，并且在期刊文章中很少被报道。表格还用于确定刺激的特征和每个条件下的被试数量，由“单元格”或个别方框表示。表6.2代表了一个假设的2 x 2因素设计，其中一个因素是照片中的人是否微笑，另一个因素是相机的距离，即特写或远景（long shot）。（依赖变量并没有在这里提及。

A screen shot of a computer screen

Description automatically generated

在设计表中，可以展示出诸如在看到照片后，关心和同情心强的受试者如何评价该人的情况。行和列代表因素；单元格代表四个不同条件之一，并在文章的文本中描述为：微笑/特写；微笑/远景；不微笑/特写；不微笑/远景。单元格内部是每个条件下的被试数量，用符号n表示。大写字母N代表总被试数量，即所有单元格n的总和。如何确定被试数量在第8章中有详细介绍。条件数量由自变量水平的组合形成，通过乘以各自水平的数量来计算。在2 x 2的设计中，共有4个条件；在3 x 2的设计中，共有6个条件。

设计表的单元格可以包括其他信息，例如在研究进行后，各组的边际均值。设计表可以根据需要添加列或行来包括两个因素的任意多个水平，如表6.2所示。然而，它只能适用于具有两个因素的设计，因为它显然是二维的；对于三个或更多因素，需要使用三维表。它能够清晰地展示随着增加水平数量，被试和刺激的数量是如何增加的。

A screenshot of a computer

Description automatically generated

选择设计

从本质上讲，并没有什么使得因素设计比单因素设计更好或更具理论性。然而，如果研究主题可能同时受到两个或更多因素的影响，并且这些因素可能相互作用，那么因素设计是更好的选择。因素设计在期刊中比单因素设计更常见，并且由于经济性更高以及能够测试交互作用，它们具有更强的优势。Bausell指出，随着处理组数量的增加，需要更多的被试，而在进行多个统计测试时达到显著性的概率就会降低。他说：“我总是建议从一个两组设计开始。”然后他继续说，如果有足够的资源，多于两个组可以通过深入研究因果因素来增加理论。

研究人员根据证据和理论对使用哪种类型的设计进行决策，这些设计是基于依赖变量所使用的概念的证据和理论预测。测试两个相互竞争的理论或相互矛盾的预测的实验需要因素设计。例如，在一项关于人们如何通过媒体了解外国国家的研究中，探索了两种关于这种影响机制的竞争性理念。逻辑也可以决定研究是单因素设计还是因素设计。例如，一篇论文研究了在当前媒体环境中工作的新闻记者面临的一个严重问题：是快速报道但存在一些不准确性，还是准确性但落后于所有其他新闻机构。速度与准确性的问题适合使用两个水平的单因素设计：快速但有些不准确，或者慢但准确。之所以没有将速度分为快速和慢速，将准确性分为准确和不准确的单独水平，是因为快速和准确或慢速和不准确的组合在现实世界中并不现实。快速和准确显然是理想的；如果新闻记者总是能够产生这种组合，就不会有问题。在现实世界中，慢和不准确的组合也是不太可能的。如果新闻记者慢慢地报道一则新闻，就没有理由存在不准确性；他们有足够的时间确保信息正确。在2 x 2因素设计中，将产生的四个条件中的两个根本没有意义。Shadish、Cook和Campbell称之为“测试对理论没有兴趣或不太可能在政策中实施的处理组合是浪费资源”。这说明了决策过程；文章并不总是解释研究人员为什么选择单因素或因素设计。然而，选择应始终基于可以向读者解释的良好实际或理论原因。此外，随着每增加一个因素和水平，被试数量也会增加，从而增加了研究的成本和时间。限制因素的天文费用是一个合理的限制因素。在选择设计时，还应考虑实际因素，如被试数量、费用、研究持续时间和退出率等。因为实验的目的是设计实验条件来控制自变量，这些因素在实验条件之间的比较上起着关键作用。因素设计的一个关键优势是能够检测自变量之间的交互作用，从而深入了解因素对依赖变量的影响方式。

回答理论提出的非常具体的问题时，狭义的研究比包含“杂乱无章”的研究更能有效地发展理论。通常的说法是，包含X、Y和Z概念“超出了本研究的范围”。

## STUDY SPOTLIGHT 6.3

Example of a Single-Factor and Factorial Design in One Study

这项组织沟通领域的研究探究了危机时发言人的非语言表达如何影响公众对公司的感知。研究者进行了两个实验，一个采用了单因素设计，另一个采用了因素设计。

在第一项研究中，单一因素是发言人的声调，研究者将声调调整为高音或低音。被试只被分配到一个条件中，即他们只听到高音或低音之一，而不是同时听到两种声调，因此这是一个两组间设计（between-subjects design）。在第二个实验中，研究者改变了与第一项研究的语言线索不同的非语言沟通形式——视觉线索，并添加了第二个因素——危机阶段。这是一个2 x 2的因素设计，其中第一个因素是危机阶段，两个水平分别是危机期间和危机后。第二个因素是非语言视觉线索，两个水平分别是有权力和无权力。请阅读该研究以了解如何在三种非语言表达中操纵权力。这样就形成了四个处理条件，如下的设计表所示：

被试被随机分配到四个条件中的一个，使得这是一个两组间设计（between-subjects design）。文献综述表明，仅凭视觉线索就已知会对人们对个体能力的感知产生影响，因此这第二个实验并不那么关注两个因素的主效应，而更关注它们之间的交互作用。实验发现了一个显著的交互作用，显示在危机期间，显得有权威会被认为具有能力，但在危机之后，权力的视觉线索对感知没有影响。该研究还测试了对真诚度的感知和其他假设，但在这里为了说明单因素和因素设计的使用，已经进行了简化。

A screenshot of a computer

Description automatically generated

*来源：Claeys, An-Sofie和Verolien Cauberghe。2014。“保持控制：危机时期组织发言人的非语言权力表达的重要性。”《传播学杂志》（Journal of Communication）64：1160-1180。*

### 被试在设计中的使用方式

设计真正实验的另一个方面是确定被试将接受多少个因素水平的影响。接下来描述的内容还涉及需要多少个被试、需要创建多少个刺激以及其他相关问题。

**组间设计**

组间设计可能是最容易理解的设计之一，也是最常用的设计之一。在这种设计中，每个被试被分配接受一种且仅一种类型的处理，代表设计表中的一个单元格。每个被试只属于一个组，而不是多个组。Morton和Williams称之为“处于一个世界状态”。有时也称为“嵌套设计”，因为被试嵌套在条件的水平之内。最简单的组间实验是单因素设计，其中包括两个组：处理组和对照组。一个假设的例子是教育研究中的一个实验组通过自学计算机教程学习数学，而对照组通过教师讲授课堂教学学习数学，即数学的常规教学方式。在这种情况下，组间设计是必要的，以避免交叉效应，即如果被试对接受第二个条件有反应，那是因为他们已经接受了第一个条件。学生只能被分配到计算机教程组或讲座组，因为他们显然会学习到信息，并会影响到处于另一组中的结果。一旦他们学会了数学，就无法“取消学习”。对于每个组需要完全不同的一组人。通过在组间设计中进行随机分配，可以控制可能混淆结果的未测量变量。

当研究者无法操纵条件，例如性别，组间因素也被用于这种情况，因为一个被试只能属于一种性别。例如，在一项研究中，研究者测量了青少年的先前药物使用情况，并将其作为一个因素，尽管这是一个他们无法操纵的因素。年龄、种族、政治意识形态和教育等因素通常作为组间因素在实验中使用。这些无法操纵的因素也被称为“测量因素”。

组间设计的一个缺点是需要的被试数量。如果功效分析显示数学研究中的两个组需要每组四十名被试，那么需要八十个不同的被试。如果添加了第二种类型的处理，比如团体学习，那么还需要另外四十个被试，总共需要一百二十个被试。

而且这只是针对单因素设计而言。对于一个2 x 2的二因素设计，每个单元格需要四十个被试，这意味着总共需要160个被试。而一个2 x 3的设计则需要200个被试，依此类推。设计越复杂，所需的被试数量就越多。

**组内设计**

解决被试数量呈指数增长的一种方法是使用组内设计。如果不存在诸如练习效应等交叉效应的问题，这是最经济的设计类型，因为同一个被试可以参与所有的处理组。以前述的两个水平新闻框架研究为例——客观和倡导，没有理由认为阅读倡导框架的报道会影响对阅读客观框架报道的反应（只要报道是不同的），因此组内设计可以让这40个人同时参与两个组，而组间设计则需要80人，每组40人。Morton和Williams将这称为“同时处于两个世界状态”。有时也称为交叉设计，因为被试与条件相交。增加因素或水平时，也会出现同样的指数增长。例如，在对微笑和相机距离的假设性研究中，2 x 2设计中，每个人可以获得四种类型的照片——一个微笑的近距离照片，一个微笑的远距离照片，一个不微笑的近距离照片和一个不微笑的远距离照片。由于每个人在四个单元格中都有表示，组内设计只需要总共40个被试，而不是组间设计的120个被试。从本质上讲，每个人“计为”四个被试。有时也称为“重复测量”设计。应该使用一种随机轮换刺激呈现顺序的系统来减少潜在的交叉效应；例如，四分之一的被试被随机分配到首先获得微笑的近距离照片，四分之一被分配到首先获得微笑的远距离照片，四分之一被分配到首先获得不微笑的近距离照片，四分之一被分配到首先获得不微笑的远距离照片。这有助于在整个样本中“消除”任何交叉效应，将在第7章中进行讨论。

组内设计的权衡是研究者需要创建更多的刺激。例如，在对客观和倡导框架报道的研究中，研究者不能使用完全相同的报道以不同方式编写。研究者需要编写等效的报道，但又要使其足够不同，以使被试不会认为他们在阅读相同的内容两次。虽然在被试方面可以节省，但在刺激的创建上需要更多时间。

不使用组内设计的主要原因是为了避免交叉效应，即如果被试可能会因为接受第一个条件而对其他条件产生反应。

这个最容易理解的例子是，如果“治疗”实际上是能够治愈一个人的东西。例如，如果一个因素是三种不同的头痛药物——阿司匹林、布洛芬和非甾体抗炎药（NSAIDs），那么研究者只能将头痛患者分配到一个单元或治疗组，因为在服用药物后，他们的头痛可能会得到缓解。换个角度来看，如果一个受试者不能恢复到最初的状态，那么就不适合采用受试者内设计。45

以社会科学的例子为例，有一项研究调查了人们对人类是否具有自由意志的信念对他们的慈善捐赠行为的影响。研究者通过让受试者阅读一篇反对自由意志的科学期刊文章来操作自由意志因素；对照组受试者阅读了同一期刊上关于可持续能源的文章，没有包含自由意志信息。依赖变量是受试者的慈善捐赠金额，研究假设那些阅读反对自由意志文章的受试者捐赠的金额较少。这是一个受试者间设计，因为一旦参与者阅读了关于人类没有自由意志的文章，他们的捐赠决策就会受到影响。他们无法恢复到之前不了解反对人类拥有自由意志的论点的状态。他们已经永久地改变了。换句话说，有些治疗方法是无法回退的。如果研究者将其设计为受试者内设计，将会污染治疗方法并混淆结果。另一个例子来自一项商业研究，其中治疗方法是一家自愿召回有缺陷产品的公司与一家没有召回的公司。受试者阅读关于这些公司的信息，然后对它们的诚信、信任和善意进行评价。受试者只能接触到自愿召回或未召回的情况，而不能同时接受两种情况的治疗，因为接受其中一种情况的治疗会影响另一种情况的评价。

研究者在担心受试者可能“猜到”假设时也会使用受试者间设计。例如，在商业研究中多次阅读关于召回的信息后，受试者可能开始怀疑研究者是否认为公司是否召回产品会影响他们对公司的看法。我们从需求特征（见第四章）中了解到，受试者喜欢让研究者满意，他们可能会更好地评价召回公司，因为他们认为这是研究者希望他们做的。如果只有将受试者分配到一种治疗方法是有意义的，那么就应该采用受试者间设计。不使用受试者内设计的另一个原因是如果进行多次测试需要很长时间，受试者可能会感到疲劳，不能给后续测试提供同样的努力和注意力。然而，如果没有理由相信受试者在接受多种治疗方法时会产生影响，那么受试者内设计的效率更高，比受试者间设计更可取。

受试者内设计的另一个好处是，每个受试者都作为自己的对照。也就是说，无论个体差异如何可能影响他们的反应，对于两种治疗方式来说，个体差异都是相同的。例如，在倡导性与客观性框架研究中，拥有大学学历的人可能会比没有大学学历的人更加批判地思考所有的信息。在这种情况下，接受过大学教育的受试者会接触到倡导性和客观性框架的信息，这样就控制了教育因素，因为每个受试者都通过接收两种类型的信息来作为自己的对照。在受试者间设计中，通过随机分配来控制个体差异，以确保实验组在教育、年龄、性别、种族和其他可能产生差异的特征上是等效的，例如政治意识形态，甚至心理特征，如教条主义或威权主义人格。但是，使用随机分配的组间等效性永远不如在每个组中具有完全相同的受试者等效性好。这将在第7章中更详细地介绍。

混合因子设计

有些研究在同一个实验中同时使用受试者间设计和受试者内设计的组合。如果某个因素由于carry-over效应、疲劳或假设猜测等原因，一个设计适合使用受试者间设计，而另一个设计适合使用受试者内设计，则可以采用混合因子设计。在这种设计中，一个独立变量被设定为受试者间因子，以便受试者只接受该治疗的一个水平，而另一个独立变量被设定为受试者内因子，以便受试者接受该治疗的所有水平。例如，在一项采用混合设计的研究中，受试者阅读了四个新闻记者可能面临的道德困境，并对他们会做出什么决策进行了判断。受试者内因子是他们必须做出的决策类型，其中两个水平是是否应该刊登一张照片以及是否应该刊登具有匿名消息来源的报道。所有受试者都在两个困境中做出了关于刊登照片的决策，并在另外两个困境中做出了关于使用匿名消息来源的决策。受试者间因子是受试者是否在困境中看到照片。对于这个因子，一半的受试者在所有四个困境中都看到了照片，而另一半没有在任何一个困境中看到照片。之所以这样做，是因为预计对于决定是否刊登照片的决策不会对受试者决定是否使用匿名消息来源产生任何影响，反之亦然；然而，有人担心在某些困境中看到照片可能会影响受试者对其他没有看到照片的困境的反应方式。受试者只接收照片因子的一个水平是有意义的；他们要么接收照片，要么不接收照片。在这个受试者间因子中，每个受试者都只有一个水平。

只接受了一种处理方式，而不是两种。而且他们同时接受两种不同类型的决策，使用照片或匿名消息来源。在这个受试者内因子中，每个受试者都同时接受了这两种处理方式。

起初，可能很难记住哪种类型的设计只包含一个组的受试者，哪种类型的设计包含两个组的受试者。可以尝试记忆这个助记短语：“在两者之间做选择；你只能选择一个。”因此，受试者间设计将受试者仅分配到一个组中。

论文应该报告所使用的受试者设计类型——受试者间设计、受试者内设计或混合设计——以及阶乘记法。以下是一些示例，其中受试者设计的类型已用下划线标出：

• “我们采用了一种2（论证质量：强/弱）x 2（个人相关性：高/低）x 2（证据：叙述/统计）的混合因子设计。”

• “第二项研究采用了一种2（危机阶段：危机期/后危机期）x 2（视觉非语言线索：无力/有力）的受试者间阶乘设计。”

有关如何在论文中报告实验设计的更多示例，请参见《如何做》框6.2，报告阶乘设计。

**不完全阶乘设计**

关于不完全阶乘设计，之前所描述的所有设计都是完全交叉的阶乘设计，有时简称为完全阶乘设计。在这些设计中，所有水平和因子的组合都被分配给受试者。但有些设计会留下一个或多个设计表中的单元为空；这些被称为不完全阶乘设计或分数阶乘设计。正如之前讨论的，某些组合可能没有意义。Shadish、Cook和Campbell给出了一个研究设计的例子，该设计没有将受试者分配到八个单元中，因为这些单元在政治上不受欢迎或成本太高，无法在扶贫政策中实施。当没有实际或理论上的原因研究某些单元组合时，将这些单元留空并不分配任何受试者给它们更为经济；然后将其报告为不完全阶乘设计。使用不完全阶乘设计的另一个原因是允许设置一个不接受任何处理的对照组。如果对照组不适合任何单元组合，它将被视为独立的单元。它将呈现为表6.5中的设计。图中的假设研究中，“视觉类型”因子有三个水平，受试者可以看到静态照片、观看一次的电视视频或观看三次的互联网视频。第二个因子代表图像中人物的摄像机距离，可以是特写、中景或远景。对照组没有图片或视频，只有其他受试者得到的故事。

A screenshot of a computer

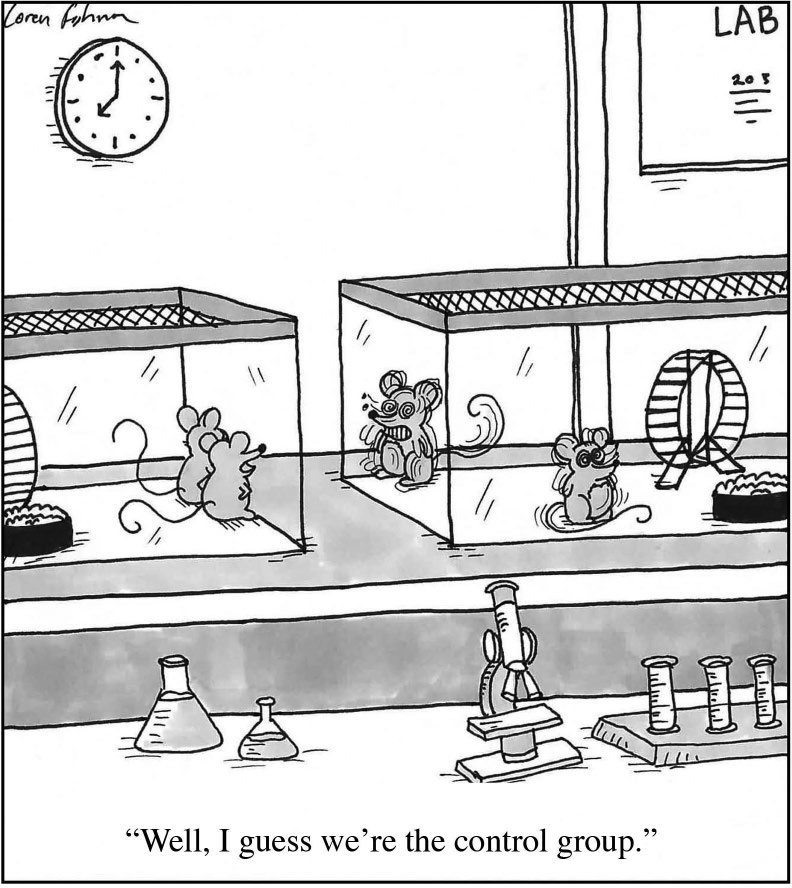
Description automatically generated

他们的图像。在设计表中没有办法“安排”这个因素，因为表中的两个因素都是照片或视频的特征。因此，对照组在表外有自己的单元格。控制组可以拥有比处理组更少的受试者。推荐至少有3:1的比例，意味着处理组的受试者数量可以是对照组的三倍。

**对照组**

无论是完全阶乘设计还是不完全阶乘设计，阶乘设计的一个优势在于可以通过与未接受处理的对照组进行比较来进行控制。在完全阶乘设计中，因素的一个水平可以代表对照条件。例如，一项研究是一个4 x 3的混合阶乘设计，其中一个因素称为社会问题，由三个不同的故事代表，一个关于洪灾受害者，一个关于少女殴打另一个女孩，一个关于卖淫行动。第二个因素是视觉类型，与前面的假设性例子相同——静态照片、电视视频、互联网视频——但增加了一个对照组：无图像。对照组的受试者阅读了关于三个不同社会问题的相同故事，但没有看到图片或视频。该设计表如表6.6所示。

从第4章中涵盖的Campbell和Stanley的实验设计分类中可以看出，对照组对于“真实实验”的所有三种类型都是必不可少的。前测-后测对照组设计和仅后测对照组设计都有一个对照组，而Solomon四组设计有两个对照组。对照组在确保不同组在受控条件下反应的重要性不可忽视。



For Review-No Commercial Use(2023)

©Loren Fishman, [www.cartoonstock.com](http://www.cartoonstock.com/)

A screenshot of a computer

Description automatically generated

另一种看待对照组的方式是将其视为尚未接受治疗或干预的基准测量，或者是一种中立的情况。如果实验组接受了治疗并发生了变化，而对照组没有接受治疗却发生了变化，那么我们就知道治疗并不是导致变化的原因。但是，如果实验组接受了治疗并发生了变化，而对照组没有接受治疗却没有发生变化，那么我们就知道治疗确实导致了变化。

在随机分配之后，对照组是进行真实实验时最重要的事情。它是实验者将接受治疗的人与未接受治疗的人进行比较的方式。对照组是确保观察到的任何效应都是由于治疗变量而不是其他因素引起的另一种方式。如第1章所指出的，医学研究人员通常会给对照组的受试者服用安慰剂，例如糖丸、生理盐水注射或其他让人们相信他们接受了某种“治疗”的物质，即使它是惰性的或不被期望具有任何益处。由于暗示的作用，人们需要认为他们接受了某种治疗或者有人对他们进行了某种操作，以准确地代表对照组。正如Campbell和Stanley指出的那样，“将对照组与无对照组进行比较是一种过于简化的方式。实际上，比较的是对照组的具体活动，这些活动填满了与实验组接受X的时间相对应的时间段。”一个适当的对照组不是一群在没有接受他们认为是治疗的东西的情况下就被测量某个结果变量的人群。对于社会科学家来说，没有“糖丸”。接下来，本章将讨论这些研究人员创建社会科学安慰剂的一些方法。

**无处理对照组**

对照组的经典例子是接受处理组中操作的缺失。例如，一项研究中的研究人员使用明确将一个问题称为“重要”的故事来作为一个处理组，而使用隐含地通过呈现该问题的后果来将一个问题标记为重要的故事作为另一个处理组。对照组看到的是完全相同的故事，但没有对问题的评估认定为重要，也没有讨论任何后果。在之前讨论的不安全性研究示例中，处理组撰写了关于他们的财务未来的文章，引发了高度不安全感。而对照组撰写了关于音乐的文章，预计不会引发不安全感。

i我要感谢匿名审稿人提出了这种清晰简洁的描述对照组的方式。

在控制组的理念中，关键是让控制组或“无治疗”组的受试者接受一种尽可能与治疗组相似但不会影响结果的处理。

**控制组的创造力**

有时，需要一些创造性的思考来确定受试者在不进行任何操作时应该做些什么。例如，研究烹饪实践的研究人员将三种类型的营销信息作为处理：关于健康益处的信息，关于节省时间和金钱的信息，以及关于两者的信息。没有信息是明显的控制组，但是那么受试者应该做些什么呢？研究人员通过将讨论小组用作控制条件来解决这个问题，使受试者与他人讨论常见的烹饪实践，而不是接受营销信息。在一项比较倡导性和客观犯罪故事的新闻学研究中，控制组阅读了与犯罪无关的主题的故事，以避免他们经历与犯罪故事引发的恐惧等相同的情绪。他们仍然阅读故事，只是不是犯罪故事。在这两项研究中，与其让控制组受试者什么都不做，他们参与了一次小组讨论，而不是看到营销信息，或者阅读了与治疗组的故事完全无关的主题的报纸报道。鲍塞尔称之为“让控制组接受有趣但无关紧要的活动”，这意味着给予控制组受试者一些与治疗组看起来相同的东西，要比让他们来填写问卷后就离开而没有得到任何治疗要好得多。控制组受试者必须执行某些任务，使他们认为自己接受了合法的治疗。他们需要感知到发生了一些事情。

**现状控制组**

在其他情况下，研究人员将控制组定义为最正常、常见或典型的情况。在这些实验中，可能会使用一个代表现状但未被标识为控制组的不同组。例如，在一项关于如何最好地教授哥斯达黎加三年级学生英语的实验中，两个处理组使用了不同的计算机辅助学习软件，而控制组则接受了教师授课的英语教学，这是学校通常采用的方式。教师授课组作为控制组。新闻学研究经常使用以中立、两面观点编写的故事作为控制条件，因为它代表了观众通常阅读的内容。

没有控制组

没有所有实验都需要控制组，在某些情况下可能并不需要。在某些研究中，提供与处理相似的控制处理可能非常困难。例如，在HIV信息的研究中，关于故事框架的因素的控制处理是什么？不存在没有任何框架的故事。或者在关于声音音调的研究中，两个水平是提高和降低声音音调，不存在没有声音音调的情况。在许多比较两个或多个处理的研究中，所定义的控制组是模糊的。在这些研究中，控制组可能并不是必需的，因为它们是将处理组彼此进行比较，而不是与没有声音音调或没有框架的故事的某个组进行比较。

无论研究是否有经典的控制组或其他类型的控制组，在变量管理方面都应该进行仔细控制。当设计实验并对每个要素进行详细审查时，请参考第5章中概述的有效性威胁。各组之间只应在干预方面有所不同。即使是如早上对比下午、不同房间或不同实验者这样简单的差异也可能影响结果。

当然，还有其他类型的因子设计，不仅限于这里概述的内容，特别是分裂区设计、部分复制设计和随机区组设计等。正如Campbell和Stanley所说的“一些令人困惑的因子设计之谜”。由于这是一本入门性的教材，不会深入探讨每个设计的奥秘，而是建议读者参考更高级的文献。

下一章将涉及作为真正实验必须进行的最重要的事情——随机分配。

• 设计过于复杂的研究

• 在可能的情况下未使用对照组

• 未给予对照组任务可执行的事项

**常见错误**

**168** Designing Experiments for the Social Sciences

1. 具有两个自变量，每个自变量具有两个或更多水平的实验称为：

a. 单因素设计

b. 单路径设计

c. 阶乘设计

d. 三路径设计

2. 对于以下假设性实验，使用哪种阶乘符号表示：一个研究人员操纵广告活动的情绪调性（积极/消极/中性）和传播媒介（电视/印刷品/互联网）。

a. 2 x 2

b. 2 x 3

c. 3 x 3

d. 2 x 3 x 3

3. 阶乘设计可以有以下哪些结果？

a. 所有因素的主效应显著，但无交互作用效应

b. 某些因素的主效应显著，但不是全部，其中可能存在显著或不显著的交互作用效应

c. 显著的交互作用效应，但没有主效应

d. 所有上述情况

4. 如果受试者在某个任务上由于练习而可能表现更好，最好使用：

a. 组间设计

b. 组内设计

c. 混合设计

d. 不完全设计

5. 在一项研究中，探讨提醒司机支付交通罚单的最佳方式，以下哪组是对照组？

a. 通过短信发送的提醒

b. 通过电子邮件发送的提醒

c. 通过邮寄发送的提醒，按照正常方式发送

d. 通过电话呼叫发送的提醒

6. 在一项关于HIV/AIDS的公益广告研究中，研究人员通过广告中的论证质量（强势/弱势），以及HIV/AIDS对受试者的个人相关性和两种不同的广告形式（故事叙述和统计数据）进行了调查。这是一个：

a. 单因素设计

b. 单路径设计

c. 阶乘设计

d. 三路径设计

Chapter 6 ■ Factorial Designs **169**

a. 单因素设计

b. 阶乘设计

c. 组间设计

d. 组内设计

7. 仅一个因素的不同水平能够导致结果变量的变化被称为

a. 混淆因素

b. 可行的替代解释

c. 主效应

d. 交互作用效应

8. 当一个因素的结果取决于另一个因素时，这被称为

a. 混淆因素

b. 可行的替代解释

c. 主效应

d. 交互作用效应

9. 在一项关于烹饪实践的研究中，包括关于健康益处、节省时间和金钱的信息以及一个讨论小组，哪个代表对照组？

a. 关于健康益处的信息

b. 关于节省时间的信息

c. 关于节省金钱的信息

d. 一个讨论小组

10. 在一项教育研究中，比较了计算机辅助学习软件和学校通常的教学方式，使用了哪种类型的对照组？

a. 没有对照组

b. 现状对照组

c. 比较两种处理方式

d. 内部对照组

答案:

1. c

2. c

3. d

4. a

5. c

6. b

7. c

8. d

9. d

10. b

**170** Designing Experiments for the Social Sciences

1. 在第1章中提出的实验想法，并在第3章中为其撰写了文献综述和假设，请绘制一个设计表格，以帮助你可视化因素、水平和每个组合。将其转化为阶乘符号表示。包括设计类型、因素和水平。请参考"How To Do It"框中的指南。
2. 撰写一页关于你选择使用组间设计或组内设计的解释。为你的选择进行辩护（例如，为什么不会出现carry-over效应）。想象一位审稿人质疑你的选择，并建议相反的选择可能更好。你会如何回应？
3. 撰写一页关于你所选择的控制组类型以及受试者在不接受处理时应进行的活动。解释你的选择并为你的决策进行辩护。

**应用练习**

《Methods in Behavioral Research, 10th Edition》（2009年）中的第10章“复杂实验设计”，作者Paul C. Cozby。出版地：波士顿，出版社：McGraw-Hill。

《Conducting Meaningful Experiments: 40 Steps to Becoming a Scientist》（1994年）中的第6章“设计实验研究以实现统计显著性”，作者R. Barker Bausell。出版地：加利福尼亚州千橡市，出版社：Sage Publications。

**建议阅读**

Chapter 6 ■ Factorial Designs **171**

**注释：**

1. Diana Mutz，《基于人口的调查实验》（普林斯顿大学出版社，2011年），125页。

2. Graeme D. Ruxton和Nick Colegrave，《生命科学实验设计，第3版》（牛津大学出版社，2011年）。

3. William R. Shadish, Thomas D. Cook和Donald T. Campbell，《广义因果推断的实验和准实验设计》（Wadsworth Cengage Learning，2002年），263页。

4. Sean Aday，《新闻框架效应：倡导框架与客观主义框架的实验测试》，《新闻与传播学季刊》，第83卷，第4期（2006年冬季），767-784页。

5. Tim Kasser和Kennon M. Sheldon，《财富与死亡：物质主义、死亡意识和消费行为》，《心理科学》，第11卷，第4期（2000年7月），348-351页。

6. Christopher D. Johnston和Julie Wronski，《个性倾向与难易问题上的政治偏好》，《政治心理学》，第36卷，第1期（2015年），35-53页。

7. T. Rogers等，《巧妙的欺骗：使用真实陈述误导他人的风险与回报》，《人格与社会心理学杂志》，第112卷（3期），456-473页。

8. Shadish, Cook和Campbell，《实验和准实验设计》。

9. Ronald A. Fisher，《实验设计》（爱丁堡：Oliver and Boyd，1937年）。

10. D. T. Campbell和J. C. Stanley，《研究的实验和准实验设计》（芝加哥：Rand McNally，1963年），3页。

11. 同上

12. Jueman Zhang等，《艾滋病/艾滋病公益广告的说服力与论据质量、个人相关性和证据形式的关系》，《社会行为与人格：国际期刊》，第42卷，第10期（2014年），1603-1612页。

13. Ruxton和Colegrave，《生命科学实验设计，第3版》。

14. Campbell和Stanley，《实验和准实验设计》，29页。

15. Ruxton和Colegrave，《生命科学实验设计，第3版》。

16. Zhang等，《艾滋病/艾滋病公益广告的说服力》。

17. Campbell和Stanley，《实验和准实验设计》，29页。

18. 同上

19. Xiaoli Nan等，《信息框架、感知易感性和非洲裔美国家长对HPV疫苗接种意愿的影响》，《健康传播学》，第31卷，第7期（2016年），798-805页。

20. 同上

21. Gabriele Esposito等，《通过“推”避免在线购买不兼容的数字产品：一项实验研究》，《PLoS ONE》，第12卷，第3期（2017年），1-15页。

22. Nan等，《信息框架、感知易感性》，800页。

23. 同上

24. Ruxton和Colegrave，《生命科学实验设计，第3版》。

25. Debby Vos，《普通议员如何登上新闻：与政治记者进行的一项因子调查实验，解释议员的新闻价值》，《大众传播与社会》，第19卷，第6期（2016年），738-757页。

26. Shadish, Cook和Campbell，《实验和准实验设计》，265页。

27. 同上

28. R. Barker Bausell，《进行有意义的实验：成为科学家的40个步骤》（千橡，加利福尼亚州：Sage出版社，1994年）。

29. Murray R. Selwyn，《生命科学实验设计原理》（博卡拉顿，佛罗里达州：CRC出版社，1996年），55页。

30. Zhang等，《艾滋病/艾滋病公益广告的说服力》，1606页。

31. An-Sofie Claeys和Verolien Cauberghe，《保持控制：组织发言人在危机时期非言语表达权力的重要性》，《传播学杂志》，第64卷（2014年），1160-1180页。32. 同上

33. Bausell，《进行有意义的实验》。

34. 同上，97页。

35. Paul R. Brewer, Joseph Graf和Lars Willnat，《启动还是框架》，《新闻界报：国际传播研究期刊》，第65卷，第6期（2003年），493-508页。

36. Angela Min-Chia Lee，《速度有多快？研究速度驱动新闻对新闻制作和受众接受的影响》（未发表的博士论文，德克萨斯大学，2014年）；Angela M. Lee，《越快越好？研究直播博客对受众接受的影响》，《应用新闻学与媒体研究期刊》，第8卷，第1期（即将发表）。

37. Shadish, Cook和Campbell，《实验和准实验设计》，265页。

38. Selwyn，《生命科学实验设计原理》，53页。

39. Rebecca B. Morton和Kenneth C. Williams，《实验政治科学与因果关系研究：从自然到实验室》（纽约：剑桥大学出版社，2010年），86页。

40. Charles M. Judd, Jacob Westfall和David A. Kenny，《多个随机因素的实验：设计、分析模型和统计功效》，《心理学年度评论》，第68卷，第1期（2017年），601-625页。

41. H. Cho和F. J. Boster，《损失框架与收益框架的禁毒广告对青少年的影响》，《传播学杂志》，第58卷，第3期（2008年），428-446页。

42. Aday，《新闻的框架效应》。

43. Morton和Williams，《实验政治科学》，86页。

44. Judd、Westfall和Kenny，《多个随机因素的实验》。

45. Ruxton和Colegrave，《生命科学实验设计，第3版》。

46. Job Harms等，《自由来帮忙？关于自由意志信念和利他主义的实验》，《PLoS ONE》，第12卷，第3期（2017年），1-15页。

47. Luiza Venzke Bortoli和Valeria Freundt，《自愿产品召回对消费者信任的影响》，《巴西商业评论》，第14卷，第2期（2017年），204-224页。

48. Aday，《新闻的框架效应》。

49. Renita Coleman，《图像对道德推理的影响：对新闻工作者在道德决策中照片的价值的研究》，《新闻与传播季刊》，第83卷，第4期（2006年冬季），835-850页。

50. 张等人，《HIV/AIDS公益广告的说服力》，1606页。

51. Claeys和Cauberghe，《保持控制：组织发言人在危机时期的非语言表达的重要性》，《传播学杂志》，第64卷（2014年），第1160-1180页。

52. Shadish，Cook和Campbell，《实验和准实验设计》，265页。

53. Bausell，《进行有意义的实验》，第265页。

54. Shadish，Cook和Campbell，《实验和准实验设计》。

55. 同上。

56. Aimee Meader等人，《数字时代的伦理：运动图像和照片对道德判断的影响比较》，《媒体伦理杂志》，第30卷，第4期（2015年），第234-251页。

57. Campbell和Stanley，《实验和准实验设计》。

58. Gerry Stoker，《探索政治科学中实验的承诺：微基础洞察和政策相关性》，《政治研究》，第58卷（2010年），第300-319页。

59. Morton和Williams，《实验政治科学和因果推断的研究：从自然到实验室》，86页。

60. Cengiz Erisen，Elif Erisen和Binnur Ozkececi-Taner，《政治心理学研究方法》，《土耳其研究》，第13卷，第1期（2013年），第13-33页。

61. Campbell和Stanley，《实验和准实验设计》。

62. 同上，第13页。

63. Kristin Bulkow，Juliane Urban和Wolfgang Schweiger，《议程设置的二重性：信息处理的作用》，《公众舆论研究国际期刊》，第25卷，第1期（2013年），第43-63页。

64. Kasser和Sheldon，《财富和死亡：物质主义、死亡敏感性和消费行为》，《心理科学》，第11卷，第4期（2000年7月），第348-351页。

65. Theresa Beltramo等人，《营销信息和分期付款对购买节能炉具意愿的影响》，《经济行为与组织杂志》，第118卷（2015年），第333-345页。

66. Aday，《新闻的构架效应：倡导性与客观性框架的实验研究》，《新闻与大众传播季刊》，第83卷，第4期（2006年冬季），第767-784页。

67. Bausell，《进行有意义的实验》，第74页。

68. 同上。

69. Horacio Alvarez-Marinelli等人，《哥斯达黎加小学的计算机辅助英语学习：一项实验研究》，《计算机辅助语言学习杂志》，第29卷，第1期（2016年），第103-126页。

70. Aday，《新闻的构架效应》。

71. Ruxton和Colegrave，《生命科学的实验设计》，第3版。

72. Nan等人，《信息框架、知觉易感性和针对非洲裔美国家长的预防HPV疫苗接种意愿》，《健康沟通》，第31卷，第7期（2016年），第798-805页。

73. Claeys和Cauberghe，《保持控制》。

74. Morton和Williams，《实验政治科学》。

75. Ruxton和Colegrave，《生命科学的实验设计》，第3版。

76. Stoker，《探索政治科学中实验的承诺》，304页。

77. Bausell，《进行有意义的实验》。

78. 同上。

79. Campbell和Stanley，《实验和准实验设计》，第27页。

**172** Designing Experiments for the Social Sciences

7

# 随机分配

正如代表性可以通过随机的方法得到保证一样...等价性也可以通过偶然性来保证。1

**—W. A. McCall**

• 理解随机分配的作用和工作原理。

• 为实验设计一个有效的随机化过程并描述它。

• 评价简单随机分配、分组、成对匹配和分层随机分配。

• 解释平衡的重要性。

• 描述拉丁方设计。

学习目标

正如房地产中的口号是“位置、位置、位置”，实验设计中的座右铭是“随机分配、随机分配、随机分配”。本书始终强调了随机分配的重要性。不厌其烦地重申，随机分配是研究者在实验中所能做的最重要的事情。与正确进行随机分配相比，其他一切都显得微不足道。随机分配是将真实实验与准实验、自然实验或预实验设计区分开来的关键。在第1章中，实验被称为黄金标准。没有成功的随机分配，研究无法真正成为实验。

**173**

然而，如果缺乏随机分配，实验很快就会变成“铜牌标准”。本章将回顾随机分配的一些优点，讨论如何进行随机分配，并探讨与平衡对照组相关的问题。

**随机分配的目的**

人们是有差异的。也就是说，他们是不同的。如果不是这样，就没有理由研究他们。每个人对不同的教学技巧、广告、健康干预和政治信息都会做出相同的反应。不需要进行实验。人们的差异提供了社会科学家进行研究的理由，也带来了他们最大的挑战。当人们在实验不关心的事物上存在差异时，这被称为随机变化或噪音。虽然这些不是研究的重点，但它们仍然可能导致实验处理结果的一些变化。也就是说，它们可能混淆结果。实验者的一个主要关注点是控制混淆因素，消除或减少随机变化带来的噪音。这样，研究关注的效果可以更清晰地显示出来。随机化可以说是科学家拥有的最强大的武器，因为它有助于确保不同处理和对照组的受试者在创建噪音的变量上几乎相同。

随机分配是一种在真实实验中将受试者分配到不同的处理和对照组的技术，目的是确保每个组的受试者具有相似的特征，即它们是等价的。等价的意思是平均相等或概率上相等，而不是完全相同。等价的目的是“公平竞争”，帮助确保两组之间唯一的系统性差异是他们在实验中接受的处理。它有助于确保一个组在结果变量上并不比另一个组更好，并且一个组不会充满那些不论是否接受处理都更有可能发生变化的受试者。这使研究者更有信心任何观察到的变化都是由于受试者接受的处理，而不是由于受试者本身的固有差异所致。在一个或多个方面有系统差异的组可能会使实验失效。例如，如果一个组只包含男性，另一个组只包含女性，那么实验就会被混淆；无法确定结果是由于处理还是由于性别造成的。随机分配有助于实现等价性的一种方式是避免选择偏差，选择偏差发生在人们自主选择所属组别或研究者以某种主观原因选择受试者，甚至可能是下意识地为了支持假设的机会。

例如，如果教育研究中，治疗组主要得到那些数学水平较好的受试者，而对照组得到那些数学水平较差的受试者，那么这个研究就是无效的。如果治疗是关于数学教学方法的改变，那么看起来治疗有效，但实际上可能只是因为那些接受新教学方法的人在开始时就数学水平较好。另一方面，如果那些数学水平较差的学生被教育者分配到治疗组，因为教育者希望确保那些最需要数学帮助的学生得到帮助，那么治疗组就会被数学水平较差的学生填满，而对照组则是数学水平较好的学生。新的教学策略实际上可能是有效的，但如果它只将治疗组提升到对照组的水平，看起来可能并没有显著差异；换句话说，治疗后两组之间没有显著差异。

另一个例子是有些人比其他人更容易改变。如果治疗组得到更多易于改变的受试者，而对照组得到较少易于改变的受试者，那么一个旨在影响公共政策立场的研究可能会显示出虚假结果——治疗并没有真正改变人们的观点；他们在开始时更容易改变观点。给那些更难改变的人相同的治疗可能根本没有效果。

**避免混淆因素**

随机分配的美妙之处在于研究者无需预先确定可能混淆研究结果的每个受试者的每个特征。例如，可以很容易预料到先前的数学知识会影响数学学习，并对数学进行预测试，然后将数学知识高和数学知识低的受试者以相等数量分配到每个组中。但对于一个研究者来说，很难预料到一个人是否容易改变立场，这也是一个混淆变量。通过随机分配，不需要对数学能力或改变立场进行预测试，因为每个组中都有相等数量的容易改变立场和不容易改变立场的受试者；或者每个组中都有数学水平较好和数学水平较差的学生。由于人类，包括研究者，对可能影响其他事物的每一个小细节都不擅长预测，并且有些事情是不可预知的，随机分配具有巨大的好处。此外，回想一下第4章中对预测试的缺点；随机分配可以消除对预测试及其对效度的威胁的需求。

随机分配虽然不是完美的，但是它是目前我们所知的确保受试者之间的系统变异不会混淆实验结果的最佳方法。它有助于消除虚假变量，包括研究者可能没有考虑到的那些变量。

当同一个人不可能同时接受治疗和不接受治疗时，研究者使用随机分配来尽量确保个体差异在每个组中分配相等。

**随机是什么？**

随机分配的原理是基于机会。为了随机分配受试者，研究中的每个人必须有相等的机会进入对照组或治疗组。第2章回顾了随机分配是如何被发现的故事。当查尔斯·桑德斯·皮尔斯和约瑟夫·贾斯特罗进行一项研究，探讨人们通过感觉和观察来准确判断物体重量的能力时，他们首先呈现了重物，然后呈现了轻物，还交替呈现重物和轻物。最后，他们洗牌一副牌，并根据牌的顺序随机分配重物。当重物按照任何有规律的模式呈现时，他们得到了截然不同的结果。对于无法根据模式猜测重量的受试者，产生了更有效的结果。

随机并不意味着杂乱无章，研究者必须小心使用适当的随机方法。它不是没有任何模式、目的或系统，而是有一种非随机的特点。在我的第一个实验中，显然我没有完全理解随机的意义，但我关注的是每组具有相等的受试者数量。我将第一个来到实验室的人分配给第一组，第二个人分配给第二组，然后将第三个人再分配回第一组，以此类推。基本上，他们被分配为：1、2、1、2、1、2、1、2。当我告诉我的导师时，她的眼睛睁大了，她说：“那不是随机的。”我最终放弃了所有的数据，重新开始实验。

正如Gueron所说，“稍微随机并没有什么帮助。”



###### 关于...盒子7.1

###### 随机分配对内部效度的威胁

当被试者和/或将被试者分配到条件的实验者对被分配到的组别不“盲”的情况下，随机分配本身以及其违反可能导致对内部效度的威胁。这四种类型包括：13

*•* 治疗扩散：当不同组别的被试者相互交流并了解非针对他们的信息时发生。例如，如果治疗组和对照组的被试者互相交谈，他们可能会了解仅针对治疗组的材料；结果可能没有差异，并且不会真实反映治疗的益处。

• 补偿均衡：当实验者或提供治疗的人试图将治疗组拥有的一些优势分配给对照组时发生。

• 补偿竞争：指的是对照组成员试图获得治疗组的一些好处。

• 愤怒的士气下降：指的是对照组的被试者表现不佳，因为他们对被拒绝接受治疗感到不满。

在某些研究中，治疗组的被试者可能感到士气低落或受到污名化，例如被视为“笨蛋”班级的一部分。因此，尽可能让被试者对自己所属的组别不知情是很重要的。

研究人员应注意在招募材料中不要将某个干预组描绘得比另一个更好或更新。如果无法让被试者不知道自己所属的组别，研究人员应测量被试者对分配到特定组别的偏好，并在计分析中加以控制。

当存在一个更偏好某个组别而不是另一个组别的被试者数量不平衡时，会对内部效度构成威胁。例如，如果两组被试者中都有60%对所在组别感到满意，40%不满意，那么偏好在各个条件下是等价的。然而，如果一个组有60%满意而另一个组只有40%满意，结果可能会混淆。只要两组具有相同百分比的满意和不满意被试者，就不会出现威胁。例如，在一项商业研究中，要求参加治疗组的人中有89%同意参与，而要求参加对照组的人中只有45%同意参与。21

真正的随机分配是一个要么做对了要么做错了的过程。为了使随机分配起作用，研究者不能因为任何原因选择将被试分配到哪个组，被试也不能选择自己所在的组。被试必须是盲目的，也就是对自己是否接受治疗不知情。确保被试

为了防止被试试图给予研究者他们认为研究者想要的结果，或者为了防止他们因未接受治疗而对结果产生偏见，让被试保持不知情是很重要的。另外，让研究者不知道被试所在的治疗组或对照组也是一个好主意，这被称为双盲研究。有关当被试和/或研究者知道自己被分配到哪个组时对内部有效性的威胁的更多信息，请参阅第7.1节的更多内容。

### 操作化随机分配

操作化随机分配有很多方法。1883年，皮尔斯和贾斯特罗使用了一副特殊的扑克牌来确定随机分配，这种方法今天仍然有效。其他方法包括掷骰子、抛硬币、从帽子中抽取数字，或使用随机数表。关于如何使用这些手动方法的详细信息，请参阅第7.2节的操作方法框。

**计算机随机分配**

如今，研究人员更常使用在线工具或电子表格和统计软件来进行随机分配。例如，免费的在线随机分配器可以让研究人员指定被试的数量和组数，并迅速返回一个显示哪个被试分配到哪个组的列表，如图7.1和7.2所示（请参阅第180-181页）。只需在互联网上搜索“随机分配生成器”即可找到这些工具。

使用GraphPad的QuickCalcs，在标有“Assign”的第一个框中输入总被试数量。在第二个框中输入组数。将“Repeat”设置为1。点击“Do it!”

它将返回一个被试的列表，编号从1到最终编号，旁边标有A、B、C等组的标签。

这个小应用程序是由一群学术研究人员专门为因子实验设计的。在第一个框中，“Number of Participants, N”输入整个研究的总被试数量。在第二个框中，“Number of Conditions, C”输入组数。点击“Compute”。

它将返回一个按顺序分配被试的组列表；它没有QuickCalcs随机分配器中的被试编号，但很容易看出第一个被试被分配到第一个组编号，第二个被试被分配到下一个组编号，依此类推。

**如何操作 7.2**

随机化被试

抛硬币方法适用于有两个组的情况。如果硬币正面朝上，将被试分配到实验组；如果硬币反面朝上，将被试分配到对照组。或者反过来。

彩票方法适用于任意数量的组。在纸条上写上每个组的编号（例如1、2、3、4）。根据你计划拥有的被试数量制作相同数量的纸条，并且每个纸条上都有相同数量的组号。例如，如果你需要160个被试，每组40个，就制作40张编号为1的纸条，40张编号为2的纸条，依此类推。

扑克牌和骰子方法需要使用只有组号的扑克牌和骰子。例如，如果你有四个组，就去掉除了1、2、3、4以外的扑克牌。可能需要多副扑克牌。洗牌后，每当一个被试到来时，抽取一张牌，并将被试分配到对应的组。对于骰子，使用只有与你的组数相同的数字的骰子。如果掷出的数字大于你的组数，就重新掷骰子。一旦某个组的被试数量达到最大值，忽略掉牌或骰子上出现的该数字。随机数表格（Book of random numbers）。这些现在已经过时，被在线随机数生成器所取代。它们由一页又一页的随机排序数字组成。信不信由你，研究者会闭上眼睛，在桌子上指向一个起始位置来决定从哪里开始。然后，研究者将被试分配到与数字相对应的组中，并跳过超出范围的数字。例如，如果数字是：3、2、5、3、7、4、1等，而研究中有三个组，那么被试将被分配到3、2、3、1等组中，跳过5和7，因为没有第五组和第七组。一旦某个组的被试数量达到最大值，也会跳过该组的编号。当某个组达到最大数量后，停止将被试分配到该组，但继续分配至其他组直至所有其他组填满。我最喜欢的随机数表格是《A Million Random Digits With 100,000 Normal Deviates》（RAND）。如果你感兴趣或是历史爱好者，你的图书馆可能还有这本书。

Excel. 关于如何在Excel中生成随机数的教程，请参考https://exceljet.net/formula/randomly-assign-data-to-groups。

SAS和SPSS. Shadish、Cook和Campbell22在书中的第311-313页给出了这些流行统计软件中生成随机数的指导方针。

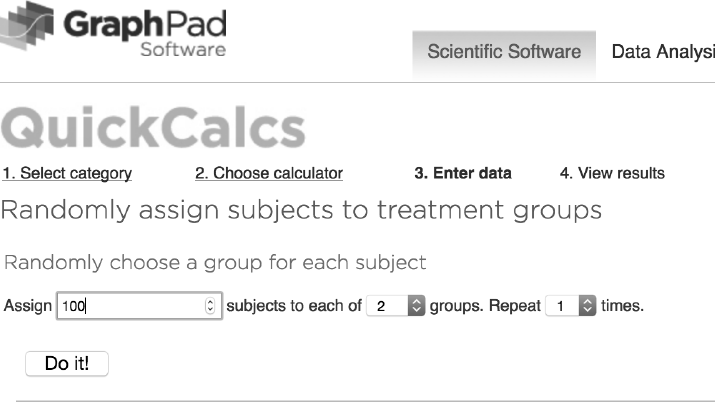
对于其他统计软件，如Stata和R，请参考相应的教程。

对于纸笔调查研究，可以使用在线随机分配器将打印的问卷按指定的随机顺序排列好，然后再去实验室或被试参与研究的地点。然后可以从堆叠问卷的顶部开始发放，无需每次有被试到来时都查看随机数的输出。i

i. 请确保在纸质问卷上做出注记，或使用其他方法来记录被试所属的组别、处理组或对照组，以防止混淆。



**FIGURE 7.1 GRAPHPAD**

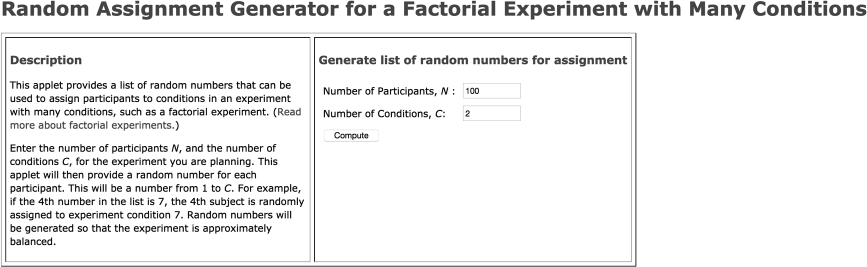


#### For Review-No Commercial Use(2023)

Source: <http://www.graphpad.com/quickcalcs/randomize1.cfm>



**FIGURE 7.2 RANDOM ASSIGNMENT GENERATOR**



#### For Review-No Commercial Use(2023)

Source: <http://www.methodologymedia.psu.edu/most/rannumgenerator>

在线随机分配器可以自动将相等数量的被试分配到每个组，前提是最大被试数量可被组数整除；例如，对于三个每组有四十名被试的组别，将总被试人数设定为120，而不是125。

**调查实验**

一种革新实验被试随机分配方式的最新技术是使用专门设计用于在线调查的软件，例如Qualtrics和SurveyMonkey等。这些调查实验，也经常被描述为“嵌入在调查中的实验”，使用随机分配功能将被试随机分配到处理组和对照组，研究者只需在启动调查前设置随机分配器即可，无需进行其他操作。计算机辅助电话访问软件也可以像调查软件一样将被试随机分配到不同条件，只是被试通过电话参与研究而不是在线参与。

有些研究者将术语“调查实验”保留给从人口中随机抽样并随机分配被试到不同条件的研究，而另一些研究者在没有使用随机抽样的情况下使用该术语，但使用了随机分配。Erisen等人区分了调查实验和实验室实验，解释说实验室实验可以控制环境因素（例如温度、一天中的时间或其他可能影响结果的因素），而调查实验可以在被试自己的家中进行，而且研究人员无法控制外部因素，例如被试是否中途停下来应门、上厕所或接电话。这在测量被试完成研究所需时间、被试可以在网上查找答案的知识问题，或者被试可能不相信自己在互联网上与真实人士进行互动的群体决策研究时尤为重要。当需要确保这种内部效度时，可以将调查软件用于实验室中的被试，既保证了研究人员的控制权，又提供了随机分配和计算机数据收集的便利性。在一项关于政治家决策的研究中，研究人员让被试使用调查软件在线完成研究，但为了保持控制，研究人员在场监督研究进行。其他研究人员已经探索了一些方法来防止查找答案以减少在线实验中这种错误的发生。

关于调查实验是否使用随机样本更好，因为具有普遍适用性，或者方便样本是否能产生类似的结果，已经有很多辩论和研究。许多研究显示几乎没有差异。这个问题在本书第5章关于外部效度的部分中得到了广泛探讨。本章的重点在于，无论被称为调查实验还是实验室实验，所有真实实验的定义特征都是被试必须随机分配到处理组和对照组。Online randomizers will automatically assign equal numbers of subjects to each group if the maximum number of subjects is divisible by the number of groups; for example, for three groups with forty subjects in each, set the total number of subjects at 120, not 125.

**面对面随机分配**

让被试通过计算机参与实验可以在很多方面更容易进行随机分配。但有时候使用纸笔进行实验可能更合适。面对面研究需要采用不同的随机分配策略。例如，政治科学家可能会利用选民离开投票站时进行实验，以确保被试确实进行了投票。如果宣传册或报纸中的信息是被测试的信息，那么使用印刷版本比在线显示更加真实。大多数道德判断的研究都是面对面进行的，因为这个主题的复杂性。与记者一起进行的道德判断实验可以在新闻编辑室进行，研究者提供午餐，并让被试在用餐期间完成研究；这样可以吸引更多忙碌的专业人士参与。其他实验可以在专业协会会议上进行，通过在大厅的桌子上进行纸笔测试。对于某些研究来说，在教室或图书馆进行传统的纸笔测试可能能更快地收集数据。在这些情况下，随机分配需要通过在线随机分配器或之前描述的传统方法进行。

**报告随机分配**

无论如何实现随机分配，都应在研究论文中进行报告，并最好说明所使用的过程。通常不需要详细介绍具体细节，但报告至少应表明被试是随机分配到条件中的，以免读者产生其他想法。以下是两个实际发表的例子：

“参与者被随机分配到一种条件。”

“然后，参与者随机获得一本小册子，其中包含研究的说明、一个测试小册子和一个回答问卷。测试小册子是关于人乳头瘤病毒和生殖器疣的四个版本中的一个。因此，参与者被随机分配到实验条件之一。”

最佳做法是提供随机分配机制的描述，例如是使用一副牌或骰子，还是使用特定软件中的随机分配功能，以及是预先生成还是现场生成。

**均衡设计和非均衡设计**

随机分配可能引起关于组别中被试数量不均等的担忧，就像在我第一次实验中遇到的情况一样。这被称为非均衡设计。非常不均衡的样本大小可能会影响组别的等价性。要求各组的被试数量完全相同并不是至关重要的。

保持组内被试数量相近即可实现近似均衡，而不需要确保每个组内的被试数量完全相同。然而，对于完全相同数量的被试分布在每个组中的均衡设计，研究将具有更高的统计力量（第8章的主题）。

有时候，非均衡设计是由于一些实际而不可避免的原因，例如某些被试退出或被清除出实验。例如，在道德判断研究中，存在一种内建的检查方法来发现试图伪造更高得分的被试。当检测到伪造行为时，这些被试将被从数据集中排除，通常导致组内被试数量不均衡。有时候，研究者可能会失去未能回答足够问题以提供有效数据的被试，或者被试在研究中有系统性的退出，这被称为脱落。对于组内被试数量不均衡的担忧绝不能成为偏离随机分配的原因，因为不正确地进行随机分配比非均衡组更具威胁性。正如在第6章对照组的讨论中指出的，对于不完全的因子设计，控制组的被试数量少至多达三分之一是可以接受的。当处理方式昂贵或难以操作时，组内被试数量较少也是可以接受的。此外，当处理方式是令人期待的时，人们可能会因为担心被分配到对照组而不愿意参与。随机分配更多的被试到处理组可以帮助克服这些顾虑。

同样重要的是报告脱落率，就像在一个商业实验的导师计划中所做的那样。在那项研究中，52%的被试退出了研究；研究者将退出的被试与留在研究中的被试进行了比较，并发现在可观测特征上没有统计上的差异。关于脱落率的经验法则说，5%到20%之间可能会导致偏差。

许多优秀的统计学书籍介绍了分析非均衡设计数据的方法。可以使用统计技术，如Levene's检验，来确定不均衡的被试数量是否导致方差不均衡。当Levene's检验具有统计显著性，表明方差不均匀时，研究者会使用更严格的差异检验，而不是假设方差相等。这是统计学课程或教材的内容。可以说，存在比偏离随机分配机制更好的处理非均衡组的方法。

**检查随机分配的有效性**

尽管随机分配是研究者获得等价组的最佳方法，并且随机化失败很少见，但仍然有人持怀疑态度。这时需要进行随机分配检查或平衡性检验，即比较组间等价性。

顺序。并不需要所有变量在各组之间均等分布；与结果或因变量高度相关的变量是最重要的关注点。一个经验法则是，如果一个变量与因变量的相关性达到0.45或更高，那么该变量在各组之间必须具有相等分布。人们普遍误解随机分配必须在每个已知变量上实现均等分布；实际上并非如此。

**聚合层级的随机分配**

与个体相比，聚合层级的随机分配检查更加重要。例如，企业、家庭、选区或教室可以作为单位进行随机分配。聚合层级分配通常用于在难以或不可能随机分配个体时使用，但这种方式不如个体随机分配理想，因为组或组织内的个体可能存在系统差异。例如，初创企业可能拥有比已建立企业更年轻的员工。一个完整的教室可能存在不同的动机水平。选区中的选民可能由于所能负担得起的居住区不同而存在系统性差异。在随机分配群体（如班级或整个学校）时，测试等价性变得更为重要。这是因为使用聚合群体时样本量通常较低于使用个体时。例如，一项研究将41所学校分配给三个条件，每组包含13至14所学校。这远低于建议的20个。

**报告随机分配结果**

并非所有期刊文章都会报告随机分配检查的结果，但这一做法越来越普遍，尽管仍存在争议。政治科学领域已经制定了指南，要求报告是否采用了随机分配，以及随机化的单位，并提供基线均值和标准差的表格或文本。

i. ii有关此详细讨论，请参阅Mutz和Pemantle的论文。

iii还有一个重要的问题是，分配单位必须与统计检验中的分析单位相同。例如，如果将学校或教室作为聚合单位分配给不同的处理组，则统计分析应基于聚合级别而不是个体级别。在一个假设性的研究中，有30所学校的2,500名学生，其中15所学校被分配到处理组，另外15所学校被分配到对照组，总样本量是30，而不是2,500。否则，由于过高的样本量，结果的精确性将被夸大。

iv要了解争议的情况，可以阅读Gerber等人（2014）的论文，以及Mutz和Pemantle（2016）对其的挑战以及Gerber等人（2016）的回应：Alan Gerber等人，“Reporting Guidelines for Experimental Research: A Report from the Experimental Research Section Standards Committee,” Journal of Experimental Political Science 1, no. 1 (2014): 81–98；Alan S. Gerber等人，“Reporting Balance Tables, Response Rates and Manipulation Checks in Experimental Research: A Reply from the Committee That Prepared the Reporting Guidelines,” ibid. 2, no. 2 (2016): 216–29；Diana C. Mutz和Robin Pemantle，“Standards for Experimental Research: Encouraging a Better Understanding of Experimental Methods,” ibid.

特定变量的偏差。为了监测随机分配的结果，研究人员可以使用均值和标准差或设计用于检测差异的统计量，如t检验、卡方检验和方差分析(ANOVA)，来检查组之间在重要变量上的等同性。这些结果会被报告在一个表格中。例如，如果性别对结果变量很重要，那么需要检查各组男女人数是否相等；这是一个情况，没有显著差异是件好事。大多数随机分配报告只需要基本信息。例如，下面是一个测试健康信息生动效应的实验中的报告方式：

对参与者的性别、年龄、性行为等进行了两因素方差分析，交叉两个操纵变量（"信息生动性"和"论点强度"）。结果显示，在四个实验条件中，参与者之间没有显著差异(p > .05)。因此，随机分配似乎是有效的。

有些作者会进一步展示。例如，一项研究包括了一个表格，展示了各组的独立变量的描述统计数据，但没有显著性检验（见图7.3）。以下是该研究中包含的叙述和表格：

在表1中报告了各个实验条件下独立变量的均值。如表1所示，所有独立变量的均值和标准差没有实质性的差异。两个条件下的性别比例也是相等的。由于在操纵之前收集了独立变量的数据，这表明随机分配确实导致了相等的组。

何和麦克劳德在这个表格中展示了随机分配在各个变量上得到了相等的组。

以下是另一项关于辅导的研究的示例，该研究包括了随机分配的显著性检验结果，并包含了一个表格：

为了测试学生在这些背景指标上是否随机分布，计算了各组的均值并进行了t检验，以检验处理组和对照组之间是否存在显著差异。显著性检验表明，在所有的背景指标中，处理组和对照组之间没有统计上显著的差异：随机分配的结果确保了处理前的等同性。

在研究完成后，如果存在较大的样本减少或主动退出情况，报告随机分配检查是很重要的，以确保退出者并没有对结果产生影响。

A screenshot of a graph

Description automatically generated

当研究中有相当数量的被试退出或主动放弃时，需要报告随机分配检查的结果，以确保退出者与留在研究中的被试没有显著差异。有时候，退出的原因可能是实验本身。实验可能太长、无聊或困难，导致被试选择退出。特别是在长时间的实验中，这个问题尤为突出。

**当随机分配失败时**

随机分配是目前为止实现实验组等效性的最佳方法，但并不能保证各组在每个个体差异变量上完全匹配。它能最小化混淆，而不是完全消除。由于机会因素，仍可能存在差异的出现。并不需要在每个可能的变量上实现组间等效；最重要的是那些与结果变量相关的变量。例如，在道德判断研究中，年龄和教育与道德判断高度相关，而性别则不是。因此，在年龄和教育上使各组等效更为重要，而对男女人数的等效性则不需要过多关注。当组间在重要特征上不等效时，内部效度降低，研究人员可能会做出错误的推断。然而，随机分配机制未能实现等效性的情况是罕见的。此外，显著性检验的α水平已经考虑到了由于机会因素导致某些变量在各组之间不均匀分布的事实。没有办法“修复”真正的随机分配失败，除非重新开始并正确地重新进行随机分配。A screenshot of a graph

Description automatically generated

以下是实现等效性的一些建议：

• 如果可以在所有组中包括所有被试而不产生carry-over效应（即使用组内设计），那么等效性就不是一个问题，因为完全相同的人在每个组中。

• 如果一个组中有两倍多的被试在某个无关变量上存在问题，与另一个组相比，则认为这些组是不等效的65，例如，如果一个组中男性人数是女性人数的两倍，而性别与结果相关；在一个组中，有十个男性和五个女性被视为不等效。

• 首先确保每个组中有足够的被试。在第8章中解释了功效分析的方法，可以使用它来确定需要多少被试。被试数量越多，实现等效性的机会就越大66。在使用多个试次的研究中，可以招募更多被试并随机分配，然后重新进行实验以实现等效性67。

• 一个很好的经验法则是每个组至少有二十个被试68。但即使是这样，少于二十个被试的组实际上也不容易出现错误的不等效结论，因为统计检验在样本数量较少时更难检测到虚假差异69。

• 可以重新进行随机分配过程，直到实现等效性为止70。显然，这仅在治疗开始之前可以检查到随机分配时才有效。

* 如果在研究进行后才能检测到不等效性，一个广泛采用的策略是在统计分析中使用协方差分析，将不等效变量作为协变量。71 这将有助于在测试组间差异之前减少该变量的影响。例如，如果治疗组中男性人数是对照组的两倍，并且性别预计会影响结果，使用性别作为协变量将有助于平衡两组的影响。也就是说，它有助于在测试治疗效果之前消除由于性别效应导致的未解释变异，从而获得更精确的估计值。72 但是，这种方法应该保守使用，只有在事先计划好的情况下才应使用协变量（下一章将详细讨论），因为这不会“控制”真正的随机分配的缺失。而且，包含大量协变量会破坏良好设计、受控实验的目的。73

研究人员不应有意地招募更多具有所需特征的被试，并将它们仅分配给具有该变量较低水平的组74。这不是随机的。研究人员也不应在给予治疗前将被试调整分配以平衡组别75。这也不是随机的。

最后，令人欣慰的是，如果组别在某些变量上不等效，差异很可能代表系统性误差，而不是随机错误，(a2nd0a2re3un)。此外，重复研究有助于纠正因不等效而产生的任何错误结论；有关更多信息，请参阅第5章。在多个研究中，事实往往会占上风。77

在等效性可能存在问题的领域，例如教育、社会工作、犯罪学和项目评估等领域，可能需要投入大量时间来实现组别的等效性。例如，在一项关于学校药物项目的研究中，78 在随机分配完成并检查等效性后，有两所学校退出了研究，导致组别不等效。研究人员不得不重新选择学校，进行随机分配，并重新检查等效性。

**分组、配对和其他策略**

降低组别不等效的一种方法是采用配对和分组的策略。这涉及到将被试根据重要变量进行配对，然后将它们作为一对或一组分配到治疗组和对照组。79 这是为了确保与结果变量相关的外生或干扰80变量不会混淆结果。组别是根据具有相同水平的分组变量的被试来创建的。例如，如果性别是分组变量，将按性别随机配对——一男一女，然后将每对随机分配到治疗组或对照组。81

在一项关于导师计划效果的企业研究中，由于雇主想要选择具有最大晋升潜力的员工参加计划，研究人员无法将被试随机分配到条件组。82 相反，该研究采用了配对设计，根据薪水相似性、绩效评级、组织工龄、在同一办公室工作以及之前未参加导师计划等五个特征，选择了与治疗组员工相似的对照组员工。该研究报告了统计检验结果，显示治疗组和对照组在这些变量上没有差异，同时注意到“治疗组和匹配对照组可能在未观察到的特征上有所不同。”83

实验者必须在进行研究之前预先预测和测量变量，因此，对于未知的混杂变量，匹配是无助的；基于这个原因，简单的随机分配仍然更可取。

分组也可以在多个变量水平上进行。例如，如果年龄是希望确保在各组间等效的变量，可以创建不同年龄组的分组；例如，对于年龄组分为18至34岁、35至54岁、55至64岁和65岁及以上的四个分组。在创建这些分组后，每个分组中的被试将随机分配到治疗组和对照组。现在，年龄不能成为实验组间任何差异的原因。

分组不如简单的随机分配可取，但在可能只有很少潜在被试可用或样本量较小的情况下很有用，例如在以学校为单位的研究中。在教育研究中经常使用匹配策略，将学校根据重要特征进行匹配，然后随机分配到各个条件组。84 另一个例子来自一项关于成长环境中是否有不同族裔朋友对刻板印象的影响的研究。85 研究者预计很难找到在多元化邻里中长大的被试，因此在随机分配到治疗组和对照组之前，他对实验志愿者进行了预测试，测量了他们与少数民族的个人接触水平，并将高接触组和低接触组进行了匹配。

在随机分配不太可能被正确实施的情况下，分组和匹配策略也可以很有用，例如在程序评估中，研究者无法控制分配。86

**分组与简单随机分配**

分组与简单随机分配相比，匹配策略并不可取，尽管从直观上具有吸引力。88 首先，当分组因素对结果变量的影响不大时，统计检验的效果会降低。89 分组是一种控制混杂变量的有力策略，但其实施成本较高且对统计分析的复杂性有一定要求。对于大多数实验研究来说，简单随机分配是实现组别等效性的首选方法，因为它能够消除混杂变量的影响，并确保在不同组别间获得可比较的结果。

更常见于生命科学中研究植物和动物。90 另一个缺点是区组随机分配需要一个两步骤的过程，首先对区组因素进行测量，然后将个体随机分配到组中，进行处理，最后进行结果测量。社会科学专家一致认为，简单随机分配是实现可比较组的首选方法。91 即使是对随机分配最严厉的批评者也没有提出其他替代方案。92 Campbell 和 Stanley 尤其批评了“对匹配的普遍和错误偏好...因为匹配并不能真正帮助克服最初的组别差异”。当混淆变量未知且无法控制时（有时称为潜在变量），随机分配是最佳策略，因为它能自动平衡这些变量。94

**分层随机分配**

一种简化对多个变量进行等价性评估的技术是分层随机分配，其中多个变量被合并为一个单一变量，类似于因子分析中创建的因子。95 这种方法不是通过特定变量进行分组或匹配，而是将许多相关变量合并为一个综合因子，在个体随机分配后对其进行等价性测量。换句话说，它允许研究人员在许多离散变量上进行综合考虑。例如，一项研究96 使用了七个关键变量，如学校所在社区的类型、学校的年级数、白人学生的百分比、每个年级的入学人数以及农村或城市环境。然后使用统计技术得到一个综合的分层变量。这样做是为了找到在这些学校中相关性较高的变量组合，作者称之为“农村性”，并解释说农村学校在这些特征上通常相似。在随机分配后，等价性是通过单一综合因子而不是七个单独变量进行检查的。实际过程比这里描述的更复杂。97 作者使用差异的方差分析 (ANOVA) 测试发现了等价性，并在表格中报告了结果 (见图 7.6)。这种分层过程的优点在于未知或未测量的变量在组间随机分布，而区组随机分配和匹配策略则无法实现这一点。98 与实现等价性一样，这种策略仅适用于与结果变量高度相关的变量，而不适用于每个变量。99 这种技术与区组随机分配类似，在某些学科中更为常见，因此了解您所在领域的标准非常重要。

**其他事项的随机分配**

到目前为止，本章主要关注将个体主体随机分配到条件中。但是随机分配不仅仅适用于个体分配到组的方式。事实上，专家建议尽可能随机分配研究中的许多步骤。101



**FIGURE 7.5 RANDOM ASSIGNMENT CHECK**

在解释随机分配等价性检查的文本中提到：“首先，我们验证了分配程序在与 CIS 分数相关的学校层面变量上提供组别等价性的有效性。我们进行了一个简单的方差分析 (SAS Proc GLM)，以程序（农村、经典、对照）作为类别变量。表3展示了整体F检验的结果。正如预期的那样，程序组成员与 CIS 分数、构成 CIS 分数的两个因子以及构成这些因子的七个项目无关。”100

|  |  |  |  |
| --- | --- | --- | --- |
|  | **Tutored students** | **Control students** | **Total sample** |
| Average reading grade, 1997-98 year-end (A = 4.0) | 1.65 | 1.61 | 1.63 |
| Average math grade, 1997-98 year-end (A = 4.0) | 1.79 | 1.85 | 1.82 |
| Average SAT-9 reading open-ended national percentile score, 1998 | 23.1 | 23.2 | 23.2 |
| Average SAT-9 math open-ended national percentile score, 1998 | 16.7 | 18.6 | 17.6 |
| Full year attendance rate, 1997-98 | 91.8 | 90.6 | 91.2 |
| % of students not promoted, 1997-98 | 9.2 | 9.0 | 9.1 |
| % of students African American | 95.9 | 96.8 | 96.4 |
| % in home receiving weFlfaroe rassRisteanvcei(eTAwNF-) No Commercia | l U63.s3 e(2 | 0620.93) | 62.1 |
| % with guardian with a high school degree or more | 68.6 | 74.5 | 71.5 |
| % with guardian currently working for pay | 57.2 | 51.9 | 54.6 |
| % with guardian reporting health problem that limits activity | 19.5 | 23.2 | 21.3 |
| % in home with both mother and father | 37.4 | 31.1 | 34.4 |
| % reporting that someone helps with homework | 73.2 | 78.4 | 75.7 |
| % reporting that someone at home reads with them | 67.0 | 69.5 | 68.2 |
| Average number of children in household | 3.39 | 3.02 | 3.20 |
| **Total study sample** | **196** | **189** | **385** |

1. The welfare assistance data were derived from the School District of Philadelphia student information system.
2. The background data related to student guardians and the numbers of children per household were derived from the baseline survey completed by the guardians who gave parental consent for the student to participate in the tutoring programme (September 1998).
3. The figures on household composition and help with reading and homework were derived from the tutee follow-up survey administered to tutees at programme completion in May 1999.

沙迪什（Shadish）、库克（Cook）和坎贝尔（Campbell）故意使用“单位”的随机分配一词，以免仅仅暗示只有人员可以进行随机分配。102 任何可能引入系统偏差的因素都应该进行随机分配。例如，如果有多个实验员

|  |  |  |
| --- | --- | --- |
| **FIGURE 7.6**  **TESTING FOR RANDOM ASSIGNMENT TABLE** | | |
| **Variable** | *F***(2,38)** | *p* |
| CIS | 0.25 | .78 |
| Factor 1 | 1.05 | .36 |
| Factor 2 | 0.08 | .93 |
| Rurality | 0.45 | .64 |
| Npergrade | 1.62 | .21 |
| Numgrades | 0.41 | .67 |
| Pctwhite | 1.60 | .21 |
| Pctlunch | 0.10 | .90 |
| Scores | 0.22 | .81 |
| Rdrugs | 0.80 | .46 |

Note: Post hoc tests with the Duncan test showed no significant differences (*p*.10) for any individual comparisons.

如果有多个实验员参与研究并进行观察和测量，应该对实验员进行随机分配，以确定他们将监督的会话和条件。103 观察和测量受试者的实验员可能通过练习而变得更好，或者相反，可能因疲劳而在测量方面表现较差。重要的是，不应该将实验员分配为先观察控制组中的所有受试者，再观察治疗组中的所有受试者（或反之亦然），以避免引入组之间的系统差异。相反，实验员应该被随机分配到每个组以及每个会话中。

通常，当有多个刺激物（例如广告、新闻报道和健康信息）时，这些刺激物应该以随机顺序呈现给受试者。在一项关于政治家决策的研究中，研究人员随机了许多事物。104 除了将受试者随机分配到条件中，他们还随机分配了三种类型的任务以及每种任务中受试者需要做出的十三个决策。该研究还以慈善捐款的形式给予受试者激励，并且激励的方式也是随机的；以下是解释：

为了使每个涉及货币回报的决策都具有相关性，但同时确保任务之间不相互影响，我们随机选择了一个任务，以确定代表参与者向慈善机构捐赠多少钱。具体而言，我们随机选择了实验的抽奖选择部分或抽奖估值部分，并随机从该部分选择了一个任务。这样避免了参与者的选择受到所谓的投资组合效应的影响（例如为平衡投资组合而进行一些保守的选择和一些冒险的选择），或者受到之前的收益的影响。105

这项研究采用了被试内设计，但由于两个条件（收益框架、损失框架）中的情境相似，研究人员不希望受试者在两个条件下阅读每个情境，因此他们随机分配每个参与者阅读一半的情境在每个条件中。他们这样解释：

对于每个情境，随机确定参与者是接收损失框架还是收益框架，因此有可能一个参与者在一个情境中接收收益框架，在另一个情境中接收损失框架。我们还随机确定了情境的顺序。106

显然，这些研究人员遵循了Bausell的建议：“当存在疑虑时，进行随机分配。”107

**交叉平衡**

随机分配或改变顺序的原因是为了避免传递效应。在第6章中，我们描述了由学习、练习、疲劳或受试者变化引起的效应。当接受一种处理影响受试者对下一种处理的反应时，就会发生传递效应。一种特殊的传递效应源于事物呈现的顺序，称为顺序效应。顺序效应在情境和记忆研究中得到了充分的证明，我们更容易记住首尾所接触到的内容。这对于每个受试者都接受所有不同处理的被试内设计来说是一个特殊的关注点，因为他们特别容易出现疲劳、练习效应、传递效应和顺序效应。108 交叉平衡是有帮助的，因为通常一个方向上的传递效应会抵消另一个方向上的效应。例如，有些受试者在最后一个处理上表现更好（后效应），而其他受试者在第一个处理上表现更好（前效应）。当数据聚合时，这两个效应互相抵消。对于所有随机分配的事物也是如此；例如，如果有多个实验员将监督研究的多个运行，不仅应将不同的实验员随机分配到治疗组和对照组，还应在会话之间进行轮换或交叉平衡。109 交叉平衡的目标是使顺序效应在不同条件下等效。和所有等效的事物一样，不进行顺序交叉平衡会降低内部效度。

**拉丁方设计**

交叉平衡可以通过简单地随机分配来实现，但在实验研究中还有一种特定的交叉平衡策略，称为拉丁方设计，它可以使每个刺激出现的位置数量相等。它确保每个实验消息或刺激物在第一个位置出现一次，在最后一个位置出现一次，并在每个中间位置出现一次。此外，每个条件或刺激物之间只有一次连续。这比简单的随机分配更高效。110 以下是它的工作原理示例：一个假设性的实验使用了

只有三个不同的广告作为刺激物。如果研究者将广告随机分配到顺序中，会有六种可能的顺序组合

|  |  |  |
| --- | --- | --- |
| A | B | C |
| A | C | B |
| B | C | A |
| C | B | A |
| C | A | B |
| B | A | C |

但使用拉丁方块法可以产生三种组合：

|  |  |  |
| --- | --- | --- |
| A | B | C |
| B | C | A |
| C | A | B |

每个广告首先显示一次，最后一次显示一次，中间一次显示。拉丁方块的两个关键特点是，行或列中不会重复出现相同的字母，并且每一行和每一列都包含所有不同的字母。这比简单的随机分配顺序要高效，因为它减少了需要接收每个顺序的受试者的数量。在大型因子设计中，拉丁方块特别有用，因为要进行所有可能的顺序组合的实验是非常昂贵的。通过随机选择代表A、B和C的刺激来实现这种类型的均衡。接下来，通过将第一个位置的刺激移到每一行的最后一个位置，其他刺激向前滑动一个位置来进行简单的旋转。这适用于三个或更多的刺激；显然，对于两个刺激，只有两种可能的顺序组合。

在报告拉丁方块时，与报告随机分配一样，通常只会简单提及。例如，在这项关于公益广告（PSA）的健康研究中，作者说：“PSA的呈现顺序根据一个平衡的拉丁方块进行了均衡处理。”

在另一项关于电视新闻编辑决策的研究中，作者描述了三个新闻情节的轮换方式：“参与者接收的新闻故事顺序使用拉丁方块设计进行了均衡处理。”

在使用多个因素的研究中，拉丁方块有助于避免因素之间的混淆。例如，在一项关于广告中使用幽默的研究中，作者让受试者收听播放不同广告的广播节目。构成每个广告的因素包括产品类型（例如，谷类食品、奶酪、电池）、品牌名称和笑话，他们称之为一句话。以下是他们描述的均衡策略：

广播节目的三个版本的轮换设计是为了确保在研究中，某个幽默话语不会与特定品牌名称或产品类型同时出现超过一次。在每种组合中，产品类型、品牌名称和幽默话语会在三个音频带的呈现顺序中进行旋转。为了安排三个解说词的三个轮换，使用了拉丁方块的均衡技术（Keppel, 1991）。

拉丁方块设计的名称源自于一个古老的谜题，即拉丁字母在方块中的不同排列方式。它作为一种旋转实验被引入，并由R. A. Fisher推广，自那以后一直是心理学中首选的方法。

**随机分配的抵制**

大多数研究人员很快就会被随机化解决各种问题的能力所说服，但对于可能需要参与研究的非研究人员来说，情况通常并非如此。进行项目评估或在实际或现场环境中进行研究的学科可能会遇到对随机分配的抵制。例如，在一项研究中，学校管理者允许一些学生绕过随机分配过程。例如，教育研究人员发现，学校人员可能会反对为一些学生提供治疗机会，而不让其他学生享受该机会。商业研究人员发现，高管可能拒绝让他们的员工随机分配，而坚持自己挑选每个组的人员。在一项健康研究中，“相当多的”受试者拒绝参与随机分配，因为他们不想参与，如果不能确保接受潜在有益的治疗。伦理关切也引发了对随机化的反对意见；例如，在一项治疗犯罪受害者的实验中，一些表现出自残倾向的受试者被从对照组移动到治疗组，破坏了随机分配。这可能导致研究人员放弃真正的实验而采用准实验的方式，这在第4章中有所讨论。关于这个话题的详细讨论可以参考7.3附录。

尽管随机分配在各种形式中是一种优雅的过程，但它并不完美。但正如Campbell和Stanley所说：“尽管如此，这仍然是唯一的方法，也是必要的方法。”在对如何分配受试者（和其他事物）有了坚实的理解之后，下一章将介绍如何对受试者进行抽样以及如何确定最佳数量。

## 更多内容……第7.3节

## 抵制随机分配

关于随机分配，遵从是至关重要的。128就像怀孕一样，不存在“有点随机”的情况，它是一个非此即彼的选择。非研究人员经常忽视这个区别；例如，在一项研究中，当告知随机分配受到妥协时，实施的工作人员表示惊讶，并回答他们认为分配实际上是“或多或少地随机的”。130直接参与该过程的一名员工甚至认为这些指示只是“建议”。131

随机分配出错的另一个例子涉及一名退出项目的被试犯下的谋杀案，这导致检察官拒绝拒绝对任何需要的人进行治疗。132并非所有对随机分配的威胁都如此戏剧性。任何损害随机化的行为都可能导致治疗组和对照组之间出现非随机差异，从而破坏实验的内部效度。下表中，第一列列出了从实际工作者那里经常遇到的关于随机分配的一些常见异议，第二列则提供了克服这些异议的建议。已经报告使用这些技术的研究人员发现，随机化的遵从性从最低19%增加到94%。133这里的建议来自教育、咨询、刑事司法和商业领域，但适用于任何学科中在实地环境中评估干预措施的尝试。

最好的方法是研究人员坚持自己进行随机分配，并在研究人员的场所而非研究场所进行该过程。134这应该始终与现场工作人员进行沟通，让他们提出关切、回答问题并参与设计过程。135监控分配过程也是至关重要的。已知被试曾试图潜入未被分配的组。136此外，观察干预措施实际执行的程度也很重要；例如，如果教师应该使用技术，要检查他们实际使用技术的程度。137

A screenshot of a computer

Description automatically generated

A screenshot of a computer

Description automatically generated

A screenshot of a cell phone

Description automatically generated

## 研究亮点 7.4

充分利用自然环境中的随机分配

**来源：Abrams, David S. 和 Albert H. Yoon. 2007. “The Luck of the Draw: Using Random Case Assignment to Investigate Attorney Ability.” University of Chicago Law Review 74 (4) (秋季): 1145–1177.**

该研究利用了自然发生的随机分配的机会，当研究人员发现内华达州的一个县将入境的重罪案件随机分配给律师池中的律师，从而实现了一个不受选择偏差影响的自然实验。

克拉克县（包括拉斯维加斯）在一名被告的死刑被推翻后开始将律师分配给案件，因为他被分配给了一位经验不足的公共辩护律师。在先前的非随机分配方法下，更好的律师可能会被分配到更困难的案件中，从而将律师的能力与案件难度混淆。这一机会使得研究人员能够研究律师的表现，而这无法通过案件特征来解释。传统观点认为，毕业于更好的法学院的律师可能会使客户获得更轻的刑罚。该研究发现，西班牙裔律师和经验更丰富的律师为客户取得了比其他人更好的结果，但性别和所就读的法学院没有影响。

对于本书的目的来说，该研究非常重要，因为它深入展示了随机分配的价值。

研究人员首先检查确保随机分配确实得到了实施而没有受到阻碍。这有助于排除其他对审判结果的解释，如案件的难度。案件被分配给律师时，法官、检察官或团队负责人都不知道被告的任何特征，甚至不知道所指控的犯罪行为，这有助于防止任何下意识地有目的性地分配案件。研究人员使用非参数检验（卡方检验）通过被告的年龄、性别和种族来检查案件是否确实是随机分配的。他们解释说，这三个变量与被告在未观测变量上的其他特征高度相关。他们说：“至关重要的是，我们假设这为证明未观察到的变量也是随机分配提供了证据（因为与可观察到的变量相关）.”181

|  |  |  |
| --- | --- | --- |
| **TESTING FOR RANDOM ASSIGNMENT** | | |
| **Case characteristic** | *p***-value** | **Observations** |
| Defendant sex | 0.851 | 10,129 |
| Defendant age | 0.253 | 9,803 |
| Defendant race | 0.098 | 7,145 |

Note: Each row reports results from a separate simulation to test for the equality of public defender fixed effects. Defendant sex is a dummy variable for whether the defendant is male. Defendant race is 0 for black defendants and 1 for white defendants.

1. 在实验中，当参与者被随机分配到治疗组或对照组，被称为 。

a. 随机抽样

b. 随机分配

**检验你的知识**

• 没有将受试者随机分配到各组，或者没有正确地进行随机分配

• 没有将研究中的其他要素（例如实验员）随机分配到各个实验组

• 没有对刺激进行平衡处理

**常见错误**

非显著结果显示这些案件确实是随机分配的（参见图7.6）。这样使得研究人员能够测试关于律师能力差异和其他预测被告辩护结果更好的变量的假设。

这些研究人员在寻找自然实验的机会时具有创造性。与大多数期刊文章一样，这篇文章没有提到研究人员在收集数据之前获得机构审查委员会（IRB）的批准。所使用的数据可能因为是公开记录的一部分而被免除知情同意，或者研究人员可能在事后进行了知情同意。

新研究人员有时会认为，由于数据已经收集，这构成了“二次数据”，就不需要获得IRB的批准。真正的二次数据是指被试已经接受了知情同意，例如在现有数据集（如美国国家选举研究或皮尤民意调查）中。收集这些数据集并向研究人员提供数据集的组织已经从IRB的人类受试者委员会获得了许可，并为被试提供了知情同意。但并非所有现有数据都是如此。例如，填写关于对课程满意度的评价的学生并未收到知情同意文件中包含的信息。在涉及从人类受试者收集信息的任何研究中，研究人员应与其IRB联系，以确定是否需要获得IRB的批准和被试的知情同意，即使数据已经收集。第11章将更详细地讨论知情同意和IRB的批准。研究人员应当对自然发生的随机分配机会保持敏感，但也应当小心确保获得IRB的许可并遵守获得数据使用人的知情同意的协议。

1. Designing Experiments for the Social Sciences

c. 随机误差

d. 选择偏倚

2. 在实验中使用随机分配的主要原因是确保以下哪一点？

a. 一个代表人群的样本

b. 参与者和实验者都不知道某个人属于哪个组

c. 组之间在已知和未知变量上尽可能具有等价性

d. 因变量在不同条件下没有差异

3. 以下哪个不是将受试者随机分配到不同组的方法？

a. 从帽子里抽名字

b. 抛硬币

c. 使用随机数发生器

d. 通过移动受试者使各组人数平衡（例如，1、2、1、2……）

4. 组之间需要在所有可测量的变量上具有等价性。

a. 真

b. 假

5. 确保系统性变异不混淆研究的首选方法是什么？

a. 预测试

b. 阻塞或匹配

c. 简单随机分配

d. 在随机分配后通过调整受试者来平衡组别

6. 除了受试者外，还应该随机分配什么？

a. 没有，只有受试者

b. 实验者

c. 所有研究的程序

d. 除A之外的所有选项

7. 拉丁方是用于 。

a. 创建等价的组

b. 控制杂项个体变量

c. 最小化顺序效应

d. 将受试者随机分配到组中

Chapter 7 ■ Random Assignment **203**

8. 完成以下内容以形成拉丁方：

A B C D

9. 如果男性和女性被配对，并作为一对分配到治疗组或对照组，这被称为

a. 阻塞或匹配

b. 随机分配

c. 分层随机分配

d. 拉丁方

10. 随机分配的一个缺点是实验者必须预先预测和能够测量混淆变量，因此对于未知的混淆变量没有帮助。

a. 真

b. 假A screenshot of a white background

Description automatically generated

1. 使用在第一章中开始制定的实验研究，并在第五章中创建的设计表和对照组，决定如何随机分配受试者到组中。撰写一页关于将使用何种策略以及原因的内容——随机数生成器、抽签等。进行“测试运行”。假设研究中每个组有40名受试者。使用所选的随机分配方式将受试者进行分组。分析结果；它们是否均衡？重复此过程，尝试不同的随机分配方式，观察结果是否保持均衡。

**应用练习**

**204** Designing Experiments for the Social Sciences

使用不同类型的随机工具进行实验，观察结果如何（例如，如果先使用在线随机工具，再使用抽签的方式）。

1. .写一页关于研究中可以进行随机分配的所有要素。除了研究对象外，还有哪些步骤可以进行随机分配？为什么？假设你的研究中将使用至少三种刺激物（例如，三种不同的治疗方式、干预措施、教学技术、广告、公益广告、故事、信息等）。设计一个计划来进行平衡处理。
2. 写一页关于可能引起混淆的变量，需要通过随机分配来使各组之间平衡。阅读与你的因变量相关的文献，看看其他人发现与你的因变量高度相关的变量。运用想象力和常识，尽可能多地识别出这些变量。

A screenshot of a computer

Description automatically generated

Chapter 7 ■ Random Assignment **205**

**注释**

1. W. A. McCall，《教育实验方法》（纽约：麦克米兰出版社，1923年），第41页。

2. R. Barker Bausell，《有意义的实验：成为科学家的40个步骤》（千橡市，加利福尼亚州：Sage出版社，1994年）。

3. Richard A. Berk，《随机实验作为铜标准》，《实验犯罪学杂志》1卷（2005年）：416-433页。

4. Graeme D. Ruxton和Nick Colegrave，《生命科学的实验设计》，第三版（英国牛津：牛津大学出版社，2011年）。

5. Bausell，《有意义的实验》。

6. D. T. Campbell和J. C. Stanley，《研究的实验和准实验设计》（芝加哥：兰德·麦克纳利出版社，1963年）。

7. Charles Sanders Peirce和Joseph Jastrow，《论感知的微小差异》，《国家科学院备忘录》第3卷，第75-83页（1885年）：8。

8. William R. Shadish、Thomas D. Cook和Donald T. Campbell，《推广因果推断的实验和准实验设计》（加利福尼亚贝尔蒙特：Wadsworth Cengage Learning，2002年）。

9. J. M. Gueron，《随机分配的政治：实施研究和影响政策》，收录于《证据至关重要：教育研究中的随机试验》，F. Mosteller和R. Boruch编辑（华盛顿：布鲁金斯学会出版社，2002年），第26页。

10. Diana C. Mutz和Robin Pemantle，《实验研究的标准：鼓励更好地理解实验方法》，《实验政治科学杂志》2卷，第2期（2016年）：192-215页。

11. Martin T. Orne，《需求特征和准对照概念》，收录于《行为研究中的人工制品：罗伯特·罗森塔尔和拉尔夫·L·罗斯诺的经典著作》，Robert Rosenthal和Ralph L. Rosnow编辑（牛津：牛津大学出版社，2009年），第110-137页。

12. Cathaleene Macias等，《随机分配中的偏好：对随机试验解释的影响》，《管理与政策心理健康与心理健康服务研究》36卷，第5期（2009年）：331-342页。

13. T. D. Cook和D. T. Campbell，《准实验研究：用于实地环境的设计和分析问题》（芝加哥：兰德-麦克纳利，1979年）。

14. Colin Ong-Dean、Carolyn Huie Hofstetter和Betsy R. Strick，《教育研究中实施随机分配的挑战和困境》，《美国评估杂志》32卷，第1期（2011年）：40页。

15. Macias等，《随机分配中的偏好》，331-342页。

16. 同上。

17. D. F. Halpern，《思维与知识：批判性思维导论》，第4版（新泽西马霍，Erlbaum，2003年）；M. Patricia King等，《参与者和医生干预偏好对随机试验的影响：系统综述》，《美国医学协会杂志》293卷，第9期（2005年）：1089-1099页。

18. Macias等，《随机分配中的偏好》，331-342页。

19. 同上。

20. 同上。

21. Sameer B. Srivastava，《网络干预：评估正式指导对工作场所网络的影响》，《社会力量》94卷，第1期（2015年）：427-452页。

22. Shadish, Cook和Campbell，《广义因果推断的实验和准实验设计》。

**206** Designing Experiments for the Social Sciences

23. Cengiz Erisen, Elif Erisen和Binnur Ozkececi-Taner，《政治心理学研究方法》，《土耳其研究》13卷，第1期（2013年）：17页。

24. Diana Mutz，《基于人群的调查实验》，（普林斯顿，NJ：普林斯顿大学出版社，2011年）；

Steven L. Nock和Thomas M. Guterbock，《调查实验》，收录于《调查研究手册》，P. V. Marsden和J. D. Wright编辑（Bingley，英国：Emerald Group Publishing Limited，2010年），837-864页。

25. Erin C. Cassese等，《社交媒体互联网调查：招募在线实验参与者》，《政治科学与政治》46卷，第4期（2013年）：1-10页；James N. Druckman，《选民启动：美国参议院选举中的竞选效应》，《政治心理学》25卷，第4期（2004年）：577-594页；Julio J. Elias、Nicola Lacetera和Mario Macis，《市场与道德：一项实验调查研究》，《公共科学图书馆综合》10卷，第6期（2015年）：1-13页。

26. Erisen, Erisen和Ozkececi-Taner，《政治心理学研究方法》，17页。

27. Scott Clifford和Jennifer Jerit，《方便是否有成本？实验比较实验室和在线研究数据质量》，《实验政治科学杂志》1卷，第2期（2014年）：120-131页。

28. Rebecca B. Morton和Kenneth C. Williams，《实验政治科学与因果关系的研究：从自然到实验室》，（纽约：剑桥大学出版社，2010年）。

29. Jona Linde和Barbara Vis，《政治家是否像我们其他人一样冒险？议员心理学前景理论的实验检验》，《政治心理学》38卷，第1期（2017年）：101-117页。

30. Clifford和Jerit，《方便是否有成本？》，120-131页。

31. Cassese等，《社交媒体互联网调查》；Kevin J. Mullinix等，《调查实验的普适性》，《实验政治科学杂志》2卷（2015年）：109-138页。

32. Renita Coleman和Lee Wilkins，《新闻工作者的道德发展：与其他职业的比较和高质量道德推理的模型》，《新闻与传媒季刊》81卷，第3期（2004年）：511-527页。

33. Renita Coleman和Lee Wilkins，《寻找道德新闻工作者：新闻从业人员道德发展的探索性研究》，《大众传媒伦理学》17卷，第3期（2002年）：209-255页；Renita Coleman和Lee Wilkins，《新闻工作者的道德发展》；R. Coleman和L. Wilkins，《公共关系从业人员的道德发展：与其他职业的比较》，《公共关系研究》21卷，第3期（2009年）：318-340页。

34. Renita Coleman，《少数族裔新闻工作者的道德判断：来自亚裔美国人、黑人和西班牙裔职业新闻工作者的证据》，《大众传播与社会》14卷，第5期（2011年）：578-599页。

35. Erisen, Erisen和Ozkececi-Taner，《政治心理学研究方法》，17页。

36. Kate E. West，《谁在做决策？电视记者、他们的老板和顾问驱动的市场研究》，《广播与电子媒体杂志》55卷，第1期（2011年）：27页。

37. Jun Myers，《在预防健康信息中追踪“生动效应”：论证质量对信息生动性有效性的调节作用》，《促销管理杂志》20卷，第5期（2014年）：634页。

38. Mutz和Pemantle，《实验研究的标准：鼓励对实验方法的更好理解》，《实验政治科学杂志》2卷，第2期（2016年）：192-215页。

39. Bausell，《有意义的实验进行》，40步成为科学家（千橡市，加利福尼亚州：Sage，1994年）。

40. Mark L Davison和Stephen Robbins，《道德发展的客观指标的可靠性和有效性》，《应用心理测量学》2卷，第3期（1978年）：391-403页；James R. Rest，《道德问题判断的发展》（明尼阿波利斯，明尼苏达州：明尼苏达大学出版社，1979年）。

41. Bausell，《有意义的实验进行》。

42. Christopher H. Rhoads和Charles Dye，《两级随机分配和回归不连续研究的最优设计》，《实验教育杂志》84卷，第3期（2016年）：421-448页。

43. Sameer B. Srivastava，《网络干预：评估正式辅导对工作场所网络的影响》，《社会力量》94卷，第1期（2015年）：427-452页。

44. D. Fergusson等，《随机分配后的排除：意图处理原则和排除病人的分析》，《英国医学杂志》325卷（2002年）：652-654页。

45. Howard Levene，《方差相等的鲁棒性检验》，收录于《概率和统计的贡献：为了纪念哈罗德·霍特林而编写的文章》，Ingram Olkin等编辑（斯坦福，加利福尼亚州：斯坦福大学出版社，1960年），278-292页。

46. Mutz和Pemantle，《实验研究的标准》，192-215页。

Chapter 7 ■ Random Assignment **207**

47. M. J. Strube，《随机分配的小样本失败：进一步研究》，《咨询与临床心理学杂志》59卷，第2期（1991年）：346-350页。

48. Shadish, Cook, and Campbell，《实验和准实验设计》。

49. John Graham等，《在药物抗性策略农村项目中将学校随机分配给组：一些新的方法学扭曲》，《预防科学》15卷，第4期（2014年）：516-525页。

50. 同上。

51. L. M. Hsu，《心理治疗结果研究中的随机抽样、随机分配和对比组的等价性》，《咨询与临床心理学杂志》57卷（1989年）：131-137页。

52. Alan Gerber等，《实验研究的报告准则：实验研究部门标准委员会的报告》，同上1卷，第1期（2014年）：81-98页。

53. Myers，《追踪“生动效应”》，636页。

54. Shirley S. Ho和Douglas M. McLeod，《面对面和计算机中介沟通中的社会心理影响对观点表达的影响》，《传播研究》35卷，第2期（2008年4月）：190-207页。

55. 同上，197页。

56. Gary W. Ritter和Rebecca A. Maynard，《使用正确的设计获得“错误”的答案？志愿辅导项目的随机分配评估结果》，《儿童服务杂志》3卷，第2期（2008年）：4-16页。

57. Merton S. Krause和Kenneth I. Howard，《随机分配的作用和不作用》，《临床心理学杂志》59卷，第7期（2003年）：751-766页。

58. 同上。

59. Bausell，《有意义的实验进行》；Hsu，《随机抽样》；Strube，《随机分配的小样本失败》，346-350页。

60. James R. Rest等，《后常规道德思维：新柯尔伯格理论方法》（新泽西州马霍，Erlbaum，1999年）。

61. Krause和Howard，《随机分配的作用和不作用》。

62. Mutz和Pemantle，《实验研究的标准》。

63. 同上。

64. 同上。

65. Hsu，《随机抽样》。

66. Strube，《随机分配的小样本失败》。

67. Graham等，《学校的随机分配》。

68. Hsu，《随机抽样》。

69. Strube，《随机分配的小样本失败》。

70. Bausell，《有意义的实验进行》。

71. 同上。

72. Donald P. Green和Daniel Winik，《利用随机法官分配来估计监禁和缓刑对药物违法者再犯率的影响》，《犯罪学》48卷，第2期（2010年）：357-387页。

73. Mutz和Pemantle，《实验研究的标准》。

74. Shadish, Cook, and Campbell，《实验和准实验设计》。

75. 同上。

76. Strube，《随机分配的小样本失败》，346-350页。

77. Krause和Howard，《随机分配的作用和不作用》；Strube，《随机分配的小样本失败》。

78. Graham等，《学校的随机分配》。

79. Campbell和Stanley，《实验和准实验设计进行研究》。

80. Murray Webster和Jane Sell，《社会科学实验室实验》（阿姆斯特丹：爱思唯尔，2007年）。

81. Bausell，《有意义的实验进行》；Campbell和Stanley，《实验和准实验设计进行研究》。

82. Srivastava，《网络干预》。

83. 同上，438页。

84. Graham等，《学校的随机分配》。

85. David Alan Free，《电视新闻中的刻板印象持续存在：异种接触对内容的影响》（未发表的博士论文，德克萨斯大学，2012年）。

86. E. W. Gondolf，《从成功和失败的随机分配测试施虐者计划创新中汲取教训》，《实验犯罪学》6卷（2010年）：355-376页。

87. Mutz和Pemantle，《实验研究的标准》。

88. Campbell和Stanley，《实验和准实验设计进行研究》。

**208** Designing Experiments for the Social Sciences

89. Ruxton和Colegrave，《生命科学的实验设计》。

90. 例如，见上述引文；Murray R. Selwyn，《生命科学的实验设计原则》（Boca Raton，FL: CRC Press，1996年）。

91. Campbell和Stanley，《实验和准实验设计进行研究》；Strube，《小样本随机分配的失败》。

92. Krause和Howard，《随机分配的作用和不作用》。

93. Campbell和Stanley，《实验和准实验设计进行研究》，第15页。

94. Selwyn，《生命科学的实验设计原则》。

95. Graham等，《学校的随机分配》。

96. 同上。

97. 同上。

98. 同上。

99. Bausell，《有意义的实验进行》。

100. Graham等，《学校的随机分配》，第521页。

101. 同上。

102. Shadish, Cook和Campbell，《实验和准实验设计》，第253页。

103. Bausell，《有意义的实验进行》。

104. Linde和Vis，《政治家是否像我们一样冒险？》

105. 同上，第107-108页。

106. 同上，第111页。

107. Bausell，《有意义的实验进行》，第78页。

108. Shadish, Cook和Campbell，《实验和准实验设计》。

109. Bausell，《有意义的实验进行》。

110. Roger Kirk，《实验设计：行为科学的程序》，第4版（洛杉矶，CA: Sage，2013年）。

111. Campbell和Stanley，《实验和准实验设计进行研究》。

112. Ruxton和Colegrave，《生命科学的实验设计》。

113. Shadish, Cook和Campbell，《实验和准实验设计》。

114. Kirk，《实验设计》。

115. 张觉满、张迪和T. Makana Chock，《HIV/AIDS公益广告对态度和行为的影响：感知威胁和自我效能的相互作用》，《社会行为和个性：国际期刊》第42卷，第5期（2014年）：799-809页。

116. West，《谁在做决策？》第27页。

117. Eron M. Berg和Louis G. Lippman，《广播广告中的幽默对新产品品牌名称的认知有影响吗？》《广义心理学杂志》第128卷，第2期（2001年）：194页。

118. 同上，第199页。

119. Kirk，《实验设计》，第671页。

120. E. L. Thorndike、W. A. McCall和J. C. Chapman，《通风与心理工作的关系》，第78卷，哥伦比亚大学教师学院教育贡献（纽约：哥伦比亚大学教师学院，1916年）。

121. Ronald A. Fisher，《研究工作者的统计方法》（爱丁堡：Oliver and Boyd，1925年）。

122. Campbell和Stanley，《实验和准实验设计进行研究》。

123. Ritter和Maynard，《使用正确的设计得到“错误”的答案吗？》

124. 同上。

125. Srivastava，《网络干预》。

126. Yun Hyung Koog和Byung-Il Min，《随机参与者分配是否会导致研究参与者获益减少？部分随机针灸试验的系统综述》，《替代和补充医学杂志》第15卷，第10期（2009年）：1107-1113页。

127. Richard A. Berk、Gordon K. Smyth和Lawrence W. Sherman，《当随机分配失败时：明尼阿波利斯伴侣虐待实验的一些经验教训》，《量化犯罪学杂志》第4卷，第3期（1988年）：209-223页。

128. Ong-Dean, Huie Hofstetter和Strick，《挑战和困境》。

129. Gueron，《随机分配的政治：实施研究和影响政策》，第26页。

130. Ong-Dean, Huie Hofstetter和Strick，《挑战和困境》，第39页。

131. 同上，第45页。

Chapter 7 ■ Random Assignment **209**

132. Gondolf，“成功与失败的随机分配的经验教训。”

133. Ong-Dean, Huie Hofstetter, 和 Strick，“挑战和困境。”

134. Gary W. Ritter 和 Marc J. Holley，“在学校进行随机分配的经验教训，”《儿童服务杂志》第3卷，第2期（2008年）：28–39。

135. 同上。

136. 同上。

137. 同上。

138. 同上。

139. Ritter 和 Maynard，“使用正确的设计。”

140. Ritter 和 Holley，“在学校进行随机分配的经验教训。”

141. 同上。

142. Bausell，“进行有意义的实验”；Ritter 和 Holley，“在学校进行随机分配的经验教训。”

143. 同上。

144. 同上。

145. Ong-Dean, Huie Hofstetter, 和 Strick，“挑战和困境”；Ritter 和 Holley，“在学校进行随机分配的经验教训。”

146. 同上。

147. Berk，“随机实验作为铜标准”；Ong-Dean, Huie Hofstetter, 和 Strick，“挑战和困境。”

148. Bausell，“进行有意义的实验。”

149. 同上。

150. Ritter 和 Holley，“在学校进行随机分配的经验教训。”

151. 同上。

152. Ong-Dean, Huie Hofstetter, 和 Strick，“挑战和困境。”

153. Ritter 和 Holley，“在学校进行随机分配的经验教训。”

154. 同上。

155. 同上。

156. Srivastava，“网络干预。”

157. Gondolf，“成功与失败的随机分配的经验教训。”

158. Ritter 和 Holley，“在学校进行随机分配的经验教训。”

159. Bausell，“进行有意义的实验。”

160. Ong-Dean, Huie Hofstetter, 和 Strick，“挑战和困境。”

161. 同上。

162. 同上。

163. 同上。

164. 同上。

165. Ritter 和 Holley，“在学校进行随机分配的经验教训。”

166. 同上。

167. Ong-Dean, Huie Hofstetter, 和 Strick，“挑战和困境。”

168. 同上；Ritter 和 Holley，“在学校进行随机分配的经验教训。”

169. Berk, Smyth, 和 Sherman，“当随机分配失败。”

170. Ong-Dean, Huie Hofstetter, 和 Strick，“挑战和困境”；Koog 和 Min。

171. Berk, Smyth, 和 Sherman，“当随机分配失败。”

172. Bausell，“进行有意义的实验。”

173. Gondolf，“成功与失败的随机分配的经验教训。”

174. Ong-Dean, Huie Hofstetter, 和 Strick，“挑战和困境。”

175. 同上。

176. Ritter 和 Holley，“在学校进行随机分配的经验教训。”

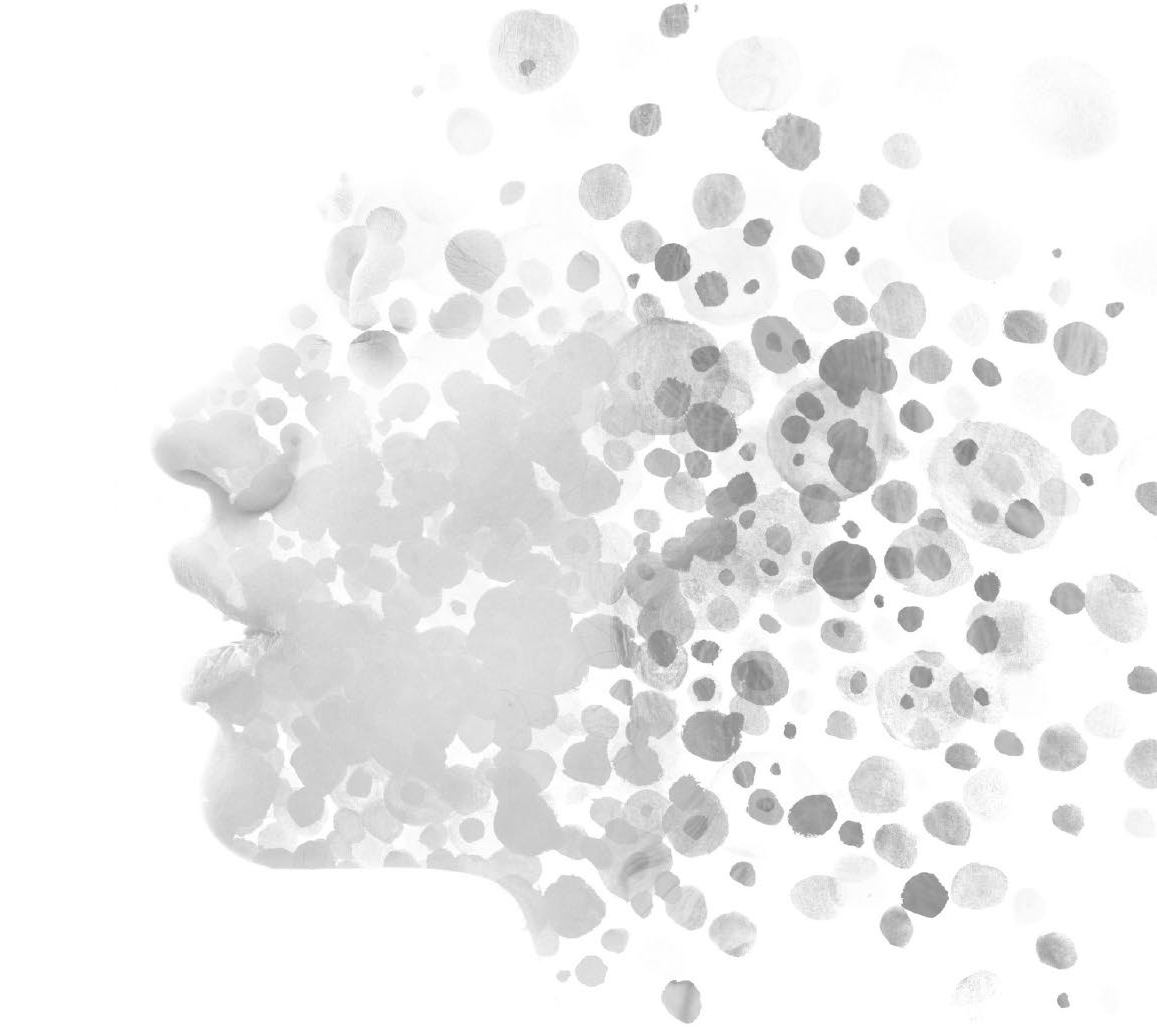
177. 同上。

178. 同上。

179. 同上。

180. Campbell 和 Stanley，“研究的实验和准实验设计。”

181. Abrams 和 Yoon，“抽签的运气：利用随机案例分配研究律师能力。”



For Review-No Commercial Use(2023)

8

# 抽样和效应量

*当有些人要求我们将大学生排除在信息来源之外时，我们可以半开玩笑地问：“当一个人的手碰触到他们的学位时，发生了什么使他或她成为一个在研究中的有效对象，而在审查和评论中不是？”*

*- Michael Basil*

*如果学生不再是我们首选和唯一选择，我们的研究结果将更加可信。*

*- William Wells*

学习目标

• 对来自不同来源的学生和研究对象的使用进行批判性思考。

• 制定招募和激励实验对象的计划。

• 描述样本量和统计功效之间的关系。

• 解释效应大小。

* • 准备并报告统计功效分析的结果。

**211**

在实验中，没有比上一章讨论的随机分配更重要的话题了。可能看起来本书的先后顺序有些颠倒，因为它在讨论如何分配研究对象之前先介绍了研究对象本身。这样安排的原因是理解随机分配对于认识为什么某些研究对象适合实验而其他方法则不适合至关重要。本章的主要关注点是随机分配的对象，即样本中的研究对象。同时，本章还探讨了统计功效（power）和效应大小（effect size）等相关问题，统计功效有助于确定所需的研究对象数量，而效应大小则评估了处理（treatment）的效果。

本章在第5章关于外部有效性的讨论基础上，结合第7章关于随机分配的理解进行了拓展。简单回顾一下，实验中使用的研究对象类型会影响结果的普适性。对实验的最严重批评之一是它们通常无法推广到研究之外的人群。正如第5章中指出的，大多数实验使用的是方便样本，也就是那些容易获得的人，而不是从总体中随机抽样的人群。与随机抽样相对应的是实验中的随机分配，它有助于确保不同组的研究对象在关键特征上具有相等性。在实验中，等效性比推广能力更重要。因此，实验依赖于逻辑推理来将因果发现扩展到研究对象之外。

实验在内部有效性和随机分配的优势方面，并不意味着研究人员可以随意选择容易获得的研究对象。虽然我们称它们为“方便样本”，但研究人员选择研究对象时不仅仅考虑方便性。研究人员应该仔细考虑每项研究中的研究对象选择，并在研究论文的方法部分解释选择的原因。本章将为这些思考和解释提供支持，并概述实验研究中一些常见的样本来源。假设读者已经相当确信，随机抽样在大多数情况下是一种“不足”的收集实验对象的方式，并且在实验中使用非代表性的方便样本是合理的，本章转向最常用和最常受批评的研究对象类型之一——学生。

**学生样本**

对于学术界的研究人员来说，没有比大学生更方便的研究对象了。研究表明，在市场营销和传播学期刊中，75%至90%的实验对象是学生。税务学科的比例约为52%

有些学系通过提供学生资源池或参与者资源池，使使用学生作为研究对象变得更加容易。在这些学系中，学生需要在每个学期报名参加一定数量的研究项目以获得学分。这在研究方法等研究重点的课程中很常见，也适用于心理学和广告等领域，学生可以亲身学习实践中的方法和程序。例如，在广告学中，决定制作何种类型的广告的许多决策都是基于广告或市场研究的结果，因此学生应该学习影响这些选择的过程。学生资源池中的学生可供教职员研究人员使用，有时也可供研究生使用。即使在没有方便的学生研究对象资源池的学系中，研究人员也会通过额外学分、礼品卡、披萨等方式吸引学生参与研究。然而，审稿人最常见的批评之一是研究受限于使用学生作为研究对象。有些人甚至认为使用学生是“非科学的”，或者仅在探索性研究和试点研究中有效。审稿人有时因为使用学生作为研究对象而拒绝接收研究。在达到这一点之前，实验研究人员有必要仔细考虑使用学生或任何人作为研究对象，选择最适合研究的对象，并为研究对象的使用提供强有力的理论依据。

**高于平均认知能力**

使用学生作为样本有时可能存在问题，不仅在外部有效性方面可能存在问题，因为它们可能不具有普适性，而且在内部有效性方面也可能存在问题，因为他们的回答可能会受到学生身份的偏见影响。显然，当学生是要推广到的人群时（如教育研究），普适性不是一个问题，但对于其他研究而言，使用学生作为研究对象的问题似乎可以归结为以下几个类别，这些类别由Sears提出。因为大学是一个旨在培养批判性思维的地方，而且参加测试和填写问卷是一种常见的活动，所以学生在研究中的表现可能会比那些没有上过大学或已经毕业一段时间的人要好。Basil指出，这可能会“改变所进行的思维过程的性质”，他举了一个调查决策过程的研究的例子，这个过程在学生中可能会有所不同。另一方面，学生更高的认知能力实际上可以降低随机误差，因为他们可以更准确地回答研究中的问题并报告他们的想法。研究人员应该仔细评估认知功能是否会对研究造成偏差；如果没有理论、证据或逻辑依据来说明为什么使用学生会有偏见，那么使用学生作为研究对象可能会引发问题。

**HOW TO DO IT 8.1**

Rationales for Subjects

以下是一些研究者在描述和证明其样本选择时的示例。

大学生样本示例

下面是一项政治研究中使用大学生的正当性证明，它超出了通常的说法，不仅表明学生样本是合适的，而且没有理由认为学生和非学生之间存在差异，而是具体引用了该研究主题的证据来支持这一观点：

“这种方法的一个批评（例如，Graber，2004）是，由于像许多其他媒体实验（例如，Druckman，2001a；Miller＆Krosnick，2000；Nelson＆Oxley，1999）一样，本研究使用大学生作为参与者，因此，研究结果无法推广到其他人群。然而，诸如议程设置、引导和框架等媒体效应在实验中已经被重复，这些实验使用了各种样本，从大学生（例如，Druckman，2001a；Nelson＆Oxley，1999；Miller＆Krosnick，2000）到研究者所在城市的成年志愿者（例如，Iyengar＆Kinder，1987）再到一般人口（例如，Nelson＆Kinder，1996），以及将媒体内容分析与一般人口调查相匹配的研究（例如，Iyengar＆Simon，1993）。因此，批评认为使用大学生样本将导致与一般人口样本有质的不同结论的观点没有得到证据支持。”10

Miller, J. M. 2007. “Examining the Mediators of Agenda Setting: A New Experimental Paradigm Reveals the Role of Emotions.” Political Psychology 28 (6): 689–717.

这是另一个超出通常范围的例子，提供了与该研究相关的具体信息：“实验通常需要一个明确定义的被试人群，可以追踪该人群，可以对正确的人进行随机分组和给予治疗，并且能够测量治疗组和对照组的感兴趣的结果。选民动员实验通过关注已注册选民来满足这些要求，其中有一个官方名单可以进行随机分组，并随后根据投票率进行更新。不幸的是，没有一个官方的未注册人名单。即使存在这样的名单，未注册人的居住迁移率要高得多，这样的名单的可靠性也会受到怀疑。在大学校园进行实验解决了定义参与者群体的问题。……大学生对于选民注册研究来说是一个有趣的人群，原因有四个。首先，大学生具有地理流动性，极有可能在最近的过去搬家，因此需要重新注册（Squire，Wolfinger，and Glass，1987）。这种流动性还可能使学生远离可以帮助他们参与公民活动的社会支持网络（如父母）。其次，大学生年轻，不太可能养成投票的习惯（Plutzer，2002；Bendor，Diermeier，and Ting，2003；Green，Green，and Shachar，2003；Fowler，2006）。养成习惯的反面是，在年轻时增加参与度将转化为未来更高的参与度。第三，大学生属于许多与选举参与度低相关的人口类别：年轻人（Wolfinger和Rosenstone，1980），缺乏兴趣

这里还有一个使用大学生的例子，因为他们在这个职业中有经验，这是对他们进行正当性证明的理由：“一所中西部大型大学的194名新闻学生参与了研究。使用学生是适当的，因为这个研究旨在查看是否存在效应；它并不关注一般化。样本限制在新闻学生中，以模拟记者的思考过程。

许多学生在入学前有着显著的职业经验。一旦被录取，这些学生作为课程的一部分被要求在报纸、杂志或网络附属电视台的新闻编辑室工作，为整个社区提供服务，但这些机构由学校管理。”12

Coleman, R. 2006. “The Effect of Visuals on Ethical Reasoning: What’s a (Photo) Worth to Journalists Making Moral Decisions?” Journalism and Mass Communication Quarterly 83 (4): 835–850.

MTurk示例

在这项关于信念回响的政治传播研究中，即虚假信息如何在纠正后仍然影响一个人的推断，作者使用了MTurk进行受试者招募。以下是其报告和证明：

“实验在2011年8月至2012年11月期间进行。参与者通过亚马逊的Mechanical Turk平台（MTurk）进行实验。2 Mechanical Turk是一个在线平台，受试者可以获得报酬来完成从给照片加注释到参加调查等任务。该研究仅限于年满18岁的美国参与者。3参与者根据参与的调查长度而获得0.61美元至0.75美元的报酬（支付根据调查的长度而有所不同），参与者完成调查通常需要8到12分钟。所有受试者经过筛选，确保没有参与多个研究。本文所描述的三个实验每个都包含一组独特的参与者。4有关每个样本的人口特征的表格在补充附录中。

几项研究表明，MTurk是传统上使用的其他便利样本的合理替代品。5尽管在许多方面类似于一般人口，MTurk样本的倾向更年轻和更自由。从理论上讲，不应该预期信念回响的形成会因年龄而有所不同，因为产生这些效应的认知过程应该在所有年龄段中是类似的。尽管党派可能会对信念回响产生影响，但实验通过随机向受试者提供与其现有党派偏好相一致或相矛盾的错误信息来考虑到这一点。每个实验遵循相似的格式。我会在对实验1的描述中详细解释这个格式，然后在对实验2和3的描述中描述任何变化。”

*(Continued)*

(Continued)

面板受试者示例

在这项研究中，关于政府资助对艺术机构捐款的影响，作者使用了一个提供研究和市场研究参与者的公司的受试者。以下是他们的描述：

“实验的参与者是CivicPanel项目的成员，该项目是一个选择加入的电子邮件面板，在本研究进行时大约有12,000名活跃的面板成员。CivicPanel是一个与大学相关的在线研究项目，旨在提供一个志愿者的普通人群，参与有关公共和公民事务的调查和在线研究。志愿者成员一直通过各种在线发布进行招募，包括Google广告、Craislist.org和开放目录Dmoz.org。鉴于对美国非营利部门的实质性关注，分析中剔除了来自非美国居民的133个回应。另外，还剔除了24个部分填写的回应，最终形成最终的分析样本n = 562。

表2显示了四个组别或处理组中参与者的数量以及他们的人口特征和政治态度。在典型的完全随机设置中，处理组和对照组在特征上应该相同，除了给予他们的处理（Remler＆Van Ryzin，2011）。表2中被随机分配给每个条件的组之间没有统计学上显著的差异，这证实了实验组的统计等价性。”14

Kim，M.和G.G.Van Ryzin. 2014. “政府资助对艺术机构捐款的影响：一项调查实验。”《非营利和志愿部门季刊》43（5）：910–925。

这里是另一个使用参与者面板的研究示例：

“参与者是由Qualtrics研究公司维护的人口普查代表性面板招募而来，该公司使用分层配额抽样方法。作为参与的回报，现金价值奖励被存入参与者的在线账户中。从面板中随机选择了2,301名年满18岁的成年公民，并通过电子邮件邀请他们参加在线研究；861名成员同意参与，768名个体成功完成研究。这表示合作率为33.4％（COOP1，AAPOR）。”15

Lee，H.和N. Kwak. 2014. “政治讽刺的情感效应：讽刺幽默、负面情绪和政治参与。”《大众传播与社会》17（3）：307–328。

专业样本

这是一个使用专业人士作为受试者的研究：

“参与者是来自美国南部和西南部地区的报纸和电视台的专业记者。研究者通过写信和电话与新闻室的管理者联系，征得他们允许研究者前往新闻室进行研究的许可。没有任何一个管理者拒绝允许研究者前往。新闻室的管理者提前通过电子邮件向员工宣布了研究的日期、时间和地点。然后，研究者前往新闻室，在会议室中让参与者完成问卷调查，并为参与者提供午餐或者本地餐厅的礼品卡作为回报。整个研究大约需要30分钟的时间。”16

Coleman，R. 2011. “记者对儿童的道德判断：要我说，不要我做。”《新闻实践》5（3）：257–271。