**社会科学中的实验设计**

**Designing the Experiments for the Social Science**

芮妮塔 科曼（Renita Coleman）著

××× 译

**目 录**

序言：作者介绍

**第1篇 发现因果关系**

**第2篇 伦理和历史上的著名实验**

**第3篇 理论、文献和假设**

**第4篇 实验类型**

**第5篇 内部和外部效度**

**第6篇 因子设计**

**第7篇 随机分配**

**第8篇 抽样和效应大小**

**第9篇 刺激和操纵检查**

**第10篇 仪器和测量**

**第11篇 伦理审查委员会和进行伦理实验**

**详 细 目 录**

**第1篇 发现因果关系**

**第2篇 伦理和历史上的著名实验**

**第3篇 理论、文献和假设**

**第4篇 实验类型**

**前 言**

有很多关于实验数据统计处理的优秀书籍和教学课程，但很少有关于如何获取足够和适当的数据以供进行统计分析的方法的帮助。

在W. A. McCall写下这句话四十年后，Don Campbell和Julian Stanley说道：“这句话在今天仍然是正确的。”

已经过去了五十多年，我可以说这句话至今仍然适用，尤其是对于社会科学而言。

这本书是关于获取足够和适当数据的。

这本书与大多数其他实验设计书籍有两个不同之处：首先，它专注于实验计划的方法论和设计问题，而不是在收集数据后使用各种统计方法进行数据分析。从一开始就进行仔细的研究设计是进行良好研究的关键，而这在实验中尤为重要。这本书着重介绍如何有效地设计实验，而不是如何分析实验数据。它关注研究人员在介入和处理实验之前做出有关实验程序方面的决策的阶段。

这本书将帮助读者从一开始就学习如何规划和执行实验，通过逐步指导决定是使用单因素设计还是因子设计，如何分配受试者到不同组别，选择和收集样本，创建刺激和工具，进行操作验证，向机构审查委员会（IRB）申请批准，并进行试验前研究，以及在实验过程中做出其他选择时要考虑的事项。它提供了决定在创建特定实验中最佳使用哪些元素以及在不可避免的权衡中需要考虑的因素的指南。这本书是实用的，注重应用。

《社会科学中的实验设计》这本书只是简要介绍了实验的统计方面，因为它们密不可分。然而，主要的重点是让读者学习如何从一开始设计实验。统计学的内部工作机制、公式以及如何手工计算或使用软件进行计算并不包含在这本书中。理解本书内容并不需要深入的统计学知识。这本书仅限于帮助读者了解哪些统计方法适用于什么类型的分析和测量水平，但它是非常基础的。我建议除了阅读这本书之外，还可以参加传统的实验设计课程、统计学课程或阅读专注于统计学的相关书籍，以获得更深入的统计学知识。

这本书与其他书籍的第二个不同之处在于它面向广泛的社会科学领域的研究人员。读者将会发现除心理学之外的其他学科领域的例子，涵盖政治学、商业、经济学、信息科学、社会工作、教育、社会学、健康领域、广告等等。我的学科是新闻学和更广泛的传播领域。我的第一门实验设计课上有来自新闻学、广告学、公共关系学、广播、电视、电影、政治学和人际交流学等不同专业的学生。随着研究人员逐渐认识到了解因果关系对于他们学科的重要性，实验在社会科学领域变得越来越普遍。随着易于使用的软件（如SPSS、G\*Power、Qualtrics）的引入和对人类受试者（如Mechanical Turk）更广泛的获取，实验设计变得更加普遍。这本书直接面向这些研究人员，提供了他们所在领域的实例。它还探讨了一些特定学科面临的挑战，例如教育领域常见的对随机分配的反对意见，政治学中关于样本响应率的报告，以及经济学中使用观察或基于绩效的测量而不是自我报告的情况。

许多社会科学领域的研究人员通过与其他实验研究者进行一对一的合作来学习。这本书是为那些以这种方式学习的人提供的补充材料，对那些独自工作的人来说是指南，对于组织的课程来说是教科书。我写这本书的初衷是当我在德克萨斯大学教授实验方法课程时，无法找到完全符合我的要求的书籍。那门课程是这本书的基础。我将其设想为独立的实验设计方法课程的核心教材。对于在实验在社会科学中占主导地位的广告学、政治学、健康科学等学科的一般研究方法课程，它也将是一个很好的补充教材。

我撰写《社会科学中的实验设计》时考虑到了研究生学生，尽管我作为一名职业记者已经有了十五年的经验，知道如何为广大读者写作。这本书使用易于理解的语言撰写，以便本科生也能够理解。然而，对于完全没有经验的人来说可能不太合适；它更适合受过教育但缺乏经验的初学者，那些之前进行过研究但没有进行过实验的人，或者曾经在他们自己的实验中提供帮助但没有进行过自己的实验的人。它至少假设读者已经修过一门科学方法的概览课程。那些具有一些研究经验但以前从未进行过实验的人不会觉得太具挑战性。对于那些从导师的指导中学习而不是参加组织的课程的人来说，这本书旨在作为参考和补充，辅助这些优秀的教师。它还可以帮助产生一些工具和策略的想法，这些想法可能不常被导师们包括在他们的教学中。其他读者可能包括具有其他研究方法经验的学者，希望扩展他们探索因果关系的能力。此外，那些在政治候选人、广告机构、教育机构或其他行业从事研究工作的专业人士也会发现这本书很有帮助。我希望这本书能激发那些没有独立的实验课程的社会科学部门去创建一个。

各章节中包含了概念性和理论性问题以及实际的实践建议。它采用了一种逐步的方法，引导学生们编写假设和研究问题，设计刺激物，并书写方法部分。很多例子都来自于我的个人经验，我可以提供关于为什么选择某些程序而不选择其他程序的背景故事，研究是如何发展的以及如何应对审稿人评论的建议。这些内容在发表的论文中经常无法体现出来。此外，我还从各种社会科学学科中其他人进行的一些优秀实验中汲取了一些经验，尽管我无法提供关于他们做出选择的内部知识。

虽然各章节大致按照设计实验的顺序进行呈现，但并非所有的教师或读者都希望按照给定的顺序使用。例如，在阅读第二章关于欺骗的内容之后，一些读者可能希望跳到第11章，其中讨论了与伦理委员会(IRB)相关的更广泛的伦理问题。为了方便非按顺序阅读，书末附有术语表，其中包含了每章的关键术语，这些术语在文本中以粗体标注并进行了定义。其他教学工具包括三种类型的方框特色：

"More About..."方框则对章节中的某个问题进行扩展，供那些希望获得更多细节的读者阅读。

"Study Spotlight"方框则突出展示了一个特定的实验，以展示它如何处理章节中的主题。

"How To Do It"方框提供了执行章节中各步骤以及将其写入论文的指导和示例。

此外，还有其他三个特点：

"Common Mistakes"是一个简短的项目列表，列出了初学者特别容易犯的错误。

每章末尾的"Test Your Understanding"问题是一个简短的多项选择题测验，以便您了解自己对材料的理解程度。

"Application Exercises"是一些较长的实践作业的指导，让您能够应用所学知识，包括逐步创建自己实验的研究提案的方法。这些作业被分解成可管理的部分，以使任务看起来不那么令人畏惧。

每个章节还包含了建议阅读的书目列表，以便深入了解相关主题。每个章节的一个显著特点是实践建议和示例，涵盖了从如何描述刺激物和仪器，到与合作者合作、提交到期刊以及回答审稿人常见问题等方面的内容。书中提供了用于创建Twitter动态或Facebook帖子、在线随机化工具以及进行功效分析教程的免费软件的URL。一些主题并非仅适用于实验，但我试图通过这种方法凸显特定问题。例如，第三章解释了理论在实验中的作用，其强项在于测试和拓展理论。它还涵盖了如何利用文献中的实证证据来支持所测试的假设和理论过程，并提供了类似于句子构图的“公式”来编写假设。

伦理道德在第二章中首次引入，自然地紧随着对斯坦利·米尔格拉姆的服从权威实验和菲利普·辛巴多的监狱实验的讨论，这两个实验引起了人们对实验科学中伦理道德的重视。在这里，一个"More About"方框深入探讨了欺骗问题，在实验中，欺骗比其他方法更具问题性。对伦理道德的讨论贯穿于每个章节中——例如，在第7章中关于随机分配的讨论中，一些将其视为不道德的反对意见被讨论到。这个主题在第11章中以更详细的方式再次出现，以伦理审查委员会(IRB)为框架，考虑到其标准与伦理道德紧密相关。由于本书采用了逐步指导的方法，在实验设计完成后，研究人员必须申请批准之前才会讨论IRB。然而，伦理实验的重要性并不是唐突加入的，相反，它被包含在第11章中，因为IRB是执行伦理行为的主要手段，所以这一章似乎是同时讨论两者最合乎逻辑的地方。

这本教科书并不能涵盖实验所需了解的全部内容，学习过程应该继续通过阅读“建议阅读”部分中的优秀著作、阅读自己学科领域的实验以及与他人合作来提升。我希望你们喜欢并且能够从中获益。

如需与本书配套的额外教学和学习资源，请访问study.sagepub.com/coleman。

教师可以找到按章节整理的幻灯片，以协助讲座准备，同时还提供了包含多项选择和开放式考试问题的Word试题库。学生可以找到书中亮点所涉及的SAGE期刊文章的全文版本。

**致谢**

我要衷心感谢那些在本书的制作和完成过程中做出贡献的人们。首先，我要感谢我的导师们，Esther Thorson和Glenn Leshner。当我还是一名研究生时，你们让我对实验充满了热情。感谢你们的指导、耐心和对知识的分享。你们告诉别人，而这也传到了我的耳中，说“她总是言出必行”，这帮助确保了我能够实现承诺。你们是一个榜样，每次新的被试者坐在椅子上时，你们都会测量屏幕到椅子的距离，让我认识到细节的重要性。你们总是抽出时间听我讲述新项目，并给予建议。此外，我还要感谢我的研究伙伴Denis Wu，我很幸运在我担任第一个教职时遇到了他，而且我们一直携手并进。他似乎擅长我不擅长的事情，并且一直是我们自我引导实验之旅中的坚实支持者。我要感谢所有与我一起进行实验的杰出的前研究生，包括但不限于Ben Wasike、Lesa Hatley-Major、Rebecca McEntee、Lewis Knight、Carolyn Yaschur、Trent Boulter、Viorela Dan、Angela Lee、Raluca Cozma、Avery Holton、Dani Kilgo、Siobhan Smith、Kate West、Joseph Yoo和Ji Won Kim等人。我还要感谢所有参加我的实验设计课程的人，尤其是来自其他部门的学生，你们让我认识到不同领域的新理论和独特挑战，也让我在撰写本书的过程中能够对你们进行实验。我从你们身上学到了很多，甚至比你们从我这里学到的还要多。

最后，这本书的完成离不开所有审稿人的时间和专业知识。他们温和地指出了缺陷和不足之处，引导我找到宝贵的资源，并耐心地帮助我改进了这本书。我还要感谢SAGE的高级编辑Helen Salmon，她冒险选择了这个提案，并提出了许多重要的补充意见；感谢Jeremy Shermak为本书创建了网站；以及SAGE的优秀员工，包括Chelsea Neve、Megan O'Heffernan、Eve Oettinger、Diane Wainwright以及研究方法、统计学和评估部门的其他团队成员，他们帮助推动了这本书的完成。

SAGE要感谢以下审稿人对他们的反馈表示感谢：

* Richard E. Adams, Kent State University Francis O. Adeola, University of New Orleans Anna Bassi, University of North Carolina
* Jacqueline Craven, Delta State University
* Sarah A. El Sayed, University of Texas at Arlington
* Janet Reid Hector, Rutgers University
* Kyle J. Holody, Coastal Carolina University
* Glenn Leshner, University of Oklahoma
* Ji Hoon Ryoo, University of Virginia
* Gerene K. Starratt, Barry University
* Michael Teneyck, University of Texas at Arlington
* Geoffrey P. R. Wallace, University of Washington

注释

1. W. A. McCall, How to Experiment in Education (New York: MacMillan, 1923), Preface.

2. D. T. Campbell and J. C. Stanley, Experimental and Quasi-Experimental Designs for Research (Chicago: Rand McNally, 1963), 1.

**关 于 作 者**

Renita Coleman拥有佛罗里达大学新闻学学士学位，以及密苏里大学新闻学硕士和博士学位。

她的研究专注于伦理、框架和议程设置，尤其关注视觉传播。她研究了照片对伦理推理的影响，健康新闻中的框架和责任归因，以及新闻和公共关系从业者的道德发展等课题。她在包括《传播学杂志》、《新闻与大众传播季刊》、《广播与电子媒体杂志》、《大众传媒伦理学杂志》、《新闻学》和《新闻学研究》在内的学术期刊上发表了40多篇同行评审文章。她合著了两本书：2015年出版的《选民决策中的图像与情感：情感议程》和2005年出版的《道德媒体：新闻从业者的伦理推理方式》。

在开始学术生涯之前，Coleman在报纸和杂志工作了15年。她曾在北卡罗来纳州罗利的《新闻观察报》、佛罗里达州萨拉索塔的《先驱报》以及佛罗里达州奥兰多的《哨兵报》等新闻机构担任记者、编辑和设计师。

Coleman教授本科和研究生课程，内容涉及伦理学、生活方式新闻学和实验设计。

**第一篇 发现因果关系**

*生活是对因果关系的永久教诲。*

*——拉尔夫·瓦尔多·爱默生(Ralph Waldo Emerson)*

|  |
| --- |
| **学习目标**   * **解释因果关系在实验中是如何起作用的。** * **比较实验方法与其他方法的优势。** * **确定实验的三个基本标准。** * **描述变异性、混杂因素、对照组和分组的要素。** * **制定一个问题陈述，并回答一个关于你自己研究的“那又怎样”的问题。** |

这本书是关于实验的，其范围差别很大。实验是在精心控制的条件下进行的一种对某个假设或原理的科学测试，旨在确定或发现某种未知的事物。实验提供了关于事物之间关系的洞察，其中一个事物的变化导致另一个事物发生变化。我们在日常生活中都做过不知不觉的非正式实验，正如拉尔夫·瓦尔多·爱默生的开头引语所示。例如，如果你在食谱中改变某种成分的量，味道会更好吗？如果你喝低热量的啤酒，你会减肥吗？只要你一次只改变一种成分，或者不增加锻炼量或吃低热量的一切，你可能会认为是那个成分或者低热量啤酒导致了味道的不同或者你的减肥。

**因果关系**

这些日常实验的例子说明了因果关系的概念。在前面的例子中，更美味的千层面或减肥是效果。"原因"是对这些事情发生的解释——更多大蒜使千层面的味道更好，低热量啤酒帮助你减肥。我们在日常语言中使用"原因"这个词，比如"死因"或"事故原因"。在实验中，这个意思没有不同，但你也会看到使用因果性或因果关系这些术语。

当然，对于本书的目的，我们更关注系统性实验，而不是所描述的简单的日常实验。在医学中，这些被称为临床试验或随机临床试验（RCTs；有关RCTs的更多信息，请参见“More About”框1.1）。在网页设计和市场研究中，它们也可能被称为A/B测试。这些领域的术语与社会科学中的实验有所不同，但目标是相同的——找出对特定问题最有效的治疗（或原因）。在医学中，问题是疾病或疾病，效果是治愈或改善。在社会科学中，我们寻求解决的问题可以是促进品牌意识的电视广告、教授唐氏综合症学生的策略，或者帮助会计师更加诚实的干预措施。

为了实现这一点，社会科学家使用实验作为“证明的基本语言”。通过展示在保持其他条件不变的情况下改变某个因素时会发生什么，实验为我们提供了因果关系的证据。通过这种方式，我们更有把握认定发生改变的因素是导致我们观察到的结果或效果的原因。

在医学领域，实验是最常见的研究类型，但在社会科学领域，实验可能是最少被使用的方法之一。在传播学期刊中，大约有12%的研究采用实验设计。在国际关系领域，这个比例可能低至4%。在特殊教育领域，实验被提倡作为对呼吁提高证据为基础的专业的质量和严谨性的回应。在政治学等领域，实验的应用也在稳步增长。随着社会科学领域对这种方法的好处逐渐被认识到、数据分析软件的易用性提高以及技术使得受试者更加廉价和容易招募，实验的使用只会增加。因此，掌握如何进行有效的实验变得更加重要。已经有许多文章讨论了在学术界和专业领域中对实验的增加使用，例如政治学和信息系统学科。其他文章讨论了在实验方法方面创建有组织的课程的日益重要性。有些人甚至认为实验是他们学科中最重要的方法。实验对于使用以证据为基础的实践的领域非常关键，比如社会工作和教育领域，其中随机临床试验被视为黄金标准。

|  |
| --- |
| **内容延展 1.1**  **随机对照试验**  随机对照试验（Randomized Clinical Trials，简称RCTs）基本上与真实实验相同，其中参与者被随机分配（R代表随机）到治疗组或对照组，并接受一个或多个实验性的程序或药物。它们也被称为随机对照试验（Randomized Controlled Trials），纯粹的语言学家使用“controlled”来指代包括对照组的研究，其中被试接受安慰剂或无治疗。当使用临床一词时，可能有或没有对照组。可以推测，临床的使用是因为这些研究是在医疗患者或其他临床环境中进行的。试验一词是因为所研究的治疗或药物正在尝试中，也就是说，它尚未获得广泛使用的批准，而该研究将确定其是否安全有效。  随机对照试验（RCTs）被认为是医学研究的金标准，就像真实实验或实验室实验在社会科学中的地位一样。  随机对照试验（RCTs）的历史可以追溯到公元前600年，当时犹太人丹尼尔比较了巴比伦王室的饮食和素食饮食。还有人将RCT的起源归功于詹姆斯·林德（James Lind），他在1747年进行了治疗坏血病的实验，这些实验在第2章中有描述。通常认为，第一次现代RCT是在1948年对链霉素对结核病的疗效进行的测试。该研究首先解释了结核病的临床试验缺乏适当的对照，导致对黄金作为治疗手段的夸大宣称。该研究提供了对方法的“完整描述”，因为规划如此严格的试验存在困难，其他人可以复制该研究。 |

**实验与其他方法相比**

实验方法的主要优势在于它提供了一个强大的工具来发现因果关系。许多其他研究方法，如调查，可以确定变量之间的相关性或共变性，但这不一定意味着因果关系。相比之下，实验可以揭示改变一件事情如何导致另一件事情发生变化。通过有意识地和系统地改变某些因素，我们可以观察到潜在的因果因素。在相关性研究后进行实验是增加我们对这些关系实际上是因果关系的信心的良好方法。三角测量，或者使用不同的方法研究同一现象，对于科学研究非常重要，因为它有助于增加我们对使用一种方法观察到的现象可以通过另一种方法进行复制或重现的信心。令人担忧的是，过度依赖只显示相关性但没有进行实验跟踪以找到因果关系证据的调查和观察方法。

在推断相关性引起因果关系的问题中，最著名的例子之一是古老的传说，认为鹳鸟带来了婴儿。上世纪30年代在哥本哈根进行的一项研究记录了鹳鸟数量较多的年份也出生了更多的婴儿，相关性达到了0.85。但仅仅因为这两个变量高度相关，并不意味着其中一个导致了另一个。相反，存在着合理的替代解释或其他可能的原因，并未进行研究。这是在第一次世界大战后不久，那些在外待了很长时间的士兵回家，导致更多婴儿出生。此外，人们从农村迁往城市就业，因此更多人生孩子意味着更多婴儿。随着人口增加，建造了更多房屋，导致鹳鸟筑巢的地方增加，从而导致了更多的鹳鸟。这种循环不断延续。

**实验的基本要求**

这个例子阐述了所有实验必须具备的三个基本特征：即原因必须在效应之前发生，在这个例子中，鹳鸟确实在婴儿之前存在；但原因也必须与效应相关，而在这个例子中并不存在相关性。没有逻辑上或理论上的原因，也没有实证证据表明鹳鸟与婴儿有关。这是为什么实验者不仅仅测试主观猜测的原因之一，而是寻找理论、逻辑或现有证据来测试联系的原因。这有助于确保原因实际上与效应相关。在鹳鸟和婴儿的例子中，实验的第三个要素也缺失了——对于效应来说，没有其他合理的替代解释。一个真正的实验必须同时具备这三个条件：原因必须在效应之前发生，与效应相关，并且没有其他合理的替代解释。如果满足了这三个条件，实验可以让我们更有信心认为某个因素导致了另一个因素，而不仅仅是存在一种可能性，这种关联性是不太可能仅仅由于偶然发生的。

虽然能够测试因果关系是一个强大的工具，但它仅仅是一种描述，不能解释为什么会发生某事。例如，在社会工作领域，许多研究已经确定了成功减少青少年犯罪的项目，但很少有人对其有效性进行了研究。Hay和他的同事注意到了这个空白，并设计了一个实验来发现解释参与项目如何降低犯罪率的机制（有关这项研究的更多信息，请参见研究重点 1.2）。在我的研究中，我假设观看照片会导致记者在做新闻决策时使用更高质量的伦理推理。我找到了我想要的效果，但那只是一种描述，而不是一种解释。仅凭这一发现并不能说明它发生的原因。为了提供因果解释，实验需要构建解释这些效果的潜在机制。如果照片确实改善了伦理推理，了解为什么很重要。类似这样的中介变量和调节变量将在另一章节中更详细地讨论。在这里可以简单地说，好的实验应该还包括解释为什么发生某种效果的机制。

在对实验与其他方法的基本讨论之后，接下来我们将转向实验的一些具体要素。所有这些要素将在后面的章节中详细阐述，但在这里介绍它们是为了提供一些基本的实验理解。

|  |
| --- |
| **学习要点 1.2**  **发现效果并解释原因**  **Hay, Carter, Xia Wang, Emily Ciaravolo, and Ryan C. Meldrum. 2015. “Inside the Black Box: Identifying the Variables That Mediate the Effects of an Experimental Intervention for Adolescents.” Crime & Deliquency 61 (2): 243–270.**  这项研究是一个很好的实验示例，它不仅发现了效果，还解释了导致这些效果的原因。研究作者首先重新确认了特定项目的治疗确实降低了青少年犯罪的风险。然后，他们研究了介入变量，也就是参与该项目和减少青少年犯罪之间的中介因素。他们检查了一共十一个风险因素，这些因素被认为是该项目有效的原因，研究人员称之为因果机制。他们说：“简而言之，如果我们不了解一个项目减少犯罪的确切机制，那么在借鉴其优势、在其他地方复制并运用于公共政策方面的努力必然受阻”（第248页）。  在可能解释该项目有效性的十一个变量中，研究人员只发现了一个显著的因素：“减少与从事不良行为并对他们施加压力的同伴的交往”（第263页）。换句话说，与不良团体交往。  研究人员解释了在追求其他十个变量时，时间和金钱的投入的重要性，而这些变量实际上并没有产生影响。用更友善的方式来表达，研究人员说：“我们的分析-一个罕见的测试考虑到中介变量-表明这些项目针对的许多风险因素可能对项目服务没有反应”（第264页）。 |

**实验的要素**

**变异**

改变或变化是进行良好实验的首要条件。显然，如果没有任何变化，那就没有什么可以研究的。必须有某种变化发生。在实验中，变化是通过操纵独立变量（IV）来实现的，即研究人员认为会引起变化的因素。这也被称为操纵、处理或干预。研究人员应该仔细控制这些变量。例如，许多研究使用专业演员扮演政治候选人，而不是真正的政治家，以免偏倚被试者的反应。在一项研究中，演员在一次采访中展示积极的非语言行为，在另一次采访中展示消极的行为，而在第三次采访中展示中立的身体语言。研究作者研究了这些行为的具体表现形式，例如，双臂交叉是消极的，向前倾身并注视面试官是积极的，并确保演员展示这些行为。仅仅要求演员表现积极或消极是不够的，研究人员需要确切知道这些行为应该是什么样子，并确保其正确展示。

良好的变异性另一个关键是一次只能变化一件事情；否则，就无法确定多个变量中的哪一个导致了结果。实际上，许多实验同时变化多个变量（稍后会详细介绍），但关键在于它们不允许这些变量共变——也就是说，它们不能同时变化。例如，有研究人员注意到，显示负面政治广告比正面广告更有影响力，并且允许广告的语气与其中包含的信息量共变。负面广告中的信息量总是比正面广告多；因此，更强烈的影响力可能是由于信息更多，而不一定是由于其负面语气。这为这些其他研究中发现的效应提供了一个合理的替代解释。为了确定是否如此，研究人员进行了一项研究，保持了信息量恒定——也就是说，负面广告和正面广告中的信息量相同。他发现，语气根本没有造成任何效应——造成效应的是信息量；负面广告通常比正面广告包含更多的信息，这就是其他研究所记录的更高水平的竞选知识、兴趣和投票率的来源。

实际上，原因很少是单一的，或者由一个变量引起的。社会科学研究人员很少期望单一因素成为另一个因素的直接原因。通常有许多因素起到了作用。在实验中，通过将其他因素保持恒定来管理这种情况，以确保这些因素不会对效果产生影响，而只有有意变化的因素才能被认为是导致效果的原因。

**混淆变量**

能够提供合理的替代解释的因素被称为混淆变量。确保没有混淆变量非常重要，因为它们可能会影响实验的准确性。成功的实验关键在于控制可能导致结果的外部影响因素。例如，一个研究人员研究报纸上方和下方放置的照片对人们的影响，作为刺激，他使用了一张垂直的照片，显示一个近距离的人物放在报纸的上方，以及一张水平的照片，显示远处的人物放在报纸的下方。他只使用了报纸下方的水平远处照片和报纸上方的垂直近距离照片。这种做法的问题在于，造成他发现的效应可能是照片的垂直或水平格式，或者其中的近距离或远处的人物，而这些不是他有意变化的变量。照片在许多层面上发生了变化，不仅仅是他感兴趣的报纸位置。这代表了一个混淆变量，即一个合理的替代解释，而这个解释没有得到控制。

为了避免混淆变量，在一项旨在确定照片中人物的种族如何影响记者道德判断的研究中，研究人员采用了完全相同的照片，并对人物的肤色、发型和面部特征进行了数字修改，以操纵他们的种族。这样，除了种族这一被设计为变化的因素外，其他一切都是相同的。背景相同，与相机的距离相同，人物的吸引力等等也都相同。这样就避免了任何混淆变量。研究人员知道，例如，特写照片比长镜头照片使人们对照片中的人物更加舒适，而吸引人的人在各种特征上的评价比较不吸引人的人要好，包括政治候选人的信任度和可选性。因此，这项研究需要的是照片在除了其中的人物的种族外，在其他方面完全相同。

在实验中，尽可能追求精确度是有帮助的。例如，研究人员可能会在每个新受试者进入实验前测量受试者坐在椅子上离电视的距离。所有的受试者应该离电视的距离完全相同，因为即使稍微靠近一点也可能产生差异。

研究人员通过运用常识和阅读其他研究的方式发现潜在的混淆变量。文献中有大量的证据显示，政党和意识形态会影响许多事物。因此，许多实验通过创建不暗示或明示虚构候选人的政党倾向的刺激来控制这些潜在的混淆变量。这代表了一种实验控制，即对参与的变量进行仔细管理。避免潜在混淆变量的另一个例子是在选举期间的实验中不使用当时新闻中的真实问题。当研究人员无法控制自变量和因变量之间潜在虚假关系时，另一种方法是测量并在统计上加以控制，即在检查操作的效果之前将其排除在外。测量人们的政治党派和意识形态，然后将它们作为协变量使用，是一种统计控制的形式。

阅读相关文献以发现需要控制的因素是非常重要的。例如，在道德发展的研究中，研究人员发现自由派或保守派与道德判断质量之间存在关联。这并不是一件直观明显的事情，因此通过阅读文献来发现这一点是关键。年龄和教育程度也与更好的道德判断有关，因此在将道德判断作为结果或因变量的实验中，通常会测量参与者的政治意识形态，以将其作为协变量纳入统计分析中，以防止随机分配没有使组别在该变量上具有等价性。协变量的作用是消除潜在混淆变量的影响，使研究人员能够看到被操作变量的真实效果。了解所研究领域的文献是非常重要的，因为并非所有可能被怀疑为因变量和自变量之间潜在虚假关系的因素都真的存在。例如，在道德判断中，性别通常并不重要，但在许多其他方面却是重要的。由于大多数效应很少是由单一因素引起的，所以了解其他可能影响结果的因素是很重要的。（脚注：在实验中，协变量并不总是像观察性研究中那样必要，因为随机分配旨在消除这些变量的影响。有关随机分配的详细内容将在第7章中进行讨论。）

同样重要的是，要控制自己的常识告诉你可能会影响结果的潜在混淆因素。即使在文献中没有看到某个因素，如果你认为它可能会产生效应，就将其纳入实验中进行测试。假设你正在研究一个你很了解的现象。如果你是一位专家，你应该对影响你所研究的现象的因素有一些想法。这是知识的产生和文献发展的一种方式。

**对照组**

另一个实验的重要要求不仅是了解接受治疗的人会发生什么，还要了解如果他们没有接受治疗会发生什么。拥有一组未接受治疗的人，也就是对照组，可以帮助我们推断如果没有治疗的话，结果会是什么样。效果是接受治疗的人实际发生情况与如果他们没有接受治疗时可能发生情况之间的差异。这样，我们可以分离和测试每个变量的效果，并更加确定这个效果是由该变量引起的，而不是其他因素。接受治疗或干预的人被称为治疗组、实验组或干预组。而没有接受治疗的人则是对照组。对照组是一组未接受干预的人，用来与“正常”、“中立”或“无干预”情况进行比较。

在医学研究中，对照组通常会接受安慰剂，它看起来像是一种药片、注射、疗法或其他治疗，但实际上只是一种糖丸、盐水注射或其他让人们认为自己接受了某种治疗的物质（有关安慰剂的更多信息，请参阅1.3节的“有关安慰剂的更多信息”框）。由于暗示的力量，仅仅对对照组的人们不做任何实际操作是不够的；他们必须感觉到自己接受了某种治疗。对于社会科学家来说，这有时可能会带来问题。例如，在测试不同类型的营销信息时，什么构成对照组呢？当然不能没有任何信息，但是那样的话，受试者该做什么呢？不能让他们只是进来填写问卷而没有接触到任何东西，否则他们会认为研究过于人为。对照组将在接下来的章节中进一步讨论；这里讨论的重点是实验需要一种比较接受治疗和未接受治疗的人之间发生的情况的方式。

**分配**

最后在这个讨论中，也许是实验中最重要的一个方面，是关于人们如何被分配到实验中不同的干预措施中去的问题。这个话题非常重要，以至于它有一个单独的章节（第7章）来讨论。简而言之，最佳的方法是随机分配ii，这是一种将受试者放入不同组的方式，以确保个体差异在不同组之间均匀分布。这一特征对于确保人们的个体特征不成为研究的混淆因素非常重要。男性和女性应该在各个组中均匀分布，年轻人和老年人也应该如此，例如。本书专门撰写了一个章节来讨论将受试者分配到实验条件中的问题，包括在无法以这种方式分配人员时应该采取的措施。

ii这本教科书使用术语"subjects"来指代实验中研究对象。美国心理学协会（APA）的最新版（第6版）《出版手册》指出，"subjects"和"participants"这两个术语都是适用的，并指出"subjects"已经使用了几百年。对于"subjects"一词的反对意见始于1994年APA手册的第四版，当时更倾向于使用"participants"。当2010年出版的当前版本问世时，"subjects"和"participants"被视为平等的术语。第73页的条目指出："事实上，超过100年来，"subjects"这个术语一直被用于实验心理学中作为描述样本的一般起点，其使用是恰当的。"由于"subjects"和"participants"同样合适，本教材使用"subjects"一词，以保持与实验设计语言中其他术语的一致性，包括between-subjects设计、within-subjects设计以及IRBs所使用的human subjects。有关这个主题的更多信息，请阅读Roddy Roediger在APS Observer 2004年4月刊上的文章"What Should They Be Called"，可在http://journal.sjdm.org/roediger.html获取。）

|  |
| --- |
| **内容延展 1.3**  **安慰剂**  在医学研究中，安慰剂的使用比简单地注射生理盐水、服用糖丸或假装进行手术来模拟对照组受试者接受实际治疗要复杂得多。事实上，这些无效但无害的安慰剂实际上已被证明具有一定的效果。被称为安慰剂效应，接受安慰剂的人实际上会报告在他们所谓的治疗的任何病症上有所改善。  “Placebo”这个词源于拉丁语，意为“我将取悦”。追溯到1801年，第一项使用安慰剂进行对照试验的记录已经找到。对安慰剂的认识视其为欺诈、虚假和不道德的观点是因为直到20世纪中叶，许多执业医生会在伪装成真正药物的情况下向患者使用安慰剂。对于在研究中使用安慰剂是否符合伦理的问题至今仍存在争议。  第一份关于安慰剂效应的报告发现，接受安慰剂治疗的五分之四的患者报告症状缓解；当他们接受实际治疗时，结果相同。该报告提到了希望、信念和想象力的力量。自那时以来，对安慰剂效应的研究一直在继续，结果显示患者在接受无效的安慰剂后出现不同程度的改善，包括与活性药物相当或更好的效果。对十五项研究的综述显示，平均而言，安慰剂在35%的情况下与活性治疗效果相当。对该综述存在缺陷的批评已经提出，但研究仍然显示安慰剂在约三分之一的情况下产生“作用”。1961年，Beecher进行了另一项综述，并发现37%的患者存在安慰剂效应。安慰剂效应这个术语被定义为无效物质的积极效应，归因于暗示的力量，或者说是“在一个公正实验中，安慰剂治疗组与未治疗对照组之间结果的差异”。  在社会科学研究中，安慰剂类似于接受一种看似治疗或操作的控制组，例如，研究的真正焦点是关于连续性和主题框架的犯罪故事，而不是读取关于一部新电影的故事的受试者。  安慰剂是给予受试者的无效或伪装的治疗；安慰剂效应是受试者对其的反应。这与霍桑效应不同，在第二章的“更多关于”框中有描述，它描述了受试者由于受到观察而改变表现的情况。在安慰剂效应中，受试者的改变是因为他们对治疗产生了期望、信念或希望，而不是治疗本身。为了有效，治疗必须通过安慰剂测试，也就是说，治疗在结果上必须显著优于接受安慰剂的对照组。  社会科学研究也发现了安慰剂效应，例如，对于声称能提高运动表现的产品、影响财务决策的因素、饮料标签与醉酒感知以及知名品牌等方面。在日常生活中也存在安慰剂效应，例如，路口的按钮让行人产生了一种控制感，但实际上并不真正影响交通信号灯。 |

**开始进行自己的研究**

现在我们对实验相对于其他方法的作用有了一些基本的了解，这一章将转向开始自己的实验的第一步：清晰而简明地陈述研究问题。在各种研究中，清晰而简明的写作都很重要，但在我个人的经验中，写实验报告尤为重要。因为实验可能比较复杂，并且许多社会科学家可能对实验方法不如对调查或其他技术熟悉，实验似乎更难让读者理解。因此，本书的一个关键焦点是帮助读者以平易近人的语言和术语编写实验，以便任何具有适度社会科学知识的人都能理解。这比你想象的要难。

我们从研究问题的陈述开始。这个问题的范围比具体的假设更广泛，尽管它们是相关的。编写假设将在后面的章节（第三章）中详细介绍。要开始一个新的实验，你必须提出一个要测试的想法。以因果关系的形式思考会对此有所帮助，测试一个事物是否导致另一个事物。例如，以下是三个（相当简化的）研究想法：

1. 看照片是否能提高道德推理能力？

2. 声音音调是否会影响广播播音员的可信度？

3. 女性政治家的身高是否会影响选民对其资质的评估和投票可能性？

在第一项研究中，处理组是向受试者展示照片，而对照组则没有照片。在第二项研究中，广播播音员的声音在处理组中被调高，在另一个处理组中被调低，而对照组则保持正常。在第三项研究中，矮个子和高个子女性的照片分别作为两个处理组，而平均身高的女性作为对照组。当然，这三个实验最终比这个简单的“A是否导致B”的方法更为复杂，但这种简单的方法是所有研究的起点。

**撰写问题陈述**

下一步是撰写一个清晰而有重点的研究问题陈述。一个好的开始方式是以“本研究的目的是……”或类似的方式开头。以下是一些来自实验的好的、清晰和有重点的问题陈述示例：

* "本文研究了媒体对微博工具Twitter的使用如何影响人们对所报道问题的看法和信息的可信度。"
* "我们通过实验研究普遍认为金钱可以购买政治影响力的观点。"
* "我们评估沟通环境（即面对面与在线聊天室讨论）对个人表达观点的意愿的影响程度。"
* "首要问题是提供关于价格和市场福利属性的一般信息是否会改变对令人反感的交易的态度。"

所有这些例子都有三个共同点：它们说明了干预、操作或原因是什么，以及效果或结果是什么，通常按照这个顺序。它们都有一个动词。在第一个例子中，Twitter是干预或原因，而效果或结果是对问题的认知和可信度。在下一个例子中，金钱是原因，而政治影响是效果。在最后一个例子中，顺序被颠倒了，效果——道德信念和判断——被列在第一位，原因——伤害关联和道德情绪——在最后给出。您可能还注意到这些是独立变量（原因、干预或操作）和因变量（效果或结果）。第三个共同点是每个例子中都有一个动词，即描述原因对结果产生的影响或一个事物对另一个事物产生的作用的动作。在这些例子中，动词分别是“影响”，“购买”，“影响”和“改变”。对于实验研究，建议避免使用不太具体的词语，如“探索”，“理解”和“检查”，而是使用更具体、因果关系的语言。其他适合使用的词语包括“不同”，“改善”等类似词语。例如，一个处理可以说“本研究的目的是测试照片是否改善了道德推理”，或者“本研究的目的是观察到广播播音员声音的音高是否会影响对可信度的评估”。

为了为一项实验撰写一个清晰而专注的研究问题陈述，首先确定因果关系（或干预、操纵、独立变量）是什么，然后说明它预期会对某个结果（效应或因变量）产生什么影响（不同、影响、修改、引起——动词）。在这里，我提供了一个填空模板，用于撰写研究问题陈述：

本研究的目的是探究（某个因素，插入原因、干预、操作、处理或独立变量）如何（做某事，插入动词——不同、影响、修改、引起等）（某个结果，插入效果、结果、因变量）。”

如果这些例子中的某些术语，如伤害关联，不熟悉，那么检查您所在领域的一些实验，寻找其他更常见的例子可能会有所帮助。还要注意到所有这些例子都是清晰、具体和聚焦的，描述了实验的目的和问题陈述。撰写这样一个明确、聚焦且易于理解的陈述，用一句话描述实验，并不容易。很可能需要多次写作、修改和编辑。请熟悉该学科的人阅读并确认是否能理解。随着项目的发展，您可能需要重写这个陈述，甚至可能需要多次。

许多实验有多个目的，因此可以编写两到三个问题陈述。如果是这种情况，请将它们放在一起，并使用诸如“本研究还旨在……”之类的短语进行衔接。例如：“本研究测试照片是否改善道德推理。如果发现这种效果，本研究还旨在确定这种改善的因果机制。”这样，主要目的与其他目的都在同一个地方呈现，而不是分散在整篇论文中。读者会感激将研究意图的所有内容都列在一起，而不是阅读一个逐渐演变的研究目的，像是等待一个谋杀悬疑小说的情节逐渐展开。

在满意研究问题陈述后，作者需要定期提醒读者研究的使命，因此请确保问题陈述保持一致。我在审阅论文时经常看到一个问题，随着论文的进展，实验的问题陈述会发生变化。

最后，一个好的做法是将问题陈述写在便利贴上，粘贴在您的计算机上，以便您在工作时可以看到。这将有助于保持专注和一致。撰写这样一个明确、聚焦且易于理解的陈述，用一句话描述实验，并不容易。很可能需要多次写作、修改和编辑。请熟悉该学科的人阅读并确认是否能理解。随着项目的发展，您可能需要重写这个陈述，甚至可能需要多次。

许多实验有多个目的，因此可以编写两到三个问题陈述。如果是这种情况，请将它们放在一起，并使用诸如“本研究还旨在……”之类的短语进行衔接。例如：“本研究测试照片是否改善道德推理。如果发现这种效果，本研究还旨在确定这种改善的因果机制。”这样，主要目的与其他目的都在同一个地方呈现，而不是分散在整篇论文中。读者会感激将研究意图的所有内容都列在一起，而不是阅读一个逐渐演变的研究目的，像是等待一个谋杀悬疑小说的情节逐渐展开。

在满意研究问题陈述后，作者需要定期提醒读者研究的使命，因此请确保问题陈述保持一致。 我在审阅论文时经常看到一个问题，随着论文的进展，实验的问题陈述会发生变化。

最后，一个好的做法是将问题陈述写在便利贴上，粘贴在您的计算机上，以便您在工作时可以看到。这将有助于保持专注和一致。

回答“那又怎样”的问题

在制定了这个明确而专注的陈述之后，下一步是阐述为什么这项研究很重要。这通常被称为回答“那又怎样”的问题。有些期刊甚至有一个特别突出的部分专门用来阐述这一点——例如，参见《美国国家科学院院刊》的“意义”部分。这是研究人员常常认为他们的研究的重要性是显而易见的，不需要指出的另一个领域；事实上，这是一项最关键的任务之一。对于这个任务，我告诉学生他们需要陈述显而易见的事实。对于你来说可能非常明显，但对其他读者来说并不一定如此。为了做到这一点，可以从三个不同的层面考虑你的研究为什么重要：（1）对其他学术界人士的重要性，（2）对该领域的专业人士的重要性，（3）对社会或一般人群的重要性。对于第一个层面，可以指出某个知识空白或研究克服的障碍。

此外，研究还可以在某种程度上对理论做出贡献，或者帮助揭示某些现象的机制或原因。有一种原因单独来看是不可接受的，那就是因为以前从未进行过这样的研究。这只是一个起点，你应该进一步阐明为什么除了以前从未进行过这种研究之外，它的重要性。否则，也许以前从未进行过这项研究是有充分理由的。

经常看到研究人员强调对其他学者的重要性，但忽视了其他两个群体：专业人士和公众。对于任何类型的研究，能够清晰地阐明你的研究结果对于学术界以外的领域的意义至关重要，这对于有实际影响的研究至关重要。R. Barker Bausell的书《Conducting Meaningful Experiments》正是基于这个前提。在考虑“那又怎样”的问题的第二步中，请问你的研究对你所在学科的专业人士是否感兴趣？它是否能帮助会计师、竞选经理、教师、公关专业人士或其他人更好地完成工作？研究结果是否能够让专业人士更深入地了解自己潜意识决策的过程？它是否能帮助他们克服一些障碍或提供他们改变实践方式所需的证据？也许它能够告诉他们哪些努力能够达到他们期望的结果，哪些努力则不能，无论是更积极参与的公民还是更多的客户。通过进行这种有针对性的努力，可以在学者和服务对象之间架起沟通的桥梁。

最后，能够说明为什么整个社会都会受益将产生更有意义的科学研究。公众往往认为学术研究人员是“象牙塔中的书呆子”，他们写的东西没有现实基础，只是为了发表论文。甚至有一些奖项因最浪费资源的研究而成为新闻。如果研究能够更好地阐明为什么看似愚蠢或显而易见的发现对于学术同行以外的其他人来说很重要，就可以避免自己的研究遭受这种嘲笑。这种关注对于推进研究事业或增加研究资金没有任何帮助，而进行对普通人有意义的研究则是有意义的。所以要回答“那又怎样”的问题：它是否会帮助任何人或改善某些社会问题？并不是每项研究都能达到世界和平的目标，但它可能有助于减少种族歧视、改变一种道德上令人反感的做法，或者为决策者提供制定法律所需的信息。并不是每项研究都能同时阐述这三个层面，但通过深思熟虑对这些不同受众的益处，实验研究人员可以设计出真正有意义的研究。有关回答“那又怎样”的问题的陈述示例，请参见《如何做》1.4节。

**内容延展 1.4**

**“So What”问题的回答示例：**

**从：Neil, Nicole和Emily A. Jones. 2015. “Studying Treatment Intensity: Lessons from Two Preliminary Studies.” Journal of Behavioral Education 24: 51–73.**

“只有最近的一些研究关注了治疗强度，并且对于特定疾病的治疗强度的研究更为有限。也许特定病因和相关特征会影响干预强度的效果。许多患有唐氏综合症的儿童表现出差的任务持续性和不一致的动机导向...对于一些患有唐氏综合症的学习者来说，可能存在适度强度的最佳水平，在此水平之后，学习者参与更多的逃避行为和获得速度的增益递减。”59

从：Coleman, Renita. 2011. “Color Blind: Race and the Ethical Reasoning of African Americans on Journalism Dilemmas.” Journalism and Mass Communication Quarterly 88 (2) (夏季): 337–351.

“这项研究具有价值，因为它提供了评估解决刻板媒体形象问题所提出的一个解决方案的重要信息 - 雇佣和提升更多的少数民族记者。全国各地的新闻编辑部主要由白人组成；根据论点，加入更多的少数民族观点应该导致对少数群体更平等的报道。因此，重要的是要考察少数民族记者在认知加工中是否确实表现出对少数群体更宽容的态度。迄今为止，没有实证证据表明黑人记者对新闻中的黑人有更积极的看法...这项研究还填补了这方面的空白，探讨了种族如何影响黑人在涉及黑人和白人的新闻报道中的道德推理。”60

从：Aday, Sean. 2006. “The Framesetting Effects of News: An Experimental Test of Advocacy Versus Objectivist Frames.” Journalism and Mass Communication Quarterly 83 (4) (冬季): 767–784. “具体而言，我们对于为什么在某些情况下发现效果而在其他情况下却没有发现效果了解还不够，而且对于探索这种效果的认知基础的研究还很少，这将使我们能够建立一个关于一些属性何时以及为何具有二级效应的理论。”

从：Elias, Julio J., Nicola Lacetera和Mario Macis. 2015. “Markets and Morals: An Experimental Survey Study.” PLoS ONE 10 (6) (6月1日): 1–13. Public Library of Science. “禁止其中一些交易是有代价的。例如，寿险合同曾经被视为违背上帝的赌博而被禁止，但现在它们为数百万人创造价值，并被视为一种‘制度化的利他主义’。同样，虽然有关其效率的论证早就提出，但在美国长期以来一直拒绝实行全志愿付费军队的想法。禁止对捐献器官的人进行支付导致器官需求和供应之间的差距日益扩大。禁止某些交易也可能导致非法市场的形成，这反过来又带来了进一步的成本，如暴力。”

从：Grober, Jens, Ernesto Reuben和Agnieszka Tymula. 2013. “Political Quid Pro Quo Agreements: An Experimental Study.” American Journal of Political Science 57 (3) (7月): 582–597.

"有充分的理由怀疑政治中的金钱不受欢迎的普遍观念背后有一定的真实性。首先，尽管被禁止，政治上的交换可以在公开观察不到的渠道之外发生。其次，对于那些经济实力强大的特殊利益集团，其中大多数是大型企业公司，以增加利润的形式给予支持更有可能是一种比政治参与更合理的解释。此外，对这些利益集团的回报，如特定的税收减免、补贴和法规，可以轻易地掩盖为经济上的必要性，因此很难量化。第三，主要候选人之间的勾结也可能采取关于某个特定政治问题的共同观点的协议形式......最后，即使公众对政治中金钱的影响被夸大了，这种信念仍可能影响公众的政治信任和行为。"63

**常见错误**

* 在论文中没有清晰地陈述研究的目的，并且没有在整篇论文中保持一致。
* 没有将研究的所有内容逐步呈现出来，而是在研究进行过程中逐渐揭示研究目标。
* 在论文中没有在开头和第3页之前清楚地陈述研究目的。
* 未能说明为什么该研究对理论、其他研究人员、职业和普通人重要。"因为从未有过类似研究"本身并不是一个研究重要性的理由。

**检测您的知识**

1. 实验需要表明因果关系先于效果，但不一定与效果相关。如果两者之间存在统计学上的显著关系，那就足够了。

a. 正确

b. 错误

2. 一位研究人员研究了吸引力对学生对教师的喜好程度的影响。吸引力较高的教师年龄为25岁，吸引力较低的教师年龄为45岁。这里的问题是：实验需要展示因果关系先于效果，但不必证明二者必然相关。如果两个事物之间存在统计上的显著关系，那就是最重要的

a. 年龄与吸引力存在混淆变量

b. 很难定义吸引力

c. 因果关系不在因果前

d. 研究人员没有控制政治意识形态

3. 你将15名员工分配到为期一天的压力管理研讨会，另外15名员工分配到为期一周的

研讨会。一个月结束时，你测量每位员工的感知压力水平。这项研究中的处理或操纵

是什么？

a. 员工的压力水平有多高

b. 研讨会的持续时间有多长

c. 如何选择这30名员工

d. 压力管理导师的质量

4. 一项研究在学生参加考试前测量他们的兴奋水平。结果发现，随着兴奋水平的增加，表现下降。这一发现表明：

a. 因果关系

b. 相关关系

c. 一个合理的替代解释

d. 混淆变量

5. 以下哪项不是实验的三个基本标准之一？

a. 因果必须先于效果发生

b. 效果不太可能是偶然发生的

c. 因果必须与效果相关

d. 对于效果没有合理的替代解释

6. 变异是通过以下哪种方式实现的？

a. 保持所有条件不变

b. 使用人口统计作为协变量

c. 系统地改变某些事物

d. 设置一个对照组

7. 可能提供合理替代解释的事物被称为：

a. 协变量

b. 混淆变量

c. 独立变量

d. 因果机制

A white background with black text

Description automatically generated

8. 在实验中，对照组的作用是：

a. 让我们了解接受处理的人会发生什么情况

b. 让我们能够推广到更多的人

c. 让我们知道如果受试者没有接受处理会发生什么情况

d. 让我们能够以特定程度的确定性说某种效果发生了

9. 在实验中，“分配”是指：

a. 受试者必须完成的任务

b. 论文中作者顺序的计算方式

c. 研究人员确保受试者相信实验是真实的方式

d. 受试者被分配到不同的干预或对照组的方式

10. 在实验中，分配受试者的黄金标准是：

a. 代表性地分配

b. 故意地分配

c. 随机地分配

d. 随意地

**应用练习**

1. 使用scholar.google.com或您学校图书馆的数据库，在您的学科中查找使用实验设计的研究。在搜索词中包含实验、实验设计或对照实验。阅读三个您最感兴趣的实验，并查找涵盖的概念。具体来说，确定治疗组或治疗组群。是否有对照组？如果有，用什么来表示“无治疗”？确定问题陈述和对“那又怎么样”的问题的答案。

2. 构思三个不同的研究，您希望通过实验来进行这些研究。也就是说，应该通过操纵或改变某些因素，以观察其对某个结果的影响。清晰而专注地陈述问题。解释为什么这对学术界、职业界和整个世界都很重要（“那又怎么样”的问题）。每个问题使用250个字来表述。这些问题不应该是简单的复制，而应该是新的想法，或者它们可以是对您复制的研究进行重要扩展的复制研究。

建议阅读：

1. Bausell, R. B. 1994.《进行有意义的实验：成为科学家的40个步骤》（原文：Conducting Meaningful Experiments: 40 Steps to Becoming a Scientist）。千橡市，加利福尼亚州：Sage出版社。建议阅读引言以及第1章和第2章。

2. Shadish, W. R., T. D. Cook, and D. T. Campbell. 2002. 《广义因果推断的实验和准实验设计》（原文：Experimental and Quasi-Experimental Designs for Generalized Causal Inference）。贝尔蒙特市，加利福尼亚州：Wadsworth Cengage Learning出版社。建议阅读第1章。

3. Thorson, Esther, Robert H. Wicks, and Glenn Leshner. 2012.《新闻与传播研究中的实验方法学》（原文：Experimental Methodology in Journalism and Mass Communication Research）。《新闻与传播季刊》（Journalism and Mass Communication Quarterly）89 (1): 112–124。

**第二篇 伦理和历史上著名实验**

*如果我比别人看得更远，那是因为我站在巨人的肩膀上。*

*——— 艾萨克·牛顿*

**学习目标**

**• 定位当今实验研究中所使用的方法、技术和概念的历史起源。**

**• 阐明创造力在解决研究问题中的重要性。**

**• 总结实验研究的无意后果，作为制定伦理准则和伦理审查委员会的基础。**

**• 解释在实验中使用欺骗的有害影响。**

**• 描述如何在实验的背景下道德地使用欺骗。**

这一章简要偏离了实际的操作方法，为了提供一些背景来理解社会科学实验中当前问题、概念和技术的作用。这个简要历史将通过讲述一些最著名和臭名昭著的实验的故事来说明实验方法论的发展。

过去的历史可以追溯到一些方法、技术和概念的起源，例如随机化或将被试分配到不同组别的过程，以及为什么实验研究者使用诸如分割区块之类的术语。一些故事突出了回答问题的巧妙方式，鼓励研究人员在自己的研究中富有创造力地思考。创造力在研究中的重要性不仅在于追求新颖性，而且在于以最佳方式回答问题。其他的例子展示了保护参与者免受伤害的重要性，以及它们如何帮助推动伦理审查委员会（IRB）的发展，这些委员会负责确保研究人员遵守伦理准则。其中许多故事非常有趣，有些甚至被改编成了面向大众的电影。更重要的是，理解历史根源可以更好地掌握当代的实践。其他的例子展示了实验研究结果如何影响现实世界的问题，例如著名的研究探讨了为什么好人会做坏事。本章介绍了一些“巨人”的故事，正如艾萨克·牛顿的引言所说，实验方法论正是建立在这些巨人的肩膀上。

实验设计的历史中充满了引人入胜的研究，这些研究常常被炒作为最邪恶、令人毛骨悚然、离奇或者出乎意料的研究。那些无法再进行的研究也是“疯狂科学”类别中的热门主题。本章将采取不同的方法，不再详述这些实验产生的发现或理论，而是着重介绍随机分配、对照组以及同伴（Confederate）的使用等重要发现的发展和早期运用。重要的是，本章还将包括伦理准则的发展和IRB的形成过程（第11章将更详细地讨论IRB和其他伦理问题，这里不完全涉及）。许多这样的研究已经通过流行传说广为人知；在这些情况下，我尽量避免重复显而易见的内容，而是突出一些不太熟悉的方面。本章并不旨在详尽无遗地涵盖所有重要的实验，必然地会有很多被遗漏的实验。特别是，传统的历史倾向于忽视女性和有色人种的实验者。本章在“更多内容……”框2.1中涉及了一些被遗忘的人物，即女性的贡献。许多历史学家将实验的发展追溯到古希腊或其他时期。在介绍了一项非常早期而重要的医学实验之后，我将本章的剩余部分献给了更现代的社会科学实验。

**坏血病研究**

早期的一些实验是在1747年针对坏血病进行的。尽管如今罕见，但在船上特别成问题，因为它会导致水手体弱贫血，皮肤出血，牙龈腐烂。船医詹姆斯·林德进行了一项最早的对照实验以寻找治疗方法。他选择了十二名患有坏血病的男子，尽量使他们“相似”。然后他将他们分成六组，每组两人。这六组人被分别给予不同的治疗方法，如柑橘水、醋、硫酸等。通过观察不同组的反应，林德发现柑橘水对于治疗坏血病效果最好。这项实验为坏血病的治疗方法奠定了基础，也标志着实验研究方法的发展。

### **MORE ABOUT . . . BOX 2.1**

**Contributions of Women**

尽管社会科学实验设计的早期先驱大多以男性为主，但也有许多重要的女性，包括有色人种女性。这些科学家在职业生涯中面临着各种障碍和歧视，包括缺乏奖学金、被拒绝进入研究生项目以及无法在允许研究和发表论文的学术职位上就业。一些女性完成了博士学位的研究工作，包括毕业论文，但被拒绝授予学位。男性获得了一些女性贡献的认可。一些女性与丈夫合作，而反亲戚主义政策阻止了她们的雇佣。种族歧视给少数族裔女性带来了更大的负担。

很少有历史包括女性，主要是指她们的理论贡献而不是她们在实验设计方法论上的进展，而这正是本书的主题。接下来是三位女性的简要介绍，其中包括一位非洲裔美国人，她们对实验方法论做出了贡献。更多的女性当然值得被认可。

玛丽·惠顿·卡尔金斯是实验心理学的先驱之一，她因发明了成对联想任务而闻名，这是一种使用配对数字和颜色进行记忆测试的方法。尽管她完成了博士学位的所有要求，但哈佛大学拒绝授予她学位，原因是她是女性。耶鲁大学和密歇根大学向她提供了录取机会，但她拒绝了，因为这些学校没有实验室，而哈佛有。对女性学者来说，在女子学院找到学术职位更容易，所以卡尔金斯去了韦尔斯利学院，并在那里建立了一个实验室。尽管没有博士学位，卡尔金斯在科学期刊上发表了四本书和100多篇论文。

以及哈佛签署的一份请愿书上有十三位重要的校友签名，但她在1927年再次被拒绝授予学位。

卡尔金斯在本章中与另一位实验研究者约瑟夫·贾斯特罗交集，贾斯特罗与查尔斯·S·皮尔斯一起帮助确定了随机分配的好处。贾斯特罗发表了一项研究，研究男性和女性在被要求迅速写出单词时产生的类型，得出结论认为女性的单词是重复的、个体的和具体的，而男性的单词是建设性的、有用的和抽象的。卡尔金斯批评了他的结论，指出环境、训练和女性社会化的混杂效应。她一生都是妇女权益的倡导者，多次拒绝接受哈佛哈佛德学位，因为她在哈佛的工作成果。

玛米·菲普斯·克拉克是一位非洲裔美国心理学家，因她在种族研究中哥伦比亚大学。从哥伦比亚获得博士学位的第一位非洲裔美国男性是她的丈夫和研究伙伴肯尼思·B·克拉克。他始终认为是她提出了玩偶测试的想法。

玛米·克拉克在实验方法方面的贡献包括向儿童展示两个完全相同的玩偶，一个是黑人，一个是白人。问孩子哪个玩偶是坏的，哪个是好的，他们喜欢和哪个玩偶玩，并且哪个玩偶最像他们自己。许多黑人孩子认为黑人玩偶是坏的，几乎一半的孩子认为白人玩偶最像他们。在种族隔离的学校里，这种情况比集成学校更为明显。克拉克夫妇的实验成为证明种族隔离对儿童有害的重要证据，并对第一个大规模生产的黑人婴儿玩偶产生了影响。克拉克从未在学术界找到职位，而是在一个儿童之家担任研究员和临床心理学家，直到她和丈夫在哈莱姆开办了一家为少数族裔儿童提供测试和咨询服务的中心。

玛丽·埃因斯沃思因“陌生情境”评估技术而被人们铭记。使用这种方法的研究有助于心理学家了解儿童与照顾者的依恋关系。埃因斯沃思的方法是让研究人员透过单向镜观察孩子在大约三分钟的八个不同情景中的行为，这些情景是母亲和孩子来到研究人员的实验室，房间里有玩具。不同的情景包括陌生人进入房间并试图与孩子交朋友，母亲将孩子独自留给陌生人，母亲和陌生人离开孩子独自待着，陌生人再次进入安慰孩子，最后，陌生人抱起孩子。每15秒记录一次观察结果，使用1至7的评分标准。该评估具有很好的可靠性，即其他研究人员可以重现研究结果。作为实验室实验的典型特征，这种方法受到批评，认为它是人为的，缺乏生态效度，并且基于伦理原因对幼儿造成压力。

还有许多其他女性通过编写或合著其他测试和测量工具为实验方法论做出了贡献，包括格雷斯·肯特-罗桑诺夫的词语联想测试、弗洛伦斯·古登努夫的画人测试以及格雷斯·费尔纳尔德的品格测试，这些测试在道德发展等其他测试之前出现，等等。然后，这些群组接受了不同的治疗，包括一夸脱苹果酒、硫酸、半品脱海水、一剂由大蒜、芥末和辣根混合而成的“一颗肉豆蔻大小”的药剂、醋，或者两个橙子和一个柠檬。林德在这项医学研究中设定了六个治疗条件，在当今社会科学中被认为是很多的，通常情况下会有两到四个条件。他只有很少的受试者数量，仅有十二人，每个治疗组只有两人。在医学研究中，小样本量比社会科学中更为常见，因为医学研究中通常存在着更显著的效应。这项研究中的另一个重要问题是如何将人们分配到不同的治疗条件中。林德认识到了这些群组中人们在重要因素上的差异，了解到人们固有的个体差异，例如年龄、体重和一般健康状况，可能会影响结果。为了得到不受这些外部因素混淆的治疗效果，他选择了在某些要求上尽可能相似的男性作为样本。在这里，他采用了后来被称为匹配策略的方法，将具有相似特征的男性进行配对。重要的是，他只操纵了这六个条件，同时保持其他一切（他能想到的）都保持不变。他的自变量——这六种不同的膳食补充剂，并非来自于理论，而是来自先前提出的治疗方法。然而，林德确实写到了关于坏血病的理论大多是从未见过该病症的研究者的猜测，并主张理论与实践相结合。他对其中一种治疗剂量的测量——肉豆蔻大小——并不十分精确；如今，这将以克或类似的单位进行测量。无论从任何标准来看，林德的研究都取得了巨大的成功。仅仅在六天之后，那些摄入橙子和柠檬的男性就恢复了健康。其他人也有所改善，但吃柑橘水果的人与其他人相比有了显著的恢复，这最终导致人们认识到维生素C才是起作用的物质。虽然没有进行显著性统计检验，但六天后的效果是显而易见的。她的成就使得哥伦比亚大学和史密斯学院授予她荣誉博士学位。尽管获得了这种认可

这项研究之所以重要，有几个原因：它认识到人们需要以一种不受个体差异影响的方式被分配到治疗组中，而林德通过精心选择具有相似特征的男性来实现了这一点。这发生在1747年。到了19世纪末，将受试者分配到条件中已经成为一个紧迫的问题，出现了不同的解决方案，尽管这些方案更多地应用于治疗方法而不是受试者，如下一个故事所示。林德的研究在任何标准下都是巨大的成功。仅仅经过六天，那些食用橙子和柠檬的男性就康复了。其他男性也有所改善，但食用柑橘水果的人相比其他人有了明显的恢复，最终导致人们认识到维生素C是起作用的因素。并没有进行统计学上的显著性检验，但六天内的效果是显而易见的。

这项研究之所以重要有几个原因：首先，它认识到人们需要以一种不受个体差异影响的方式分配到治疗组中，林德通过精心选择具有类似特征的男性成功地实现了这一点。这是在1747年。到了19世纪末，将受试者分配到条件中已成为一个热门话题，并出现了一种不同的解决方案，尽管这种解决方案是应用于治疗方法而不是受试者，这将在接下来的故事中说明。

**查尔斯·皮尔斯的贡献**

查尔斯·桑德斯·皮尔斯和他的学生约瑟夫·贾斯特罗（Joseph Jastrow）据说是第一批使用随机分配的人。他们在1885年进行的一项研究中首次采用了随机分配，以观察人们是否能够通过感觉和观察来判断物体的重量。皮尔斯关注的理论构建是判断错误的源头，从而产生了诸如“刚刚能察觉的差异”之类的概念，这些概念在商业、营销和广告学者中得到了认可。在第一项研究中，皮尔斯始终从最重的物体开始和结束实验。在第二项研究中，他尝试将最重的物体与最轻的物体交替使用。最后，他描述了如何使用一副扑克牌进行随机分配为了随机分配重物的顺序，皮尔斯的结果与前两次试验完全不同，这显示了这种技术的重要性。他使用了一副扑克牌，简单地洗牌后抽出一张牌来决定下一个重物的顺序。如今，我们倾向于使用随机数生成器，但一副扑克牌仍然可以同样有效地使用。他提到这种方法有时会产生“长时间以一种特定的变化方式运行，这种变化有时会出现凭借机会”，但他指出这比被试知道没有这样的模式更可取。对于大多数人来说，“随机”意味着不可预测；但对于皮尔斯来说，随机意味着“在长期运行中，任何一个个体被选中的次数与其他个体一样频繁”。当以这种方式从较大的类别中抽取较小的样本时，皮尔斯说较小的群体会显示出与较大群体相同的特征。他称之为“归纳规则”。第7章将深入探讨社会科学中的随机分配；如今，我们更常常考虑将人们随机分配到不同的处理组，但正如皮尔斯和其他人的早期研究所示，随机化尽可能多地包括实验中的刺激是很重要的。

顺便提一下，为了展示皮尔斯的能力范围，他还因发展了符号学和象征学理论而闻名——这是一种经典理论，今天仍然在批评/文化研究中使用。

**罗纳德·费雪的茶**

几十年后，罗纳德·艾尔默·费雪重新提出并推广了治疗的随机分配的想法，从而导致其被广泛使用。费雪是20世纪20年代和30年代伦敦外的一位农业科学家。他被认为是实验设计的发明者，包括今天仍在使用的许多统计学、概念和程序。例如，费雪开发了方差分析，其统计量F检验以他的名字命名，并且还提出了拉丁方阵的概念，一种用于对治疗进行排序以补偿系统误差或控制意外变异的程序（关于这一点将在第7章中详细介绍）。费雪还被认为提出了0.05和0.01的概率值作为统计显著性的标准，这是科学家判断某个事物是否偏离机会的方法。0.05的p值意味着“结果错误的概率为二十分之一”。费雪的工作是在英国罗瑟姆斯特德农业实验站上对各种肥料进行测试。因此，许多今天在实验中仍然使用的术语都具有农业基础，例如分割剧情设计，实际上是指将田地分割成不同部分，以便每个部分可以接受不同的处理。费雪到达实验站时，一直以来的惯例是每年测试一种不同的肥料。但是费雪意识到，每年带来了降雨量、温度、杂草生长、排水和其他因素的固有差异。当他意识到无法将肥料的效果与这些无法控制的其他条件分离时，他创造了混杂这个术语。为了确定结果是由肥料还是其他因素引起的，他决定将所有的处理方式都纳入到同一个实验中，对他来说，这些处理方式是许多不同种类的肥料。通过将田地划分为小区，将小区划分为不同的行，并给每一行分配不同的处理方式，他推理出每年的差异将适用于所有的处理方式，有效地“控制”了这些混杂因素。此外，费雪还被认为推广了随机分配的方法。这也源于他在努力控制各种条件时意识到，将不同的肥料有系统地分配给田地不能排除与土壤和田地本身相关的混杂因素。为了解决这个问题，他随机分配田地接受不同的处理方式。在当今社会科学中，我们认为随机分配适用于被分配到不同处理方式的受试者；对于费雪来说，田地就是他的受试者。从随机分配的意义上讲，“随机”意味着按照机会程序，例如抛硬币，确保每个参与者有平等的机会被分配到任何一个治疗或对照组。这有助于确保任何系统性差异在不同组之间平均分布，从而可以将任何差异归因于治疗，而不是组内人员固有的差异。第7章将详细讨论随机分配及其实现的更多细节。费雪进行的一个较为引人注目的随机分配实验是现在著名的“品茶女士实验”。它还说明了零假设的概念，即认为接受治疗和不接受治疗的人之间不会有差异。故事是这样的，一个声称她可以通过品尝告诉奶先加还是茶先加的女士被给予了四杯先加茶的杯子和四杯先加奶的杯子，其顺序是随机的。零假设是她无法将它们区分开来。有人全部正确猜测的概率是七十分之一；据说这个研究中的女士全部都答对了。这个品茶实验据说是一个夏日的娱乐活动，并不是一项具有出版结果的科学调查。费雪在他的著作《实验设计》的第2章中描述了这个实验。他没有说明是否实际进行了实验或给出了结果，所以其中的许多内容都是民间传说。在费雪之前，已经进行了几百年的实验，但它们是非常个人化的，每个实验者的方法都有所不同。费雪的书是第一本系统总结如何进行实验的著作。

**B. F. 斯金纳：小样本，高技术**

费舍尔的原理在B.F.（缩写自Burrhus Frederic，但朋友们都叫他Fred47）斯金纳开始研究老鼠的行为时才开始受到欢迎。48斯金纳以创建斯金纳盒子而闻名，这是他用来训练老鼠推动杠杆以获取食物或停止电击的装置，他在开发他的操作性条件反射理论时使用了这个设备。在这里不对他的理论或研究进行详述，但基本的前提是动物学会对刺激做出反应，而强化某种行为会使动物重复这种行为。尽管斯金纳的想法有时会引起争议并被误解，但他的工作和理论仍然在激励着今天的研究，包括在人类心理学、动物行为、教育、营销、健康、社会工作、运动训练和其他领域。49这里的讨论是专门针对他对实验研究方法的观点，因为这就是本书的目的。

斯金纳被认为是开发了一种他称之为行为的实验分析的心理学实验研究。50这与当时大多数心理学实验所使用的演绎方法形成了鲜明的对比，该方法首先形成一个假设，然后测试看看它能否被证伪——这种方法仍然是现在最常用的方法。斯金纳的方法是归纳的和以数据为驱动的，特征是观察和实证测量行为。他从未使用过正式的假设，他说：“如果我参与了实验设计，那只是为了完成或扩展已经观察到的某种秩序的证据。”51他在他的书《生物的行为》52中详细地发布了他的方法，这比当时的常规更详细。53

他的大部分研究都是以个案进行的，一次只进行一个主题或在小组中进行；在《生物的行为》中，大部分研究都使用了四只老鼠。54他为他使用小样本的做法进行了辩护，并描述了它们可能比大样本更好的情况55——例如，当研究招募大量同质性受试者困难的群体时，如特殊教育学生、精神病患者或残疾人。58今天，这种方法被称为单一案例实验设计，仍在用于此类人口，尤其是在临床工作中。59

斯金纳研究大群体的一个困难原因是每个受试者都需要一个斯金纳盒子的限制。他的实验室里只有四个这样的盒子（他更喜欢称它们为“操作性室”）。60他描述了他曾经如何获得一项资助，用以制造足够的盒子同时测量24只老鼠，并将数据集中在一个平均性能曲线中。他和他的同事在一项研究中能够运行95只老鼠，斯金纳写道“使用大群体的动物的可能性大大提高了以前报告的方法，因为提供了显著性的测试，并且在单一案例中不明显的行为特性可能更容易被检测到。”61但是，他最常见的做法是运行个别受试者或样本太小以至于无法进行任何统计分析。他依赖功能分析来确定干预是否有效。例如，当八只鸽子中有六只执行了他们被训练进行的行为时，“所产生的反应定义得如此清楚，以至于两个观察者可以完全一致地计算实例。”62他觉得当真正需要的是更严格的控制和技术时，统计被过分强调。63他将“控制你的条件，你将看到秩序”这个洞见归功于巴甫洛夫。64

斯金纳还将实验方法从研究者的观察推进到更客观的测量工具。斯金纳从小就是一个发明家，当他还是个男孩的时候，他就建了一个设备提醒自己把睡衣挂起来。65为了配合斯金纳盒子的工作，他设计了一个“累积记录器”，用笔和纸卷自动记录当老鼠按下杠杆或鸽子啄键时的情况。其他研究者广泛采用了反应率的使用。66他继续设计了他所称的“小工具”或“设备”来进行许多其他用途。67例如，他使用秒表记录老鼠跑动和发出声音的行为。68另一个设备使用了一个绕在纺锤上的细线来记录数据的曲线。“我知道科学大量利用曲线，尽管根据我所了解的，在聚集图（polygram）上的小纹理（pips）使用得非常少。”69他在谈到他以前的测量方法时写道。“结果显示，曲线揭示了反应速率以及该速率的变化中的事物，这些事物在其他情况下肯定会被忽略。”70响应曲线在许多学科中仍然是一个重要的原则。71他另一项发明是改装一台药剂师制药机，制作出大小统一的老鼠食粮颗粒，解决了现有老鼠食品大小不受控制的问题。当他意识到为了强化每一个行为，他需要制作多少老鼠食粮时，他开始每分钟只强化一次，这启动了一项关于“周期性强化”的新的研究项目。72

这些实际问题驱动了他的大部分研究，他开始将这些问题视为意外的幸运，而不是烦恼。例如，当老鼠食粮输送装置卡住时，他写道：“我最初将此视为一个缺陷并急于修复”它。73然后，他意识到看似挫折的事情实际上提供了发展“消退曲线”的机会。“毫不夸张地说，一些最有趣和令人惊讶的结果首次出现是因为类似的意外。”74

斯金纳将科学称为一个“混乱且偶然的过程”，没有“明确定义的开始和结束”。75《生物的行为》76描绘了科学的混乱性和偶然发现的角色，也就是在寻找一件事物时发现另一件事物。（关于现代的例子，参见2.2研究聚光灯的“混乱”研究。）在斯金纳的例子中，他乐于看到这种情况。他用自己的话说：“这里有一个科学方法学家们并未正式认可的第一原则：当你遇到一些有趣的东西，把其他所有事情都放下，去研究它。”77

这仍然是今天的好建议。

#### 斯坦利·米尔格拉姆震惊世界

斯坦利·米尔格拉姆的研究是一些最为人熟知的，无论是学者还是普通大众都知道。俗称为“电击实验”，斯坦利·米尔格拉姆关于顺从的实验旨在探究人们为什么会服从权威人士的命令，即使这些命令违背了他们自己的良知。在20世纪60年代，

**学习要点 2.2**

**“混乱”的研究示例**

巴巴德、皮尔和贝纳约恩（Babad, E., E. Peer, and Y. Benayoun）在2012年的《应用社会心理学杂志》（Journal of Applied Social Psychology）发表的研究论文《一个情境中是否可能存在多重偏见？从媒体偏见研究中寻找证据》（Can Multiple Biases Occur in a Single Situation? Evidence From Media Bias Research）是研究混乱性的当代案例。这篇文章中罕见地坦率解释了研究如何偏离轨道、产生了意想不到的结果和违反直觉的发现。然而，这种情况比它看起来的频率更高，发现在论文中能找到这样坦诚的讨论令人耳目一新，也让人感到安心。这篇文章还展示了计划之外可能出现的意外结果。

这篇论文的第一句话承认了这个想法是由于实验研究的意外结果而出现的。这项研究涉及影响人们判断和行为的认知和情绪偏见。在这些研究人员的领域中，这意味着当电视记者采访政治候选人时，他们的非语言偏见（NV行为）是无意识的。这些研究人员研究的问题是实验设计中一些最好的技术导致的，具体来说，设计出不受干扰影响的实验与现实生活中发生的情况不同。此外，实验研究者倾向于一次研究一种偏见，而不像在日常生活的复杂性中，同时存在多种偏见的情况更有可能。作者们承认，最初他们没有想到这个想法，也没有发表的研究指出可能存在多重偏见的可能性。他们说：“我们必须承认最初我们没有构思出多重偏见的概念，并且也没有研究表明可能存在多重偏见的可能性。”78

在得出两种不同的偏见同时起作用的结论之前，研究人员进行了自己研究的艰难复制，涉及了七个不同的样本。这篇论文报告了这些研究的元分析结果。在研究结束时，他们坦率地指出哪些发现是出乎意料的、违反直觉的、令人惊讶和令人失望的。他们承认当发现难以解释时。例如：“在这项研究中，光环效应并不是真正预期的，因为实验并没有被设计成产生光环效应...最令人意外和违反直觉的结果，也许具有最广泛的概念和应用意义，是两种偏见现象的独立性，这一点通过拒绝替代光环效应假设得到了证明...很难解释这种效应，确实是令人费解的。”79

这些研究人员并没有试图掩盖和伪装，他们展示了实验研究的困难和混乱，这种情况比发表的文章给研究人员带来的印象更常见。

**斯坦利·米尔格拉姆震惊世界**

斯坦利·米尔格拉姆的研究是一些最为人熟知的，无论是学者还是普通大众都知道。俗称为“电击实验”，斯坦利·米尔格拉姆关于顺从的实验旨在探究人们为什么会服从权威人士的命令，即使这些命令违背了他们自己的良知。在20世纪60年代，

纳粹战犯审判正在进行中，因此对这个问题的兴趣非常浓厚。米尔格拉姆的实验因引发心理伤害和欺骗性而受到心理学家的批评。这些批评引发了对社会科学研究伦理的更广泛讨论。这些研究也被认为有助于建立IRB（伦理审查委员会），以确保研究不会伤害志愿参与的人（有关IRB的更多信息请参见第11章）。

虽然这些研究主要被记作关于伦理研究的警示故事，但其中还有其他方面适合作为实验设计教训的内容。米尔格拉姆的所有研究的主要结构是让一个“天真”的受试者（即不知道研究目的的人）“教授”另一个人一组词语配对。受试者被告知目标是观察惩罚（电击）是否能提高学习效果。在这种情况下，“学习者”是实验者的同谋或共谋者，他知道研究的目的，只是假装受到了电击。在那个时候，这样使用共谋者和欺骗的实验很受欢迎。米尔格拉姆在每个实验结束时进行了解释说明会，告诉受试者实验的真实目的，即他们实际上并没有伤害任何人，他们的反应是正常的。在发表了两篇文章并开始出现批评之后，米尔格拉姆开始报告他采取的措施来最小化对受试者的伤害以及他进行的事后调查以了解是否存在伤害。戴安娜·鲍姆林德（Diana Baumrind）驳回了米尔格拉姆的事后实验自我报告，这些报告显示超过80％的受试者对他们的欺骗性实验表示高兴，她说这些报告是“事后加上的”，“毕竟，如果自我报告可以被视为衡量实验条件影响的准确指标，我们完全可以摒弃实验操纵和行为测量。”如今，共谋者已经不再流行；然而，在涉及欺骗的实验中，解释说明仍然是最佳实践，尽管像米尔格拉姆的研究这样的实验今天很可能不会获得IRB的批准。然而，值得注意的是，在电视节目中，如ABC的秘密摄像系列节目《你会怎么做？》中，这种类型的“实验”仍然很受欢迎。当记者或娱乐制片人进行此类实验时，无需获得IRB批准。

第1章解释了因果关系的概念。米尔格拉姆感兴趣的效果是顺从，以最大电击的依赖变量来衡量，其范围从0到30.124。他所研究的原因是各种独立变量，包括受试者与受害者的距离、受试者是否能听到抗议和哭喊声、是否必须接触受害者、给予指示的人的权威性等等。

### **内容延展2.3**

**欺骗**

虽然社会科学实验中的欺骗可能不再像斯坦利·米尔格拉姆的电击实验那样引人注目，但它并不是过去的事情。它在不同学科中使用的方式各有不同；在主要协会的规范下，社会学和社会心理学中很常见，但在经济学中被禁止，尽管隐瞒信息不被视为欺骗，所以是被允许的。

欺骗被定义为有意提供虚假信息和隐瞒信息以误导他人。这并不妨碍研究人员隐瞒假设、实验条件或其他可能导致受试者改变行为的研究方面的信息。然而，如果隐瞒此类信息会导致受试者不同意参与，则超过了底线。

在可能导致身体疼痛或严重情绪困扰的研究中，禁止使用欺骗。指导方针要求研究人员在没有其他可行的方法或者研究价值得到充分证明时，不得使用欺骗。然而，正如鲍姆林德所说：“很容易让研究人员或同行评审委员会相信，一种巧妙的欺骗操纵的长期效益超过了被欺骗者的短期成本。”

欺骗性研究并不局限于斯坦利·米尔格拉姆的电击实验。在该实验之后，菲利普·辛巴达罗（Philip Zimbardo）在考虑使用其他方法和雇佣客观观察员来监督研究时进行了反思。然而，在随后的几年里，辛巴达罗和同事们使用催眠诱发部分失聪，导致受试者出现偏执症状，对受试者提供错误信息并隐瞒研究目的和他们将要经历的事情。就像监狱实验一样，这项研究和另一项涉及误导受试者的研究都获得了IRB的批准。1977年，斯坦利·米尔格拉姆的电击研究被复制，并对年龄只有六岁的儿童进行了实验。这项研究是在一个没有IRB的国家进行的，但获得了系主任和其他教师、校长的批准。研究报告称参与的儿童表现出“大声紧张的笑声，咬嘴唇，颤抖”。在文章的结尾，作者们说，“确实令人惊讶的是，相对较少的社会心理学家对米尔格拉姆在顺从方面的开创性工作进行了后续研究。”这项研究当时和现在都发表在顶级社会心理学期刊上，该期刊也发表了米尔格拉姆的另一篇论文。这并不意味着这些研究人员是不道德的，甚至是无视问题的，而是指出检测欺骗问题可能比人们想象的更困难。

内容分析显示了欺骗性研究的发展轨迹。从1921年开始，在领先的社会心理学期刊《人格与社会心理学杂志》中很少有欺骗，直到20世纪30年代。随着实验方法的兴起，它逐渐增加，尤其是在20世纪50年代至70年代之间。从20世纪80年代到1994年，欺骗性研究有所减少。研究发现，电脑和其他虚假设备部分取代了同谋者以误导受试者。在1969年之后，误导同意和对受试者的错误反馈有所减少，但在1992年又有所增加，研究主题的变化解释了在1978年至1986年间欺骗性研究的下降。对所检查的文章，发现其中高达66％存在欺骗性实践。研究显示类似的结果。在1971年至1974年间，54%的心理学实验使用了欺骗，比1948年的18%和1963年的37%有所增加。最近的数据显示，在社会心理学研究中，约有33%的研究存在欺骗性。欺骗可能对受试者造成伤害是显而易见的。除了身体疼痛或心理困扰之外，欺骗还会损害研究对象的自主权，违反他们选择参与的权利。

但是，欺骗性研究还具有对研究行业整体产生影响的潜在危害。正如许多社会科学家指出的，参与研究的自愿性已经在下降。这在很大程度上归因于公众对研究的怀疑，这是基于研究在历史上造成伤害、欺骗和滥用权力的情况。这不仅仅影响了召集受试者的能力，还可能削弱公众对研究结果的信任和自信，以及整个学术界的声誉。社会变得不愿意支持研究。此外，一些实验者使用欺骗的意愿也有可能引导其他人也这样做，从而削弱研究人员自身的道德机构性。对整个社会的伤害包括整体信任的损失和越来越多的怀疑，因为违背了讲真话和信守承诺的社会契约。

以合乎伦理使用欺骗的最常见原因是需要控制需求特征，即受试者改变自己的行为以适应研究的目标的倾向。鲍姆林德认为，使用欺骗会促使其本来打算避免的事情发生：因为人们知道欺骗在实验中经常被使用，受试者不再天真，也就是他们不再相信研究是真实的。与此同时，其他研究发现，欺骗不会影响研究的外部有效性。

除了完全避免使用欺骗之外，合乎伦理使用欺骗的最常见技巧包括：

正如米尔格拉姆所写：“该研究的关键在于系统地变化被认为会改变对实验命令服从程度的因素。”他总共进行了十八个不同的实验，分别发表在国际期刊文章中，并在他的著作《对权威的顺从》中报道了整个系列实验。

• 向受试者提供知情同意书，其中包括研究的真实目的，而不揭示假设，并描述研究的程序和他们将要做的事情。

• 如果有任何信息被隐瞒，在研究结束后立即向受试者进行解释说明。同时，应该询问受试者是否同意推迟接收所有信息直到解释说明会。然而，要认识到解释说明无法消除伤害。

• 允许受试者在任何时候退出研究，包括在解释说明后撤回他们的数据。

实验者应该意识到，他们可能倾向于选择欺骗性的方法，因为这些方法比非欺骗性的方法更容易。相反，研究者应该考虑非欺骗性的方法，比如自然实验，其中没有涉及操纵，或者使用不引人注目的工具，使被试无法改变他们的反应，比如心率的心理测量或潜伏期反应，用于测量人们对刺激作出反应所需的时间。自然实验在第4章中有所介绍，不引人注目的工具在第10章中有所介绍。在第11章中有关伦理和欺骗的内容，涉及IRB（机构审查委员会）的情境。其他关于非欺骗性实验的建议包括让研究者成为自己的研究对象，探究他们的动机和行为，122或者使用“代理被试”——即告诉人们关于研究目的和程序，看看他们是否接受。123最后，我个人的建议是应用黄金法则——就是像你希望别人对待你一样对待他人。将自己置于你的研究对象的角度，问问自己是否会对被欺骗、被说谎或了解一些关于自己令人痛苦的事情有任何反感。考虑后果，包括如果你的雇主、家人或朋友发现你宁愿保密的事情会是怎样的情况。

第1章还介绍了“为什么”问题，即解释为什么一个实验很重要。米尔格拉姆通过引用他的研究可以揭示的实际问题回答了“为什么”问题，从战争罪行到员工服从上司。他还想出了一个聪明的办法来回应对仅确认我们已知信息的研究提出的批评，而不是揭示新的或意外发现的研究。他设计了一个程序来确定他的研究结果是否显而易见和符合直觉，即进行了一项单独的研究，邀请受试者参加关于他研究主题的讲座，并描述了实验而不透露结果。讲座结束后，受试者被问及他们自己会怎么做以及他们认为其他人会怎么做的几个问题。所有110名受试者都表示他们会在某个时候拒绝给予电击，并且几乎所有其他受试者也会如此。而在实验中的情况恰恰相反，65%的受试者一直进行到最高水平的电击。因此，米尔格拉姆能够显示他的结果并不是显而易见的，实际上与人们预期的相反。

就像林德在他的坏血病实验中所做的那样，米尔格拉姆运用了理论和观察相结合的方法来测试服从的原因。例如，他在第8个实验中使用女性受试者，就引用了关于女性的矛盾理论作为原因。尽管在今天看来有些过时，但当时的一些理论和证据指出，女性比男性更服从和更少攻击性，这应该导致她们比男性更频繁地给予电击。而矛盾的理论和证据则表明，女性比男性更具共情能力，这预示着她们会比男性更少给予电击。实际上，女性和男性在电击实验中的表现大致相同，但在事后的访谈中，她们对给予电击的原因使用了不同的表述方式。第3章将涵盖理论对实验的重要性以及观察和经验如何推动实验问题的产生。

米尔格拉姆还在他的研究中讨论了抽样和参与者的问题。他明确拒绝使用本科生作为受试者，尽管这在今天相当普遍，但也引发了激烈的辩论。他指出，使用学生会更容易，因为他们可能已经从已经参与过研究的其他人那里听说过这些研究，并且他们太同质化，年龄、智力和对心理学研究的了解都相似。相反，他希望从不同社会阶层背景的广泛个体中选择。因此，他在报纸上刊登广告，从周围社区招募参与者，并以4.5美元的现金激励来吸引他们。米尔格拉姆在书中写到，每个条件下他都使用了四十名受试者进行所有的研究，但没有提到他是如何确定这个数字的。今天，会使用统计学中的功效分析（power analysis）来确定一个特定研究需要多少受试者。这将在本书的第8章中介绍。他也没有随机分配受试者到不同条件，这是为了确保等价性的特征，已经被其他人发现，但主要用于排序处理方式而不是分配受试者。相反，米尔格拉姆使用了一种类似于林德在坏血病研究中使用的配对技术。他提到平衡每个条件中的受试者在年龄和职业上的分布很重要，所以他将他们分成熟练工人和非熟练工人的类别。

在所有的研究中，米尔格拉姆还通过使用受害者的抗议录音来控制混淆因素，确保每个受试者和每个电击水平的抗议声都是相同的。

米尔格拉姆的研究采用了多种测量方法，这是第10章的主题。他的主要依赖变量是行为性的，即受试者所施加的“电击”水平，范围从零到最高的30级，代表着惊人的450伏特。实际上，最低的电击水平是20伏特。他还测量了电击的持续时间，即受试者按下开关的时间和按住开关的时间。

其他的测量指标包括受试者在实验结束后通过问卷报告的自我报告，包括他们在研究过程中经历的冲突、紧张和紧张感，以及他们对受害者所感受到的疼痛程度的估计。米尔格拉姆还使用了多种回答选择，包括开放式问题、投射性测试和态度量表。所有这些内容将在第10章中详细介绍。在米尔格拉姆的前四个实验中，还询问了受试者对实验者、受试者和受害者对于一个人被迫接受电击的责任分配情况。问卷还询问了通常的人口统计变量，如政治党派、宗教信仰、教育水平以及在这种情况下的兵役年限，因为这在1960年代是常见的，并且被认为对于一个人是否能很好地遵守命令很重要。

此外，米尔格拉姆还记录了实验过程中受试者和实验者之间的对话，并且研究人员记录了自己的观察。除了个别访谈和实验后的小组讨论，这些定性的测量方法还与实验后的个别访谈和小组讨论一起使用，这些内容在第1章中讨论了三重性的重要特征。

在米尔格拉姆的研究中，我们还发现了试点研究的例子，这是第11章的一个主题，以及操纵检查，这是第9章的主题。试点研究在正式研究启动之前用于测试实验，以解决问题。在使用本科生进行试点研究时，米尔格拉姆发现受试者需要练习读给学习者的单词，因此在正式研究开始之前，他加入了一个训练环节，受试者读十对单词。此外，试点研究中的另一个重要经验教训是“声音反馈”。最初，米尔格拉姆让受试者在另一个房间给受害者施以电击，而受害者保持沉默。他发现受试者在给予电击时没有变化，每个人都进行到最高电击水平。由于结果没有变化，没有人违抗命令，因此无法确定违抗的因果机制。因此，米尔格拉姆引入了受害者的呼喊、抗议和拍打墙壁的声音，这导致受试者在不同的时间点停止电击，从而获得了第1章中讨论的变化。

第9章的一个主题是如何确定操纵是否被视为真实。例如，如果实验要求受试者阅读用生动、描述性语言或非生动、抽象风格编写的信息，那么就有必要知道得到生动写作的受试者是否认为它比得到非生动写作的受试者更生动。

为了确定这一点，需要进行操纵检查。在米尔格拉姆的研究中，他想知道受试者是否相信受害者实际上接受到了痛苦的电击，并且是否相信他们自己是施加电击的人。只有两个受试者在实验后的问卷中回答这些问题时选择了“否”，因此认为操纵有效。

**菲利普·辛巴多：在斯坦福地下室中唤起良知**

斯坦利·米尔格拉姆的电击研究不久之后，菲利普·辛巴多进行了一项同样引人注目的研究。就个人经历而言，米尔格拉姆和辛巴多在1949年在布朗克斯的高中相遇过。他们在1960年作为助理教授重新联络，米尔格拉姆在耶鲁大学，辛巴多在纽约大学。巧合的是，辛巴多最初构建了一个实验在耶鲁大学的实验室用于他的电击研究之后，米尔格拉姆搬出了“优雅互动实验室”，而辛巴多在全国另一个地方的地下室进行了一项备受关注的研究，其中一个引人关注的问题是伦理问题。

1971年夏天，辛巴多在斯坦福大学心理学楼的地下室中建造了一个模拟监狱，用于一项由海军研究办公室资助的研究，该研究关注囚犯和军队狱警之间冲突的原因。在这里，来自全国各地的21名大学生扮演囚犯或狱警的角色，进行了为期六天的实验。然而，与1961年米尔格拉姆进行实验时不同，20世纪70年代已经有了伦理审查委员会(IRB)的存在。实际上，辛巴多和他的研究生们获得了IRB的批准。辛巴多的著作《魔鬼效应》描述了他采取的措施，以避免对受试者造成伤害，包括禁止使用身体暴力。所有参与者都经过知情同意，其中包括他们可以随时退出实验的信息。但是，与米尔格拉姆的研究不同，这项研究没有涉及欺骗。然而，研究人员给予了狱警一些建议，告诉他们应该如何行为，而不是让所有行为自然发展。这项研究被认为是有史以来最有害的研究之一。例如，它使用心理学技术引发了无聊感，剥夺了受试者的隐私、个体化和权力感。受试者对研究的反应非常强烈，扮演狱警和囚犯的学生以惊人的程度内化了他们的新身份，因此实验在计划的两周而不是六天后被终止，因为出现了心理和情绪创伤。有些囚犯情绪失控，尖叫不止，进行绝食抗议，而一些狱警表现出虐待行为，甚至用灭火器袭击囚犯。为了预防问题，研究团队在实验结束后对每个受试者进行了详细的解释，并对每个受试者进行了一年的跟踪观察以寻找不良影响。作为对批评的回应，辛巴多指出，美国心理学协会唯一收到的调查该研究的请求来自他本人，并且该研究得到了完全的批准。一项后续研究的提案被IRB拒绝，辛巴多在2007年的书中向参与者道歉。

监狱研究是一项实验室实验，通常受到高度人为因素的批评，因此辛巴多和他的同事们竭尽全力使其尽可能真实，就像米尔格拉姆一样。例如，他们说服了帕洛奥尔执法部门让真正的警察“逮捕”受试者。他们给狱警穿上与真正监狱制服类似的服装。囚犯整天呆在有钢铁栏杆的牢房中，

同时，监狱看守们按照常规班次工作。他们安排了真实的亲友探访日，并进行了模拟假释听证会。151 这项研究被誉为至今为止最具真实感的典范。152 研究人员指出，尽管他们希望实验尽可能真实，但在实际监狱中进行实验并不可行，因为“现实世界或社会科学家所称的‘领域’中有太多无法控制的变量。这就是实验室研究的便利之处：实验者掌握主导权。”153 他们指出，已经在现有的刑事机构进行了许多研究，他们称之为自然实验。154 实验室实验、自然实验和野外实验之间的区别将在第4章中探讨。辛巴多解释说：“社会学家和犯罪学家已经对实际监狱生活进行了研究，但它们存在一些严重的缺点。这些研究人员无法自由观察监狱生活的所有方面......他们只能看到他们被允许看到的。”155 他们解释选择进行尽可能真实的实验室实验，是因为他们的研究问题无法使用真实的监狱和囚犯进行研究，因为需要将监狱环境的影响与其居民的特征分离开来，即真实世界中存在的混淆变量，如第1章所述。在这种情况下，进行这项研究是因为他们认为监狱中存在的暴力和残酷行为是由囚犯的反社会个性和看守的威权特征所致。因此，他们决定选择那些没有已被认为是监狱环境的原因而具有反社会和虐待个性的人（即200人），这个模拟监狱“完全由在所有基本维度上与社会其他人没有区别的个体组成。”156 因此，通过保持被认为对监狱环境负责的情境因素不变，该研究表明监狱中的行为可靠地归因于情境，而不仅仅是其中人的个性。

为了获得“正常”的受试者，研究团队结合了米尔格拉姆基于特定标准的有目的选择和费希尔的随机分配。他们通过报纸广告招募受试者，承诺每天参与可获得15美元的报酬。与米尔格拉姆不同，他认为大学生会对研究造成偏见，辛巴多有意选择了相似或同质化的学生。回应广告的七十五名学生不仅来自斯坦福大学，还来自全国各地的大学，碰巧在夏天在帕洛奥图度过。招募对象必须回答有关他们的家庭、身心健康和犯罪参与程度的问题，并接受研究人员的面试。研究人员选择了最稳定和成熟的受试者。与米尔格拉姆和当时的大多数其他研究一样，受试者都是男性。尽管他们寻找与美国人口相似的学生，但除一人外，其余受试者都是白人。

在经过慎重而有目的的选择程序后，所选的受试者随机分配到囚犯或看守的角色。随机分配到条件的目的是确保扮演看守的受试者在任何重要特征上与扮演囚犯的受试者没有差异。研究团队通过给受试者进行一系列心理测试来检查这一点，结果显示两组之间没有显著差异。随机分配起到了作用。159 第7章将详细介绍如何进行随机分配。

正如米尔格拉姆进行了前期研究一样，辛巴多和同事们在进行这项研究之前也进行了相关研究。这个想法起源于辛巴多一门课上的一个学生的项目。辛巴多随后进行了一项“野外实验”，在加利福尼亚州帕洛奥图和纽约州布朗克斯放置被废弃的汽车，并记录人们对其进行破坏的情况，就像拍摄隐藏摄像头一样。蓝领的布朗克斯和白领的帕洛奥图都是白人，衣着整洁，一些人还鼓励他们的孩子，他们都对这些汽车进行了破坏。通过这些前期研究，辛巴多得出结论，使人们感到匿名的环境可能促进反社会行为。161 第11章将讨论如何进行前期研究。

原始研究文章将监狱实验设计描述为“相对简单”，只有“一个处理变量，随机分配到‘看守’或‘囚犯’条件”。这种设计相对于如今典型的实验设计要简单得多，比如米尔格拉姆的设计，每个实验中只有一个条件或因素水平，但比起现今的实验设计来说要复杂得少得多。实验设计的类型将在第4章中详细讨论，而因子设计将在第6章中讨论。这项研究甚至没有一个正式的假设，研究人员解释这是由于研究的“探索性质”。他们报告了多个统计检验来支持提出的一个总体假设：“被分配到‘看守’或‘囚犯’的处理条件会在行为互动的行为测量、情绪状态和病理的情绪测量、对自我态度以及对这种新情境的应对和适应的其他指标方面产生显著不同的反应。”164 米尔格拉姆在他的初始论文中也没有提出假设。

和米尔格拉姆的实验一样，监狱研究使用了大量的测量工具，包括对受试者实际行为的视频和音频记录、研究人员的观察、使用自我报告的问卷和对受试者的面试。165 问卷包括情绪清单、人格测试、测量威权主义和马基雅维利主义的量表，以及包括诚实、有序、服从、活动、稳定、外向、男性化和共情的人格量表。166 第10章将更详细地讨论构建此类测量工具的方法。研究人员还使用了定量和定性技术的组合来进行数据三角化。例如，他们使用归纳法来对视频和音频记录以及观察的数据进行分类，然后将其转化为分数。

其中他们进行了统计检验。他们还定性地报告了研究结果，包括受试者的引用。研究结果的一个优点是，“在各种测量工具和不同观察者之间出现的关系模式的一致性”。167这篇论文的研究结果至今仍然是如何撰写实验论文中的这一部分的一个良好示例。大量的篇幅用于讨论研究中独特的发现、它们的意义以及背后的原因。它们与其他研究有关，而不会让整个部分看起来像另一个文献综述。168讨论还包括研究的局限性，包括样本量较小（N = 21），模拟监狱的条件与真实监狱相比较为简化，以及需求特征，在本例中，受试者知道他们正在被观察，并希望取悦研究者。该论文展示了研究人员如何通过他们所拥有的数据克服了其中的一些限制，例如分析受试者不知道自己被观察但他们的牢房被窃听的时刻。能够使用数据展示研究的局限性并不否定结论的能力是一种有用的技能。

最后，这项研究最重要的一点是全面而明确地回答了“为什么”问题。美国心理学会将其誉为心理研究如何应用于解决现实世界问题并为非学术人士所理解的典范，使研究与普通公民相关联。169作为对研究不道德性质的批评的回应，辛巴多报道了对社会的影响，并列举了公众对该研究表现出的兴趣，包括媒体报道、来电、来信、演讲请求以及对法律和政策的变更等等。170虽然辛巴多是在回应不道德处理的指责，但所有实验都应该对研究改善社会状况和解决现实世界问题的能力提出强有力的理由。

**总结**

过去这些著名的实验有一些共同之处，其中包括创造力。这些社会科学家愿意不遗余力地确保他们的实验找到真实的东西，而不仅仅是支持一个假设。菲利普·辛巴多亲自请来了真实的警察协助他的实验。B·F·斯金纳发明了自己的测量设备。在他们的探索过程中，一些人忽视了保护被试者的责任，从而使今天的研究人员更好地理解了伤害与真实之间的权衡。实验者仍然应该努力使他们的研究条件尽可能逼近现实；例如，如果一项研究将测试电视信息的影响，那么被试者应该通过电视观看，而不是以书面脚本的形式阅读。但是，防止伤害是至关重要的。第11章将提供更具体的建议，以避免或减少实验中的欺骗行为，以及其他伦理问题。

这些科学家并不害怕尝试新事物。重要的知识进步是通过知道何时借鉴他人的经验，以及何时独立开展研究而取得的。太多人害怕偏离已经做过的事情，或者以与已有测量方法不同的方式进行测量。这导致了研究只是确认他人研究结果而没有其他进展，而科学的目的是创造新知识。

这些历史上的例子还表明研究是混乱的。它并不总是像期刊上的文章所展示的那样整洁。大多数研究者在他们的著作中记录了这个混乱的过程，而不是在期刊文章中。斯金纳、米尔格拉姆和辛巴多的著作是很好的例子。

最后，这些研究显示重要的是不要伤害受试者，并尽量避免欺骗，同时遵循伦理指南（参见有关欺骗的2.3小节）。在平衡保护受试者的责任的同时，寻找有效的答案至关重要。如果可能的话，应避免使用欺骗，并在不可避免时尽量减少使用。在研究结束后向参与者进行说明，并提供研究结果的报告有助于以尊严和尊重对待人们。这些步骤还将有助于确保自己的研究因正确的原因而载入史册。

**常见错误**

• 不使研究具有现实性

• 在可以避免的情况下使用欺骗

• 复制他人的研究结果而不是创造新知识Not making studies realistic

**测试你的知识**

1. 在他对坏血病的研究中，詹姆斯·林德使用了以下方式将受试者分配到治疗条件中：

a. 随机分配

b. 匹配策略

c. 代表性策略

d. 雪球策略

2. 据某些说法，最早使用随机分配的实验家之一是:

a. 詹姆斯·林德

b. 艾萨克·牛顿

c. C. S. 皮尔斯

d. 阿尔伯特·爱因斯坦

3. 罗纳德·费雪对于一位女士无法分辨茶和牛奶的先后顺序的预测是一个例子，这是一种：

a. 另一种假设

b. 零假设

c. 显而易见的假设

d. 理论假设

4. B. F. 斯金纳以使用：

a. 小样本

b. 大样本

c. 随机样本

d. 目的样本

5. 当B. F. 斯金纳的装置卡住无法工作时，他将其认定为：

a. 需要纠正的问题

b. 科学的混乱和机缘巧合的作用

c. 破坏行为

d. 控制混杂因素的方法

6. 斯坦利·米尔格拉姆在他的服从研究中使用了延迟测量、自我报告以及受试者的行为。这说明了：

a. 测量

b. 反应选择

c. 控制混杂因素

d. 最大化变异

7. 斯坦利·米尔格拉姆还使用了态度量表。这说明了：

a. 测量

b. 反应选择

c. 控制混杂因素

d. 最大化变异

8. 预备研究的目的是：

a. 发现操作是否按预期工作

b. 发现受试者是否认为实验是真实的

c. 通过试运行实验来确定需要进行哪些改变

d. 确保等效组

9. 在以下哪项研究中存在欺骗行为？

a . C. S. 皮尔斯关于重物的研究

b. 罗纳德·费雪的品茶实验

c. 斯坦利·米尔格拉姆的服从权威实验

d. 菲利普·辛巴达的监狱实验

10. 在监狱实验中，菲利普·辛巴达使用了：

a. 目的抽样

b. 随机分配

c. 既有目的抽样又有随机分配

d. 以上都不是

**Answers**

1. b

2. c

3. b

4. a

5. b

6. a

7. b

8. c

9. c

10. c

**应用练习**

1. 本章并没有涵盖社会科学领域所有创新实验设计的先驱者。请识别一些其他人，并对他们进行自己的研究，例如库尔特·莱温、利昂·费斯廷格、所罗门·阿什、伊万·巴夫洛夫、保罗·拉扎斯菲尔德、穆扎弗尔和卡罗琳·谢里夫等。特别关注女性和有色人种的科学家。阅读他人对这些实验学家及其工作的评论，同时查阅并阅读这些社会科学家自己撰写的原始研究论文或书籍。

2. 历史上著名的实验似乎最近吸引了电影制片人的想象力。选择其中的一部电影并撰写两页论文，探讨伦理研究的重要性。

• 《实验者》（The Experimenter），关于斯坦利·米尔格拉姆著名的电击实验，2015年制作。可在Netflix、Amazon和Microsoft Movies and TV上观看。该电影在2015年的圣丹斯电影节上放映。米尔格拉姆本人也拍摄了实际的实验过程，你可以在YouTube和其他视频服务上搜索“Stanley Milgram experiment video”以观看真实的录像片段，其中包括米尔格拉姆本人。

• 《斯坦福监狱实验》（The Stanford Prison Experiment），关于菲利普·辛巴达的实验，2015年制作。可在Amazon、Microsoft Movies and TV以及iTunes上观看。辛巴达拥有一个网站，上面提供了更多关于这项研究的信息：http://www.prisonexp.org。

**推荐阅读**

没有什么比阅读原始研究更能启发人了，而不仅仅是别人对其的评论。因此，我推荐以下内容：

• Haney, Craig, Curtis Banks, 和 Philip Zimbardo. 1973. “模拟监狱中的人际动态。”《犯罪学与刑罚学国际期刊》 1: 69–97.

• Milgram, Stanley. 1964. “群体压力与针对个人的行动。”《异常与社会心理学期刊》 69: 137–143.

• Milgram, Stanley. 1965. “群体压力的解放效应。”《个性与社会心理学期刊》 1: 127–134.

• Milgram, Stanley. 1965. “服从和违抗权威的某些条件。”《人际关系》 18 (1): 57–76.

• Peirce, Charles Sanders, 和 Joseph Jastrow. 1885. “有关感觉的微小差异。”《美国国家科学院的纪念》 3: 75–83.

• Skinner, B. F. 1948. “鸽子的迷信行为。”《实验心理学期刊》 38: 168–172.

对于有关研究创新的利弊的深入讨论，请阅读：

Voosen, Paul. 2015. “对于研究人员来说，风险正在消失。”《高等教育纪事报》。 www.Chronicle.com。这是一篇高级文章，意味着需要付费或订阅。请查看您所在的大学或学院图书馆是否有相关资源。

对于那些对欺骗历史特别感兴趣的人，请阅读这本书：

Korn, J. H. 1997.《现实的幻觉：社会心理学中的欺骗历史》。奥尔巴尼：纽约州立大学出版社。

对于一些展现这些实验学家更人性化一面的有趣阅读，请阅读：

Slater, Lauren. 2005.《揭开斯金纳的盒子：20世纪伟大的心理学实验》。纽约：诺顿出版社。

**第三篇 理论、文献和假设**

*如完全成功的干预措施从未仅仅通过头脑风暴而出现。它们总是基于理论、先前的研究或丰富的临床经验提出的。*

*—**R.巴克尔·鲍赛尔*

**学习目标**

**• 解释理论和文献在实验中的作用。**

**• 准备一份文献综述，将先前的工作与你自己的研究联系起来，并做出理论贡献。**

**• 描述假设和研究问题在实验中的作用。**

**• 总结何时预测假设中的差异或方向。**

**• 使用给定的图表制定原创假设。**

根据Bausell的说法，我们应该补充一点，即实验并非用于探索性研究，即在问题处于初步阶段时使用。相反，实验需要相当丰富的理论和先前研究的证据。因此，在开始任何实验之前，进行充分的准备工作非常重要，其中包括对文献进行彻底的搜索，包括理论方面的文献。

本章在前一章的基础上，通过考察前人所发现的角色，或者说像艾萨克·牛顿所说的“站在巨人的肩膀上”的概念，进一步探讨了这一角色的重要性。尽管本章的内容也适用于其他研究方法，但实验特别强调通过建立在先前发现的基础上来创造新知识。本章还涵盖了“良好执行的实验的七个属性”中的两个，包括理论解释或对正在测试的理论进行解释，以及明确确定研究问题和假设。本章从如何解释正在测试的理论，将其与实验中使用的变量联系起来，并将所有假设与该理论联系起来的角度来审视论文的文献综述部分。假设在实验中是特别强大的分析工具，因此，了解如何撰写文献综述以及由此产生的假设对于这种方法尤为重要。

正如第1章所介绍的，设计实验的第一步是明确要测试的因果关系，并在目的陈述中清晰表达出来。但这些关系不应仅基于直觉或好奇心。最好是，要测试的想法应该建立在对领域和其中的文献和理论的理解之上。衡量因果关系需要理论思考；否则，实验将成为一个凭运气的试错过程。(关于理论的更多信息，请参见3.1更多关于理论的内容框。)

社会科学知识发展的一种常见途径是通过使用观察、访谈、人类学研究等方法进行探索性工作。这些类型的定性方法明确设计用于描述和生成理论。它们揭示了潜在的含义——即在表面上不容易看出的含义——或者帮助我们理解现象的本质并预测其发生的条件。但由于其固有特性(样本小、非随机抽取、缺乏控制等)，它们不适合测试或完善生成的理论。这就是实验的主要目的之一。在通过定性或其他方式生成理论之后，研究可以进一步采用测试它们的方法。调查可以很好地获取关于行为的意见、态度和自我报告。内容分析可以揭示各种类型的消息、文本或视觉特征。虽然这些方法产生的研究增加了新的知识并扩展了理论，但它们也提供了可能的原因，并提供了相关性证据，为实验工作奠定了基础。随后，实验研究能够确定在自然环境中观察到的现象是否可以在受控环境中重复出现，

**内容延展3.1**

**理论**

理论是对某一特定现象的概括性解释，或者说是一个“组织起来的概念和原理体系”。这种学术说法实际上是指理论是关于现实世界中实际发生的事情的。引用门禁机制之父库尔特·莱温的话来说，“没有比一个好的理论更实用的了。”虽然理论可能看起来神秘而难以理解，但它实际上只是一种理解生活中发生的事情的工具。当持续支持假设时，理论会不断发展。一个好的理论不仅应该描述现象，还应该解释现象并能够对其进行预测。例如，框架理论的一个方面描述了新闻报道如何构建为离散的事件——比如一起犯罪案件的报道——或更普遍地作为增加背景的主题——比如关于犯罪的报道，而不仅仅是某一个具体案例。它解释了人们如何感知这些类型的框架，并且还可以预测人们对其的反应——例如，在事件以离散的方式被框架时，将个人视为责任人；而在事件以主题方式被框架时，将责任归咎于社会因素。实验通过测试假设和扩展我们现有的理解来建立理论。

然而，在生物学和相关学科中，有一些关于“探索性实验”的讨论。这些研究人员所指的探索性实验并不是无目标、无理论基础的数据收集项目。相反，探索性实验是指在理论指导或基础下进行的实验，但不一定通过正式的假设来测试它。这些研究具有理论背景，而不是完全缺乏理论框架。理论仍然“对探索性实验的成功至关重要”。以“探索性”为借口来证明缺乏理论、功效不足、受试者非随机分配、未控制变量或其他弱点是不合适的。在这些情况下，研究人员应将其视为试点研究，吸取经验教训，并进行改进以消除弱点，而不是试图将其作为探索性研究发表。

在实验中，需要排除所有可能干扰结果或对现实世界中的可行替代解释负责的许多变量。当确定实验是下一个合乎逻辑的方法时，第一步是进行彻底的文献和特定理论的搜索。

这并不是说研究人员永远不能通过其他方式提出实验的想法，比如观察并对为什么发生特定现象产生好奇，就像詹姆斯·林德在他对坏血病的研究中所做的那样。通过这种方式产生的想法当然可以通过实验进行测试。但是，进行测试时应该有一些基于理论或证据的原因。在追寻由研究问题产生的时候，理论或证据应该为测试这些问题提供支持。在进行观察时，找到可以帮助引导和完善它们的经验证据和理论非常重要。这一目的可以通过将两个变量联系起来的相关研究来实现，也可以通过使用不同的受试群体或不同的因变量(DV)的实验来实现。例如，在一项研究中，探讨了通过展示照片是否可以提高记者的道德判断力，以往没有使用道德判断作为因变量(DV)，照片作为自变量(IV)的研究。然而，照片已经被证明可以改善其他认知过程，如细化、移情和参与度。所有这些因变量都与道德判断相关，这是其他研究所证明的。这些理论和经验的基础为自变量和结果变量提供了联系，虽然间接。这个想法源于对于当涉及到照片时记者的伦理推理如何改善的文章和其他非经验性作品的大量猜测。

#### **文献综述**

在论文的文献综述部分，可以对使用的理论及其与研究变量之间的关系进行解释。一篇完成良好的文献综述为整个实验提供了坚实的基础，然而，Bausell指出，“大多数初学者最常违反的单一规则”是对相关文献了解不够充分。有经验的研究人员可能不觉得有必要在进行实验之前撰写正式的文献综述，因为他们始终关注最新的研究发现并熟悉他因为他们始终关注最新的研究发现并熟悉他们所了解的理论。但是对于其他人来说，在形成任何可测试的假设之前，应该进行对相关文献的彻底搜索，并熟悉相关理论。这个过程的第一步是进行文献搜索，第二步是在论文的文献综述部分进行撰写，这也是本章的主题。

阅读他人的研究可以成为实验构想的重要源泉。首先，阅读现有的研究可以了解是否已经有人做过类似的研究，以避免只是简单地重复别人的工作。它还可以揭示已经完成的工作中存在的空白以及需要进一步开展的工作，以完善理论。每篇研究文章末尾的“局限性”和“未来研究”部分也可以提供更多的思路。这包括需要复制的研究以及对理论进行扩展和改进的想法。阅读文献还可以提供方法论的帮助，包括他人是如何测量变量的，可能需要多少被试以及新颖程序的想法。

这本书提供了如何进行文献检索和撰写文献综述的指导，以及可能已经进行和撰写了一篇或多篇文献综述的读者。由于实验在设计阶段需要投入大量工作，并且一旦收集到数据就无法更改，因此对实验的文献检索尤为重要。我建议在开始收集数据之前，实际上先起草一份文献综述，无论你对该主题有多熟悉。我发现写作的过程会减慢速度，让我有时间以一种阅读时不会有的方式思考文献。它还迫使我综合各种不同的研究结果，有助于更好地分析意义并找到可能需要填补的空白。制作一张图表可以更清晰地展示事物，有时这些图表最终会出现在发表的研究中，例如表格3.1。

在实验中，需要隔离或控制除了被操纵的变量之外可能影响结果或因变量的所有因素，因此了解所有这些因素尤为关键。然后可以对它们进行控制，或将其用作中介变量、调节变量或共变量。如果研究人员不知道它们是什么，并且未能测量或控制它们，整个实验的结果的有效性将受到威胁。

有许多关于如何进行文献检索和撰写文献综述的好书和文章，因此在这里不会详细描述它们。相反，下面是一些建议，以避免文献综述中的一些常见问题。实验设计。这些建议并不仅适用于这种方法，但它们解决的问题在实验中经常发生，因此值得一提。

**实验文献综述写作的建议**

**这不是一篇书评**

重要的是分析和综合文献搜索中找到的各项研究的结果。很多时候，文献综述读起来像是书评。它们只是对之前研究结果的总结，但是没有对这些结果进行分析，展示它们之间以及与即将进行的研究之间的联系。不要假设读者对这个主题已经很熟悉；因此，需要明确说明所综述的研究之间以及与当前研究的关联。

对他们来说并不像对你一样显而易见，所以需要详细说明。此外，实验是基于先前的研究提出假设并进行测试的，因此对这种方法来说，对这些研究结果进行分析尤为重要。

与其简单地串联摘要的总结（例如，“Smith和Jones发现...”；“White、Brown、Green和Black展示了...”；“Tinker和Taylor研究了...”），一个好的文献综述应该是对这些研究的批判性分析，综合研究结果，并提出自己的新观点。这种方法也避免了看起来像是显眼地提及他人的名字，或试图提高作者的引用次数或期刊的影响因子（衡量期刊重要性的指标）的行为。

作为分析和综合的一个例子，我会讲述一个同事和我进行议程设置研究的故事，我们想要了解新闻议题或候选人特质对选民的影响哪个更为强大。我们进行了文献搜索，注意到媒体议程与公众议程之间的相关性（议程设置理论的基本前提）在特质方面要高于议题方面。这成为我们得出以下假设的基础：特质议程设置比议题议程设置更具影响力。这并非我们在文献中发现的，而是通过对理论的批判性思考和比较所有研究结果而发现的。

**不要成为维基百科**

许多文献综述都像维基百科条目一样。相反，一个好的综述应该只覆盖与研究相关的文献。例如，如果实验是关于二级议程设置的，可能需要简要提及一级议程设置是什么，并说明它与二级议程设置的区别，但不需要详细解释一级议程设置，总结每个相关研究，描述所有的中介变量、调节变量或条件限制情况——即它起作用的时候和不起作用的时候。相反，提及一级议程设置并说明它与本实验要测试的内容有何不同，然后继续进行。换句话说，如果研究的是B，就只关注B，即使还有A。确保涵盖所有相关的文献；有些作者在另一个方向上做得太过头了，忽略了重要的工作。关键是找到平衡点——涵盖适量的内容，不要遗漏任何真正具有重要影响的研究或经典作品。这也适用于文献综述中涉及的理论。有许多重叠的理论可以用来支持一个实验，但一个研究应该着重于最相关的理论，从而使其与实验的目的和问题紧密联系起来。综述中的理论部分应该解释为什么选择这些理论，并阐明它们如何为实验提供理论基础和指导

**学习要点 3.2**

Shoemaker, P., J. Tankard, and D. Lasorsa. 2003. 《如何构建社会科学理论》. Thousand Oaks, CA: Sage.

初学者经常很难理解理论的概念，特别是如何构建理论。期刊审稿人常常会问作者：“你的理论贡献是什么？”、“这项研究提供了哪些理论观点？”以及“这项研究在理论构建方面有何贡献？”找到回答这些问题的方法常常让许多新兴社会科学研究人员彻夜难眠。当这本关于如何在社会科学中构建理论的书面世时，它以其逻辑严谨、循序渐进的方法直接吸引了我作为一个内心深处的实验者。它兑现了它教授“以最小困难进行理论构建这一挑战性活动”的承诺。这本书首先向读者保证他们并不孤单，承认自己对于如何构建理论和提供理论贡献的理解不足并没有什么可羞愧的。

这本书的章节逻辑清晰，首先描述了假设的书写。然后解释了理论概念，称其为“理论的构建基石”。它演示了如何将概念组合成理论陈述，广泛描述变量之间的关系。它展示了如何综合不同的部分并建立关联，无论是理论上的还是操作上的。然后它展示了如何描述和解释概念之间的关系，以提供解释力。章节涉及如何关联两个变量、三个变量、四个及更多变量，以及如何在完整理论之前构建模型。

这本书提供了社会科学学者可以理解的例子。其中一个例子是乔治·格伯纳和同事们对“培养理论”的发展，展示了它是如何起源于解释电视对观众的影响的一种方式。其中一个假设提出观看电视塑造观众的信念、意识形态和世界观。随着研究的进行和假设的支持，该理论逐渐增加了其他概念，如“共鸣”和“主流化”，指定了产生这种效应所必需的条件。进一步的研究导致理论的修订和补充，包括区分基于现实和信念的一级和二级效应的感知。

《如何构建社会科学理论》强调了创造力在构建理论中的重要性，这也是本书的一个重要观点。它提供了具体、实用的方法来为理论做出贡献。在我阅读这本书之前，我对如何发展理论从未真正理解过。

通常只使用一两个理论。例如，有至少十种理论可以用来解释通过观看电视引起的攻击行为。在使用其他研究方法的研究中，可能需要回顾许多相似的理论，而不是在旨在测试特定理论的实验中。找到最适合实验目的的理论，并为选择该理论而写出充分的理由，而不是回顾每个可能的理论。有时，两种理论会导致相反的预测，因此可以在一项研究中并列使用它们。在这个过程中，应该建立起自变量、因变量和中介变量之间的紧密联系，即治疗和结果之间的因果关系。

**进行理论贡献**

迄今为止，许多研究无论是哪种类型，都会在总结一个理论之后，然后进行作者打算进行的研究，但最终却没有回到理论这一点上。他们通常没有明确说明研究意图对该理论或多个理论做出何种贡献。对于实验研究而言，这是一个特别严重的疏忽，因为进行实验的主要原因之一就是为了发展、验证、完善或以其他方式对理论做出贡献。仅仅将理论作为基础而后弃之是不够的。必须明确指出这个实验将为哪个理论或多个理论做出何种贡献，并且在论文的讨论和结论部分对此进行扩展。（关于如何建立理论的书籍概要，请参阅研究重点 3.2。）

为实现这一目标，重要的是清晰地将所有变量（因变量、自变量、因果机制和个体差异变量）之间的关系进行链接和说明。首先，说明自变量将如何影响因变量，然后解释这将通过哪些理论过程发生。如果有助于理解，可以制作一个模型来展示这一过程，并用图表形式来可视化（例如，请参阅表格 3.1）。举个例子，前面提到的有关照片对道德判断的影响的研究感兴趣的是建立记者暴露于照片时与其道德判断水平之间的因果关系，该研究推测这是因为照片已被证明能够增加移情能力，进而加深参与感，并鼓励中心路线加工，而所有这些因素都被发现会增加道德判断。该研究的理论贡献不仅在于照片有能力提高道德判断——这是一项以前未经测试的干预措施——还在于展示了为什么照片有效的理论解释。

**将各种研究结果联系起来**

最后，一个好的文献综述的另一个重要特点是将其他研究与你自己的研究联系起来，我称之为连接各个点。在审查共同主题的每一批研究之后，明确地说明它们与即将进行的研究的关系。它将以相同的方式进行测量吗？它们是否存在需要填补的空白？这些研究中的哪些连接将有助于这个实验的开展？始终将其他研究的发现与你的研究联系起来，将其置于更大的研究领域的背景之中。读者不应该在超过一两个段落的距离内才能找到为什么要审查特定研究以及它们如何与当前研究相关联的答案。例如，在前面提到的研究中，研究了各种议程设置研究的相关性，论文解释了它们与当前研究的相关性，如下所述：“所有这些二级相关性与最高这些联系永远不会显得太过明显。良好的连接句包括：“这与我们的研究相关，因为……”、“我们在研究中利用这一发现，将其作为中介因素……”、“这些研究引发了关于……的问题，而这正是本研究的目的”，以及类似的表达。存在许多良好的连接句的示例；其中一些在"How To Do It"框3.3中提供。在阅读其他研究时，寻找这些连接句并适应到自己的工作中。

概念定义与操作化定义

**如何实施3.3**

**连接句子的示例**

在文献综述中，将研究与其他研究以及自己的研究联系起来，并将自己的工作置于其他研究的背景之中是非常重要的。以下是一些研究者如何在他们的实验中明确表达这一点的示例。

• 在回顾了层次效应模型中的三个序列之后，这项研究指出：“我们对序列并不特别关注，而是关注这些模型的前两个组成部分——知识和态度——对行为结果（投票）的影响强度。换句话说，我们试图确定不是哪个先发生，而是哪个产生更强的影响：对问题的认知（第一级）还是对属性的感觉（第二级）。"31

• 在回顾了名人在Twitter上的使用和信息来源可信度后，作者说：“承认名人在社交媒体中产生的信息的新奇性和重要性，本研究测试了名人在Twitter上的电子口碑对消费者对信息来源可信度、传播电子口碑意图、与品牌相关结果、在线社交资本搭建和与名人的社会认同的影响。”32

• 在回顾归因样式之前，这项研究指出：“本研究深入探讨了消费者不同的归因样式的角色，特别是不同的归因样式如何在负面名人信息的评估中产生不同的影响。”33

• 在一项关于数学学习的研究中，作者回顾了文献并总结了研究结果，并将其与该研究旨在回答的问题联系起来：“简单地说，随着数学焦虑的增加，数学成绩下降。然而，有一个明显的问题：数学评估/问题上的糟糕表现是由于数学焦虑还是由于内容掌握不足？”34

• 在回顾了有效帮助学生学习英语的不同方法的研究后，这项研究以这样的方式总结：“对于英语学习者，通过技术丰富的课程所呈现的显著特点和好处强调了计算机辅助语言学习在第二语言教学和学习中的关键作用。”35

• 在总结文献之前，一项测试退出过度拥挤市场的企业的研究指出了各项研究之间的相似之处，并指出了他们的工作填补了一个空白：“尽管耗竭战在博弈论文献中占据重要地位，但与之直接相关的实验研究令人惊讶地很少。”36

文献综述也是研究中所有重要概念和变量在概念上进行定义的地方。概念定义是使用一般的定性术语对某个事物进行抽象的理论描述。学生们经常将概念定义与操作定义（或操作化定义）混淆，而操作定义则将概念以具体、具体的方式进行定义，以测量其具体特征。概念是关于某个事物的一般想法，具有许多具体特征。例如，在一项研究中，可信度在概念上被定义为“公众对新闻质量的感知”37，然后通过十三个具体特征对其进行操作化定义，其中包括新闻报道的公正性、完整性、准确性、可信度、可信度、信息性、有趣性、受欢迎程度、深度、重要性和写作质量，以及信息源和信息的可信度。这是两种不同类型的定义，它们出现在论文的两个不同部分。仅提供一系列用于测量概念的术语列表并不足以作为概念定义。概念定义应该基于理论描述和解释概念。另一个例子是道德判断的概念定义，即人们“用来判断一种行为在道德上是对还是错的理由”38。方法部分中给出的操作化定义是关于在面对特定困境时主题认为重要的事物的十二个陈述（有关示例，请参见脚注ii）。

之前引用的可信度研究提供了一个很好的例子，说明不同研究者对一个概念的定义可能存在差异，其中包括将其定义为“对沟通来源的态度”，“对消息来源可信度的整体评估”，包括可信度、专业性和善意在内的组合，以及对媒体的质量、盈利能力、隐私、社区福祉和可信度的综合公众感知。对于某些概念的概念定义缺乏共识并不罕见。在这种情况下，重要的是在文献部分承认这一点，回顾不同的概念定义，然后捍卫你选择的定义并解释为什么它最适合你的研究。重申一下：在研究中使用的每个变量必须出现在三个位置：

• 在文献综述中进行概念定义，

• 在假设或研究问题中加以运用，以及

• 在方法部分进行操作化。

追踪每个变量贯穿整篇论文，确保每个变量都在这三个位置出现。不应该有孤立的概念或变量。

一个很好的方法是将文献综述视为一个大纲，列出所有的因变量、自变量、中介变量（中介者、调节变量、因果机制）和协变量（如果有的话）。使用与大纲相对应的副标题来帮助读者理清思路，也有助于组织写作。这样确保了所有内容都有定义，涵盖了经验证据，并且关于一个变量的所有讨论都在同一个地方。经验丰富的读者期望在同一个概念中找到所有相关信息，并且当某个概念在论文中的不同位置出现时可能感到不满。

涉及多个实验的文献综述

许多研究人员在一篇文章中报告了多个实验，原因包括复制发现和排除合理的替代解释。有时这是事先计划的，而其他时候则是在第一项研究结果知道后才出现的需求。这个现象将在第5章关于有效性中探讨。当一篇论文中报告多个实验时，可能有一个部分涵盖了两个研究的所有文献，或者可能有两个文献综述，第一个最长，第二个只包括与第二个研究相关的新文献。通常，它们将被标记为“研究1”和“研究2”，如果有更多的话还会有更多的标记。实验。一般的模式是先报告第一个研究的文献综述，然后在类似于“研究1”的子标题下报告其方法，然后是结果和讨论。第二个研究将在“研究2”或“实验2”的子标题下进行，首先描述该研究，如果需要阐明任何新变量，可以进行简短的文献综述。例如，如果为了排除第一个研究中可能出现的混杂因素或合理的替代解释，引入了新的自变量、中介变量或调节变量，那么它们显然不会在第一个文献综述中涵盖。不需要重复主要文献综述中的所有先前信息，只需添加新的概念并将其与先前的文献和理论联系起来。一个很好的例子来自一项关于议程设置中认知加工的研究。第二个实验是在第一个实验显示出作者预期的效果在某些受试者上有作用而在其他受试者上没有作用之后构思的。他们推测这可能受到了他们使用的故事内容的影响，并写道：“我们决定进行后续研究，以测试这样一个假设，即相较于其他人更加中心地加工信息的人可能会受到文章内容方面的影响，例如记者对问题重要性的评估。”然后他们用“研究2”作为子标题，并以一页纸的文献综述开始，该文献综述仅涉及处理新闻故事内容的研究。

最后，还有一些来自文献综述的其他重要提示：

• 包括最新的文献来源。除了经典或重要的研究，确保包括最新的作品。如果研究是在一段时间之前开始的，这一点尤为重要。回顾一下最近发表的期刊，看看是否有遗漏的研究。在这个过程中，包括计划提交的期刊。

• 使用不同的搜索方法。使用数据库中的关键词进行搜索并不能总是找到所有重要的文献，即使使用多个数据库。为了避免遗漏重要的研究，浏览过去两三年可能会发表关于该主题的文章的期刊，看看是否有遗漏的内容。在这个过程中，包括计划提交的期刊。

• 包括相互矛盾的发现。很少有研究项目始终产生一致的结果。寻找那些与预期不符的研究。不仅诚实是重要的，了解可能导致你的实验不按计划进行的原因也很重要，这样你就可以尽可能地进行控制。回顾那些不支持你的预测的研究，并解释为什么它们可能以这种方式出现。

• 适度引用他人的观点。过多或过频繁地引用他人的观点会引起读者的警惕，因为这可能会让人觉得你没有理解所引用内容。

不要滥用引用。除非没有更好、更清晰、更简洁的表达方式，否则要养成不使用直接引语的习惯。学术论文中符合这一标准的句子很少。使用直接引语时，请适当引用，并注明页码。

• 注意避免抄袭。请参考前面的观点。许多抄袭行为是由于粗心而不是恶意导致的，例如，通过从原始文章中复制和粘贴来做笔记，然后忘记这些笔记是否已经改写成自己的话。从其他文章中重复自己的写作也是一种抄袭行为。建立一个系统，了解哪些是改写的内容，哪些不是，例如通过高亮显示、使用引号或使用不同的字体。

• 始终阅读论文。追溯并阅读其他人文献综述中引用的原始论文可能需要耗费时间，但依赖于他人对论文的解释而不是自己阅读可能是危险的。学者们有时会不加批判地一遍又一遍地报告同一项研究的摘要，甚至可能是错误的。42我发现这种情况尤其常见，特别是在书籍或其他文章的文献综述中。

**假设与研究问题**

文献综述部分也是假设和研究问题通常被纳入的地方。一种方法是在文献综述的末尾列出所有的假设。这种风格涉及在方法部分之前使用一个假设模块，如下所示：

“假设1（H1）：当个人捐赠者知道非营利组织接受一定数量的政府资金时，他们的捐赠金额会减少。（分类排挤假设）

假设2（H2）：当非营利组织从政府获得更大比例的资金时，个人捐赠者的捐赠金额会减少。（连续排挤假设）

假设3（H3）：知道政府资金来自竞争性的基于功绩的计划将增加个人捐赠的意愿。（排挤内化假设）”43

术语“捐赠”和“给予”可能具有不同的定义，因此在编写假设时重要的是不要使用同义词。尽量使用平行的措辞来表达假设。

为了避免读者在回顾整个文献综述以查找每个假设的证据时不得不翻阅许多页面，一些期刊在假设之前或之后提供了一份关于每个预测所依据的文献的简要摘要。在文献综述部分详细描述研究时，再在结尾重复一份摘要可能效率不高，另一种方式是将假设嵌入到与各个假设相关的文献综述部分中。例如，关于气候观念判断的研究比较了明示线索和隐含线索的相对权重，44作者首先回顾了明示线索并提出了以下假设：

“H1：调查作为明示媒体线索会影响接收者对气候观念的判断，使其朝着调查结果的方向发展。”45

接下来，他们回顾了隐含线索，并提出了以下预测：

“H2：如果媒体报道中的论点支持（反对）某种观点，则公众对该观点的认同会增加（减少）。46”

然后，他们回顾了媒体中这两种线索的说服效果的文献，并给出了以下假设：

“H3a：如果提供有关未知问题的调查信息，接收者倾向于追随多数意见。”47

还有几个假设，所有这些都在模型中以图表形式总结（见图3.1）。

在假设之前引入概念而未在文献综述中详细阐述该概念可能是一项挑战。例如，不要在文献综述中未讨论过某个概念的情况下引入包含该概念的假设。在探讨概念之前就引入假设可能会让组织文献回顾变得棘手，因为它并不假设读者对这些概念有先前的了解。这使得将所有的假设都放在最后成为一种诱惑，但是读者在到达H1时可能已经很难回想起关于这些概念的内容。相反，将假设分解成较小的陈述，一个自变量一个，按照过程发生的顺序进行，这是关键。

在前面的示例中，假设被正式表述，并带有“H”加上一个数字的标识（H1，H2，H3等）。小写字母（H5a，H5b，H5c）表示相关假设的子集。但是，假设以叙述风格表达也非常常见。例如，在一项关于威权政府是否愿意将公民偏好纳入他们制定的政策中的研究中，中国的学者以以下方式提出他们的假设：

因此，根据以前的研究以及中国共产党的声明，我们有理由相信当国家与社会关系和谐时，领导人可能对通过互联网和正式渠道表达的意见同样持开放态度。然而，当领导人认为国家与公民之间存在对立时，我们预计他们将对正式渠道和互联网渠道的意见持较少开放态度。48

假设应该用将来时态（例如，“将会有”，“可能是”），而不是现在时态（例如，“是”），因为它们尚未经过测试；现在时态听起来像是已经成为事实。

需要多少个假设？学生经常问这个问题，但我只能给出“金发女孩”式的回答：不要太多，也不要太少，恰到好处。一个假设可能表明研究规模太小，或者需要将一个过于庞大的假设拆分为更小的部分。超过十个假设可能意味着研究试图做得太多，应该分成两个独立的出版物。让理论、逻辑和读者的注意力来决定。让同事们阅读文献回顾，看看他们是否能够清楚地理解，而不会迷失或感到厌倦。

比假设的数量更重要的是它们的清晰性。如果指定了一个过程，假设应该在逻辑上相互补充。有时，实验探索的因果机制涉及复杂的理论过程中的多个步骤。在这种情况下，将事物分解为一系列特定的假设，并建立一个连贯的模型是有意义的。例如，在研究照片如何提升道德判断的研究中，该过程被分解为三个假设：

“H1：看到照片的参与者在利益相关者方面的详细程度明显高于没有看到照片的参与者。

H2：道德推理与利益相关者的详细程度显著相关。

H3：看到照片的参与者的道德推理水平明显高于没有看到照片的参与者。”49

第三个假设是研究最感兴趣的，但它在因果事件链中处于最后。

Mutz在关于近距离镜头视角和不文明电视言辞的研究中以类似的叙述方式提出了一组假设（在此处为了阅读方便，它们被分成了单独的段落）：

“第一个假设是近距离镜头视角和不文明行为都会增加情绪激发的程度。

第二个假设是，由第一个假设得出，情绪激发的增加会增加回忆的程度。在这个例子中，第一个预测涉及近距离摄像透视和不礼貌行为导致情绪激发。第二个预测将来自第一个假设的激发与回忆联系起来。顺序并不是随机选择的；研究者按照这个顺序提出它们，是因为假设它们按照这个顺序发生。对于更复杂的过程，可以通过图表来展示这些路径，以便更容易进行可视化。

**假设基础**

**零假设与备择假设**

在解释如何撰写假设之前，让我们回顾一下零假设和备择假设的概念。零假设表明没有差异或方向性，备择假设表明存在差异或方向性。这些概念的提出是为了减少实验者的偏见，或者说是为了避免寻找支持自己观点的证据。很久以前，科学家们提出了一种基本上反对自己立场的观点，也就是研究者不应该试图证明某种处理引起效应，而是试图证明它没有。换句话说，试图证明两个事物之间没有关系或两个组之间没有差异。零假设用H0表示，备择假设用Ha或H1表示。通过否定零假设，我们可以认为存在关系或差异的可能性，因此我们可以说我们的假设是正确的。技术上讲，这就是备择假设。换句话说，我们从零假设（无差异或无关系）为真的立场出发，直到有证据表明相反为止。如果拒绝零假设，我们说我们找到了支持我们的假设的证据。我们从不使用“证明”这个术语，因为技术上是不可能的。相反，应该使用适当的语言来支持我们的假设或拒绝零假设；拒绝假设或“未拒绝”；或者说假设“未被否定”。这在研究论文中很少明确说明，但却是所有论文背后的思维过程。

假设是独立变量和因变量之间关系的预测。可以将其视为“如果...那么...”的陈述，例如，“如果记者看到照片，那么他们将使用更高水平的道德判断”。假设应该以概念为基础，而不是具体的操作定义。例如，在商业中，有一个概念称为管理信任度（MTW），在概念上定义为“对主管的信任度”。通过九个项目进行测量，包括能力、胜任、关心和强烈的正义感等。假设涉及管理信任度，而不是胜任、正义等。例如：

“H1: 强调与设定具有挑战性但可行的目标相关的内部管理对个体对MTW的感知有积极影响。”

**差异和方向的区别**

假设分为两种类型：一种是预测因果方向的假设，即哪个变量的影响比另一个更大；另一种是仅预测差异而不提出哪个变量将比另一个更大。后者也被称为无方向性假设。一个预测差异的假设的例子是：

H: 看到照片的记者在道德判断上会与未看到照片的记者显著不同。

这个假设预测了看与不看照片之间的差异，但没有指明结果会更高或更低。根据这个假设，看到照片可能会导致道德判断更好或更差。相比之下，这个假设的另一个版本预测了一个方向：

H: 看到照片的记者在道德判断上会显著比未看到照片的记者更高。

第二个假设明确指出了两种条件之间的差异，即看与不看照片的差异。上述管理信任度研究的另一个方向性假设的例子是：

“H1: 与设定具有挑战性但可行目标相关的内部管理强调对个人对管理信任度的感知将产生积极影响。”56

该假设使用了“积极”一词，预测了一个因果方向，即研究者预期目标设定条件下的管理信任度水平更高。一种无方向性的表述方式是，目标设定会对管理信任度产生不同于其他条件的影响，但不指明该影响是积极还是消极。

选择预测方向或差异的假设取决于可用的理论和经验证据。如果有足够的证据表明因果方向，那么预测这个方向是合适的。如果没有足够的证据，或者理论不提供方向性预测，则更适合预测差异而不指定方向。例如，在第一次研究照片对道德判断的影响时，根据理论和经验证据的限制，可能更适合预测差异而不预测具体的方向。

在测试道德判断方面的照片效应时，存在理论上的原因使人们相信照片可能会改善或恶化道德判断。在这种情况下，合适的做法是提出看到照片会导致被试者行为不同的假设，而不预测他们的道德判断是否比那些未看到照片的人更好或更差。在研究发现某个方向的效应之后，从那时起，使用方向性假设是合适的。

如果理论和先前的证据非常有限，以至于研究者无法进行任何预测，那么应该使用一个研究问题。研究问题代表了关于理论和证据的更初步的情况。因此，继续上面的例子，如果没有足够的理论和证据来预测照片对道德判断会产生某种影响，那么提出问题“看到照片是否会影响记者的道德判断？”是合适的。作为一个经验法则，根据现有的证据和理论进行尽可能具体的预测是最好的，因为这为我们提供了更强大的分析工具。

研究者是否对方向性进行预测对于如何在统计上测试假设是重要的。对于那些没有接受过统计学入门课程的读者来说，下面的内容可能没有意义；对他们来说，重点在于所使用的假设类型在统计测试中很重要。对于基本统计学知识的读者来说，方向性假设适合使用单尾检验。对于无方向性假设，使用双尾检验。（为简单起见，这些例子假设只有两组，即一个处理组和一个对照组。）这在正态分布图中可以找到图形化的说明。无方向性或差异性假设使用双尾t检验进行测试。因为没有指定方向，只要t值落在正态分布的任一端，即图3.2中蓝色阴影区域内，就支持这个假设。

只有当t值落在假设指定的正态分布的一端时，才支持方向性假设。在这个例子中，假设是照片会导致道德判断水平更高，所以t值必须落在图3.3中表示的蓝色阴影区域内才具有显著性。实际上，如果存在显著性，使用方向性假设更容易找到它；然而，如果预测的方向是错误的，那么可能得到错误的结论。

如果统计测试发现显著性，但结果与预测的方向相反，那么假设就不被支持。

**假设写作公式**

既然这本教科书已经讲解了何时以及为什么要提出明确规定方向的假设，我提供一个基本的公式来编写明确而重点明确的假设。对于每个假设，该公式包括以下内容：

• 预测要么是具有统计显著差异的预测，要么是差异的方向（用下划线表示）。

• 自变量（用粗体表示）。

• 因变量，以概念术语表述（用双下划线表示）。

对于使用两组实验，即处理组和对照组的实验，假设使用“between”一词。使用将来时态也更为常见。可以类似填空的方式书写如下：• 看到Twitter上的报道和在网站上看到的报道之间的可信度评分（因变量）将有显著差异（自变量）。58

• 看到长篇或短篇报道（自变量）的受试者对重要性评分（因变量）将有显著差异。59

• 暴露于倡导框架（自变量）的观众会将犯罪的重要性（因变量）与暴露于客观报道（自变量）的观众显著不同地排名。60

这些假设都没有明确指出差异是更高还是更低，更大还是更小。它们也可以以指定方向的方式书写，例如：

• 看到Twitter上的报道的可信度评分（因变量）将显著高于在网站上看到的报道的评分（自变量）。

• 看到短篇报道的重要性评分（因变量）将显著低于看到长篇报道的评分（自变量）。

• 暴露于倡导框架（自变量）的观众将将犯罪的重要性（因变量）显著提高于暴露于客观框架（自变量）的观众。

如果一个实验涉及三个或更多组，例如两个处理组和一个对照组，使用“among”而不是“between”，如下所示：

• 看到Twitter上的报道、网站上的报道或报纸上的报道的可信度评分（因变量）之间将有显著差异（自变量）。

请注意，这个例子只是将自变量改为包括三个条件——Twitter、网站和报纸，而不是两个。

在所有这些基本假设的例子中，每个假设都被准确地表述。它们使用清晰、简单、单一的因果预测。使用的是相同的术语而不是同义词。这些术语应该与文献综述中使用的术语相同。每个假设都预测了特定自变量和因变量之间的明确因果关系。如果可以通过理论和/或先前的证据支持，更喜欢使用有方向性的假设。

如果实验中包括协变量或统计控制变量，这些变量已被证明影响因变量，但在随机分配中分布不均等，则在假设末尾添加“控制年龄、教育和性别”等变量的短语：

H1：看到Twitter上的报道、网站上的报道或报纸上的报道之间的可信度评分将有显著差异，在控制年龄、教育和性别的情况下。

多个自变量的假设

到目前为止，在本书中，我们已经讨论了逐个操纵一个自变量的情况，例如在前面的假设中，信息的格式是被实验对象所看到的Twitter、网站或报纸。在这个例子中，独立变量是框架：宣传性或客观性。这些被称为单因素设计，并将在后面的章节中详细解释。然而，在实际研究中，许多实验会同时操纵多个独立变量，例如在同一研究中操纵信息的格式和框架。当操纵超过一个独立变量或因素时，应该为每个因素编写单独的假设，也称为主效应，以及两个因素的交互作用或考虑两个因素一起产生的效应。因此，例如，在一个操纵信息格式（Twitter、网站）和框架（宣传性、客观性）的实验中，应该有两个假设，分别针对每个独立变量或因素的主效应：

H1：读Twitter上的报道的受试者的可信度评分将显著高于读网站上的报道的受试者的可信度评分。

H2：读客观框架的报道的受试者的可信度评分将显著高于读宣传性框架的报道的受试者的可信度评分。

最后，还应该有一个关于信息格式和框架之间交互作用的假设或研究问题，例如：

H3：读网站上的客观框架报道的受试者的可信度评分将高于其他信息格式和框架组合的任何情况；

或者

RQ1：信息格式和框架之间是否存在交互作用？

当然，选择是预测还是提出研究问题取决于理论和证据；尽可能使用假设。下一个交互作用假设的例子展示了两个实验操作或独立变量——恶心感和伤害感——与受试者先前的政治观点相互作用。作者预测，根据受试者是否支持或反对某一政治做法，独立变量的效应将不同：

“H4：在反对一项政治做法的人中，偶发的恶心感和伤害感联系将增加道德坚定和导致更严厉的道德判断，而在支持者中将产生相反的效果。”62

对于涉及多个独立变量或因素的研究，这些假设还包含了一个逻辑的推进，从一个自变量到另一个自变量，然后到两个自变量的相互作用。如果提出了中介变量，那么它们也会按照逻辑顺序逐个考虑。为了更清晰地思考这些更复杂的假设，创建一个表格列出独立变量或因素及其水平、中介变量和因变量可能会有所帮助。可以使用这些表格来可视化假设，如图3.4所示。

下面是一个包含中介变量的假设的例子：权力感概念。这个实验研究了应对食品污染组织危机的发言人。这个实验中的独立变量是声音音调，有高音调或低音调两个水平。中介变量是权力感，它在低音调的声音中被感知到。因此，假设是：

“声音音调降低会导致对能力的更高认知，与声音音调升高相比，这是由于被感知到的权力感的中介效应（H2）。”62

在发表的期刊文章中，假设的变化还有很多；这旨在向初学者介绍假设写作的基础知识。在阅读已发表的实验和开展自己的研究时，请根据每个研究的需要调整你的假设写作方式。

**如何操作 3.4**

**假设写作表格**

编写假设时，为了更容易地绘制出步骤，可以通过以下表格来可视化变量。在左侧的列中列出独立变量或因子，并列出每个变量的水平。在右侧的列中列出因变量。如果存在中介变量或因果机制，请在中间列中列出它们。然后，绘制线条来展示将成为假设的过程和预测。下面的表格是简化的，但可以根据需要扩展以适应任意多的因子、中介变量和因变量。

**当存在两个或多个因素时的交互作用效应**

H1: 在 Twitter 上阅读故事的受试者的可信度得分明显高于在网站上阅读故事的受试者。（平台因素的主要效应。）

H2: 阅读客观框架故事的受试者的可信度得分明显高于阅读倡导性框架故事的受试者。（框架因素的主要效应。）

H3: 阅读网站上的客观框架故事的受试者的可信度得分将高于其他消息格式和框架的组合。（平台和框架因素的交互作用。）

**对于基于彼此构建的因果机制的假设**

H1: 观看照片的参与者在利益相关方方面的详细阐述程度将明显高于未观看照片的参与者。（因素导致因果机制。）

H2: 对利益相关方的更多详细阐述将与更高的伦理推理水平显著相关。（因果机制导致因变量。）

H3: 对于观看照片的参与者，其伦理推理水平将明显高于未观看照片的参与者。（因素导致因变量。）

在撰写假设时，最后一条建议是要思考每个假设的重要性。问问自己这些假设是否有趣、有意义或重要。它们是否不言而喻或已经经过充分测试？尽量避免提出我所谓的“显而易见的假设”，即那些答案显而易见的假设。例如，媒体最关注的议题与公众认为最重要的议题是相同的这个基本议程设定假设已经经过充分测试。没有必要把这个问题当作第一次提问一样去测试，不会有什么惊人的发现。许多议程设定的研究从未提出这个问题。当它们确实提出时，通常 是为了更具体的预测，或者看看未经测试的条件是否会导致与过去四十多年反复显示的结果有所不同。

还要问问如果以下三种情况之一发生，会意味着什么：

• 如果假设得到支持，那意味着什么？

• 如果没有支持（没有显著性），那意味着什么？

• 对于定向假设，如果显著性发现与预测的方向相反，那意味着什么？

这个练习旨在帮助你思考所有可能的结果以及每种结果的含义。这将帮助你思考研究的意义，以及你如何在讨论部分解释结果。如果在这些情况下你无法想象出结论，或者它们似乎显而易见或不重要，那么就是重新考虑研究的时候了。此外，考虑在研究中可能包括哪些证据来支持解释。例如，如果数据不能支持假设，或者发现显著性与预测相反，那可能是什么原因？这个练习还涉及到我经常在初学者中发现的一个现象。

实验研究者往往会感到自己的假设得不到支持就“失败”了。即使在不能支持假设的研究中，仍然能够学到一些有价值的东西。在某些情况下，不支持一个假设比支持它更有趣。毕竟，进行研究的目的首先并不仅仅是支持所有假设，而是为了获得有价值的发现。正如坎贝尔和斯坦利所说，“理论测试的任务主要是拒绝不充分的假设。”请放心，不能支持假设并不意味着研究不能被发表。

**研究问题**

除了用于探究两个自变量之间的相互作用，研究问题有时也用于主效应。当没有足够的证据或理论支持一个预测时，或者在测试模型时，会提出这些问题。在实验中编写研究问题与其他方法并无不同。除了之前展示的正式研究问题风格外，研究问题也可以用叙述风格来书写，例如：

总的来说，本研究涉及四个主要问题。第一，这两种教学干预在CVS学习的整体效果上是否有差异？第二，这两种教学干预在学习转化到新领域的程度上是否有差异？第三，学生从这两种干预中学到的内容是否有所不同？具体来说，他们是否从基于无效设计的干预中更好地理解控制无关变量的必要性？第四，这两种教学干预在不同学习环境中的学生（例如成绩较高与较低的学校）之间是否有不同的相对效果？64

设计实验的过程不一定按照本书的章节顺序进行。有时，研究人员需要重新审视已经完成的事情。编写良好、清晰的假设就是其中之一。此外，通常需要写、改写、编辑、再写、再编辑、再重写，直到对假设的准确性和清晰性感到满意。

下一章将讨论实验设计，包括本书重点关注的真实实验的经典版本，以及准实验、自然实验和现场实验。与假设编写一样，一旦您思考了自己的设计，可能需要回过头来重新修改假设和研究问题。这是正常的，研究是一个杂乱的过程。

**常见错误：**

• 没有充分了解相关文献

• 孤立的变量 - 在假设或研究问题中没有将所有变量在概念上定义，并在方法部分中进行操作性定义

• 没有解释变量之间的假设关系及其与理论的关联

测试你的知识：

1. 哪个是概念定义的例子？

a. 权威主义被定义为与厌恶不确定性和对威胁敏感性密切相关的个性特征。

b. 权威主义通过尊重长者、有良好的礼貌、服从和表现良好来衡量。

c. 权威主义在7分制度上进行衡量。

d. 权威主义在不同研究者中有不同的定义。

2. 以下假设属于哪种类型？

“具有高水平专业知识的来源将积极影响对可信度的感知。”

a. 方向性假设

b. 无方向性假设

c. 交互作用假设

d. 零假设

3. 制定方向性或无方向性假设取决于你对你的假设被支持或不被支持的风险有多少。

a. 真

b. 假

4. 当你有多个假设时，总是从代表你测试过程最终结果的那个假设开始。

a. 真

b. 假

5. 文献综述的目的是 \_\_\_\_\_\_。

答案:

a. 发展理论贡献

b. 解释变量的测量方法

c. 提供足够的细节，使其他研究人员能够复制研究

d. 在前人的研究基础上进行进一步构建

6. 实验的想法应该来源于：

a. 直觉和传统智慧

b. 反复试验和错误的过程

c. 理论和现有证据

d. 以上都不是

7. 假设应该 \_\_\_\_\_\_。

a. 被描述为已被证明或未被证明

b. 用具体操作化的术语陈述

c. 放在方法部分

d. 在逻辑上相互关联

8. 当没有足够证据进行预测时，应该使用研究问题代替假设。

a. 真

b. 假

9. “看到照片的记者在道德判断上将与未看到照片的记者显著不同。” 这个陈述属于以下哪个选项？

a. 定向假设

b. 无方向假设

c. 研究问题

d. 零假设

10. 是否制定定向或无方向假设取决于 \_\_\_\_\_\_。

a. 理论

b. 证据

c. 使用的统计检验方法

d. 既有理论又有证据

1. a

2. a

3. b

4. b

5. d

6. c

7. d

8. a

9. b

10. d

**应用练习**

1. 从第一章的任务中选择你的三个想法之一，并在其中增加至少五页的文献综述，包括发展、测试或拓展该理论以及相关文献。至少使用25篇文章。确保对文献进行分析和综合，而不仅仅是总结，并将文献与你的研究联系起来。
2. 为你正在开发的实验编写两到三个明确而简明的假设。使用本章的公式。如果你的实验中有多个因素（自变量），为每个因素的主效应编写一个假设，并为交互作用效应编写一个假设。让两位同事阅读并检查是否清晰明了。

**建议阅读**

From the University of Washington’s Psychology Writing Center, this white paper on how to write a literature

review:

https://depts.washington.edu/psych/files/writing\_center/litrev.pdf

Shoemaker, P., J. Tankard, and D. Lasorsa. 2003. How to Build Social Science Theories. Thousand Oaks, CA:

Sage.

Chapter 2, “Manuscript Structure and Content,” in the Publication Manual of the American Psychological

Association, 6th ed. (2010). Washington, DC: American Psychological Association.

Chapters 2, 3, and 4 in Bausell, R. Barker. 1994. Conducting Meaningful Experiments: 40 Steps to Becoming a

Scientist. Thousand Oaks, CA: Sage.

**第四篇 实验类型**

*真相与结论无关，而与方法论有关。*

*- 斯蒂芬·莫林纽*

**学习目标**

• **总结使用Campbell和Stanley的分类法的不同实验类型。**

**• 推荐何时使用三种真实实验设计。**

**• 评价带有前测设计的优缺点。**

**• 描述准实验、自然实验和现场实验。**

**• 解释准实验、自然实验和现场实验彼此之间以及与真实实验的区别。**

在进行文献综述的基础上，至少在草稿形式中，正如前一章所讨论的，下一步是决定要进行的基本实验类型，即本章开头引用中所指的实验的方法或系统。在这个领域，有一本权威的参考书是Donald T. Campbell和Julian C. Stanley的经典著作《实验与准实验研究设计》。这部作品不包括参考文献只有71页。它最初是1963年的研究手册中的一章，但因为非常受欢迎，所以后来以单行本的形式继续出版。书中描述了16种不同的实验方式，至今仍然有效，尽管有些设计比其他设计更受欢迎。该书是为教育领域撰写的，因此示例都来自教育研究领域。接下来是对Campbell和Stanley的实验设计分类的总结，重点介绍社会科学中流行的六种设计以及对每种设计的评价。今天实际进行的大多数实验都是这里描述的设计的微小变体。除了Campbell和Stanley的分类中的这些设计，本章还将简要介绍准实验、自然实验和现场实验。

**坎贝尔和斯坦利的实验分类法**

坎贝尔和斯坦利使用图形简化来描述实验，对于足球迷来说可能很熟悉，但在这里有着不同的含义。

X表示处理，通常是自变量（IV）。

O表示观察或测量的结果变量，通常是因变量（DV）。这些数据是记录下来的，可以是由研究人员记录，也可以是通过技术仪器（如心率监测器）或被试自行报告的。如果进行多次观察或记录数据，则使用下标表示，如O1、O2，表示第一次和第二次观察。

如果有一个空格（ ），没有X或O，那意味着没有给予处理或进行观察。

R表示将受试对象随机分配到不同条件的实验组。这将在第7章中进一步讨论。基本上，这是将实验区分为“真实”实验和“准实验”的要素。

前三种设计类型被归类为“前实验设计”，根据坎贝尔和斯坦利的说法，它们“几乎没有科学价值”。我在回顾的文章中发现，这些设计倾向于被报告为“探索性”的，但“前实验”是一个更准确的术语。由于缺乏随机分配，它们也可以被归类为准实验设计。在这里，我简要回顾它们，但建议使用其他描述的设计。

**三种前实验设计**

**单次案例研究**

有时候，这种情况在新闻报道中可以看到，当有人想将某种原因归因于某种效果时，例如，在某种干预措施后，学校的旷课率与之前的年份相比下降了。可能是某所学校实施了一种短信系统，在上午9点前提醒学生准时到校。图形表示如下：

X O

短信系统是干预措施（X），旷课率是观察指标（O）。然而，这种设计存在一些问题，其中包括观察的学生没有进行随机抽样，没有对比组的学生没有收到短信，也没有控制措施排除其他解释。例如，一些学生的父母可能开始早上做华夫饼，其他人窗外的新建筑工地吵醒了他们，有些人可能买了新车并兴奋地想要去学校展示，而其他人可能看到了有关大学重要性的演讲，并突然觉得自己需要去上学。所有这些因素（约束条件未找到）都可能解释旷课率的降低，并不仅仅是短信系统的作用。由于这些问题，单次案例研究不是一个适合进行真正社会科学实验的良好设计。

**单组前测-后测设计**

略好于单次案例研究的是在给予处理前增加一个观察或测量，称为前测。在介入干预后进行相同的观察或测量（后测），以便研究者可以将任何变化与基准测量进行比较。

O1 X O2

例如，研究人员可能在前测中测量学生的数学水平，然后进行干预，如计算机辅助辅导，然后再次测试学生，以查看他们学到了多少。前测和后测分数之间的差异代表了效应。然而，有时添加前测并不一定会使情况变得更好，因为人们倾向于在一次考试后学会如何更好地应对考试，或者当他们意识到自己正在被观察时。例如，那些多次参加智商测试的人被证明在第二次尝试时变得更聪明。在研究中，这种测试的效应被称为“测-测增益”。还可能存在其他解释。反应性是指测量某物会改变它；仅仅知道自己被观察，人们倾向于表现更好（详见第4.1个信息框）。

### **内容延展4.1**

**霍桑效应**

霍桑效应是一种特定类型的反应性，即人们对被观察并改变行为的反应。这在研究中很重要，因为有些人在参与研究时不会给出真实的回答。这种效应不一定是故意的；有时，人们会在不知不觉中发生变化。霍桑效应是一个流行的、尽管有人认为不适当的名称，它指的是研究对象因为知道自己正在被观察而改变行为的需求特征。它起源于霍桑工厂的工人生产力研究，该工厂是西部电器公司在芝加哥外的一个工厂，从1924年到1932年进行研究。这些研究也被称为照明研究，因为测试了不同的照明水平对工人生产力的影响；实际上，还研究了许多其他变量，包括工作时间、休息时间、地板和工作站的清洁度等。研究结果的简化版本是，当研究人员改变了工人的照明和休息时间时，他们的生产力有所提高。尽管改变保持不变，但当研究结束时，工人的生产力下降了。最初的结论是，关注工人会导致更高的生产力，但后来被重新解释为人们在被观察时会改变自己的行为。并不是所有的研究都显示出霍桑效应，研究至今仍在探讨它，特别是在健康科学、人际关系和组织行为等领域。霍桑研究还在实验设计中做出了另一个重要贡献，它显示了操作性检查的重要性，操作性检查将在第9章中介绍。在霍桑研究中，影响并不是那么重要，而是工人对这些影响的解释，理解受试者的解释是操作性检查的目的。

类似的现象被称为需求效应或需求特征，即某些人改变他们的答案或行为以取悦实验者的想法。在研究实验情境时，马丁·奥恩发现研究对象试图猜测目的并根据自己的解释改变行为。类似的现象还包括光环效应或社会期望效应，即研究参与者试图展示自己的积极形象。霍桑研究还在实验设计中做出了另一个重要贡献，它显示了操作性检查的重要性，操作性检查将在第9章中介绍。在霍桑研究中，影响并不是那么重要，而是工人对这些影响的解释，理解受试者的解释是操作性检查的目的。

如果观察者记录数据，还可能出现测量标准的变化。还有一种现象被称为回归平均，即那些得分非常高或非常低的人在下一次测试中趋向于回到中间水平。此外，如果第一次观察和第二次观察之间有时间间隔，可能发生其他事情。例如，如果记录的观察是一个人对飞行的恐惧程度，而在O1和O2之间发生了一起飞机失事事件并被报道在新闻中，这可能会改变结果。

**静态组比较**

在这种设计中，添加了一个未接受处理的第二组（O2）。

X O1

O2

例如，比较观看总统候选人辩论的人与未观看的人，或者比较接受反欺凌培训的学生与未接受培训的学生。但是这种设计也存在问题，包括无法确定两组人在重要个体特征上是否相同，例如，也许在观看辩论的群体中，共和党人比未观看的群体多，或者一组中的学生比另一组中更具攻击性。这将在第7章中的随机分配一章中详细讨论。人们的政治认同和学生固有的攻击倾向可能会影响结果。

由于我不推荐这三种设计，所以不会详细讨论它们的优缺点。相反，本章将重点讨论下面三种真实实验设计，这些设计是我推荐的。

**三种真实实验设计**

**前测-后测对照组设计**

这个设计是社会科学中最常用的设计之一。它增加了一个关键的特征，即将受试者随机分配到治疗组或对照组，从而确保两组在可能导致任何变化的重要特征上是相等的。

R O1 X O2

R O3 O4

这个设计的缺点与单组前后测设计相同，即被观察或测量两次可能会导致受试者的表现、态度或其他结果发生变化。由于这种威胁被证明是如此普遍，前测的流行度逐渐下降，实际上在真正的实验设计中，前测并非必不可少的。19

**所罗门四组设计**

常常将所罗门四组设计描述为黄金标准。它消除了前面所描述的所有缺点，此外，研究人员可以通过设立没有前测的组来判断是否存在测试效应。

R O1 X O2

R O3

O4

R X O5

R O6

但是由于成本较高、复杂且相对困难，相对而言使用它的研究并不多。20 Bausell甚至称其为“浪费”。21 在这种设计中，有四组，受试者被随机分配到所有组中。有两个对照组，其中受试者不接受处理或干预，它们作为与接受处理组进行比较的基准。一个处理组进行前测，另一个处理组不进行前测；一个对照组进行前测，另一个对照组不进行前测。这使得研究人员不仅可以比较处理前后的差异，还可以看到前测是否影响了结果。这提供了Levy和Ellis所称的“对大多数对立假设的有力回应”。22 然而，这也需要更多的受试者参与实验，而这种成本有时超过了好处。（参见研究重点4.2，了解所罗门四组设计的一个例子。）

**后测唯一对照组设计**

在Campbell和Stanley的实验设计分类中的最后一个设计是不包括前测的设计，这是我最常使用的设计，也是目前最为推荐的设计。23

R X O

R O

### 

**学习要点 4.2**

一项使用Solomon四组设计的研究SAGE期刊文章 study.sagepub.com/coleman

Genç, M. 2016. “六年级学生合作学习过程评估。”《教育研究》95(1)：19-32。

这项研究采用了所罗门设计来评估特定教学策略对六年级学生科学知识的影响。这种教学策略被称为合作学习，教师将学生组织成小组，在小组内独立进行研究，共同学习信息和解决问题。合作学习的教室被视为处理组或实验组；而以传统方式向学生传授信息的教室则被用作对照组。

在这个所罗门四组设计中，有两组学生接受了合作学习干预，另外两组没有接受干预，成为对照组。其中一组对照组和一组干预组进行了前测；另外两组没有进行前测。作者称所罗门设计的使用是为了确定前测是否产生了影响，因为学生可能仅仅通过两次测试来学习。学生被随机分配到了这四个组中。下面是对所罗门设计的描述：

“所罗门四组设计是为了消除前测的可能影响而采用的。它涉及将受试者随机分配到四个组中，其中两个组进行了前测，而另外两个组没有。一个经过前测的组和一个未经前测的组接受了实验处理。尽管每个组都经历了实验后评估，但前实验评估仅在两个组中进行，一个是实验组，另一个是对照组。前两个组被视为前测-后测对照组设计，而另外两个组被视为后测-对照组设计”（第22页）。

这项研究并没有详细讨论前测是否影响了结果；然而，从平均分可以看出，接受前测的组在后测中的得分要高于未接受前测的组。然而，该研究没有报告任何显著性检验结果来确定仅前测是否有显著影响。接受前测的学生在后测中的平均分为27.70（实验组）和22.50（对照组），而未接受前测的学生在后测中的平均分为26.74（实验组）和21.91（对照组）。为了说明设计并便于比较前测组和非前测组的后测得分，我重新制作了作者的表格，以下是重新制作的表格：

需要注意的是，前测和后测使用了不同形式的测试，而不是相同的测试，这有助于防止通过两次测试来学习。

合作学习处理有效，合作学习教室中接受前测和未接受前测的学生在实验结束时的科学测试中表现明显更好。

一些社会科学学科比其他学科更常使用所罗门设计，因此了解同事对此的看法非常重要。

在一项比较前测-后测设计和仅后测设计的研究中，Gorard（引文24）发现仅后测设计的结果“更少误导性”。即使是Campbell和Stanley（引文25）也表示前测在真实实验设计中并非必需，并解释了人们对前测的误解。当前测的目的是确保被分配到不同条件的受试者之间的等价性时，随机分配是一种足够的预防措施，因此不需要前测。他们指出，几乎所有罗纳德·费舍尔（Ronald Fisher）的农业实验都没有前测。

前测仍然被某些领域视为重要；这是在该领域中的标准。前测与处理之间的交互作用在大多数情况下并不大，但特别是在教育、心理学和社会学等领域，效应较大，不应忽视。一些研究则处于前后测的界限，例如进行了使用前测的试点研究，如果排除了测试效应，那么进行实际实验时就不再进行前测（引文27）。 （参见"如何做"4.3框中有关如何描述仅后测对照组设计的示例。）

这些设计代表了一种基本结构；实际的实验可能会有细微的变化，但仍然是可接受的。到目前为止，本章回顾了三种被归类为“真实”实验和三种被归类为“准实验”的实验设计，根据Campbell和Stanley的分类法。Campbell和Stanley还描述了其他十种被称为“准实验”的设计。本书不会详细介绍它们，而是总结准实验的基本特征，并将读者引向Campbell和Stanley的书籍以获取更多详细信息。

**准实验研究**

真实实验和准实验之间的关键区别在于，在准实验中，受试者没有被随机分配到条件中，群体也不一定是

**如何操作 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。

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，

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。

**常见错误**

• 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。

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

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

Answers

1. b

2. c

3. b

4. a

5. d

6. b

7. a

8. d

9. c

10. b

**应用练习**

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？

**建议阅读**

Campbell, Donald T., and J. C. Stanley. 1963. Experimental and Quasi-Experimental Designs for Research.

Chicago: Rand McNally.

Crasnow, S. 2015. “Natural Experiments and Pluralism in Political Science.” Philosophy of the Social Sciences

45 (4/5): 424–441.

Gerber, Alan S., and Donald P. Green. 2012. Field Experiments: Design, Analysis, and Interpretation. New York:

Norton.

Levy, Y., T. J. Ellis, and T. Cohen. 2011. “A Guide for Novice Researchers on Experimental and Quasi-

Experimental Studies in Information Systems Research.” Interdisciplinary Journal of Information,

Knowledge and Management 6: 151–161.

**第五篇 实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。*

*— 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

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。"

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**

**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。

"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。

**内容延展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。

**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。

**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。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。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。

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。

**常见错误**

• 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

**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

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

**应用练习**

1. 继续发展你的实验，写出你将如何在你的研究中确保外部和内部的有效性。写一到两页关于你将如何通过使你的研究更加真实和可推广来提高外部有效性。以本章的主题为指导。再写一到两页关于你将如何最小化对内部有效性的威胁，特别是考虑本章中描述的七种威胁。同时，包括对权衡的讨论，因为你很少能有完美的内部和外部有效性。哪个更重要，你会放弃什么；为什么？假设评审者在这些问题上对你提出了质疑，并对他们的担忧进行回应。

2. 找出你领域内的三个已发布的实验，并对每个实验的作者如何处理内部和外部有效性问题写一页的批评。他们做出了哪些权衡？是否令人信服？你是否有其他的论点可以使用？他们有没有没有考虑到但是应该考虑的其他威胁？"

**推荐阅读**

第2章，“统计结论有效性和内部有效性”，以及第3章部分内容，“外部有效性”，页码83-102，出自Shadish，W. R.，T. D. Cook，和 D. T. Campbell的作品。2002年出版的《实验和准实验设计：用于推广因果推断》。你还不需要阅读建构效度部分，那将在下一章中出现。

想要阅读一篇关于实验现实性的好例子，请阅读Iyengar，Shanto和Donald R. Kinder的作品。1987年出版的《新闻影响力：电视和美国舆论》。这些研究尽可能地模拟实际的国家新闻广播。

Stoker，Gerry。2010年的《探索政治科学实验的承诺：微观基础的洞见和政策相关性》。这篇论文讨论了争论在政治科学中需要更多实验的有效性问题，但这些论点适用于所有的社会科学。

**第六篇 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*

**学习目标**

**• 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。**

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。

**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。

**内容延展6.1**

**阶乘表示法**

在阶乘表示法中，数字的数量告诉我们实验中有多少因子或变量。数字的值告诉我们有多少层次的因子。表6.1解释了一些最常见的阶乘设计。

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。

**如何操作6.2**

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.

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。）

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。

**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”。

**"学习聚光灯6.3**

**一个研究中的单因子和阶乘设计示例"**

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。

**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。

**常见错误**

• 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

**测试你的知识**

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

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

**Answers:**

1. c

2. c

3. d

4. a

5. c

6. b

7. c

8. d

9. d

10. b

**应用练习**

1. 选择你在第一章中提出的实验思路之一，并在第三章中写下文献回顾和假设，然后画出一个设计表来帮助你将因素、层次以及每个因素和层次的组合进行可视化。将其转化为阶乘表示法。包括设计类型、你的因素和层次。请参阅如何做到它的框。

2. 写一页解释你选择了组间设计还是组内设计。捍卫你的决定（例如，为什么不会有连带效应）。假设一个评审者质疑了你的选择，认为相反的选择会更好。你会说些什么？

3. 写一页关于你的控制组类型以及受试者在没有接受治疗的情况下应该做什么。解释你的选择并捍卫你的决定。

**推荐阅读**

R. Barker Bausell (1994)的《进行有意义的实验：成为科学家的40个步骤》中的第6章，“设计实验研究以达到统计显著性”。千橡树，加利福尼亚：Sage出版社。

Paul C. Cozby (2009)的《行为研究方法》（第10版）中的第10章，“复杂的实验设计”。波士顿：麦格劳-希尔。

**第七篇 随机分配**

*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*

**学习目标**

**• 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。**

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。

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。”

**内容延展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。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。

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。

**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。

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。

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。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

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但使用拉丁方块法可以产生三种组合：

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。

A screenshot of a computer

Description automatically generated

**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 cell phone

Description automatically generated

**常见错误**

•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

检测你的知识

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

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

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

**应用练习**

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。

**建议阅读**

Chapter 5 in: R. Barker Bausell. 1994. Conducting Meaningful Experiments: 40 Steps to Becoming a Scientist.

Thousand Oaks, CA: Sage.

Chapter 8, “Randomized Experiments: Rationale, Designs, and Conditions Conducive to Doing Them,”

pp. 246–278, in: Shadish, William R., Thomas D. Cook, and Donald T. Campbell. 2002. Experimental and

Quasi-Experimental Designs for Generalized Causal Inference. Belmont, CA: Wadsworth Cengage Learning.

Read these two articles, one critical of random assignment and one more supportive, for a comparative perspective:

• Krause, Merton S., and Kenneth I. Howard. 2003. “What Random Assignment Does and Does Not

Do.” Journal of Clinical Psychology 59 (7): 751–766.

• Strube, M. J. 1991. “Small Sample Failure of Random Assignment: A Further Examination.” Journal

of Consulting & Clinical Psychology 59 (2): 346–350.

**第八篇 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。**

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。

**如何操作？**

**8.1 科目选择的理由**

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。”

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。

有许多方面使大学生与其他人不同，例如，他们更有可能更多地使用社交媒体而较少使用传统媒体，34并且更不保守。由于这些差异可能对研究结果产生影响，研究者应该仔细考虑是否需要将大学生作为研究对象。然而，需要注意的是，并非所有研究都受到这些差异的干扰。在一些情况下，大学生可能是适当的研究对象，尤其是当研究的现象或过程对大学生与其他人不会有差异时。21只要现象或过程对大学生与其他人没有差异，研究结果就应该是有效的。22

**服从权威**

因为遵循指示和取悦权威对在大学中取得好成绩很重要，学生可能更有可能试图给研究者他们期望的结果。23这一现象在第2章中涉及了需求特征。请回忆一下，Stanley Milgram因为这个原因拒绝在他关于服从权威的研究中使用学生。学生也可能更容易察觉到研究的假设，因为他们有参与研究的经验，24这导致了更多的偏见回答。25其他研究表明，了解研究的受试者可以有意识地控制他们的回答并破坏结果。26研究者应该始终隐藏假设，并使用减少各类受试者猜测假设的测量方法。27关于这些测量方法的更多信息将在第9章中介绍。此外，学生与权威人物（教授）的正常互动可能会使学生的回答与非学生不同。一位研究人员在测试两种压力缓解治疗方法时发现，一种治疗方法对学生样本有效，但对来自社区的人（2023)没有效果，而另一种治疗方法对学生没有效果。28在某些情况下，更高的服从度也可能支持将学生作为研究对象，这可能导致更准确的答案和更少的非响应误差。29

**自我意识较弱**

自我意识较弱不仅适用于大学生，也适用于大学外的其他18至21岁的年轻人，因为在生命周期的这个阶段，年轻人通常在发现自我身份。然而，就研究而言，这会导致意见更加易变，并更容易受到有说服力的论点的影响。30这反对了在态度变化研究中使用学生的做法，例如。31同样，研究者应该考虑这种特征是否与结果变量有关，如果是，就应避免使用学生作为研究对象。32Basil举了一个态度变化研究不适合学生样本的例子，然而，意见形成的过程可能并没有不同，因此学生可能是适当的研究对象。33

**其他种种差异**

大学生在许多方面与其他人存在系统性差异；例如，他们可能更多地使用社交媒体而较少使用传统媒体，34并且更不保守。由于这些差异可能对研究结果产生影响，研究者应该仔细考虑是否需要将大学生作为研究对象。然而，需要注意的是，并非所有研究都受到这些差异的干扰。在一些情况下，大学生可能是适当的研究对象，尤其是当研究的现象或过程对大学生与其他人没有差异时。21只要现象或过程对大学生与其他人没有差异，研究结果就应该是有效的。

而且教条主义。35与所有研究一样，必须仔细考虑样本的系统差异。当研究不声称推断到更大的人群，并且结果变量不受他们之间的差异影响时，大学生通常是合适的。36关于差异对研究产生偏见的证据或反证应该来自于理论、现有证据和逻辑。如果没有这样的证据，研究人员应该在有充分理由使用大学生而不是其他人的情况下继续使用大学生。37使用大学生样本的研究有能力增加我们的知识，当没有异议可以确认时，应该优先使用他们。

**同质性**

作为研究对象，大学生的一个不基于研究人员简便或成本较低的论据是，他们是同质的，也就是说，在几个特征上彼此相似。在实验中，受试者的同质性有助于内部效度，通过帮助控制混杂因素，即受试者的年龄、教育水平、技术使用等方面的差异，如果这些因素没有变化，就不能解释结果的差异。38可能的干扰因素被最小化，测量的变异性减少，这有助于拒绝零假设，并确定理论的真伪。39在追求对照组和实验组的等效性时，如第7章所述，拥有在与结果相关的变量上相似的受试者是关键。大学生在某些变量上比非大学生更同质，但在其他变量上与普通大众一样多样化。41大学生在年龄、教育、媒体使用、42生活经历、政治参与43和政治知识44等方面彼此相似。就政治知识和参与度而言，他们往往低于平均成年人，因此研究人员应避免在大学生中进行政治问题的研究。45同样，大学生对纳税46、利率和抵押贷款等方面缺乏经验，因此在这些问题的研究中应避免使用大学生作为对象。47这个论点适用于任何一组同质的受试者；例如，政治参与、知识和积极性高的人不代表普通人。48因此，从政治组织（例如妇女选民联盟）中选择的受试者具有同质性的好处，但不应推广到其他人。在某些领域，如跨文化研究，例如，大学生的相似教育水平可以很好地估计代表性样本。49

**基本心理过程**

使用学生作为研究对象的主要论据是，当研究涉及到适用于所有人的基本心理过程时，使用学生的结果具有普遍性。51如果没有证据或理论依据表明所研究的心理过程在学生和非学生之间存在差异，那么使用学生作为受试者是合适的。52这个论点在研究基本的认知过程、感知、注意力、记忆和学习等方面尤为重要，因为预期这些基本过程在各个人群中是相似的。53研究人员在决定是否使用学生样本时应考虑研究结果的推广性以及所研究心理过程的特性。

总而言之，在许多研究中，使用大学生作为研究对象是常见且合理的。然而，研究人员应仔细考虑与学生样本相关的潜在局限和偏见。他们应评估学生与非学生之间的系统差异是否与研究问题和结果变量相关。当所研究的现象或过程不受这些差异影响且没有证据表明存在差异时，大学生可以作为合适的受试者。此外，学生样本的同质性可以增强实验设计的内部效度。然而，研究人员应始终对学生样本的适用性进行批判性评估，并在研究方法和报告中清晰地阐明选择的理由。

在没有理由相信学生的认知过程与非学生不同的情况下，使用学生样本是合适的。52其他人进一步表示，如果研究旨在得出理论性结论，例如进行基础研究和理论测试，则样本并不重要。53然而，也有人认为即使如此也值得怀疑，因为已经证明无法从一个大学的学生推广到另一个大学，54甚至从一个专业的学生推广到另一个专业，以及在不同学年中的学生。55

其他研究指出，受试者在研究中执行的任务对于确定学生是否适合作为受试者至关重要。对于具有简单决策的分析任务，学生比具有复杂决策和道德或伦理困境的任务更合适。56

**防止学生偏见的方法**

有一些步骤可以帮助实验研究人员减少与使用学生作为受试者相关的问题。当学生是研究的对象时，进行学生研究是合适的57，例如教学技术研究或在广告或消费者研究中作为目标市场。58

研究人员应该对学生样本与其他人群之间的差异进行比较。59文献是开始的好地方；在许多领域（如会计、护理和教育等）中可以找到许多证据表明，学生在许多方面与其他人没有差异。60但也有很多研究表明，学生不适合作为其他人群的代表，其中一些研究结果与同一领域的发现相矛盾。61这是第3章“了解文献”的另一个重要领域。研究人员应该至少进行初步的经验性证据综述，以了解文献对于学生样本的观点。

学生样本还可以使用百分比和平均可变性或均值方差等统计技术与人口数据进行比较，从而增加对预期变异性的了解。62（公式可以在Cochrane 1977中找到。）63

在同一研究中同时使用学生和其他人群的受试者，并进行统计分析以展示差异，这些差异可以在正文或脚注中报告，是另一种好方法。64当没有发现差异时，这有助于向读者保证学生受试者对于特定问题是合适的。如果存在差异，那么研究表明效应可能取决于样本特征。65此外，使用不同来源的样本是内部和外部效度的关键。66

研究人员还可以使用非学生样本复制自己的研究（以及他人的研究）以查看是否有差异。67正如Lindsay和Ehrenberg所说，“如果一项研究的效果只在一个样本中发现，那么它可能是一个特定样本的效应。”68最后，研究人员可以使用多个样本来源来进行复制研究，以增加对样本特征对效应的影响的了解。69

如果一项研究值得进行，那就值得进行两次。”68一些人呼吁在顶级期刊上发表论文的要求是进行至少两次不同受试者的研究复制。69在进行复制研究时，请遵循第5章的建议以某种有意义的方式扩展研究结果。

确保学生受试者不会混淆结果的另一种方法是使用统计控制变量或协变量。这些变量应与结果变量相关或在学生与其他人之间存在很大差异，例如年龄和教育等人口统计学特征。70只要可以预测和控制学生与其他人之间的差异，这些差异就不成问题。71在论文中报告这些差异，并在确实影响结果时诚实地讨论研究结果。

最后，一些人建议测试受试者对所使用的理论和方法的先前知识以及对研究的经验，以确定猜测假设的可能性。

正如第5章所指出的，最好使用与研究对象尽可能相似的受试者；如果研究涉及教学现象，而教师可以合理地预期同意成为受试者，那么这比使用学生更好。然而，使用教育专业学生可能是一个合理的替代，因为他们将成为未来的教师。正如Basil在本章中所指出的，在获得学位前一刻和获得学位后一刻，一个人在研究中的资格并没有差异。例如，有人认为，在被录取为新闻学专业本科生之前，精英新闻学院的本科生已经在学生媒体方面有了重要经验，可以作为新闻工作者的替代。73还有人使用过先前在该领域担任过专业人员的公共管理硕士专业学生作为受试者。74对于实验来说，真正关心的是样本是否适合所研究的内容，而不是样本是如何选择的。75将调查和其他方法的标准应用于衡量实验设计中受试者的质量是不恰当的方式。

最后，论文应该始终讨论为什么学生，或者任何类型的受试者都适合这项研究。76彼得森表示，这应该通过受试者在测试特定研究问题上的理论相关性来证明。77在某些领域，这是罕见的；例如，Marriott发现在过去二十年里的十本税务学期刊中，80%的论文没有为他们的样本选择提供理由或承认与使用学生有关的任何局限性。78根据一项研究，市场营销和消费者研究中有63%的论文忽略了样本的局限性。79

此外，限制使用学生样本的研究的一般化，并且不要从这些研究中提供任何政策变更或对专业人员的实践建议。

例如，将学生的研究结果推广到纳税公众或提出税收政策建议是不合适的，尽管研究通常会这样做。80相反，结论不应集中在“人”，“个体”或研究针对的专业人员上，而应集中在“本研究的参与者”，“这些受试者”或“本研究中的专业人员”上。

这也是论文中重申实验的目的的好地方-建立因果关系，而不是推广到整个人口，正如第5章中所解释的那样。本章的大部分内容都专注于学生作为受试者的使用，因为这是实验研究中最重要的问题之一。下一节将讨论实验的其他类型样本以及对它们提出的异议。

使用非学生组成的样本并不会自动解决研究者的所有问题。早些时候描述的许多问题也适用于其他类型的样本。一般来说，方便性样本存在偏见，因为受试者是自愿参与的。他们之所以这样做，可能是因为他们对所研究的主题非常关心，或者是因为他们受到了激励，无论是金钱还是其他东西。81虽然没有完美的样本，即使是从人口中随机选择的样本，也有其他好的受试者来源可供实验研究使用。下面将讨论一些最流行的来源。

**亚马逊的机械土耳其（Mechanical Turk）**

机械土耳其（Mechanical Turk）是进行实验的最新受试者来源之一，它是亚马逊的在线众包平台，人们可以通过执行计算机无法完成的各种任务来获取报酬。82亚马逊于2005年开发了MTurk，并用于自己的需求83，但也向其他需要只有人类才能提供的服务的人开放，例如撰写书摘、评论或标题，识别照片中的物品，转录音频甚至是购物收据，84等等。亚马逊将这些任务称为人工智能任务（Human Intelligence Tasks，HITs）。86最不寻常的任务之一是让工作者在卫星图像中寻找失踪人员。87迄今为止还没有通过这种方式找到过任何人。机械土耳其的名字来自18世纪的一种棋类游戏机器，它在集市上很受欢迎，设计成像土耳其巫师。实际上，里面藏着一个小人，他扮演（并通常战胜）对手-因此，该平台的名称旨在隐藏人类工作者。88在2010年，社会科学家发现它是招募人员参与调查和实验的手段。89其优点包括易于操作，成本低廉，潜在地能够代表整个人口。

**MTurk是否具有代表性？**

早期对MTurk受试者的批评是他们的年龄较小、教育程度较高，并且女性比例较高。90但情况已经发生了变化，人口统计学数据更加符合美国人口的代表性。91现在，许多研究表明，MTurk样本与美国工人、标准互联网样本和其他受试者群体相当，92在多样性方面也比学生更加多样化，特别是在年轻的亚洲人和年轻的西班牙裔女性方面。94但并不是所有的研究都显示出这一点；在一项研究中，学生实际上比MTurk工作者更能代表整个人口。95其他研究将MTurk工作者的代表性与国际人口进行比较。96

随着MTurk的普及，对其外部有效性的挑战也越来越多。97探索MTurk与其他样本来源之间的差异已经成为一种几乎成为一种行业。有关部分文献的评论，请参见第8.2框中的更多关于MTurk代表性的研究。

与使用学生作为受试者一样，研究人员应该在考虑是否与其他受试者群体和感兴趣的人口有所不同时对使用MTurk工作者进行批判性思考。110此外，仅依赖一种类型的样本，无论是学生、MTurk工作者还是其他来源的受试者，都会带来自己的缺点。111在同一项研究中使用多种来源是实现实验的更大外部有效性的一种重要方法。112

**优势和缺点**

MTurk在社会科学领域广受欢迎，因为它具有一些优点，但也存在一些缺点。除了拥有大量的普通人群外，低成本可能是研究人员使用MTurk的最大原因之一。工作者按任务付费。一些研究人员支付61美分到75美分，以换取8到12分钟的工作时间。113具有来自其他请求者的高批准率（MTurk用于发布HIT的人）的工作者会获得更高的报酬。例如，在一项研究中，具有95%批准率的工作者获得了3美元的报酬。114平均时薪在1美元到1.38美元之间。115亚马逊还会收取一定比例的费用。即使以较高的费率计算，MTurk的报酬可能远远低于学生和成年人所需的报酬。使用MTurk样本所实现的节约可以用于运行更多的受试者或增加更多的条件。

在MTurk中招募受试者和收集数据也非常快捷和简便。发布HIT后，数十万名工作者会自动收到通知。与此相反，走遍班级宣传研究，或在图书馆或咖啡店花费几天时间寻找受试者。数据收集同样迅速。研究人员报告称，在几个小时或不到一天的时间内就能够收集到足够数量的受试者，116例如

## 

## **更多信息......方框8.2**

## **关于MTurk代表性的研究**

许多研究都围绕着MTurk工作者与人口的代表性问题展开。这对于使用MTurk工作者作为受试者的研究的概括性具有重要意义，特别是对于调查研究和实验研究而言。以下是关于这个问题的一些研究的简要概述，包括正面和负面的观点。由于不断有新的研究在进行，读者应该寻找更多和更新的文献进行了解。

Bartneck和同事发现，通过MTurk招募的受试者与在校园或在线招募的受试者有所不同，但他们认为这种差异“没有实际意义”。98 Huff等人将MTurk与合作选举调查进行了比较，并发现了强烈的相似性，在年龄、性别、种族以及政治、职业和地理信息等变量上的最大差异不到7个百分点。他们表示：“研究人员有充分的理由考虑使用MTurk对许多更广泛的受试者群体进行推断，并鼓励研究人员不要因为研究使用了MTurk受试者而自动将其排除。”99

然而，Levay等人复制了2012年美国国家选举研究（ANES）系列中的问题，并与MTurk工作者进行了比较，发现存在显著差异。通过使用九个变量的协变量（年龄、性别、种族和族裔、收入、教育程度、婚姻状况、宗教信仰、意识形态和党派倾向），这些差异得到了缩小。作者得出的结论是“MTurk的回答者在不可测量的方式上似乎与基于人口的回答者没有根本的不同”，并建议将这九个变量作为协变量进行控制。对于实验研究人员来说，他们表示“MTurk的代表性在大多数情况下不会对实验推断构成威胁”，但他们对可能存在的不可测量的差异变量提出了警告。

Berinsky等人在2008年至2009年的ANES中得出了类似的结论，发现MTurk受试者的代表性不如ANES受试者或互联网调查的受试者，但比当面方便样本更具代表性。102

就个体的人口统计学和政治变量差异而言，MTurk受试者可能与美国人口有所不同。103然而，研究人员仍然声称，从MTurk样本中得出的结论在许多情况下比学生样本更好，只要我们了解这些差异。104

比较不仅限于人口统计学上的相似性，研究还检查了实验中的任务类型和结果变量，以比较MTurk和其他受试者之间的差异。例如，Paolacci等人和Birnbaum发现，在决策实验中，MTurk受试者与其他受试者之间仅有轻微差异。105Buhrmester等人发现，在各种心理测量测试中，MTurk和其他受试者之间没有实质性差异。106Bartneck等人发现，在使用MTurk、在线服务或学生作为受试者的研究中，研究结果没有差异。他们表示，MTurk更适合于对整体人口而不是特定子群体的研究。107Crump等人在一系列任务中，包括潜意识启动和学习任务，得出了不同的结果。108Krupnikov发现，当要求MTurk受试者阅读文章时，他们产生的结果与学生和全国代表性成年人样本显著不同。109

虽然所有这些相互矛盾的结果可能令人困惑，但关键是要知道MTurk工作者在个体变量、结果和任务上与人口的差异，并相应地采取行动。

然而，这些优点也伴随着一些缺点。低报酬会导致工作者尽快完成任务，这可能导致他们不专注，反复按同一个按钮，或者随机选择答案。在网上进行实验的受试者，无论是在MTurk还是其他平台上，都比实验室中的受试者更容易受到干扰。为了控制这一点，可以加入一个注意力检查问题或两个。例如，在研究的几个环节中，可以问一个关于受试者刚刚阅读内容的事实问题，或者插入一个指示，要求他们在下一个问题中选择特定的答案，无论他们实际的回答是什么。例如，“请在下一个问题中选择‘4’作为回答，无论你的实际回答是什么。”在说明、同意书和研究开始时清楚地说明，研究中包含用于确保他们注意力集中的问题，未能正确回答这些问题将导致无法获得报酬。工作者非常希望得到报酬，这不仅是因为钱，而且因为未能获得报酬会导致差评，从而减少他们未来能够参与的高报酬任务的数量。在研究结束时还可以问一些关于刺激细节的问题。为了减少回答错误，建议研究人员视觉上检查模式，例如重复按同一个按钮，或者使用对同一个问题以不同措辞的成对一致性和其他机制来检测受试者匆忙填写问卷。研究表明，MTurk和传统调查在注意力方面没有区别。可以通过其Amazon ID识别工作者，以确保他们不会多次参与同一项研究。

报酬较高似乎与完成的任务数量相关，但数据的质量与报酬率无关。工作者表示他们参与研究既是为了娱乐也是为了报酬。

其他缺点包括研究人员无法像在实验室中那样控制MTurk受试者的环境。此外，收集人口统计信息和测量回应时间需要额外的软件，如Qualtrics调查平台或AdobeFlash，因为Amazon平台不支持这些功能。MTurk既可以用于招募受试者，也可以直接托管问卷，或者使用诸如Qualtrics、SurveyMonkey等调查工具包来托管问卷。

对于学生较少的研究人员来说，这些缺点可能会被权衡。例如，小学院的研究人员、没有受试者库的研究人员或非学术环境中的研究人员。

**其他受试者的来源**

其他更传统的获取受试者的来源可能成本更高，但通常不会带来之前提到的关于学生和MTurk样本的担忧。其中包括商业调查和市场研究中的选择参与者面板，例如Qualtrics、Survey Sampling International和Knowledge Networks（现为GfK KnowledgePanel）。其他包括SurveySavvy、Harris Poll Online、YouGov的PollingPoint Panel、Polimetrix和Survey Spot。一些研究人员使用Craigslist。还有很多其他选择，但这些似乎经常出现在发表的学术论文中。这些公司拥有一批同意以报酬参与研究的人，并选择他们愿意参与的研究项目。有些还提供定向样本，可以代表某个人口群体或具有特定特征，例如注册选民。费用取决于问题数量、受试者类型和其他细节。研究人员报告的付费范围从每个受试者3美元到15美元不等。

当使用来自特定人口的受试者进行调查实验时，例如由选择参与数据平台提供的人口，一些学科和期刊要求报告响应率。事实上，美国政治学会（APSA）的实验研究部门对每一项使用调查数据收集方法的实验都提供了具体要求，甚至包括针对大学生样本。关于是否适当或必要存在一些争议，一些研究人员表示，如果研究不声称具有普适性，那么响应率就没有意义。目前，除了政治学以外，社会科学学科中很少有要求调查实验的响应率；然而，研究人员应该遵循自己所在领域和投稿期刊的标准和指南。

更传统的方法是通过报纸广告、城市张贴传单和邮件列表来招募受试者。这些方法已应用于新媒体，如社交媒体和在线论坛，现在也成为受试者的来源。例如，一项研究使用互联网上的乐高论坛进行了一项关于判断乐高玩具上面的面部表情的研究。作者注意到这种方法的潜在偏差在于受试者在对乐高的兴趣和经验上是同质的。另一组研究人员开发了一种从社交媒体招募受试者的方法，可以针对特定主题进行定向，例如政治参与者-这是学生样本的一个特定弱点。它吸引了在线意见领袖，如博客作者和讨论论坛版主是招募受试者的另一种传统方式。这种招募方式特别适用于研究特定样本。例如，对于政治学家来说，能够找到政治上精明的受试者（他们在整体人口中占很小比例）可以更准确地估计效应并提高外部有效性。另一个例子是一项使用仇恨团体聊天室中的受试者研究触发攻击性信息类型的研究。作者写道：“通过使用具有全国代表性的概率样本、传统学生样本或其他在线自愿参与方法（如MTurk），无法获得这种洞察力。”（请参阅“研究重点8.3”以获取有关社交媒体作为受试者来源的更多信息。）研究人员还可以寻求特定亚群的组织帮助，例如加入美国公共关系协会的公共关系专业人士。

其他研究人员在学校职员、陪审团名单（由随机抽样公民组成，更加多样化）和选民名册中找到受试者。有些研究人员能够在大规模调查中免费包含实验性问题，例如社会科学的分时实验；而其他一些研究如合作的国会选举研究可能会带来成本。

**招募**

另一个招募受试者的方法是在公共场所如公园和户外咖啡馆进行招募。一位研究人员在咖啡馆等场所的桌子外，得到业主的许可后，通过提供该场所的礼品卡来招募过路人。另一位研究人员发现消防部门对研究人员的访问持开放态度，因为消防员在等待出警时有很多空闲时间。受试者还可以通过诸如狮子会和教堂团体等社交组织来招募，这些组织邀请演讲者参加他们的会议。在会员花费十五到二十分钟用纸和铅笔参与研究后，研究人员可以提供有关他们研究的引人入胜的讲座。这种方法也在新闻社进行伦理研究时被使用；主编通常乐于邀请教授给他们的记者进行伦理小研讨会。其他策略包括在会议上设立展台，以招募职业人士，例如在互联网域名投资者的会议上。毫无疑问，创意展示，如装满钱的公文包，有助于吸引受试者。当受试者需要代表特定人群，例如职业人士时，这种方法非常有效。

**研究重点8.3**

**陪审团名单和社交媒体作为受试者来源**

Murray, G. R., C. R. Rugeley, D-G. Mitchell, and J. J. Mondak. 2013. "便捷但不是方便样本：陪审团名单作为实验受试者来源."《社会科学研究》42 (2013): 246–253.

这些作者发现了另一个便捷、多样且廉价的实验受试者来源，而不是依赖大学生——那就是被传唤参加陪审团的人。本文讨论了实际问题，如获取访问权限和实施实验，以及包括回应率、样本质量和代表性在内的证据。

其他研究人员也使用过潜在的陪审团成员作为研究对象，包括Niven、Sigelman等人、Terkildsen和Lewis等人。

使用陪审团作为研究对象存在一些问题，包括获取法庭官员和法官的合作。这些作者自2009年以来一直与陪审团名单进行实验，他们报告称没有遇到问题，并表示该县与大学之间的良好关系有所帮助。并非所有人都有相同的经历；例如，一个使用陪审团名单的全国性研究只有15%的合作率。

其他问题包括法庭管理员担心工作负担增加，法官担心法庭程序被打乱。法官还担心陪审员能够得到公平对待，并且研究不会对他们产生偏见。其他与机构审查委员会（IRB）的要求相一致的担忧是，受试者必须自由回答（问题）或停止研究。

法官对研究内容表示关注，希望避免与法律问题有关的内容；政治话题是可以接受的。法官会审查研究并批准，这些作者在他们的所有经验中只有一个问题被删除。

在程序上，研究人员使用纸笔格式，但他们认为手持电子设备也可能可行。由于潜在陪审员的有限空闲时间，他们建议将研究时间限制在10至15分钟。他们在陪审员刚到达并等待选人过程开始时进行研究。他们在陪审员到达之前在座位上放置夹板、带有铅笔的封面信、说明和问卷。他们在夹板和奖励上花费了350美元。应该有一位研究人员在场监督研究并回答问题。他们发现没有必要提供奖励，并估计每个受试者的成本为57美分，与MTurk的成本相当，包括复印和用品成本，但不包括前往法院的交通费。

在这些研究人员所在的县，陪审团名单每周召集两次，每次有大约75至200人参与。在10万人口的县中，他们估计每月平均完成的研究数量为400份。他们的回应率是参与陪审员中的79%；根据所有被传唤者进行调整，回应率在39%至68%之间。

由于陪审团名单是根据选民登记和驾照名单等来源进行抽样的，陪审团名单是多样且代表社区的。这些作者将陪审员与学生样本和同一研究所使用的电话调查进行了比较。您可以在表格中阅读他们的结果。他们得出结论：“总体而言，我们认为陪审团名单数据在成本效益社交媒体网络和互联网调查：通过社交媒体招募在线实验受试者."《PS: Political Science & Politics》49 (4): 1–10.

这项研究报告了作者们在六个研究中使用的一种方法，即通过社交媒体招募实验受试者的“社交媒体网络调查”（SMIS）方法。该方法通过请求博客作者、讨论论坛管理员以及Facebook和Google+页面管理员——所有这些人都是社交网络中的“核心节点”——帮助招募其网站的读者和访客作为研究对象。该方法的优势之一是可以招募特殊亚群体，如对特定社会问题有政治见解的人士和活动分子。这些研究人员所属的学科——政治学常常对高度参与、知识渊博和政治活跃的公民感兴趣，但这样的人在普通人群中并不常见。使用传统的招募方法将导致很多不符合政治熟悉者要求的受试者。而通过这种方法，他们可以有针对性地招募到这些人。例如，可以通过“与亚利桑那站在一起（反非法移民）”Facebook页面招募对移民问题感兴趣并积极参与政治的人士，该页面在2013年初拥有60万关注者。

各个学科的研究人员都可以使用这种方法。例如，关于疫苗的健康信息的研究可以通过Vactruth.com和Anti-Anti-Vaccine-Campaign的Facebook页面招募受试者。

研究人员使用SMIS方法首先确定吸引目标亚群体的社交网络，然后通过请求博客作者和页面管理员来支持他们的研究并争取他们读者、联系人和关注者的参与。他们建议使用Technorati（www.technorati.com）和博客名单（例如，通过BlogHer.com可以找到女性博主的目录）来识别博客和论坛。由于参与研究的邀请来自一个被认可的意见领袖而不是一个未知的研究人员，这种个人联系有助于增加参与的可能性。

作者们提供了与社交媒体协作者建立信任和合作关系的建议，包括与协作者联系以获得许可并回答项目相关问题，指出这种联系需要获得伦理审查委员会的批准。他们提供了一个在线补充附录的链接，其中包含他们使用的材料。在这篇论文中报告的六项研究中，研究人员联系的每八位博客作者中大约有一位同意推广研究。他们表示，平均每项研究他们招募了1,569名受试者，远远超过他们通过学生受试者群体能够招募到的数量。每个社交中介人平均招募了104名受试者。与MTurk不同，这些受试者是自愿参与的，并且不需要提供激励或支付报酬。

研究人员将他们的SMIS样本与MTurk、学生和2008年ANES（American National Election Studies）的样本进行了比较，并得出结论称，尽管它们不具代表性，但在人口统计学和地理方面比学生样本更加多样化。

这种方法效果不错。然而，研究人员应该意识到，在社交俱乐部和组织的情况下，样本将仅限于自愿参与的群体，他们有时间或对该主题感兴趣，并且共享特定特征，这些特征可能不代表整个人口，例如退休人员或家庭外没有就业的妇女。因此，从多个来源进行抽样以避免系统偏差可能是一个好策略。

**激励措施**

虽然一些研究人员发现他们不需要提供激励措施，但许多人发现即使是小小的奖励也很重要，以向被调查对象传达他们在研究中所付出的时间和努力是受到重视的。激励措施可以帮助激发参与，提高回应率并改善数据质量。激励措施的类型和金额取决于研究的性质、目标人群和可用资源。常见的激励措施包括现金支付、礼品卡、代金券或参与奖项抽奖。在使用激励措施时，重要的是要遵守伦理准则和机构规定，确保其不会对参与者产生不当影响或损害研究的完整性。研究人员还应考虑提供激励措施可能引入的潜在偏差。例如，特定的人口群体在对不同类型或金额的激励措施可能有不同的接受度，根据激励措施的高低，可能吸引主要受财务收益驱动而非对研究主题真正兴趣的个体，这可能会影响其回答的质量。

总之，尽管激励措施可以是招募和激励被调查对象的有效工具，但研究人员应仔细考虑其使用，同时考虑伦理问题、潜在偏差以及目标人群的特定需求和特征。

是有价值的。除了金钱激励外，还可以使用书籍、铅笔、文件夹和邮票等作为激励措施。礼品卡比现金更具成本效益；例如，相比于10美元现金，更多人可能因为5美元的礼品卡而选择参与。最好将激励描述为“一张（咖啡连锁店的名称）的礼品卡”，不必提及具体金额；很少有人会问具体金额是多少。提前打电话确认商店是否有足够的礼品卡存货，或者是否需要在总部批量购买。

另一种方法是告诉参与者他们将有机会参加一个更大奖品的抽奖活动，比如iPad。有研究人员进行过一次书籍抽奖活动，吸引了29名学生参与。研究人员还可以抽奖赠送有限数量的10美元礼品卡，这意味着他们可以购买少于受试者数量的礼品卡。并非所有的伦理审查委员会都允许使用抽奖激励措施，请事先咨询。

除了解释使用被调查对象的研究理由外，研究人员还应报告激励措施（如果有的话）和其他有助于读者评估研究的细节。在政治科学领域，美国政治学会（APSA）实验研究部门的标准委员会已经将一系列指南编码化，包括谁有资格参与、招募日期、研究进行日期和每组分配的受试者数量等内容。有关完整的指南清单，请参阅Gerber等人的文章。其他学科和学术期刊也可能发布自己的报告标准。最广泛使用的标准可以在《美国心理学协会出版手册》中找到。

毫无疑问，还有许多其他可以作为实验对象的好来源和招募方式，只受限于想象力的限制。在阅读研究时，注意受试者的身份、来源和招募方式。还要注意使用的受试者数量，这将是下一节的主题。

**样本大小和功效**

“我需要多少个受试者参与我的实验？”这是学生们最先问的问题之一。好消息是，相比于调查等方法，实验通常所需的受试者数量要少得多。但这个问题的答案并不快速或简单。

样本大小和功效的主题是相互关联的。样本大小，或者研究中的受试者数量，在实验中被表示为N或n，大写的N代表研究中的所有受试者，小写的n代表某个条件下的受试者。在设计实验的过程中估计样本大小非常重要。

为了确保统计结论准确可靠。如果样本太小，统计检验可能由于功效不足而无法显著，而不是因为假设错误。155 如果样本太大，研究人员将在时间和资金上花费很多，却获得很少的额外好处。因此，确保受试者的数量恰到好处非常重要.iii

这就是统计功效的作用。功效是支持假设的概率，或者在备择假设成立时拒绝错误的零假设的技术性概率。也就是说，功效是在实际存在效应时实现统计显著性的概率。功效也是第二类错误的概念，当零假设 不被拒绝而本应被支持时会发生这种错误；也就是说，备择假设应该被支持，但实际上没有被支持。这表明p值不低于0.05。当达到0.05或更低的显著水平时，我们有更多理由相信结果是“真实的”，而不是由于偶然性。当由于第二类错误而未达到统计显著性时，除了对研究人员来说非常令人失望，这还意味着某些可能有效的治疗现在被报告为无效。重要的干预措施可能永远不会在实践中采用。

研究人员往往专注于第一类错误，这是第二类错误的相反情况，也就是零假设被拒绝，但实际上不应该被拒绝。156 但第二类错误在发表的研究中也很常见。157 研究人员应该注意减少错误并提高功效，并且防范第一类和第二类错误。158

增加样本量是增加功效的最简单方法之一，从而减少第二类错误。159 较大的样本意味着更高的功效，因为它们更准确地代表了所抽样本的总体。160 以一个例子来说明，假设整个人口总数为一百万人。要测量所有一百万人是困难的，因此科学家从人口中抽取样本，并估计某个变量的均值。如果样本由十个人组成，可能会抽到特别低或高分的人。将样本增加到一百可以增加抽到分数较中等的人的概率。基于一百人的样本均值比基于十个人的样本均值更接近总体均值。

在这个计算中的另一个重要统计量是标准误差，它估计样本与总体统计量之间的平均差异。161 它是衡量样本误差的标准。

抽样统计量的标准误差也是重要的统计指标之一。当样本较大时，标准误差往往较小，因此，较大的样本可以增强统计效果。然而，由于时间和经费在大多数情况下都是有限的，无法无限制地增加样本量。因此，研究人员需要在需求较大样本和可用资源之间进行平衡。

同时，样本数量也不应过多，以免在所有变量之间发现关系。如果样本过大导致统计效果过强，则几乎可以确定会发现统计显著性。尽管保证假设得到支持似乎是一个好主意，但关系可能如此微小，以至于在实际上并不重要。因此，研究结果在实质上并不重要，这是下一节关于效应量的主题。读者、审稿人和编辑越来越善于发现当研究主要依靠过多的受试者来支持假设时。此外，尽量使用最少的受试者是更加符合伦理原则的做法，以减少潜在的伤害。

还有一点很重要，就是要了解在您的学科中可能会影响样本量的特定问题；例如，在广告学中，存在一种被称为“广告失明”的现象，受试者在潜意识中忽视广告。广告研究人员可能需要多采样约30%的受试者，以弥补那些无法回忆起广告内容的受试者。在道德发展研究中，约有8%的受试者未能通过旨在检测伪装高分的问题，因此被清除出样本。

**效应量**

效应量也是与统计功效相关的一个概念，它是描述自变量（IV）与因变量（DV）之间关系强度的统计量，或者说解释或占据因变量变异程度的独立变量的量。效应量统计量是对p值所代表的统计显著性的补充。效应量统计量涉及关系的实际意义。换句话说，p值可以告诉我们治疗是否有效，而效应量可以告诉我们治疗有效的程度。"8.4效应量的更多信息"解释了不同的效应量统计量。

效应量可能与p值有关，也可能无关。例如，极显著的p值可能伴随着较低的效应量统计量，这表明治疗对结果的影响很小，几乎不可能是偶然的；最理想的情况是p值显著且效应量较大，这意味着治疗效果不仅是真实的，而且是重要的。治疗效果在实际意义上是有意义的。

**关于...方框8.4**

**效应量统计学**

不同的统计学方法代表着不同的效应量大小。效应量用于估计实验中由自变量解释或占据的变异程度。第一列包含了统计学方法，第二列包含了效应量统计以及对应的符号和解释。这些值的范围通常从0到1。

效应量是一个重要的统计指标，它可以让读者看到治疗效果在样本大小的干扰下的影响。而统计显著性在很大程度上受到样本大小的影响，效应量的影响较小。例如，一个样本量为500的研究可能具有很高的统计显著性，但效应量可能很小。反之亦然，一个样本量为50的研究可能没有统计显著性，但效应量可能很大。在研究中同时报告这两个统计指标可以让读者自行判断治疗的实质意义。在心理学中，报告效应量已成为标准做法，其他学科也认为这是良好的做法。许多期刊现在要求报告效应量。过去，一项未达到统计显著性的研究往往会被拒绝发表，但现在，即使一项研究达到了统计显著性，如果效应量非常低，也不能保证被接受发表。

报告效应量只需要在结果部分的显著性测试结果列表的末尾添加正确的统计指标。以下是一些例子，效应量用下划线标示：

• "使用条件（行动示范、不行动示范和对照组）作为自变量，负性情绪作为因变量，单因素方差分析（ANOVA）报告了一个显著效应，F（2, 192）= 9.24，p < .001，η² = .09。"

• "在三种议题条件下对候选人的‘形象’评估的影响结果呈现出显著发现（F = 6.92，df = 6, 450，p < .01，η² = .16）；对候选人的‘能力’评估也呈现出显著发现（F = 11.80，df = 6, 450，p < .01，η² = .25）。"

这些例子使用了η²作为效应量的指标，因为进行的是方差分析。下面是一个使用r²作为相关性测试的例子：

"H3得到了支持：实验网站上的故事受欢迎程度与公民参与之间存在显著正相关（r = .811，p < .001，r² = .66）。"

Cohen根据这个约定对效应量进行了分类，也被称为“T恤尺寸效应”，并在大多数社会科学学科中使用。虽然Cohen提供了一个经验准则，但效应量的大小取决于上下文。在某些学科中，例如社会科学，一个0.2至0.35的效应量可能被认为是大效应。当你阅读自己领域的研究时，注意报告的效应量，以了解你的学科和特定变量的效应量。在任何实验的规划阶段，确定一个能指示治疗或操作的重要性的最小效应量是非常重要的。

样本大小、效应量、统计功效和显著水平之间是相互关联的；如果你知道其中的三个，第四个可以确定。如何通过统计功效分析来确定这一点，将在下一部分中介绍。

**统计功效分析**

确定样本大小是统计功效最常见的应用。在收集数据之前进行功效估计和分析被称为先验统计功效分析。还有一些经验准则方法可以确定研究所需的样本量，但免费的在线功效分析计算器几乎已经取代了这些方法。（请参见表8.1关于样本量的经验准则。）所有的统计功效分析都依赖于要进行的统计检验。例如，使用回归分析或结构方程模型的实验所需的样本量比使用t检验或方差分析的实验要多。到目前为止，在本书中，我们主要关注的是测试组间差异的研究，这些研究适合使用t检验等统计方法进行检验，当有两个以上的组别时当比较两个组时使用t检验，比较三个或更多组时使用方差分析（ANOVA），如果有多个因变量，则使用多变量方差分析（MANOVA），它是ANOVA的一种形式。进行这些测试需要另外学习一门统计学课程，本书并不旨在取代这些课程。但读者应该意识到所选择的统计检验会影响统计功效，从而影响所需的样本大小。

为了确定样本大小N，需要至少三个以下四个数据：

•显著性水平α—社会科学中的惯例是使用0.05或更低。与0.01相比，α水平为0.05时，测试的功效更高。

•功效—建议至少为80%，即0.8或更高。191

•效应大小d—Cohen的规则提供了一个经验准则。阅读现有研究，确定特定主题研究的典型效应大小。

•均值和标准差—显然，尚未进行的研究不会提供均值和标准差，因此可以从类似的先前研究、研究人员在该主题上进行的先前研究或试点研究中获取这些数据。元分析是找到多个不同研究的均值的另一个好方法。

以道德发展研究为例，使用了两组数据——一个处理组（关于儿童的故事）和一个对照组（关于成年人的故事），因此在G\*Power中指定了t检验。没有预测方向性假设，因此使用了双尾检验。（如果有一个预测方向性，例如，对于具有儿童的故事，道德判断水平较高的假设，那么将使用单尾检验。）指定了先验模式，因为功效分析是在进行研究之前进行的，以确定使用多少个受试者。插入了来自先前研究中的均值和标准差；处理组（照片）的均值和标准差为5.27（1.18），对照组（无照片）的均值和标准差为4.25（1.50）。指定了功效为0.8；研究人员可以在0.8到0.95之间变化以观察其对样本大小的影响。α错误概率应为0.05，这是社会科学中的标准。当输入这些数据后，G\*Power命令将计算结果。在这种情况下，G\*Power指示需要总共94个受试者，每组37个。由于使用了均值和标准差，它计算出的效应大小为0.75，根据Cohen的规则，这被认为是大效应。如果不知道或无法估计均值和标准差，可以输入所需的效应大小。改变效应大小可以显著增加样本大小；如果对于研究主题来说常见的话，则指定较小的效应大小。

回顾一下，计算第四个数据需要三个信息中的三个。已确定α水平（0.05）和功效（至少0.8）；通过检查文献确定效应大小或均值/标准差。使用G\*Power分析进行实验，看看需要多少个受试者才能获得较大的效应大小。进行功效分析可以帮助您决定是否需要转向更廉价的受试者形式，例如MTurk，或者是否负担得起从普通成年人群体中获取样本。

这个G\*Power分析的示例相对简单；实际上还有更多内容。例如，G\*Power允许使用F检验，包括单因素方差分析、主效应和一种单因素交互作用，以及组间或组内的方差分析等等。这需要更高级的统计学理解，许多来源提供更深入的G\*Power教程。一些教程可以在“建议阅读”部分中找到。

功效分析的报告可以从一个句子到包含所有细节的附录不等。以下是一些例子：

•“基于Symons和Johnson在大规模元分析中报告的自我参照编码与语义编码的效应大小（d = 0.65），进行了先验功效分析。功效分析表明，在功效为95%、α为0.05的条件下，样本容量为33足以检测到自我参照编码的效应。”193

•“最终样本包括20名参与者（平均年龄为19.70岁，15名女性）的数据进行了分析，样本容量的选择基于先验功效分析（详见补充材料）。 ”194

•“进行了先前G\*Power 3的功效分析，结果显示，大约60名参与者的总体样本容量可以实现可接受的功效（即1-β > 0.80），用于测试中等效应大小。”195

**观察功效**

最后，有一种被称为观察功效、事后功效或后验功效的概念，一些期刊要求研究人员在统计检验结果不显著时包含这个概念。这是指在研究完成后进行的功效测试，旨在了解为什么无法拒绝零假设的原因是否归因于功效不足。这是一个有争议且复杂的话题，在这里不会详细讨论。简单来说，不显著的统计检验总是具有低功效。事实上，功效与p值之间存在一一对应关系。许多研究不恰当地报告低功效作为不支持假设的可能解释。尤其是在报告SPSS或G\*Power提供的观察功效值时。相反，事后功效应该基于独立感兴趣的总体效应大小，或者结果应该使用p值、效应大小和置信区间进行描述。由于这超出了本书的范围，建议读者参考统计学的书籍和课程。

下一章将讨论在实验研究中受试者将接收到的刺激以及用于测量其反应的仪器或观察手段.

**常见错误**

• 使用学生样本而不考虑它们与感兴趣人群的差异

• 没有提供使用任何类型受试者的合理理由

• 对研究中的受试者进行推广

• 在执行研究之前没有进行功效分析来确定所需的受试者数量

• 没有报告每个组的效应大小、功效、显著性水平和受试者数量

**考验你的知识**

1. 大多数实验使用哪种类型的样本？

a. 随机样本

b. 代表性样本

c. 概率样本

d. 便利样本

2. 以下哪个不是使用学生作为实验对象时的关注点？

a. 高于平均认知能力

b. 同质性

c. 对权威的服从

d. 自我意识较弱

3. 使用亚马逊的机械土耳其作为实验对象的以下哪个是不利之处？

a. 费用昂贵

b. 数据收集较慢

c. 实验对象不注意问题

d. 实验对象缺乏多样性

4. 第二类错误是什么？

a. 在应该拒绝零假设时未拒绝零假设

b. 在应该拒绝零假设时拒绝零假设

c. 在不应该拒绝零假设时拒绝零假设

d. 以上皆非

5. 功效（Power）与下列哪个相关？

a. 样本大小

b. 效应大小

c. 显著水平

d. 以上皆是

6. 下列哪个与大样本相关的问题存在？

a. 增加了研究所需的成本和时间

b. 检测所有变量之间的关系

c. 发现在实际上并不重要的显著关系

d. 以上皆是

7. 效应大小是指什么？

a. 解释或涵盖因变量方差的程度，由自变量解释或占据

b. 与显著水平相同

c. 受样本大小的很大影响

d. 不应在研究中报告的内容

8. 在社会科学中，效应大小为0.35被认为是：

a. 非常小

b. 小

c. 中等

d. 大

e. 非常大

9. 在研究进行之前使用功效分析估计样本大小被称为：

a. 后验功效

b. 先验功效

c. 后验分析

d. 观察功效

e. 功效滥用

10. 功效不依赖于将要进行的统计检验

a. 真

b. 假

**Answers**

1. d

2. b

3. c

4. a

5. d

6. d

7. a

8. c

9. b

10. b

**应用练习**

1. 使用与您正在进行的研究类似的研究数据进行功效分析。使用G\*Power或类似的功效计算器。您可以在http://www.gpower.hhu.de/找到G\*Power的免费下载、手册和教程。

2. 撰写您正在创建的实验的方法部分的抽样部分。遵循本章和与您的研究相似的其他研究的示例。回答以下问题：您将使用哪些参与者，他们与目标人群的关系如何，您需要多少参与者，您将如何招募他们，您将使用什么激励措施，以及其他与您的研究相关的话题。将您的功效分析纳入其中。为您的样本使用提供合理的理由。撰写1到2页内容。

**推荐阅读**

关于学生受试者：

Basil, Michael D. 1996. “The Use of Student Samples in Communication Research.” Journal of Broadcasting and Electronic Media 40 (3) (Summer): 431–440.

关于效应量：

Levine, T. R., and C. R. Hullett. 2002. “Eta Squared, Partial Eta Squared and the Misreporting of Effect Size in Communication Research.” Human Communication Research 28: 612–625.

关于G\*Power：

Mayr, S., E. Erdfelder, A. Buchner, and F. Faul. 2007. “A Short Tutorial of Gpower.” Tutorials in Quantitative Methods for Psychology 3 (2): 51–59.

关于事后功效：

O’Keefe, Daniel J. 2007. “Post Hoc Power, Observed Power, A Priori Power, Retrospective Power, Prospective Power, Achieved Power: Sorting Out Appropriate Uses of Statistical Power Analyses.” Communication Methods and Measures 1 (4): 291–299.

**第九篇 刺激和操作检查**

*寻找一个大的物体而不是一个小的。*

*- R. Barker Bausell*

**学习目标**

**• 解释刺激是如何代表理论兴趣的类别。**

**• 创建逼真的刺激。**

**• 判断在刺激中需要控制的变量与需要在统计上进行控制的变量。**

**• 设计最大化比较并利用信息差异的刺激。**

**• 定义固定因素和随机因素。**

**• 撰写一份详细说明实验刺激和操作检查的报告。**

任何实验的核心是刺激（stimulus）。这也是最有趣的部分。实验者很少对刺激这个术语进行定义，而是依赖于字典对它的定义，即作为引起反应或反应的东西。例如，弗洛伊德称刺激为一种激励力量，与字典的定义相呼应。最常见的使用情况可能是在经济领域，讨论“刺激计划”的话题也是如此。

**研究重点9.1**

**创建逼真的刺激**

在2013年进行的Gina Masullo Chen的研究中，作者花了很多精力设计逼真的刺激和进行操作检查，为大学生创建了一个虚构但可信的社交网络网站。她首先召集了七名大学生组成焦点小组，并在他们集思广益时采用了“大声思考”的方法（更多详情请参阅第10章），让他们为网站的主题（例如“裤袜不是裤子”，“他们怎么能指望我在早上8点钟学习，当我还喝醉了”）、相关问题（例如“我iPod上的前5首歌曲是……”，“周末，你最可能在哪里找到我……”）以及加入此类网站时可能收到的消息类型进行头脑风暴。

她使用了一个名为Ning的可定制在线平台来创建虚构的社交网络网站，名为“The College Network”。参与者被告知他们正在测试一个专门为大学生设计的新社交媒体网站的测试版本。

为了开发这些实验刺激，陈采用了Graesser（1981年）提出的一个方法。有关该方法的更多信息可以在本章末尾的“建议阅读”部分找到。这样做的目的是确保该研究中使用的刺激对于大学生这个目标人群是真实和相关的，这对于在传播和社交媒体领域进行有意义的研究至关重要。

当经济不景气时，刺激很普遍。当政治家和经济学家谈论刺激时，他们指的是能够增加就业和鼓励消费的东西。他们希望做一些能够使经济产生积极反应的事情。在社会科学实验中，术语刺激指的是实验者用来传递研究假设会引起某种效应的操作的工具。正如开篇引用中Bausell的建议所说，一个“大的物体”——在这种情况下，一个大或强烈的刺激——比一个小的更适合产生反应。

实验中的刺激（刺激的复数形式）可以采取任何形式，以最好地传递操作、处理或干预给被试对象。对于伊凡·巴甫洛夫来说，刺激是一声响亮的铃声。斯坦利·米尔格拉姆构建了一个“电击发生器”作为刺激，实际上只是一个带有开关和标签的盒子。菲利普·津巴多则创建了一个整个模拟监狱作为他的刺激。在社会科学的许多不同学科中，研究者通常将信息作为主要的刺激来源，例如在政治研究中使用竞选网站、在市场研究中使用广告、在心理学研究中使用电影片段来引发情绪。刺激通常是按照其物理特征来构思的，例如网站的功能或组织，这构成了研究兴趣的概念。另一种形式的刺激是指示——例如，告诉被试对象列出他们对移民等问题的担忧。刺激的形式可以从非常简单的文本，如一张纸或计算机屏幕上的文本，到复杂的虚构电视新闻节目或网站。许多刺激介于两者之间——海报、网页、新闻稿、推文、社交媒体帖子等等。在某些评估项目或干预措施有效性的学科中，刺激就是这些项目或干预措施，比如计算机辅助学习课程或为警察提供的培训研讨会，教他们如何更好地处理患有精神疾病的人。在这些情况下，刺激已经存在，研究者可能只需要为对照组创建其他内容。

创建高质量的刺激是成功实验的关键，因此本章将致力于提供实现这一目标的建议和建议。

**刺激示例**

最好的方法来说明社会科学实验中的刺激是通过实例：

• 在政治学研究中，旨在确定哪些政治家出现在新闻中以及原因的研究中，刺激是关于议员的虚构新闻稿。

• 在一项市场营销实验中，旨在了解哪种信息最有效地促使乌干达的人们改用节能灶而不是使用火炉，实验者设计了宣传新灶具的海报。10

• 在一项公关危机研究中，通过虚构的新闻发布会广播，操作了发言人的声调。11

• 在一项教育研究中，学生们通过计算机而不是老师上课；计算机辅助课程是刺激。12

在所有这些例子中，刺激都被设计为代表研究的理论兴趣类别。它们是实验者操纵因素水平的工具，这些因素被假设会引起某种效应或结果。具体来说：

• 在政治学研究中，新闻稿是为了代表使政治家值得报道的特征而撰写的，比如他们所属的政党和他们专业化的问题。13 这些代表了在刺激新闻稿中操作或改变的因素水平。

• 在用于炉具的市场实验中，海报是代表因素水平（如健康利益和节省时间和金钱）的刺激。14

• 在公关危机研究中，通过广播一个学生扮演一个组织的发言人录制的声明，然后对其进行操作，创建一个高音调版本和一个低音调版本；音调代表了研究的因素水平。15

• 在教育实验中，计算机辅助课程是操纵语言学习理论中概念的刺激，通过改变视觉和听觉信息的呈现方式，以及不同的问答策略，来测试该理论。16

在新闻学中，故事通常被创作得像是出现在报纸或网站上。例如，在一项关于速度与准确性对读者对新闻组织可信度的影响的研究中，创造了看起来像是新闻组织博客的刺激，其中显示了不准确信息被纠正的故事随着更新而更新。17

在这项研究中，关于速度和准确性对读者对新闻组织可信度的影响，研究者创造了刺激，设计成看起来像新闻组织的博客。这些刺激在计算机屏幕上呈现给被试者，就像它们在实际新闻网站上出现的方式一样。

她模仿真实新闻组织的博客页面来创建刺激，以使被试者在观看时产生真实感。图9.2a和9.2b展示了快速条件下的一个带有错误的故事示例，以及慢速条件下没有错误的故事示例。在快速条件下，通过术语“CORRECTION”实现了操作。

创建逼真的海报、广播、广告和网站需要相当多的工作；然而，并不是所有的刺激都需要这么复杂。在一项关于企业社会责任的商业研究中，研究者通过描述企业是非营利组织还是盈利组织来操作其社会责任，参与者阅读的文本情境即为刺激。18 纸张或计算机屏幕上的文字是最简单的创建方式之一，许多实验都使用这些作为刺激。例如，在一项政治学研究中，被试者收到有关虚构候选人在问题上的立场信息，然后被要求进行投票，信息以计算机屏幕上的文字形式呈现，而不是被标榜为来自候选人的网站或新闻来源，因此无需模拟一个网站或报纸社论。19在另一个实验中，被试者被要求阅读一篇文章，激发其对恶心感或伤害担忧的情绪，然后对同性恋领养做出道德判断。20 这篇文章并未被呈现为来自杂志或报纸的社论，因此不需要使其看起来像社论。然而，如果实验需要某些通常通过电视广告或Twitter等媒介传达的信息，那么让被试者阅读描述电视广告或推文的文字远比创建一个虚构的电视广告更不真实。

这也是为什么对实验缺乏外部效度的批评如此之多的原因之一（详见第5章有关效度的内容）。（有关创建虚假社交媒体账号，如Twitter、Facebook等的在线工具，请参见方框9.2中的更多信息。）

有时，刺激本身的创建并不复杂，但实验的附属部分可能很复杂。例如，在一项经济学实验中，如果操作嵌入在书面指令中，被试者执行了一个任务，并被告知他们的捐款将要么给予组织，要么不给予组织，那么就不需要创建海报或广告等刺激。书面指令即为刺激。然而，研究者需要创建被试者将要执行的任务——将字母解码为数字。因此，他们使用软件创建了一个包含一个字母列和一个相应数字的相邻列的表格，并在计算机屏幕上向被试者展示。24

刺激并不总是虚构或由研究人员创建；一些研究使用真实的刺激。例如，在一项旨在研究真人秀节目是否能促进利他行为的研究中，研究者使用了真实的真人秀节目片段。25 这些电视节目成为刺激，被选定代表游戏式真人秀（《生存者》、《极速前进》）或有意义和鼓舞人心的真人秀（《秘密老板》和《超级保姆》）。在另一项研究中，测试互联网信息对政治效能影响的三个刺激是由明尼苏达公共广播电台创建的在线问卷，还有一篇来自电视新闻的报道。

**更多关于...盒子9.2**

**创建逼真的社交媒体刺激**

越来越多的信息通过社交媒体如Facebook和Twitter传递给人们，尤其是千禧一代，他们代表了大学生群体，往往被用作学术实验的对象。有在线工具可帮助研究人员创建逼真的社交媒体内容。例如，Twister、Tweeterino和Simitator都提供了创建虚构推文的工具，而Simitator还能创建虚构的Facebook帖子、聊天内容、SnapChat消息和iPhone短信。PrankMeNot是另一个用于创建推文、Facebook帖子和消息聊天的工具。Simitator网站宣传其“恶作剧你的朋友”的能力，但教育工作者也发现它在学校中作为教学工具非常有用，尤其对那些禁止学生访问互联网的学校，或者用于创建学生相关的教材。

有一位研究者使用Simitator来为实验创建逼真的推文。他首先随机抽样了《纽约时报》、《华尔街日报》和《华盛顿邮报》等实际发布的推文。然后，他使用该工具重新创建了这些推文，但让它们看起来像来自虚构记者的。他还模拟了记者一天中多次报道不同新闻的情况，以实现真实的Twitter跟踪体验。

更为雄心勃勃的项目是使用Ning创建一个完全可工作的社交媒体网站，尽管它没有互动功能。它允许创建者添加新闻报道、图片、标题，并操纵点赞和粉丝的数量。另一个允许自定义编辑网页的网站是CloneZone。当然还有其他工具，其中一些是免费的，但有些需要收费。

另一种适应实验的方法是使用Web浏览器中的“检查元素”工具。只需打开一个网页，然后右键单击，滚动到框的底部，选择“检查元素”。一个源代码面板会在底部打开。在上面的面板中点击并拖动以突出显示您希望更改的文本，然后再次右键单击并选择“检查元素”。文本将在底部面板中被突出显示。只需在旧字词上键入新字词，新的字词将自动出现在网页上。点击右侧的X关闭底部面板，然后截图并将其作为（非互动的）网页在实验中使用。在Mac上，您可以使用Command+Shift+4，然后使用准星指针编辑浏览器的顶部和底部。在PC上，使用Alt+PrtScn或剪切工具。更详细的教程可以在网上找到。

可以在以下网址找到这些工具：

Ning: https://www.ning.com/

CloneZone: clonezone.link

Prank Me Not: www.prankmenot.com

Twister:http://www.classtools.net/twister/

在上文提到的这个例子中，为了进行一项关于推特上记者的可信度受到暴露和性别影响的实验，

研究者使用了来自《纽约时报》的原始推文文本，通过虚构的记者刺激来创建逼真的推文。

相关链接：Boulter, Trent. 2017. “Following the Familiar: Effect of Exposure and Gender on Credibility

of Journalists on Twitter.”（《跟随熟悉的人：暴露和性别对推特记者可信度的影响》（学位论文，

德克萨斯大学奥斯汀分校）），https://repositories.lib.utexas.edu/bitstream/handle/2152/62635/BOULTER-DISSERTATION-2017.pdf?sequence=1&isAllowed=y。

Tweeterino: https://tweeterino.com/ Simitator: www.Simitator.com

20/20节目和MyBarackObama.com视频等实际新闻照片也可以作为传递处理的刺激。在教育研究中，软件公司创建了计算机辅助课程的刺激。在教育、社会工作和其他领域中，经常使用现有的刺激来测试干预措施或评估即将在现实中实施的项目。

**创造刺激的建议**

**保持真实**

当刺激是虚构的时，研究人员应该努力使它们尽可能逼真。一种方法是从实际的新闻文章、广告、海报、学习指南、网站等开始，然后重新编写或编辑它们以反映处理的变化。例如，在一项研究中，研究人员使用真实的报道，然后重写它们，将某一种框架的所有实例替换为另一种正在研究的框架。研究人员还可以使用实际的刺激，如学习指南、电影或电视广播，并在其中编辑不同的文字或视频以反映处理和对照条件。

以下是一些作为刺激在实验中使用的广告示例。研究人员收集了实际的广告，以使刺激尽可能真实，并然后修改它们以满足处理和对照条件的要求。

（见图9.4中使用这种方式创建的广告示例）

实现更逼真的刺激的另一种方法是让专业人士（或曾经的专业人士，通常是研究生）撰写新闻稿、创建网站、修改学习指南等。研究人员可能认为他们可以创建一个逼真的报纸剪报，但最终成品的标题可能是居中的并且全大写字母，而真实的报纸往往不会这样做。在创造一个专业外观的产品时，有很多微妙的东西，非专业人员并不总是意识到。在一项研究中，两位研究人员承认这一点，他们创建了新闻文章，写道：“我们努力调整新闻文章的常见布局和日常荷兰新闻报道的新闻风格。”在某种程度上，受试者可能会意识到业余的努力，而不会做出反应。

同样的方式，就像他们相信刺激是真实的一样。这并不总是需要大量资金，因为通常可以招募具有专业知识的学生来帮助完成这些任务。

**尽可能进行控制**

在创建刺激时，利用文献和常识来认识需要控制的内容。在实验中，通常使用虚构的事件、人物和地点，以避免受试者对真实人物、地点或事件的先验知识造成干扰。因此，在声调研究中，研究人员使用了一个虚构的组织和虚构的健康危机。为了防止受试者怀疑它们是虚假的，研究人员将事件设定在国家的不同地区，因为受试者可能会认为它是真实的，只是因为不在他们所在地理区域内，所以他们还没听说过。这代表了一种控制形式，称为实验控制，即刺激本身被设计成不受混淆的状态。另一个例子是在一项研究中，对新闻中的政治家进行了精心选择，以避免选择与特定政党明显相关的政治问题。避开任何对共和党或民主党有利的问题，避免受试者根据政党而不是被操作的事物来评价虚构的候选人。这是一种实验控制，因为刺激中的某些东西被控制着。

在其他情况下，使用虚构的刺激可能会损害研究结果。一组研究人员认为在研究什么使政治人物有新闻价值时可能会出现这种情况，因此使用了真实的荷兰议会成员作为刺激。为了控制受试者对真实政治人物的先前了解，研究询问受试者认为每个议员多久会成功地出现在新闻中，然后将其用作分析中的协变量，从结果中消除先前知识的影响。这是一种被称为统计控制的控制形式；当不可能或不希望在刺激中控制某些因素时，研究人员会对其进行测量，并在统计分析中将其作为协变量。如果有任何可能从刺激中学到这些内容的可能性，请确保在给受试者进行处理之前收集这些信息。作为协变量使用的变量都应与结果变量有关；如果依赖变量是可信度评估，而信息来源的性别与可信度无关，则可能无需进行测量和控制。性别和种族是被认为经常与被操纵的变量相互作用并干扰结果的变量，然而，这取决于所研究的内容。请参考相关文献。

总之，有两种方法可以控制混杂因素：实验控制（在操作或刺激中进行控制）和统计控制（通过测量并将其作为协变量包含在分析中）。例如，如果种族在评估警官与市民之间的互动时很重要，则使视频中一半警官是白人，一半是黑人，这是实验控制。如果受试者的种族很重要，则在分析中测量并包含受试者的种族，这是统计控制。如果可能，应始终优先选择实验控制。

**只变化被研究的事物**

在所有控制和处理版本中，除被操纵的因素外，其他所有因素都应保持相同。在一项关于声音音调的研究中，这意味着保持所有不同录音的语速、强调某些单词、停顿、重音、发音和措辞相同。另一项研究使用了八个新闻稿，所有这些稿件都是相同长度，都有相同类型的引用，并且都是负面框架。这通常会被描述为某些内容"被保持恒定"；例如，在政治研究中，负面框架被描述为： "我们保持了......"。

要控制新闻价值的冲突等因素不变...，所有议员都批评政府。”34 当一些不打算变化的因素发生变化时，该研究被称为被混淆。这在第一章中有解释过，可能是无心的，比如为处理组设计一个横向广告，为对照组设计一个纵向广告，或者将一张图片放在处理组的顶部，对照组的底部。在这些情况下，我们无法确定是横向还是纵向广告或图片放置位置实际上导致了任何影响。最好的方法是将所有广告设计为横向，并将所有图片放在顶部。

混淆不仅可能在设计刺激时出现，也可能在实验进行时出现。例如，使用两个不同的研究人员进行研究，或者让一些受试者在早上接受处理，而另一些在下午接受处理，也可能是混淆的一种形式。研究人员通常会注意将刺激的元素保持不变，但可能会忽略程序组成部分。任何无法避免的差异可以通过随机分配来补偿（参见第7章），例如通过将两个研究人员随机分配给处理组和对照组的受试者。但是最好的方法是完全避免混淆。Bausell建议对研究设计的每个元素进行“微观检查”，以发现任何可能的差异，无论是在刺激还是在程序方面。这涉及对实验的每个方面进行仔细审查，并可能进行一项称为预实验或pilot study的前期研究。

与实验的大多数其他方面一样，常识和文献应该作为指导，确定什么可能与结果变量相关，因此应该受到控制。

**最大化比较**

在为实验创建刺激时，重要的是使操作或处理足够强以检测真正的效应。这意味着最大化处理组和对照组之间的比较，Bausell称之为寻找大对象而不是小对象。36 这是如此重要，甚至被称为实验的第一规则。38 弱操作是一个常见的问题，可能会产生重要的后果。例如，如果一项研究旨在确定报纸声誉对读者对可信度的评估的影响，研究人员最好使用声誉差异极大的新闻机构，如纽约时报和《全国询问者》。如果在声誉最高和最低的报纸之间没有发现效应，那么更微妙的差异很可能也不会产生效应。然而，如果在最高声誉和最低声誉的报纸之间发现了效应，研究人员可以继续探讨声誉差异较小的报纸之间的差异。研究照片对道德判断的影响39使用了获得国家级奖项的照片，这些照片被证明代表了新颖性概念的最高水平。如果研究没有使用最强烈的操作进行发现效果——获奖的新颖照片——那么在较小的操作中也不太可能发现效果，比如日常照片。获奖照片并不是人们在日常报纸中通常看到的。因为在获奖照片中发现了效应，所以逻辑上可以继续测试在当地日报中更常见的照片，最终测试最不新颖的照片类型，比如头像照片。

斯坦利·米尔格拉姆在他对服从权威的研究中采用了这种方法。40 在第一项研究中，他让实验者穿着白大褂在实验对象旁边的同一房间里继续下达电击命令。当他在这种近距离下发现效应时，他开始将实验者逐渐远离实验对象。最终，实验者完全离开了房间。随着权威人物与实验对象的距离越来越远，服从程度下降，实验对象拒绝执行电击治疗。41

使处理组和对照组接收的处理尽可能不同是为了确保获取实际存在的积极结果的一个重要步骤。42 例如，如果研究要求在故事中进行框架操作，那么可以使半数或更多故事包含所研究的框架。这比真实故事中包含的要多得多，但如果在故事的大部分时间都采用某种方式进行框架，却没有发现效应，那么在故事的更少部分采用该方式进行框架，也不太可能产生效应。许多研究人员低估了产生结果所需的操作强度，因此，看似过分的操作实际上可能对受试者没有任何影响。

**利用消息变异**

到目前为止，本章主要关注如何根据处理或操作来创建刺激的差异——即对研究中感兴趣的独立变量的理论水平进行处理方差的部分。这被称为处理变异。44 但在创建刺激时还需要考虑另一种变异，即消息变异。消息变异指的是用于表示每个处理水平的消息或刺激的数量。45 例如，在研究人们的种族如何影响态度和判断时，通常使用照片作为操纵独立变量种族的刺激。与其使用一张黑人的照片和一张白人的照片，使用三张每种族的照片是更好的策略。这将导致一种更具有统计能力的设计，并且有助于检测到更细微的效应。例如，如果人们对某种独立变量的影响有着不同的观点，使用多种消息或刺激可以更好地捕捉这种多样性。同时，多样性的消息也可以提高实验的外部效度，因为它更好地模拟了现实世界中的情况。

本章主要关注如何根据处理或操作来创建刺激的差异——即对研究中感兴趣的理论独立变量的不同水平进行处理方差的部分，这被称为处理变异。然而，在创建刺激时还需要考虑另一种变异，即消息变异。消息变异是指用于表示每个处理水平的消息或刺激的数量。例如，在研究人们的种族如何影响态度和判断时，通常使用照片作为操纵独立变量种族的刺激。与其使用一张黑人的照片和一张白人的照片，使用三张每种族的照片是更好的策略。这将导致一种更具有统计能力的设计，并且有助于检测到更细微的效应。同时，多样性的消息也可以提高实验的外部效度，因为它更好地模拟了现实世界中的情况。

多信息设计是因为在每个处理水平上使用了一个以上的“信息”——在本例中是照片。如果每个种族只使用一张照片代表，那就是单信息设计。当一个处理或因素的每个水平都有多个信息或刺激时，也可以称之为重复因素。使用多个刺激来代表一个因素非常重要，因为研究人员很少关注特定的刺激或信息，比如一个特定的人物、电视节目或广告，而是实际上对该人物、节目或广告所代表的类别或水平感兴趣。例如，在对现实电视节目的研究中，研究人员对两种类型的节目感兴趣——游戏风格和鼓舞人心的节目。他们在多信息设计中使用了两个节目在每个水平上：两集《幸存者》和《极速前进》代表游戏风格，两集《超级保姆》和《老板之下》代表鼓舞人心的节目。这是重要的，因为刺激或信息的效果会有所不同，无论研究人员多么小心地试图使它们相似。刺激的多个版本有助于确保效果不是单个刺激的个体属性造成的。

与此相反，单信息设计在让研究人员推广到一个特定的信息或刺激方面更为薄弱。可靠的推广需要多个刺激。如果在现实电视节目的研究中，研究人员只使用一个节目来代表每个因素水平，那么他们只能说任何效果是由《幸存者》或《老板之下》的一个特定集所引起的，而不是这两个节目所代表的更广泛的类别。也就是说，他们将无法说某种类型的节目产生了效果，只能说用于实验的那个具体节目产生了效果。知道《幸存者》的一个特定集引起了某种结果通常没有什么用，而且无论如何，可能有未知的因素导致了这个结果，从而使结果混淆不清。例如，其中一集可能会让被试感到生气或受到冒犯，与商业广告的过渡更顺畅，或者在语气上更与广告相符。

使用一个刺激是不可能解释所有混淆的。通过使用多个刺激来代表每个因素的水平，减少了任何除了感兴趣的自变量对结果负责的可能性。作为对为什么使用多个刺激或信息很重要的例证，一个关于如何呈现健康故事的研究使用了代表四种不同健康问题的信息，并对其中三种做了类似的效果。然而，其中一个问题——吸烟——没有产生相同的结果。关于这个特定问题或故事的特殊性导致人们对其产生不同的反应。如果研究只使用了这个故事，就永远不会知道这些框架实际上在其他情况下有效。

**需要创建多少个刺激？**

需要创建多少个刺激？显然，创建多个刺激对研究人员来说更费力。但究竟需要多少个信息或刺激来进行实验？答案是“取决于情况”。有些研究，比如使用反应时来测量，即被试的反应时间，如果不使用大量不同的刺激，结果可能不可靠。Thorson等人建议使用尽可能多的刺激，但要确保被试不会因为疲劳或厌倦而注意力不集中。在一项关于电视公益广告(PSAs)的研究中，研究人员对每个因素水平使用了六个刺激。一个好主意是，如果刺激不需要被试花费太长时间来处理，比如一个简短的故事或照片，可以从四个或更多个刺激开始。对于处理时间较长的刺激，可以从三个刺激开始。之所以建议“从...开始”，是因为有时不同的刺激并不总能充分代表研究人员试图创建的自变量的概念或水平。这是本章下一节的主题。当然，永远不可能在每一个可能的类别上使多个刺激完全相同，因此在测试处理水平之间的差异时，信息变异被视为随机误差。

尽管创建更多的刺激需要更多的工作，但从长远来看，这是值得的，因为使用单信息设计的实验最终将需要再次进行，并需要使用不同的刺激来进行泛化分析。

**固定因素与随机因素**

当使用多个刺激来代表因素的单个水平时，处理各种响应的标准方法是将每个个体被试的响应值平均成代表每个水平的所有刺激的响应值。因此，例如，如果一个被试看到了三张黑人的照片和三张白人的照片，并被要求做出伦理决策，那么被试对三张黑人照片的伦理得分将被平均，从而得到该水平的一个得分，对三张白人照片也做同样处理。然后使用传统的统计技术，比如方差分析(ANOVA)进行分析。这在因素的水平被认为是固定的因素而不是随机的因素时是适用的。固定因素是那些水平是有限的，或研究人员选择了水平的因素，而随机因素则是代表一个更大的总体中的样本，而不是该总体内所有可能的水平。最常见的随机因素使用是在实验的被试中，他们是来自更大人群的样本，并且不耗尽该人群内所有潜在的被试。

有时，刺激也应被视为随机因素。以一个研究的例子，其刺激是黑人和白人的照片为例，参与实验的个体是来自一个更大总体的样本，而不是该总体内所有可能的水平。在这种情况下，刺激应该被视为随机因素。在使用多个刺激时，需要对因素和刺激进行适当的统计分析，以确保实验结果的有效性和准确性。

照片显然不是世界上唯一的黑人和白人，而是从更大的总体中进行抽样。如果研究人员对这些特定的照片进行抽样，并将它们用于推广到所有这两类种族的人口，那么这些照片将被视为随机因素。在这样的研究中，参与研究的被试和刺激照片都被视为随机因素，应使用特殊的混合效应模型的统计技术。这允许将结果推广到刺激的总体，也可将结果推广到被试的总体。

使用随机效应模型还有助于避免第一类错误，如第8章所述，即“虚假阳性”，或者是当发现一个并不真实的效应时。这些特殊的统计技术在一些社会科学领域，甚至临床和社会心理学中并不常用。当因素有无限多个水平或远多于被测试的水平，并且因素的水平是抽样的时，建议使用随机效应模型。这允许对实验中特定水平之外的水平进行推广。

相比之下，固定因素具有有限的水平，比如只有特定剂量的药物，比如10毫克、20毫克和40毫克。如果只有两种干预措施存在，比如只有两种针对暴力倾向客户的社工培训计划，那就是一个固定因素。当涉及到多种不同测试的教育研究比较时（如在不存在其他测试的情况下比较两种不同测试），这也是一个固定因素的例子。固定因素的另一个常见用途是面板数据，用于纵向比较同一批人、公司、州、学校或其他实体的数据，或跨越一段较长的时间。研究人员也可以将研究中所选的因素水平视为固定因素。例如，当研究人员只对推广到实验中使用的特定刺激感兴趣时，通常将诸如黑人和白人的照片之类的刺激视为固定因素。这可能导致需要使用不同的刺激进行复制研究结果，以增加推广性。

关于使用固定与随机效应模型的适当讨论是复杂的，超出了本书的范围。有兴趣的读者可以在“推荐阅读”部分找到更多的资料。

**操纵检查**

一旦刺激被创建，通常建议进行操纵检查，以确定被试是否按照研究人员的意图来感知刺激。如果未能确保刺激在具体意义上是符合预期的，可能会导致实验失败，尤其是第二类错误，即不支持一个应该得到支持的假设。科学知识将不能得到推进，对理论的质疑可能会产生。

实际上，当发生这种情况时，研究人员更有可能正确地进行发表研究结果。此外，研究人员更不太可能发表得出无效结果的研究。已有建议将操纵检查添加到众多社会科学实验的要求清单中，有些学科如社会心理学要求在论文被接受发表之前进行操纵检查。然而，这是一个有争议的问题，有人认为操纵检查很少或者从未必要。

**操纵检查、预试验或试验前研究？**

操纵检查是确定刺激是否确实按研究人员意图操纵自变量的方法。术语“试验前研究”和“预试验”通常与操纵检查交替使用。通常当研究人员在实际研究之前进行操纵时，他们会称其为预试验。本书将这三者视为不同的程序。在本书中，“预试验”是实际研究的一部分，在处理或操纵之前给出，与事后测试相当或相同，以便进行前后比较。预试验与操纵检查不同，因为它们的目的不同，操纵检查旨在确定研究人员对潜在构念的操纵是否有效，不论何时进行。第11章详细讨论了试验前研究，简而言之，试验前研究是一个更大的项目，旨在对整个实验进行“测试运行”。

操纵检查旨在帮助确保刺激是否符合研究人员的意图，以便最大程度地影响理论上相关的因变量。例如，在一项旨在确定生动写作是否对道德判断产生影响的研究中，操纵检查被设计为确定研究人员编写的三个故事是否“生动”。生动在理论上被定义为帮助人们形成心理图像的具体、描述性写作。操纵检查的目的是确定普通人是否发现生动的故事比传统新闻风格的对照故事更具体、更具描述性，更有助于形成心理图像。为了实现这一目的，要求实验外的受试者回答有关故事生动性的特定问题，而不是对整个研究进行参与。一些问题包括：“这个故事非常描述性”，“这个故事非常生动”，“这个故事非常具体”，“这个故事有很多细节”等，旨在测量文献所定义的生动性理论构念的特征。在这个操纵检查中，有一个故事与非生动故事没有显著差异，因此在用于实际实验的受试者之前，它被重写并再次检查。在这种情况下，操纵检查的结果实际上指示了刺激需要进行修改和进一步确认才能在实验中使用。

操纵检查可以发现问题，让研究人员及时修正，以免为时已晚。操纵检查不是对刺激对因变量产生的效应的测量，而是对刺激本身的测量。

**直接操纵与间接操纵**

在生动写作的研究中，生动性这一构念只能通过间接方式进行操纵，例如通过对人物和场景的描述。对于这种潜在的自变量，需要进行操纵检查，以确定研究人员是否成功创建了生动的故事。但并非所有操纵的变量都需要进行操纵检查；例如，在一项研究中，操纵是故事中的信息来源数量，如果一个组的故事包含五个信息来源，而另一个组只有一个来源，那么就不需要进行操纵检查。在这种情况下，操作化与构念是相同的，无需进行操纵检查来查看某些故事有五个来源，而其他故事只有一个来源，这是不容置疑的。类似地，如果姓名和外貌等信息不含糊，那么性别可能不需要进行操纵检查。例如，在一项研究中，研究人员在虚构的政治候选人简介中使用了“约翰”和“莎拉”这样的名字。政治派别的标签也不需要操纵检查，只要明确标示即可。刺激的具体属性或内在结构属性并不总是需要进行操纵检查。例如，如果两个故事一个是500字，另一个是100字，就不需要询问受试者它们的长度。不管受试者怎么说，这两个故事显然是不同长度的。

**中介变量与心理状态**

操纵检查应该针对刺激所引起的心理状态，而不是刺激本身的不容置疑的结构特征。例如，在生动性研究中，除了确保故事内容是色彩丰富、描述详细的，操纵检查还要确保这些故事在心理上产生了期望的效果，即帮助受试者形成心理图像，这是一个称为“形象化”的理论构念。操纵检查还包括与此相关的问题，例如：“我从这个故事中形成了心理图像”，“很容易想象这个场景是什么样的”，“我可以轻松地想象出这个场景”。

这些心理状态代表了受试者对故事是否生动的感知，这被理论化为诱导了心理状态的形象化或在脑中绘制图像。心理状态代表了第1章中提到的因果解释。在许多情况下，研究人员在操纵检查中使用了心理状态，但在实际研究中未对其作为潜在的中介变量进行分析。当出现这种情况时，研究人员错失了解释效应的机会。在研究中包括潜在中介变量使得研究人员能够解释为什么会产生某种效应（参见“更多关于中介变量和调节变量的信息”框）。例如，在一项研究中，研究人员考察了网络搜索中使用的词语数量对用户对控制或掌控感（称为支配感）的感知的影响。支配感是因变量（DV），搜索词是自变量（IV），有两个水平，高和低，通过三个和一个搜索词或短语来操作。搜索词或短语是刺激的物理属性。预期搜索词数量在受试者中激活的心理状态是感知相关性；三个搜索词预期被受试者感知为高度相关的结果，而一个搜索词则预期返回受试者感知为相关性较低的结果。研究人员进行了操纵检查，以确保三个搜索词导致受试者感知结果为高度相关，而一个搜索词导致结果被感知为低相关。在实际研究中，研究人员还测量了受试者对感知相关性的感知，以便将其作为中介变量进行分析；即，假设感知相关性是自变量（搜索词数量）与因变量（支配感）之间的中介关系的因果机制。在这种情况下，提出了三个假设：一个是关于自变量（搜索词数量）对因变量（支配感）的直接影响；一个是关于中介变量（感知相关性）对因变量（支配感）的直接影响；还有一个假设是自变量（通过中介变量）对因变量（支配感）的间接影响。进行操纵检查的一个理由是，如果没有进行操纵检查，科学界将丧失宝贵的信息。在一项未支持假设的研究中没有进行操纵检查，研究人员无法知道是否需要修订理论，或者是否是处理失败了。假设甚至可能被抛弃。正如Mutz和Pemantle所说：“忽略操纵检查会阻碍科学知识的增长。”94

**内容延展9.3**

中介变量与调节变量

Chapter 1强调了在实验中包含因果机制的重要性，即解释为什么实验发现的效果会发生。在那里给出的例子是一个实验，用于确定照片是否能提高道德判断。在该研究中，照片确实改善了道德判断，这是效果。所考察的因果机制是涉入和详细阐述，即对故事感兴趣或被吸引，并深入思考其中的人物。这些因果机制被发现是中介变量，即它们是心理状态，作为自变量影响因变量的途径。有关何时会发生中介和何时会发生调节的情况经常会引起混淆。本补充内容旨在帮助澄清这一点。

中介变量

中介变量是解释观察到的效果如何以及为什么发生的心理状态或条件。它们代表了一些效果发生的机制。在詹姆斯·林德在第2章描述的坏血病实验中，他发现类似于酸橙和柠檬的水果能治愈坏血病。在该研究中，他没有研究中介变量或机制，但后来发现维生素C是酸橙和坏血病之间的中介变量。在之前的例子中，中央路径处理或对照片进行深思熟虑是导致更好的道德判断的认知过程或心理条件。恐惧可能是在看到公益广告后导致戒烟意图的心理状态。除了代表心理状态外，中介变量是在给予处理或操纵后发生的。继续之前的例子，恐惧是在观看公益广告后产生的，而详细阐述是由看到照片引起的。最后，中介变量必须与自变量相关。因此，公益广告本身传达了可怕的信息，与受试者的恐惧感相关，而照片则具有新奇和不寻常的元素，与更高的认知详细阐述相关。处理可能有许多中介变量或因果机制，但研究人员只能了解他们测试的那些。

调节变量

与中介变量不同，调节变量解释了效果在什么情况下可能会更大或更小，或者在何时和对谁会发现某些效果。调节变量是第三个变量，当与自变量结合时，会对与因变量的关系强度产生影响。调节变量的例子包括年龄较大减少找工作的可能性；上幼儿园调节母亲年龄对儿童认知发展的影响；个性特质，如过度男性化，会在观看暴力内容后增加攻击性。

A screenshot of a computer

Description automatically generated

A screenshot of a computer

Description automatically generated

A screenshot of a computer

Description automatically generated

**在调查实验中进行干预检验**

使用调查实验的趋势也支持了进行干预检验的必要性。在调查实验中，受试者可能在研究者无法观察到的远程地点进行实验，因此无法确定受试者是否真正接触到了刺激。在实验过程中嵌入干预检验可以帮助研究者识别那些注意到刺激的受试者和那些没有注意到的受试者。

随机勾选答案的问题。筛选问题不能代替调查实验中的干预检验，因为筛选问题是用来测试受试者是否仅仅接触了刺激，而不是刺激是否有效。

进行干预检验的另一个原因是确定刺激在意图不变的因素上是否相似，并且可能混淆结果。例如，在设计多个广告实验时，可能会使用不同的模特，他们在外貌和吸引力上应该是相似的；故事应该在趣味或喜好上相似；政治广告应在情感上相似，无论是负面还是正面的。干预检验可以证实这些相似性，以确保混淆因素不会导致结果错误。

许多实验并未报告干预检验的结果，这使读者认为并未进行任何检验。有些文章可能会说进行了干预检验，并且取得了效果，但没有提供支持性的证据。如今，越来越多的文章在附录或在线补充材料中提供关于干预检验的详细信息。如果刺激或干预在之前的研究中已经使用过，可能就不需要再次进行干预检验，但这一点应该在文章中提及。

关于干预检验，有三个实际问题需要考虑：何时进行、使用多少受试者以及如何测量。

关于何时进行干预检验，有两种选择：在实际研究之前独立进行，或将其嵌入到实际研究中，并在实验结束时进行评估。事实上，还有第三种选择，即在实际研究之前和之后都进行干预检验。这可能是首选的选择，下面将对此进行解释。

在实际研究之前进行干预检验是设计成功刺激的有益方法。这样可以帮助研究者在对实际受试者进行测试之前发现问题并调整刺激。例如，有两位研究者在政治研究中就这样做，他们使用学生进行干预检验，以验证刺激的有效性，但在实际研究中没有将其应用于真正的政治人物，以避免过长的实验过程。对于此类在刺激开发期间的干预检验，需要的受试者较实际研究较少，但仍足够能够在处理组和对照组刺激之间发现显著差异。例如，在关于生动写作的研究中，进行了26名受试者的干预检验。但是，这可能是最少的数量，少于20名受试者可能会使进行显著性检验变得困难。

另一种社会科学研究人员常用的报告干预检验的方法是将其嵌入到实际研究中。这样做的优势在于可以确保实际实验受试者感知到刺激的干预效果与预期一致。在实际研究中，应该在给出依赖变量测量后，将干预检验的评估放在研究末尾进行，以免告知受试者关于干预效果或研究目的。这还有一个好处，即在实际研究中测量预期导致结果的心理状态，使研究者能够测试中介变量。

第三种进行干预检验的方法是将这两种方法结合起来：首先对不参与实验的另一组受试者进行干预检验，然后再在实际研究中进行干预检验，以确保实际实验受试者也接受了它。这样做的好处有三点：这样可以给研究者时间来修改刺激，确保实际实验受试者感知到刺激的预期效果，并将中介变量纳入研究中。唯一的代价是这样会延长研究的时间，因为受试者需要回答干预检验问题。为了解决这个问题，可以根据内部一致性分析（例如Cronbach's alpha或因子分析）来减少实际研究中的干预检验问题数量，从而从中选择与实际研究相关性最强的问题。先前描述的生动写作研究在第一个干预检验中使用了九个评估生动性的项目，并将问题数量减少到实际研究中最相关的三个问题。许多研究人员的经验表明，刺激在干预检验中失败并需要修改，甚至在实验前的干预检验，这些都在幕后进行，并不会在期刊文章中报告，因为文章的字数限制。读者不必失望；研究可能会有些混乱，而印刷出来的内容不一定是全部的故事。

在测量干预检验时，不能像在实际研究中那样问所有相同的问题；干预检验的目标是确定操作化是否反映了理论构建，因此评估应该集中在这一点上，而不是测量最终的依赖变量。同样，常识和文献应该指导编写干预检验问题。如果其他研究人员已经对该研究中的构建进行了干预检验，那么在尝试创建自己的评估之前，这是一个自然的起点。但是如果找不到其他的干预检验，可以根据文献为基础创建一个新的。

如果在实际研究中嵌入的干预检验失败——即受试者没有按预期感知它——那么如果无法修改并重新运行，则应该将该刺激从分析中剔除。

**报告刺激和干预检验**

在实验写作的论文中，方法部分的一部分通常被标记为"刺激"、"刺激材料"或"程序"，用于呈现与刺激相关的信息和任何可能存在的问题。这部分应该出现在方法部分的早期，并对刺激进行足够详细的描述，让读者能够想象它们。还应解决任何可能引起关注的问题，比如虚构的人物、组织等。

对于刺激细节不太复杂的实验，应该在论文正文中报告它们，包括所有信息的确切措辞和任何视觉图像的描述。例如，在关于炊具的市场实验中，海报的措辞很简短，因此全部都包含在文本中，其中一部分如下所示：

“炊烟是有毒的。它让你感觉头晕目眩，让你咳嗽，并且会因为烟雾引起眼睛或喉咙疼痛。炊烟导致严重的疾病，包括肺炎和支气管炎。乌干达的儿童因炊烟导致的这些疾病死亡的数量和因疟疾死亡的数量一样多。”103

文本还描述了海报上的图片是一个婴儿在吸烟。

对于有许多刺激或需要更复杂描述的实验，可以适当地给出一个刺激的全面描述作为示例，并引导读者参考附录或网站了解其他刺激。104 同时，最好使读者能够清楚地看到干预效果，可以用加粗或斜体标记变化的部分，而常规字体用于保持不变的部分。例如，在关于进入新闻的议员的研究中，其中一个二十四份新闻稿是这样描述的：

“绿党议员瓦特·德夫林特回应了有关Kleine Brogel核武器现代化的最新消息，他表示：‘B61核弹不需要现代化，而是应该被销毁。核武器是危险和无用的。此外，核弹现代化、储存、维护和监视非常昂贵。政府需要采取措施，开启并完成全面的核裁军。’ 德夫林特希望通过向授权内阁成员提交书面问题，获得关于比利时核武器措施的明确答复。”105

最后，对于非常复杂的刺激或刺激数量较多的情况，文本应简要描述刺激，然后引导读者参考附录或补充性网站，从中阅读和查看所有刺激内容。例如，在企业社会责任（CSR）的研究中，刺激部分只给出了场景的一般描述以及它们代表的维度，比如公司是否对环境有害，是否是一个良好的商业伙伴，以及对待员工的方式。106 由于有六个篇幅较长的文本场景，将其放在附录中比放在正文中更加高效。正文中的描述足够详细，大多数读者不会有太多问题。以下是文本的内容：

“在这个任务中，我们创建了六个不同的场景，并将它们分发给参与者。在每个场景中，参与者得到一个虚构组织运营大学食堂的信息。根据Bardsley（2000）的做法，参与者没有被告知调查对象是虚构的，因为这不构成对参与者的欺骗，而且这些场景可能是真实的。在场景1、3和5中，该组织被描述为“盈利组织”，而在场景2、4和6中被描述为“非盈利组织”。这些场景还在企业的CSR绩效方面存在差异。在场景2、3中没有提供关于CSR绩效的信息。场景3和4提供了关于组织积极CSR绩效的信息，而场景5和6提供了有关负面CSR绩效的信息。CSR绩效是通过提供有关组织行为的信息来操作的，这些行为涵盖了CSR的三个基本维度：环境（环保和环境有害废物处理）、社会（对员工尊重和不尊重的处理）和经济（愉快和不愉快的商业伙伴）。关于每个场景的原始文本可以在附录中找到。”107

刺激描述中要包含的元素，没有特定顺序，包括：

•创建的刺激数量和每个条件中的数量；

•刺激是真实的还是虚构的：

•如果是虚构的，请描述如何让受试者相信它们是真实的，并

•如果是真实的，请描述如何确保受试者的先前知识不会干扰结果；

•如果涉及欺骗，如何向受试者解释；

•如何将感兴趣的理论概念操作化；

• 刺激的物理属性，据信会导致因变量的变化（例如图片、格式、长度、字数等）；

• 与自变量相关的任何潜在混淆因素的讨论，无论是涉及刺激还是程序，以及它们是如何被控制的（例如统计上或在处理中）；

• 如何最大化差异；

• 以理论角度解释所有决策，例如性别是否与特定结果相关，以及操作化如何反映理论构建。

有关在方法部分编写刺激部分的更多示例，请参见"How To Do It Box 9.4"。

编写刺激部分的建议是在刺激创建后立即进行。在完成研究后，很难记住所做的所有事情或背后的决策原因。同样的建议也适用于撰写刺激检查结果；重新运行刺激检查数据可能会耗费大量时间（特别是在研究阶段）。养成一种习惯，即在完成每个部分的方法后立即进行编写，即使只是以笔记形式。

**刺激检查写作**

有些期刊将刺激检查放在方法部分，紧接着刺激描述之后；而另一些期刊则将其放在结果部分，特别是如果刺激检查是在实际研究的受试者上进行的。有些研究会两者兼而有之，将刺激检查所用的测量工具在方法部分中进行报告，然后在结果部分中报告这些测量工具的测试结果（见"How To Do It Box 9.5, Writing Up Manipulation Check Results"）。如果期刊没有偏好，那么应该将刺激检查结果放在最合乎逻辑的地方，并且不会让读者产生疑问。另一种选择是对于非常长的刺激检查，或在实际实验之前和期间进行了检查，可以将结果放在附录中或引导读者查看补充性网站。

**如何进行第9.4部分？**

**编写刺激材料**

操作检验结果会根据期刊的偏好和最合理的位置在研究的不同部分进行报告。以下是一些示例。

示例1：

在这项研究中，操作检验结果在方法部分中报告，出现在刺激材料后，内容如下：

“在一个七点量表上（1 = 强烈不同意，7 = 强烈同意），参与者被问及文章在(1) 积极或(2) 消极的问题后果方面处理的程度。操作检验显示成功的操作，并略微提高了标准差。当被要求评估积极后果时，机会条件下的参与者（M = 5.94，SD = 1.63）认为他们的文章比风险条件下的参与者（M = 2.35，SD = 1.93；t(612) = 2.75，p < .001）更积极。当被要求评估消极后果时，风险条件下的参与者（M = 5.87，SD = 0.146）认为文章比机会条件下的参与者（M = 3.24，SD = 2.09；t(611) = 4.53，p < .001）更消极。”

来源：Lecheler, Sophie, and Claes H. de Vreese. 2013. “What a Difference a Day Makes? The Effects of Repetitive and Competitive News Framing Over Time.” Communication Research 40 (2): 157.

示例2：

在这项研究中，操作检验结果在结果部分中报告：

“将条件（行动示例，不行动示例和对照组）作为自变量，将负性情绪作为因变量，一元方差分析（ANOVA）报告了显著效应，F(2, 192) = 9.24，p < .001，η2 = .09。Bonferroni事后检验报告，行动示例条件下的参与者（M = 3.56，SD = 1.2）报告的负性情绪得分比对照组的参与者（M = 2.85，SD = 1.33，p < .01）和不行动示例条件下的参与者（M = 2.62，SD = 1.2，p < .001）更高。因此，行动风险示例的操作是有效的。然而，不行动示例条件下的参与者产生的负性情绪得分与对照组没有显著差异，p = .95，表明不行动示例条件的操作未达到预期的效果。”

来源：Dixon, Graham N. 2015. “Negative Affect as a Mechanism of Exemplification Effects.” Communication Research 43 (6): 768

示例3：

在另一项研究中，操作检验结果包含在一个附加的在线网站中，文本给出了以下结果总结：

“损害提示产生了与方便相关的关联，而厌恶和悲伤提示分别比其他三个提示显著产生了更多相关的情绪。此外，各提示在感知负面性方面总体上是可比较的（唯一一对显著差异的文章是厌恶和损害；详见在线支持材料以获取完整的预测试结果）。”

来源：Ben-Nun Bloom, Pazit. 2014. “Disgust, Harm and Morality in Politics.” Political Psychology 35 (4): 501.

示例4：

最后，在这项研究中，部分操作检验结果在方法部分报告，部分在结果部分报告，内容如下：

在方法部分：

“测量使用7点李克特量表进行，除非另有说明。

操作检验。

在预测试和后测试中，使用来自之前研究的项目进行感知威胁的评估(Rimal et al., 2009; Tyler & Cook, 1984; Witte, 1994)。为了衡量感知易感性，参与者被问及他们自己或周围其他人可能会暴露于HIV/AIDS的可能性。参与者还被问及HIV/AIDS的严重程度和威胁程度，以及HIV/AIDS预防对个人和社会的重要性或紧迫性。回答被平均成感知严重程度指数（预测试α = .86，后测试α = .95）。”

在结果部分：

“从预测试到后测试，与主要伴侣的参与者在个人层面上感知的易感性更高，M = .50，SD = 1.34，t(157) = 4.68，p < .001；在社会层面上感知的易感性更高，M = .37，SD = 1.23，t(157) = 3.88，p < .001。从预测试到后测试，与非主要伴侣的参与者在个人层面上感知的易感性更高，M = .47，SD = 1.38，t(111) = 3.62，p < .001；在社会层面上感知的易感性更高，M = .43，SD = 1.30，t(111) = 3.49，p < .001。感知严重程度没有增加。因此，假设1在某种程度上得到了支持。感知威胁的操纵差异表现在感知易感性上。”

来源：Zhang, Jueman (Mandy), Di Zhang, and T. Makana Chock. 2014. “Effects of HIV/AIDS Public Service Announcements on Attitude and Behavior: Interplay of Perceived Threat and Self-Efficacy.” Social Behavior and Personality 42 (5): 802–803.

不管它出现在哪里，操作检查描述中应包含以下要素：

• 是否在实际研究之前、实际研究内部或两者都进行了操作检查；

• 操作是否成功或失败的陈述；

• 人数（Ns）、统计测试以及表示显著差异和差异方向的值；

• 使用的测量方法和反应集；

• 如果项目是由其他人创建的（引用他们）还是由研究人员创建的。

通过操作检查创建和确认成功的刺激后，实验设计即将完成。接下来的章节将涉及构建用于测量因变量、协变量、中介变量和调节变量的工具。当然，实验设计的步骤可以颠倒过来，工具设计可以先于刺激构建，或同时进行。

**常见错误：**

• 未使用逼真的刺激材料

• 仅使用一个刺激材料而非多个

• 在必要时未进行操作检验，或者未报告操作检验结果

• 低估达到实验结果所需的操作力度

• 在实际研究中未包括心理状态作为中介因素

**测试你的知识**

1. 在一项研究中，记者阅读了突出政治家不同特征的新闻稿，比如他们的政治派别和喜欢的议题，并被问及他们有多大可能写一篇关于他们的报道。在刺激材料中，哪个理论兴趣类别被操作了？

a. 记者撰写报道的可能性

b. 使政治家成为新闻价值的不同特征

c. 新闻稿的长度

d. 记者所在新闻机构的规模

2. 刺激材料代表理论兴趣的类别，因为它们。

a. 是实验者操纵假设会导致某种效应或结果的因素的手段

b. 是复杂的，需要很多工作来创建

c. 控制混杂变量

d. 测量和进行统计控制混杂变量

3. 在使用真实政治人物作为刺激材料的实验中，控制受试者对政治人物的先有知识是很重要的。

a. 真

b. 假

4. 统计控制是。

a. 无法在刺激材料中控制的事物

b. 被测量并进行控制的事物

c. A和B都是

d. 类似ANOVA或卡方检验的统计技术

5. 统计控制应该在哪时候进行测量。

a. 在给予刺激材料之前

b. 在给予刺激材料之后

c. 在刺激材料被执行时

d. 在研究结束后两周

6. 为了避免受试者先有知识的影响，实验中常常使用虚构的人物、地点和事件。这些称为。

a. 统计控制

b. 实验控制

c. 对照组

d. 处理

7. 选择与特定政党没有明确关联的政治议题代表了一种形式的。

a. 统计控制

b. 实验控制

c. 对照组

d. 处理

8. 协变量应该。

a. 总是包括人口统计学信息

b. 包括在操作检验中

c. 被随机分配

d. 与结果或依赖变量相关联

9. 固定因素是 。

a. 其水平是不限制的

b. 其水平由研究人员选择

c. 代表人口样本的因素

d. A和B都是

10. 随机因素是 。

a. 其水平是不限制的

b. 其水平由研究人员选择

c. 代表人口样本的因素

d. 那些是随机分配的因素

**Answers**

1. b

2. a

3. a

4. c

5. a

6. b

7. b

8. d

9. d

10. c

**应用练习**

1. 在你正在设计的实验中，描述你将如何创建你的刺激材料。回答以下问题：

- 你将如何使它变得真实？例如，你会从一个实际的网站或刺激材料开始并进行改编吗？你是否有设计广告的专业经验？你会雇佣一个有经验的人吗？你会使用软件或在线工具吗？或者你会使用现有的刺激材料，比如电视节目或课程计划？

- 你的刺激材料如何代表理论感兴趣的类别？例如，在在线故事的结尾的评论是否代表读者参与的理论类别？

- 你的刺激材料如何最大化比较？你是使用最强的操纵还是最大的差异？

- 你将如何使用信息变异？也就是说，你将创建多少个刺激材料？为什么？

一旦你计划好刺激材料的所有细节，开始创建它们。

2. 设计一个用于你创建的刺激材料的操纵检查。你将使用哪些问题来确定操纵是否被受试者按照你的意图感知？你将招募多少名受试者，以及从哪里招募？需要多长时间？一旦你创建了刺激材料和操纵检查，邀请一些受试者参与测试，看看你的刺激材料是否起作用。如果你的机构要求，先获得IRB批准。将这部分写入论文中。

3. 阅读你感兴趣领域的四篇研究的方法部分，并标记所报道的项目。这一章节中包含了什么？没有包含什么？报告了哪些特殊问题？识别在文章中哪些地方报道了操纵检查——是在方法部分、结果部分、脚注、附录还是在线？

**第十篇 工具和测量**

*我一次又一次地被测量对于改善人类状况的重要性所震撼。*

*—比尔·盖茨*

**学习目标**

**• 总结问卷作为实验仪器的优缺点。**

**• 讨论何时使用单一项目指标较为适当，何时指标较为优越。**

**• 辨识一些不显著性仪器，并能将它们与测量的心理或生理过程联系起来。**

**• 描述在实验中如何使用观察，并确保可靠性。**

**• 解释实验中独立变量（IVs）和依赖变量（DVs）所需的测量水平。**

**• 设计一个实验仪器，包括适当的测量水平和回答选择，并验证其构造的有效性。**

在准备实验材料时，处理刺激只是其中的一半工作。另一半工作涉及创建用于测量刺激引发的响应的工具。仪器与刺激不同，它是实际用于测量响应的设备，而刺激是传递操纵或处理的载体。换句话说，仪器是研究人员用于收集对刺激产生的数据的测量设备、工具或程序。测量不仅是改善人类状况的重要因素，正如开头引述所说，对于实验的成功也起到至关重要的作用。

**仪器**

仪器可以分为三个主要类别：（1）作为实验对象完成的自报告，例如问卷调查；（2）作为技术设备，例如B.F.斯金纳的测振仪，用于记录运动；（3）由研究人员记录的观察结果，例如课堂中儿童的行为。本章将讨论这些类型仪器的优缺点，并随后讨论测量问题。

在当今许多社会科学实验中，仪器通常是类似于调查的问卷，旨在测量人们对某种操作或处理的自我报告响应。问卷是用于收集和记录受试者响应的一系列问题。问卷可以通过多种方式进行调查，包括计算机、纸笔、面对面访谈或电话等方式。罗纳德·费希尔的品茶实验就是采用面对面访谈作为仪器，被试者只需告诉实验者茶或牛奶先来。

撰写问题是一门复杂的艺术形式，涉及诸多决策，如措辞、内容和布局等问题。关于这个主题有很多优秀的文献和文章，因此在此不再详述。相反，我建议参考“建议阅读”列表中的资料，并在本章的“更多信息”框中强调我常见到的实验问卷问题的要点。

重要的是，实验的问卷需要以具体方式捕捉因变量、协变量和中介变量（中介者和调节者）的抽象概念。受试者所回答的实际问题是文献综述中概述的理论概念的应用转化，也称为操作化。例如，关注概念的一种测量方法是根据受试者分配给特定任务的时间量。

**内容延展10.1**

**编写问卷调查**

• 一个问题只涉及一个事物。注意避免使用“和”这样的词，比如：“请在以下评分表上评价您对黑人和西班牙裔的态度。”人们可能对每个群体持有不同的态度。让他们只给出一个答案会引发错误。

• 确保问题对普通人来说有意义。例如，“您在这个困境中有多深入参与？”本意是将“参与”操纵为全神贯注于某事，但可能被解释为与问题本身有个人相关性。

• 问题应简短直接，并且要写得清楚明了。避免使用研究人员的术语，如“操作化”。

• 检查每个问题是否必要，以及需要多少详细信息。每个问题都会增加研究时间，可能导致受试者失去兴趣或感到疲倦，从而引入错误。例如，如果收入不会在分析中使用，那么就没有必要询问收入情况。

• 避免出现有偏见和引导性的问题。去掉形容词和副词。

• 在可能的情况下，使用在其他研究中已经使用过的问题。方法部分应给出准确的措辞，并提供参考文献中的概念定义。

• 仔细检查假设中的所有变量是否在方法部分得到操作化，并且在文献综述中有相应的概念定义。如果有任何“孤立问题”，请问自己是否真的需要这个问题。

受试者记住。2焦虑这一概念可以通过问题来测量或操作化，例如“您感到有多焦虑？”为了测量虚构企业的可信度，一项研究向受试者提出了四个陈述：“我发现信任（该企业）很容易”，“我完全信任（该企业）”，“（该企业）是一个值得信赖的组织”，“我信任（该企业）的管理层”。3问题在研究中使用的确切措辞可以在方法部分、脚注、附录或补充网站中找到。有些研究引用了以前使用过相同问题的另一项研究，研究人员需要阅读被引用的研究以获取确切的措辞。

问题也可以使用比直接使用“焦虑”来测量焦虑，以及“信任”来测量可信度的措辞，就像前面的例子一样。例如，为了测量受试者对某个产品的信任程度，可以使用如下的问题：

为了衡量社会营销效果，一项实验使用了问题来询问受试者是否有意改变、行为意图和愿意沟通。这三个测量指标都是其他研究中使用过的，研究人员进行了引用。研究人员还加入了一个新的自行设计的测量指标，询问受试者是否有意通过点击“喜欢”按钮来表达态度。

正如这项营销研究所示，使用其他研究人员开发和测试过的问题，也称为项目，应该是确保可靠性和效度的首选。还经常会对现有问题进行修改以适用于特定的使用场景。当以前的研究缺乏良好的测量指标，或研究人员认为有另一种或更好的方法来测量一个概念时，开发原始问题（如“喜欢”按钮示例）是合适的。这些新项目应进行充分的测试，以确保它们能够正确测量预期的内容。稍后在本章的“测量问题”部分将详细讨论这个问题。

**单一项目指标 vs. 指标**

只使用一个问题来测量——称为单一项目指标——应该仅用于具体的概念，比如年龄、教育水平和其他人口统计信息，以及一些简单的概念，例如学生对教师整体有效性的评价和整体工作满意度。了解哪些变量可以作为单一项目指标是有益的，因为较少的问题可以缩短研究时间。然而，如果研究中使用了不恰当的单一项目指标，由于测量效度和可靠性问题，这些研究更有可能被期刊编辑和审稿人拒绝。对于抽象和多维概念，如可信度、说服力和焦虑，应该由多个问题构成一个指标，就像社会营销研究中用于形成四个问题的指标一样。仅用一个问题来测量一个抽象概念往往较为薄弱和不稳定，而较强的测量指标更有可能在存在的情况下显示出显著性。这可以通过一项研究来说明，该研究使用了两个问题来测量宗教信仰：“您是否认为自己非常信仰宗教（7）到非常不信仰宗教（1）？”和“您对宗教信仰的位置在哪里？”选项包括极端原教旨主义者（7）到极端自由主义者（1）。当在分析中使用这个两个问题的指标时，它在p < 0.001的水平上显示出显著差异。当审稿人提出原教旨主义者到自由主义者的问题可能会对分析产生偏见时，这个问题被去除了，分析只使用询问宗教信仰的单一项目指标进行。该分析得出了与之前相同的结论，但显著性水平从p < 0.001下降到p < 0.05，可能是因为单一项目指标较原始的两个问题指标较为薄弱。这就是为什么单一项目指标仅适用于最明显、具体和单维的变量的原因之一。

相反，复杂的理论概念应该通过多个问题构成的指标来进行测量。一个很好的例子是自我效能概念，它描述了一个人对自己成功执行任务的信念。人们从四个来源中评估自己的能力，包括自己过去的表现、观察他人、表扬和批评以及情绪状态。因此，测量自我效能的问题应该涵盖这四个来源。许多自我效能指标，有时也称为量表，已经针对特定领域进行了定制，从教育领域到医疗保健等等。甚至还有专门用于色情内容避免和中国互联网信息检索的指标，以及儿童、父母、教师等的综合指标等等。（有关构建自我效能指标的指南，请参阅Bandura，2006年的文章。）例如，在一项研究中，教育自我效能指标包含了先前由其他研究人员开发的九个项目，其中包括：“我预计在这门课上表现非常好”，“我确信我能在课堂上做得非常出色”，以及“我知道我能够学会课堂上的知识。”稍微改变已开发的指标的措辞可能会产生不同的结果。例如，询问一个人“能做什么”与询问一个人“有能力做什么”是不同的，而在自我效能研究中询问一个人“将要做什么”也是不同的。重要的是阅读有关指标开发的文章以获取相关指导。

使用自报告作为测量实验操纵结果的工具具有优势和劣势。问卷主要依赖于自报告，即受试者自己回答问题，报告他们自己的态度、观点和行为。自报告快捷、廉价且简单。受试者可以私下进行，匿名性被认为鼓励了诚实回答。在线进行的调查实验可能会减少需求效应，因为受试者不受实验者在场的影响。问卷是社会科学实验中广泛使用的工具，用于测量各种事物，如记忆、态度、情绪、个性特征、行为和行为意图。然而，自报告本质上是脆弱的，可能导致不准确的结果。

自报告存在许多不准确的原因。人们可能在管理自己的自我形象或以社会上希望听到的答案回答问题。例如，很少有人会承认自己持有种族偏见。自报告试图通过询问人们不是他们认为什么，而是他们认为别人认为什么（例如，“但我们并不询问您对这些事情的意见。在这项研究中，我们关注的是对意见项的感知。”）来克服这一问题。然而，自报告在克服社会上希望性偏见或印象管理方面并不成功，因此应考虑其他的测量工具。此外，人们可能不愿诚实回答有关非法行为或涉及道德问题的敏感问题。

另一个导致依赖自报告的问卷可能产生错误答案的原因是有些事情可能超出一个人的内省能力来评估。

例如，有些感受对人们来说是无法获取的，就像有些自己的个性特征一样。在这些情况下，人们并不是故意回答错误，而是根本没有意识到他们真正的感受或特质。在这种情况下，其他测量工具可能会产生不同的结果。例如，研究显示，当使用自报告进行测量时，保守派比自由派更快乐，但当使用不显眼的测量方法时，比如观察微笑和使用积极情感语言时，自由派表现出更快乐的行为。

此外，人们可能会错误地回答、错误回忆、夸大或淡化他们在自报告中的答案。另一个解释可能是对问题的误解；在在线进行的调查实验中，受试者不能向研究人员询问他们可能不理解的问题。同样，人们可能会对问题的含义产生不同的理解，与研究人员的意图相左。例如，我曾经通过询问几个标准问题来测量受试者的参与度，其中包括“您对这个故事有多少参与？”这个问题的目的是测量他们对该故事的参与度或兴趣。研究的初步测试中使用的研究生们回答的方式是符合预期的；而在实际研究中的受试者，即本科生，在回答时写下了笔记或者在事后解释时说：“我对（毒品/卖淫/无家可归/老年虐待）一点兴趣都没有！”该问题被从指标中删除了。同样，一些人对响应量表的解释也有所不同，有些人会一直给出最极端的评分（例如1和7），而其他人则会坚持中间的评分（例如3和4）。

通常会进行效度研究，将自报告与其他测量工具进行比较。例如，在犯罪学研究中，关于犯罪行为的问题可以与刑事犯罪记录进行比较。教育研究经常关注学生的自报告准确性；例如，一项研究将高中音乐课程的成绩单与学生自报告的音乐参与情况进行了比较，发现学生自报告的参与音乐合奏的情况非常准确，但学生通常会夸大他们练习音乐的时间。教师的报告也没有更准确。

最后，成功进行实验的一个小提示是，在问卷中提供开放式回答选项可能非常有启发性和有帮助。这些问题没有固定的答案或量表供受试者选择，而是允许他们用自己的话回答，例如“告诉我关于……”或“为什么选择这个答案？”受试者的评论作为例子来说明统计结果，可以使结果“栩栩如生”。同时，它们可以用来支持主张。例如，一位审稿人曾经问过在一项研究中的受试者是否符合了仔细阅读研究材料的标准。

细察问题的各个方面而不仅仅是试图支持他们自己预先设定的立场在决策实验中是非常有价值的。能够从开放式回答中提取例子来证明他们做到了这一点是非常有帮助的。在定性数据分析方面有着丰富的文献，但这些开放式回答也可以被归类、计数和测量，将定性数据转化为定量数据。但需要注意的是，它们需要由人类进行分类（参见本章的“观察”部分），因此并不像封闭式问题那样容易使用。

关于在实验中创建成功的仪器和测量的更多提示，参见本章的10.2节。

有些学科比其他学科更倾向于使用自报告；例如，经济学更喜欢行为观察或基于绩效的测量。这些类型的仪器以及其他替代自报告的方法将在下面进行介绍。

**非干预性和无反应性测量**

除了之前描述的自报告的许多弱点外，它们还受到反应性和需求特征的影响，这是在第二章中介绍的概念，解释了受试者对于被研究的反应。为了克服这些偏见，研究人员设计了一些非干预性的仪器。非干预性仪器是确保受试者不知道正在测试什么的仪器，以便测量的行为不会影响其回答。另一个经常用来表示相同意思的术语是无干扰的。无干扰和无反应性测量存在于一个连续体上；在最无干扰的研究中，受试者甚至不知道他们正在参与一项研究，甚至研究人员也不知道研究假设。更常见的情况是，受试者知道他们正在参与一项研究，但不知道研究的目的，因此无法为或反对假设而影响他们的回答。接下来将描述的心理学仪器在某种程度上是无反应性的，因为受试者无法歪曲他们的回答，但他们肯定知道自己正在参与一项研究。

**技术仪器**

衡量实验处理的结果或反应的另一种方式是使用技术工具、机械或设备类别的仪器。一些研究人员也将其称为设备、硬件和软件。B·F·斯金纳发明了自己的仪器，比如“累积记录仪”，它使用笔和纸卷自动记录老鼠按动杠杆或鸽子啄键的次数。虽然计算机管理的仪器提供了更多的选择，但技术仪器在一些研究中仍然有很大的用武之地。例如，视觉实验可以使用眼动仪来记录受试者的注视点。其他技术仪器包括脑电图仪（EEG）、功能性磁共振成像（fMRI）和生物反馈仪。

## **更多信息... 盒子 10.2**

## **成功的测量方法10.2**

•在受试者接受处理刺激后立即进行测量。例如，受试者阅读一个故事后立即回答问题或列出所有的想法，然后再阅读下一个故事，再回答问题或列出想法。不要在所有故事都被阅读后再一次性给出所有问题或想法列表。

•使用多种因变量的测量方法，例如情感温度计、六个特质评估和两个同意-不同意的陈述。这样可以进行测量的比较，减少任何特定仪器驱动结果的机会。其中一种测量可能比其他测量更好。

•另一种克服社会期望偏见的方法是引入无关的问题或任务，以使受试者不能轻易猜测研究的内容。这种方法也被称为介入项、缓冲项和干扰任务。有关示例，请参阅Emily Thorson在信仰回响方面的研究。

•记住对那些措辞与大多数问题相反的项目进行反向编码。例如，1应变成7，2变成6，依此类推。

•实验受试者的数据不一定要一次性收集。许多实验，特别是使用心理学和生理学仪器的实验（例如，EEG和fMRI等），通常需要在一段时间内进行。空间和设备限制是最普遍的原因之一。例如，在实验室设置中向受试者展示电视节目的研究人员可能没有足够大的房间；许多心理学和生理学仪器的价格昂贵，并且只有一个或几个可用，因此需要一个一个地运行受试者。实验在一段时间内进行的其他原因包括在从一项实验中学到结果后需要加入额外的处理。然而，研究人员应该意识到可能威胁到效度（参见第5章）的因素，比如可能会影响结果的事件（例如，历史因素）。

•谨慎地包括协变量。协变量的目的是减少与因变量相关的干扰变量引起的方差。这些协变量需要事先计划，并且在给予处理前进行测量，如果处理可能会影响协变量的话。正确使用协变量可以增加实验的效果。当包括太多与因变量无关的变量时，它们实际上可能会降低实验的效力。协变量不应该用于预测自变量，而是用于预测因变量。

• 只有当人口统计学特征是因变量的强预测因子时，才需要将其包含为协变量。只需要在问卷中包括用于描述样本或将用于分析的人口统计学特征。例如，有关收入的问题可能对受试者来说特别敏感；除非收入对于分析是必要的，否则最好不要询问。

虽然问卷调查也需要技术，但它们与下一节中的高科技仪器有所不同，例如功能性磁共振成像（fMRI）机器可以检测大脑血流的变化，让研究人员看到哪些脑部区域被激活；眼动追踪设备可以记录眼睛的运动；心率监测器可以测量注意力、唤醒状态、情绪和努力；以及测量皮肤（如出汗）变化作为情绪、唤醒状态和注意力的代理的皮肤导电反应受体。50 当今许多这些仪器确实包含技术和软件；例如，测量认知处理的问题的响应速度（即主观反应到问题的快慢）是使用测量毫秒级响应速率的计算机软件，而网络跟踪软件则记录一个人点击的链接和在某个网站上停留的时间。这些仪器与使用软件记录的问卷调查分类不同，因为受试者不是自我报告他们的反应。

这些以及未在此列出的其他仪器，因其种类繁多，代表了隐式测试，即受试者并不一定有意识地意识到自己的反应。其中许多记录了心理或生理反应，例如心率。在社会科学中，有些仪器被更频繁地使用，而有些则更多地用于社会科学的特定学科领域。接下来，我们将重点介绍社会科学中一些较为流行的高科技仪器。由于篇幅限制，无法对每一种仪器的优势和劣势进行全面讨论，本文仅作简要介绍，感兴趣的读者应寻求更全面的文献来了解这些问题以及如何解决它们。

**功能性磁共振成像**

功能性磁共振成像（fMRI）是社会科学领域最新流行的仪器之一，它是一种非侵入性的脑部扫描仪，可以生成大脑的横截面图像，捕捉大脑活动而非静态大脑结构。它被用于识别与特定心理过程相关的大脑区域，比如信任，或者在经历恐惧等情绪时。在社会科学领域，对fMRI和类似的神经成像工具越来越感兴趣，特别是在市场营销和消费者研究方面，有很多关于fMRI的书籍。

使用fMRI进行市场研究的例子包括：观察人们看到自己喜爱的品牌标签时的认知过程，选择啤酒和咖啡品牌时的行为研究，了解愿意支付的程度，观察专家陈述的效果，以及基于销售广告预测销售变化。在政治学中，fMRI被用于预测对候选人态度的变化，研究情绪与投票之间的关系等。在健康传播领域，fMRI被用于研究对健康信息的记忆和测试有关危险饮酒的公共服务公告的有效性。在经济学中，fMRI被用于研究交换中的声誉和信任。这只是社会科学中使用fMRI的一些例子，几乎每个学科都至少有一些研究，其中一项研究甚至使用fMRI来观察艺术和流行音乐之间的认知差异。

使用fMRI作为研究工具不像使用问卷这样简单。这需要相当多的培训、时间和精力来进行这样的研究。甚至可能需要一名受过训练的实验室技术员，以及可能需要放射科医生来寻找大脑异常，如果由机构审查委员会要求。设备昂贵且庞大，包括一个供受试者平躺的全身管道。受试者可以通过护目镜或屏幕观看刺激，并通过口头或键盘作出回应。他们的头部必须保持完全静止，并且不能有任何金属穿孔、植入物或医疗问题。扫描仪噪音较大，实验任务需要限制在四十五到六十秒之间。由于高昂的成本和每次只能一个受试者，样本量通常较小，通常每个研究有十到二十名受试者。目前，社会科学中的fMRI主要用于补充其他研究工具，例如与自我报告或面试进行三角测量，或测量受到前面讨论的偏见影响的概念。

**心率监测仪**

心率监测仪是另一种测量生理反应的仪器。与fMRI相比，它相对便宜且易于使用，尽管需要特殊的安全培训，因为涉及电流。通过心电图（缩写为ECG或EKG）测量心率，用于评估注意力、唤醒状态、认知努力和情绪。它可以实时记录一个人在执行某项任务时的情绪和认知变化，例如观看政治广告或听音乐。心电图使用通过电极附在被试皮肤上的电脉冲来测量心跳间的时间，并需要特殊的硬件和软件。它已在社会科学中使用，包括用于了解考试焦虑、研究听现场音乐与录音音乐的差异、测试不同早餐饮料、进行公共服务广告和健康信息测试，观看电视和电影中的抖动画面以及名人广告，阅读新闻，观看政治广告，观察政治领导人做出不当表情，以及儿童使用信息技术等等。

皮肤电导变化也可以提供情绪、唤醒状态和注意力的测量。这些仪器被称为不同的名字，包括皮肤电导反应（GSR）和皮肤电导反应（EDR），它们测量皮肤包括出汗、电压和电阻等波动。皮肤电导仪特别适用于商业广告、销售与营销和消费者研究，以克服自我报告的问题。除了这些领域，皮肤电导还被用于政治科学等领域，例如研究焦虑对政治信念的影响，以及与媒体暴力、电视剧、健康信息、营销、教师压力和游戏等相关。皮肤电导仪需要将电极连接到被试的手指或手掌。与大多数高科技仪器一样，被试必须逐个进行，计算机或软件故障可能会进一步减少小样本的可靠性。

除了已提及的仪器外，还有测量人类几乎所有心理和生理反应的仪器。其中包括响度启动反射，由大声噪音激活的反射，唾液测试用于测量压力，以及用于测量优势的声音频率分析和音高仪器等。在一项健康传播研究中，使用了CPR训练模型。

记录了压缩质量的测量数据，包括频率、深度、完全释放以及被试手离开模型的时间。105

**眼动追踪**

与像fMRI这样更难且昂贵的仪器不同，眼动追踪设备相对便宜且易于使用。眼动追踪用于无干扰地捕捉被试的视觉注意模式和方向。106 眼动追踪仪器测量眼睛的注视点或位置（即眼睛停留的地方）、注视频率（注视的频率）、持续时间（注视的持续时间）107 和移动。108

频率（多久一次）、持续时间（多长时间）107 和运动。108 它用于提供关于一个人看什么、以什么顺序看以及持续多长时间的信息。109 它已被用于解释认知负荷效应110，并使用瞳孔扩张来测量唤醒状态。111

在社会科学实验中使用的眼动追踪设备包括由公司如Compumedics、SensoMotoric Instruments、Tobii Technology、Applied Science Laboratories和EyeTribe制造的设备。112 成本可以低至200美元一套。117 追踪设备可以嵌入到头戴式耳机或眼镜中，由被试佩戴，也可以安装在移动设备或计算机屏幕的底部，或者在设备的上方安装。121 红外摄像头可以用于识别角膜反射的位置。122

屏幕上的手指触摸，例如缩放和滑动，也可以用这些设备来测量。123 除了眼动和注视数据外，还可以生成对刺激物最多注视区域的热图等其他图形。124

眼动追踪仪器的缺点与fMRI类似，样本大小较小，约为20人，因为运行每位被试所需的时间非常密集。127 例如，为每个被试校准设备需要大约十分钟。错误校准设备可能导致被试的损失。128

眼动追踪仪器已被用于研究社会科学中各种主题，包括决策过程；129 新闻照片；130 政治候选人的非语言行为；131 冲动购买；132 购物时的消费者行为等。

移动设备上；133 广告；134 电视；135 经济学；136 教育，包括老师戴着眼动追踪设备的眼镜观察学生的时间；138 以及所有在线事物。139

眼动追踪已经与其他仪器结合使用，例如与脑电图一起用于三角测量目的，140 也用于验证其他仪器——例如识别问卷调查中的问题。141 一项研究使用眼动追踪和面对面访谈作为视觉注意力的衡量，研究不同形式的问题如何影响回答。142

**屏幕视线**

一种类似于眼动追踪但专门用于电视的测量工具称为“屏幕视线”。143 这个仪器用于测量注意力，捕捉被试在电视屏幕上看了多少，持续多长时间，何时以及看的频率，也被称为“定向”。它需要一个位于单向镜后面的摄像头来记录被试观看电视时的情况。它用于研究电视节目和广告，并常常作为自我报告的补充或替代。144 与自动收集被试注视位置数据的眼动追踪设备不同，

屏幕视线工具需要被试的视线位置由人工观察记录下来（2023），记录何时，多久，多频和视线落在电视屏幕上的时间。145 关于这一点更多信息请参阅“观察”部分。

**网络追踪**

另一种类似于眼动追踪的仪器是网络追踪，它记录被试在网站各个部分的每次点击和停留时间。网络追踪或监控软件已在各种社会科学学科中使用，包括在线新闻使用，146 网站设计，147 以及政治信息。148 网络监控软件允许研究者在得到被试知情同意的情况下观察他们的实际在线行为，包括进入网站的时间、方式和位置，他们的导航路径，149 以及在搜索中使用的查询。150 这种类型的仪器的优势在于用户不知道他们正在被观察，它不受记忆失误的影响，可以实时记录真实世界中的行为。151 数据可以在实验室或被试的自然环境中收集。一些研究人员设计自己的软件，152 还有些使用商业软件，153 如Camtasia，154 还有一些分析日志文件155 或浏览器历史记录。156

**延迟反应**

反应时或反应时间仪器是一种经过时间考验的隐含测量方法，源自心理学。它测量人们对问题或刺激作出反应的速度，以隐式衡量记忆、隐含态度、态度可及性和认知加工。157 它以毫秒为单位，衡量被试记忆某事物、将刺激与预先形成的态度联系起来或处理新信息所花费的时间。158

反应时在测量对于敏感话题的态度方面广受欢迎，例如种族偏见，使用隐性联想测试（IAT）等测试方法。161 这些测试旨在捕捉人们可能不愿意承认的思想、态度和感情，例如刻板印象。162 它们被称为“隐性”，因为被试不知道正在被测量的内容，就像他们对明显的问题所知道的那样。例如，IAT将单词与面孔配对，询问该单词是否积极或消极。不喜欢某些面孔（如少数族裔）的被试对与其配对的消极单词的反应速度明显快于对积极单词的反应速度。163 这些面孔刺激本质上是潜意识的，只在毫秒级时间内展示给被试，时间如此短暂，以至于人们甚至没有意识到看到它们。164 但在潜意识中，人们对自己喜欢或不喜欢的面孔有积极或消极的反应，当单词的语气与面孔不一致时，需要更长的时间来克服对不喜欢的面孔的消极反应，并决定该单词是积极的。可以在哈佛大学的隐性项目网站上进行隐性联想测试：https://implicit.harvard.edu/implicit/takeatest.html。反应时的方法也已经被改编用于对敏感主题进行社会科学研究。

社会科学中反应时的另一种用途是测量记忆。例如，被试可能会看到某个电视节目的一部分，然后在后续的测试中被问及是否看过，要尽快回答。165

been used as a

反应时在广告效果方面是一种记忆的测量方法，经常出现在传播学研究中。167 它在教育研究中也很有用，例如，多媒体学习。169

另一种使用反应时的方式是测量态度可及性，即从记忆中自动检索某种东西的速度，较强和更易访问的态度比较弱的态度更快地被检索出来。170 例如，政治态度和行为的研究常常以这种方式使用反应时。

潜伏期也常被用作信息或认知加工的测量指标，较长的时间表示深思熟虑或中央路径认知加工，而较短的时间则表示更自动化或外围路径认知加工172 例如，通过照片解决伦理困境比没有照片时需要更长的时间，这导致了更谨慎的认知加工，从而得出更好的道德判断。在一个有关求职者在人事测试中伪造答案的商业研究中，研究者也观察到这种现象；制造伪造答案需要更多的心理加工，因此需要更多的时间来思考，而诚实回答则不然。同样，在一个有关信息技术的研究中，受试者评估了网页，更具吸引力的网页与更短的响应潜伏期相关。人们判断网页吸引人所需的时间比判断网页不吸引人所需的时间短。该研究还将这种潜在指标（潜伏期）与关于网页吸引力的明确问题进行了比较，并发现两者是一致的。

同样在这一方面，潜伏期被用来评估说服知识，即人们对被展示的说服信息的认识程度。该理论解释了认知容量在意识到说服知识方面的作用，当认知容量较低时，人们较不可能意识到说服知识。例如，在观看需要大量认知容量的电视节目时，间接测量，如反应时间，比直接测量，如明确问题，更适合测量说服知识。

潜伏期反应可以通过研究人员使用秒表来简单测量，也可以需要特殊设备和软件。例如，Meyer在面对面的访谈中，研究人员在提问结束后立即启动计时器，然后在受试者开始回答时停止计时器。潜伏期也可以通过电话和实验室测量，使用包含计算机和软件的设备来传递刺激，收集反应，并内置精密计时器。在实验室环境或调查实验中，计算机软件如Inquisit，SuperLab或MediaLab测量潜伏期反应，从而减少测量误差。在被试内设计中，用于控制个体差异，因为有些人阅读和回答比其他人更快。在被试间设计中，得分应标准化。由于可能存在设备故障或受试者未能回答等错误，数据应进行清理。此外，与特殊响应设备相比，计算机键盘可能不太准确。

**次级反应任务**

次级反应任务（SRT）仪器与潜伏期反应类似，它测量了刺激和反应之间的时间；然而，次级反应任务还要求受试者在进行主任务（例如回答问题）的同时执行一个辅助任务。这种辅助任务需要认知资源，因此导致了主任务的潜伏期延长。次级反应任务被用于评估认知负担、潜在记忆和其他心理过程。它在心理学和其他社会科学领域广泛应用。

在次级反应任务中，增加了一个“干扰”任务。次级反应任务测量的是注意力而不是记忆。在这种类型的研究中，受试者在执行某个任务的同时，同时处理一条信息，实际上是同时做两件事情。例如，受试者可能会被要求在看电视的同时，当看到闪光灯时按下按钮。次级反应任务测量认知加工和一个人资源的限制，以“提供关于信息使用了多少容量的线索”。这种想法是第一个任务（观看电视，阅读故事等）的需求或难度增加将减慢受试者对第二个任务（当听到声音或闪光灯时按下按钮）的反应。次级反应任务可以回答有关注意力、唤醒和参与的问题。次级反应任务测量从声音或光线到受试者按下按钮之间的毫秒时间长度。通常情况下，如果第一个任务更困难或需要更多的努力，人们在第二个任务上会犯更多错误或需要更长的时间来作出反应。次级反应任务对于理解人们何时以及为什么不从信息中学习很重要。

由于测量的时间单位是毫秒，线索的时机必须准确，因此需要像Inquisit、SuperLab或MediaLab这样的特殊设备。必须进行实践运行来验证设备是否正常工作，并且每个受试者必须熟悉使用设备。多种测量和刺激对于控制基线反应时间的个体差异至关重要。例如，通常使用十二个到二十个刺激。这也可以防止非典型信息，如上一章中讨论的信息变异部分所述。可以一次运行几个受试者，例如每次运行六到八个受试者。参与者可以使用常规计算机键盘或特殊的反应板来回应，后者可能比键盘更准确。由于设备故障等技术原因，受试者可能会丢失，也可能因为受试者从未按下按钮等其他原因而丢失。

**连续反应仪器**

通过问卷报告的态度、情绪和其他感官反应的自我报告具有在某一时刻测量的缺点。但是，人们通常会持续体验事物，他们的认知状态会发生变化。例如，观看电视节目时，情绪可能会从恐惧转变为快乐、悲伤再到快乐。问卷只能捕捉受试者回答问题时的情绪。为了克服这一问题，研究人员使用连续反应仪器，让受试者能够一刻不停地报告他们的感受或思想和态度的价值。连续反应仪器通常是一个手持盒子，上面有一个与7点量表对应的刻度。当受试者观看刺激物，例如电视节目时，他们可以根据每时每刻的感受来调整刻度。

电影或广告时，他或她可以转动刻度来对应他或她的评估、态度、情绪或其他因变量。刻度硬件连接到具有刺激物的电视或屏幕上，以便数据叠加显示，准确显示受试者的反应。社会科学研究人员使用此技术来研究公益广告、政治候选人的面部表情和辩论表现，以及教师的有效性等等。

所有技术仪器都需要注意校准和衰减问题。在所有这些仪器上，按钮在使用过程中可能会变得更容易按下，从而对内部有效性构成威胁。如果使用多个仪器，例如两个心率监测仪，应在同一时间对同一受试者进行检查，以确保两个设备记录相同的值。研究人员应该在每个受试者之前检查仪器的功能和设置，以确保收集的数据准确无误。研究助手应接受设备正确使用的培训。

**思维记录和大声思考**

最后，最后一个需要简单仪器——录音设备的内隐测量方法是“大声思考协议”。这里将其与其另一个测量方法“思维记录”一起讨论，因为它们非常相似。这些技术测量人们思考和解决问题的方式，以及他们的思维过程随着录音设备（大声思考）或低技术的纸和笔，或者用于在线调查实验的计算机软件（思维记录）的使用而发生的变化。思维记录避免了使用明确的问题，而是让受试者接触一些刺激或让他们完成任务，然后要求他们在此期间写下所有的想法。大声思考版本实现了相同的目标，但要求受试者在暴露于刺激或执行任务时大声说出他们头脑中的一切。受试者的发言被记录下来，研究人员在场提示他们继续讲话，说出一切想到的东西。这些技术可用于研究人们在使用网站、搜索互联网、评估游戏设计和作出道德决策等方面所经历的心理过程，以及人们如何理解试图说服他们的企图或如何理解矛盾的政治信息。

这些协议避免了要求受试者内省和分析他们的思维过程的问题，但并未涉及人们没有自觉意识的自动加工。这些程序可用于测量数量和类型的认知加工，即受试者是否深入思考刺激物并付出努力，或者仅仅是表面思考，以及他们思维的内容和价值。

思维记录相对容易获得，只需要受试者写下所有的思维。研究人员必须根据思维的类别或类型进行计数和编码，这将在“观察”部分进行详细介绍。简化处理这些数据的一种方法是使用计算机软件进行自动分类和分析。研究人员还可以将思维记录与其他测量方法结合使用，以获得更全面和准确的认知加工信息。

思维记录和大声思考是研究人们心理过程的有用工具，帮助理解他们如何思考、解决问题和对刺激物做出评价。这些方法不仅可以帮助揭示人们对广告、政治信息和其他刺激物的认知反应，还可以在教育研究、心理学和其他社会科学领域中提供有价值的洞察。

任务是要求受试者将每个独立的思维写在一个单独的框中。要求他们在框内还勾选一个指示思维是积极的、消极的还是中性的选项，可以减少解释数据时的猜测。当然，这也容易受到偏见的影响，包括受试者为了加快研究进程而未记录一些思维。

大声思考需要更多的时间和资源来收集，但可以克服受试者在思维记录中未真实记录所有思维的问题。使用大声思考，只能同时测试一个受试者，每位实验者只能测试一个受试者。受试者必须首先接受培训，因为大声地说出所有想到的东西并不是正常的行为。培训可以通过简单的数学问题或词语混淆来进行。一旦受试者习惯了大声地说出所有的思维，刺激物被呈现，实验开始。研究人员在受试者附近但不被受试者看到的位置，提醒受试者在停下时继续说话。重要的是，研究人员不直接与受试者交谈、回答问题或探究他们做出某些决定的方式或原因；研究人员只能提示他们“继续说话”或“大声地说出一切想到的东西”。受试者的回答应该被记录下来并转录，以便像思维记录一样，可以使用内容分析方法进行分析。

观察是社会科学实验中另一个常见的“仪器”，例如，研究人员观察学生玩耍和与他人互动，并记录他们所看到的情况。菲利普·辛巴多的监狱实验利用研究人员对受试者在扮演囚犯和警卫时的行为进行的观察。斯坦利·米尔格拉姆将他的电击实验受试者的评论和非语言行为与受试者愿意给予的电击强度的数字数据相结合。

观察通常被认为是定性领域的方法，例如，参与观察或民族志学，但在实验中也可以使用。教育是一个常常使用这种技术的领域，因为这样比询问孩子们可能无法回答的明确问题更可靠。例如，在一个旨在评估体育教育计划对儿童道德发展影响的实验中，教师观察了儿童在体育课上、校内体育运动和课间休息时的表现。教师对儿童进行了十个亲社会行为的评估，包括争吵、抱怨、嘲笑、分享、违反规则和不轮流等。

观察也可以用作其他领域的仪器。例如，阿尔伯特·班杜拉关于暴力和媒体内容的波波娃娃实验就是一项著名的研究。在这个实验中，班杜拉向儿童展示了电视节目，其中的角色打人并扔球，同时大声喊出“砰”的声音，这些都是针对一个充气波波娃娃的。

在孩子们看过电视片段后，他们被放在一个类似的娃娃房间里。研究人员通过双向镜子进行观察，并记录下儿童是否模仿了他们在电视片段中看到的任何行为。在这项研究中，观察者都不知道哪些儿童处于哪种处理条件下，也就是说，这是一项双盲研究。在其他学科中，观察可以用于研究课堂中的干扰性行为、学生在任务上花费的时间、护士的洗手行为等等。实验经济学是另一个更喜欢观察行为的领域，例如，受试者选择竞争性报酬还是按件报酬进行有偿任务。

**观察者间一致性**

每当人类观察者用于评价、打分、判断或编码某种内容时，包括使用“眼睛注视屏幕”、“思维清单”和“大声思考”等仪器测量的数据时，都应该评估可靠性，以确保进行评估的人以相同的方式测量事物。例如，如果一个观察者将一个孩子的陈述评判为“嘲笑”，那么其他观察者也应该将其视为嘲笑，而不是“争吵”。这最初是在第5章关于效度和可靠性的讨论中提到的。通常，这些观察者间一致性检查应涉及两名或更多名观察者对同一面部表情、行为、思想或其他行为进行评判，并进行一致性计算。例如，在一项健康传播研究中，观察者记录了受试者开始进行心肺复苏术所用的时间；不同观察者对时间的观察结果进行了一致性计算。通常情况下，需要使用不包含在实际实验中的数据进行培训，以达到可接受的一致性水平，通常在0.80左右（有关观察者间一致性的更多信息，请参阅更多信息框10.3）。与技术仪器一样，人类仪器也可能随着时间的推移而发生变化，例如在准确编码方面变得更加熟练，因此需要定期检查。有许多良好的指南可以进行观察者间一致性检查，使用的方法与内容分析方法中的编码者间一致性相同。在“建议阅读”部分中有几个推荐。

与培训评判者或编码者相关的问题是培训可能正在进行实验的研究助理。受试者必须相信实验刺激物甚至是实验情景是真实的，研究是重要的，并且对其至少有些兴趣，才能得到良好的结果。进行实验的研究助理或扮演同伴的人（即他们在表现时假装成真实的受试者，就像米尔格拉姆的电击实验中的学习者一样），必须接受扮演角色的培训，要做更多的表演而不是简单地读一份脚本。

**更多信息......方框10.3**

**观察者一致性. 10.3**

根据测量的性质是否为名义、序数、间隔或比率，有不同的统计工具用于分析观察者一致性。例如，微笑可以编码为二分（微笑或不微笑），也可以使用评分尺度，例如1 = 皱眉，7 = 微笑，5 = 中性口部表情，中间的点表示中间水平。当测量是名义级别时，适用Scott的Pi或Cohen的Kappa作为统计测试来衡量一致性。这些计算考虑到了两位观察者因为偶然的巧合而达成一致的概率，而不是因为他们实际上解释事物的方式相同。与不考虑偶然巧合的Holsti's公式相比，这被认为是更优越的。对于具有比率或间隔级别数据的实验，例如1到5点的评分尺度，可以使用Krippendorf的Alpha和Fleiss' Kappa。这些方法还允许超过两名观察者的情况。

一般来说，80%（.80）或更高的一致性被认为是可以接受的，虽然有时70%（.70）也可以接受。存在免费在线的观察者一致性计算器，比如：

•ReCal2，GraphPad QuickCalcs和VassarStats适用于两名评分员（http://dfreelon.org/utils/recalfront/recal2/和https://www.graphpad.com/quickcalcs/kappa1.cfm；http://vassarstats.net/kappa.html）；

• ReCalF3适用于三名或三名以上评分员（http://dfreelon.org/utils/recalfront/recal-oir/）。

**其他不显眼的仪器**

其他不显眼的仪器中有许多被俗称为“奇怪措施”的方法。例如，去汽车修理店看看收音机调谐到哪些频道，是一种不显眼的方法来评估人们最喜欢的广播电台。衡量态度的一种更巧妙的非反应性仪器是“丢失信件”技术。在一项经典研究中，寄给不同宗教团体和组织（如基督教教堂、清真寺和犹太会堂）的贴上邮票的信件被散布在一个社区中。地址实际上是研究人员的邮箱，而好心人寄回给特定宗教团体的信件数量成为衡量社区居民偏见的指标。大多数这些创意性的测量方法已经不再流行，但新的方法正在取代它们。有关列表实验和实际努力任务的示例，请参阅第10.4节的研究焦点框中的详细信息。

**不显眼的问题**

虽然本节着重介绍非反应性仪器而非问卷调查，但也有可能构建不显眼的问题，这意味着受试者无法（或不太可能）修改他们的回答以符合社会期望或给出支持他人的答案。用于衡量道德判断的问题就属于这类问卷。虽然受试者知道该研究是关于做出伦理决策的，但他们被告知没有“正确”答案，研究人员关心的是在做出决策时哪些问题最重要。向受试者呈现的道德困境被设计成两种行动都可以作为道德行为进行辩护。然而，研究人员仍然能够衡量一个人的道德判断高低，因为受试者需要在重要性尺度上对十二个陈述进行评分，这些陈述代表了道德判断的六个阶段。选择较高阶段的陈述会得到比选择较低阶段的陈述更高的分数。除非受试者熟悉科尔伯格的理论，否则他们不太可能选择高阶段的陈述，除非他们已经发展到那个水平。事实上，该仪器还包括一种方法来检测受试者是否试图伪造较高的得分。当受试者无法刻意改变对问题的回答时，它们被认为是不显眼的。

**测量问题**

测量是另一个涉及到许多书籍、教科书和文章的主题，在本文中并不尝试涵盖所有内容。相反，这里只涉及对实验特别重要的主题。

**研究亮点 10.4**

**不引人注目的工具 - 列表、选择和真实努力任务实验**

实验研究者不一定要避免使用问卷，以使工具不引人注目。已经发展了几种方法，仍然使用问卷但避免直接提问，增加了真实的回答并避免了不回答的情况。其中三种方法在政治学家中很受欢迎，但适用于任何社会科学领域。事实上，下面描述的两个选择实验研究来自市场营销和食品政策领域。本篇聚焦于三种类型的不引人注目的工具 - 列表实验、选择实验和真实努力任务，但还有其他方法，如认可实验。还应该指出，这些工具可以用于调查和实验。

列表实验

Kramon展示了如何在选举购票这样敏感的议题上，列表实验在外部有效性方面更具反映现实性。列表实验的优势在于避免了社会期望偏倚，该偏倚可能促使参与者给出社会上可以接受的答案而非真实答案。它通过向受试者提供一份敏感和非敏感的活动列表来实现这一点，例如以选票换取贿赂，或者在他们居住区看到政治标志。受试者被明确告知不要指明哪些具体项目对他们来说是真实的，而只需简单地指出有多少项目对他们来说是真实的。这假设受试者在被要求给出数字而不是直接回答问题时更加真实。控制组只收到非敏感项目，而处理组除了收到与控制组相同的项目外，还有敏感项目。组间均值之差给出了敏感态度或行为的存在估计。Kramon的结果显示，在实验中报告的选举购票情况显著多于调查中的报告，与实际的选举购票事件相一致。

阅读研究文献:

Kramon, Eric. 2016. “Where Is Vote Buying Effective? Evidence From a List Experiment in Kenya.” Electoral Studies 44: 397–408. 10.1016/j.electstud.2016.09.006

选择实验

在选择实验中，受试者基于权衡和偏好做出决策，例如在薪水和福利之间进行权衡的工作选择，或其他类型的消费者决策，包括对某些属性的愿意支付。受试者根据所提供的不同产品或服务的特点做出选择，然后研究人员利用这些选择推断受试者所持有的价值观，例如更高的薪水是否比养老金更重要。

受试者被要求在选择之间进行选择，而不是对它们进行排名或评分。通过观察受试者选择的物品或服务的属性，而非直接问受试者他们的偏好，避免了社会期望偏倚。例如，在一项关于回收利用的研究中，受试者可能会选择回收还是不回收，而不是直接问受试者他们的偏好。

以上是关于不引人注目的工具的一些介绍，其中包括列表实验、选择实验和真实努力任务，但还有其他方法，例如认可实验。这些方法可以用于调查和实验。

自我报告可能会声称即使包装需要清洁，他们仍会回收，但在选择实验中，他们可能只会选择不需要清洁的包装。这就是在一项选择实验中发生的情况，其中受试者需要在两种三明治容器之间做出选择。230 这些权衡包括包装在回收前需要多少清洁，它需要分成多少部分（纸或塑料，或两者兼有），以及回收需要多长时间。研究发现，当包装需要分离时，受试者更有可能进行回收，而当包装需要清洁或回收时间更长时，受试者更不可能进行回收。

另一项关于食品政策的选择实验发现，受试者更重视标明牛肉安全性的标签，而不是牛肉来自哪个国家、嫩度或食源性疾病溯源能力。231 受试者被要求在两种肋眼牛排之间进行选择，这些肉排带有不同的每磅价格标签、食品安全检查、原产地国家、是否保证嫩度以及是否能追溯到原产农场。

专家建议在选择实验中将选择属性限制在六个以内，以避免过多地压倒受试者并增加他们的决策难度。232 例如，食品安全研究使用了五个属性。233 同时，建议研究人员使用25到90个选择集或情境，以估计主要效应和双向交互作用。234 回收研究使用了二十个不同的选择情境，每个受试者评估了五个情境。235 这类实验还应包含无选择选项，因为在实际的消费决策中常常存在这种选项。236

阅读这两项研究：

Klaiman，K.，D.L. Ortega和C. Garenc-Nelson。2017。“感知的包装回收障碍：来自美国消费者选择实验的证据。”《食品控制》73, 291–299。

Loureiro，M.L.和W.J. Umberger。2007。“牛肉选择实验模型：美国消费者对食品安全、原产地标签和溯源的相对偏好。”《食品政策》34 (4)：496–514。

真实努力任务

另一种有助于克服实验室环境中需求效应的实验仪器是真实努力任务实验。在一项研究中，研究人员试图了解增加慈善捐赠的最佳提醒频率，但又希望避免实验中的受试者在慈善捐赠方面的表现比他们在没有被观察时更多。237 这项研究持续了三个月，向参与者支付报酬，让他们执行一个任务，即计算5 x 5矩阵中的数字1的次数，并根据正确答案在在线账户上获得报酬。研究人员向受试者发送提醒，让他们查看在线余额，并告诉他们可以通过在线账户向慈善机构捐款。在线网站链接到慈善机构的网站，以便受试者可以了解更多信息。处理条件包括三种提醒间隔：无、每周一次或每月一次。结果发现，每月一次的提醒效果最好，并且捐赠的资金实际上被捐赠给了慈善机构。

阅读这项研究：

Sonntag，A.和D.J. Zizzo。2015。“关于提醒效应、退出和优势：来自在线慈善捐赠实验的证据。”《PLoS ONE》10 (8)：1–17。

验证研究

另一项研究特别引人关注，因为它测试了三种不引人注目的方法与直接问题的对比，以确定每种方法的优势和劣势。238 该研究发现，在处理敏感问题时，所有三种不引人注目的仪器（列表实验、选择实验和随机回答方法）都比直接问题更好，可以减少非响应和提供更接近实际行为（在该研究中为投票）的结果。该研究发现，随机回答技术表现最佳，但除非受试者练习过，否则他们很难完成。所有三种不引人注目的仪器在避免社会期望偏倚和非响应方面都明显优于直接问题。列表实验的缺点包括天花板效应和地板效应。该研究还提供了如何构建这三种类型仪器的详细信息。

阅读这项研究：

Rosenfeld，B.，K. Imai和J.N. Shapiro。2016。“关于敏感问题的流行调查方法的实证验证研究。”《美国政治科学杂志》60 (3)：783–802。

**测量水平**

无论收集的变量是连续的还是分类的，在实验中，变量都必须以适当的水平进行测量。对于将使用差异检验（例如t检验或方差分析）来分析数据的实验，因变量必须以连续方式进行测量，即在间隔或比率水平上进行测量，而自变量必须是分类的。由于自变量通常是处理组和对照组，因此这个问题相对较小，重点在于因变量的测量。干预变量可以以连续或分类的方式进行测量；对于这些变量，我通常推荐使用连续的测量方法。一个原因是连续的测量可以被合并成分类变量，并且可以用作自变量；例如，在一项研究中，参与度既作为连续测量的干预变量，又被分为高、中、低三个分类，并作为自变量来测试交互作用效应。尽管连续变量可以转化为分类变量，但分类变量不能成为连续变量。

**响应选择**

在问卷调查中，研究者为受试者提供的回答选择被称为响应量表。这些问题的答案选择范围从方向和强度的连续性中选择，使研究者能够为受试者分配分数。例如，Likert量表243提供了一种固定选择的回答格式，用于测量受试者对于某种观点的程度、频率、重要性或可能性，例如“非常同意”、

在问卷调查中，研究者为受试者提供的回答选择被称为响应量表。这些问题的答案选择范围从强烈同意（7）到强烈不同意（1），中间还包括一些选择，例如同意（6）、有些同意（5）、既不同意也不同意（4）、有些不同意（3）和不同意（2）。在5分量表上，"有些"的选择会被省略。回答选择的措辞会根据问题的情况进行变化，以反映最合适的回答。例如，对于“您体验以下情绪的程度有多少”，回答选择范围将从从不到很多。与问题措辞一样，推荐跟随以前的研究成果；通常，已经被多次使用并显示出可靠性和有效性的问题会附带一个固定的回答量表。量表比二元的回答选项（如是/否、对/错）或其他分类回答（如态度、意见、信念、情感等问题的三种选择：是/否/可能）更可取。尤其在政治方面的研究中，研究表明，量表比二元的测量更能捕捉人们对于自己信念和态度的不确定性。244 分类问题，例如一个人所属的政治党派，可以用分类来回答，但通常也会通过党派倾向的强度（您认为自己是坚定的共和党人/民主党人还是不那么坚定的？）和与哪个党派更亲近（民主党、共和党、无党派）来进行衡量。

另一方面，情感温度计的范围从0到100；例如，典型的政治学情感温度计如下：“请在下面的温度计上对您读到的两个候选人进行评分。温度计从0到100度。50分以上表示您对这个人有好感和亲切。50分以下表示您对这个人有反感和冷淡。”245

奇数的响应选择数量允许设立一个中立点。对于这是否好还是不好存在争议，这也取决于所测量的内容。例如，使用中立选项可能会遗漏真实的态度。246 有些问题不需要中立选项，或者会允许真实观点的受试者不回答问题。例如，“您对政治和公共事务有多感兴趣”这个问题在逻辑上就没有中立选项；非常感兴趣、有些感兴趣、稍微感兴趣和一点都不感兴趣这四个选择涵盖了所有可能。

避免天花板效应和地板效应也是很重要的，需要足够的响应选择。天花板效应是指变量未被高度测量，因此分数聚集在顶部，因为受试者不能再作出更高的回答，例如收入调查的最高回答是10万美元。现在，拥有10万美元或以上收入的人数明显增多，因此如果继续使用此限制，可能会导致天花板效应。

天花板效应和地板效应还可能发生在Likert量表回答和分类回答中；例如，一项研究使用了5分量表，并发现了天花板效应。247 或许选择7分量表可能更合适。

**构造效度**

如第5章所述，效度是测试一个工具测量其所声称的内容程度的能力。那一章专注于效度的理论概念，而这一节则讨论了如何测试效度。

使用之前使用过的问题的一个好处是它们很可能经过了测试，以确保它们测量了它们所声称的内容，这被称为构造效度，它在第5章首次介绍。其他类型的效度包括面对效度、收敛效度和区分效度，读者可以参阅“建议阅读”部分的其他文献了解更多内容。

构造效度表示背后的理论观念或概念，例如“人格”、“幸福”或“抑郁”。测试必须测量其声称的内容，而不是类似的内容。构造效度可以通过专家评估来衡量，看仪器测量构念的程度，或通过与测量相同或相似构念的其他工具进行比较。

构造效度可以通过统计方法来评估，以显示是否存在一个共同因素，它潜在地贯穿组成复合指数或由多个问题组合而成的情况。为形成一个指数，通常将采用相同回答集的问题求和和平均，或者除以总问题数。然而，在构建指数之前，必须确定所有问题是否"相互关联"或与彼此相关，并且代表相同的概念。一个简单的方法是看测量一个概念的多个问题是否相互关联。然而，相关系数很少用于确定这一点；相反，用于确定效度的特殊统计方法包括因子分析等。克伦巴赫的阿尔法是衡量同一构念的两个测试之间的相关性，用作内部一致性可靠性的估计。这些统计工具的优势在于它们通过识别不应该包含的项目来提高指数的强度。

**研究焦点10.4**

**不引人注目的工具 - 列表实验、选择实验和真实努力任务实验**

实验研究者不一定必须避免使用问卷才能使工具不引人注目。已经发展了几种方法，仍然使用问卷但避免直接提问，从而增加了诚实回答并避免不回应。225其中有三种流行的方法适用于政治学家，但也适用于任何社会科学领域。实际上，这里描述的两个选择实验研究来自市场营销和食品政策。本焦点主要关注三种不引人注目的工具 - 列表实验、选择实验和真实努力任务实验，但还有其他一些，比如认可实验。值得注意的是，这些工具既可以用于调查，也可以用于实验。

列表实验

Kramon226证明了列表实验在外部有效性方面更具反映性，特别是在涉及敏感问题的选举购票问题上。列表实验的优势在于避免了社会期望偏见，这种偏见可能会促使参与者给出社会上可接受的回答，而不是真实的回答。它通过给被试一个既包含敏感内容又包含非敏感内容的列表，例如是否接受贿赂换取选票，或是否在自己居住区看到政治标志等，然后明确告知被试不要标识哪些项目适用于自己，而只需简单地指出有多少项目适用于自己。这样假设被试在被要求给出数量时比直接回答问题时更为真实。对照组只接受非敏感项目，而处理组除了接受与对照组相同的项目外，还包含敏感项目。组间均值之间的差异给出了敏感态度或行为存在的估计值。227 Kramon的研究结果显示，在实验中购票报告显著多于调查，这与实际选举购票的情况一致。

阅读研究内容：

Kramon，Eric。2016年。“选票购买在哪里有效？肯尼亚列表实验的证据。”《选举研究》44：397-408。 10.1016/j.electstud.2016.09.006

选择实验

在选择实验中，被试基于权衡和偏好做出决策 - 例如，在薪水和福利之间做出选择的工作机会，228或其他类型的消费者决策，包括愿意为某些属性付费。被试根据不同产品或服务的特征做出选择，研究者随后使用这些选择来推断被试所持有的价值观 - 例如，较高的薪水是否比养老金更重要。

被试被要求在不同选择之间进行选择，而不是对它们进行排名或评级。229 这避免了社会期望偏见，因为不是直接询问被试的偏好，而是观察他们选择的事物与未选择的事物之间的属性。例如，在回收研究中，受访者可能会选择是否回收。

自我报告可能会声称即使包装需要清洁，他们仍会回收，但在选择实验中，他们可能只会选择不需要清洁的包装。这就是在一项选择实验中发生的情况，其中受试者需要在两种三明治容器之间做出选择。230 这些权衡包括包装在回收前需要多少清洁，它需要分成多少部分（纸或塑料，或两者兼有），以及回收需要多长时间。研究发现，当包装需要分离时，受试者更有可能进行回收，而当包装需要清洁或回收时间更长时，受试者更不可能进行回收。

另一项关于食品政策的选择实验发现，受试者更重视标明牛肉安全性的标签，而不是牛肉来自哪个国家、嫩度或食源性疾病溯源能力。231 受试者被要求在两种肋眼牛排之间进行选择，这些肉排带有不同的每磅价格标签、食品安全检查、原产地国家、是否保证嫩度以及是否能追溯到原产农场。

专家建议在选择实验中将选择属性限制在六个以内，以避免过多地压倒受试者并增加他们的决策难度。232 例如，食品安全研究使用了五个属性。233 同时，建议研究人员使用25到90个选择集或情境，以估计主要效应和双向交互作用。234 回收研究使用了二十个不同的选择情境，每个受试者评估了五个情境。235 这类实验还应包含无选择选项，因为在实际的消费决策中常常存在这种选项。236

阅读这两项研究：

Klaiman，K.，D.L. Ortega和C. Garenc-Nelson。2017。“感知的包装回收障碍：来自美国消费者选择实验的证据。”《食品控制》73, 291–299。

Loureiro，M.L.和W.J. Umberger。2007。“牛肉选择实验模型：美国消费者对食品安全、原产地标签和溯源的相对偏好。”《食品政策》34 (4)：496–514。

真实努力任务

另一种有助于克服实验室环境中需求效应的实验仪器是真实努力任务实验。在一项研究中，研究人员试图了解增加慈善捐赠的最佳提醒频率，但又希望避免实验中的受试者在慈善捐赠方面的表现比他们在没有被观察时更多。237 这项研究持续了三个月，向参与者支付报酬，让他们执行一个任务，即计算5 x 5矩阵中的数字1的次数，并根据正确答案在在线账户上获得报酬。研究人员向受试者发送提醒，让他们查看在线余额，并告诉他们可以通过在线账户向慈善机构捐款。在线网站链接到慈善机构的网站，以便受试者可以了解更多信息。处理条件包括三种提醒间隔：无、每周一次或每月一次。结果发现，每月一次的提醒效果最好，并且捐赠的资金实际上被捐赠给了慈善机构。

阅读这项研究：

Sonntag，A.和D.J. Zizzo。2015。“关于提醒效应、退出和优势：来自在线慈善捐赠实验的证据。”《PLoS ONE》10 (8)：1–17。

验证研究

另一项研究特别引人关注，因为它测试了三种不引人注目的方法与直接问题的对比，以确定每种方法的优势和劣势。238 该研究发现，在处理敏感问题时，所有三种不引人注目的仪器（列表实验、选择实验和随机回答方法）都比直接问题更好，可以减少非响应和提供更接近实际行为（在该研究中为投票）的结果。该研究发现，随机回答技术表现最佳，但除非受试者练习过，否则他们很难完成。所有三种不引人注目的仪器在避免社会期望偏倚和非响应方面都明显优于直接问题。列表实验的缺点包括天花板效应和地板效应。该研究还提供了如何构建这三种类型仪器的详细信息。

阅读这项研究：

Rosenfeld，B.，K. Imai和J.N. Shapiro。2016。“关于敏感问题的流行调查方法的实证验证研究。”《美国政治科学杂志》60 (3)：783–802。

**测量水平**

无论收集的变量是连续的还是分类的，在实验中，变量都必须以适当的水平进行测量。对于将使用差异检验（例如t检验或方差分析）来分析数据的实验，因变量必须以连续方式进行测量，即在间隔或比率水平上进行测量，而自变量必须是分类的。由于自变量通常是处理组和对照组，因此这个问题相对较小，重点在于因变量的测量。干预变量可以以连续或分类的方式进行测量；对于这些变量，我通常推荐使用连续的测量方法。一个原因是连续的测量可以被合并成分类变量，并且可以用作自变量；例如，在一项研究中，参与度既作为连续测量的干预变量，又被分为高、中、低三个分类，并作为自变量来测试交互作用效应。尽管连续变量可以转化为分类变量，但分类变量不能成为连续变量。

**响应选择**

在问卷调查中，研究者为受试者提供的回答选择被称为响应量表。这些问题的答案选择范围从方向和强度的连续性中选择，使研究者能够为受试者分配分数。例如，Likert量表243提供了一种固定选择的回答格式，用于测量受试者对于某种观点的程度、频率、重要性或可能性，例如“非常同意”、

在问卷调查中，研究者为受试者提供的回答选择被称为响应量表。这些问题的答案选择范围从强烈同意（7）到强烈不同意（1），中间还包括一些选择，例如同意（6）、有些同意（5）、既不同意也不同意（4）、有些不同意（3）和不同意（2）。在5分量表上，"有些"的选择会被省略。回答选择的措辞会根据问题的情况进行变化，以反映最合适的回答。例如，对于“您体验以下情绪的程度有多少”，回答选择范围将从从不到很多。与问题措辞一样，推荐跟随以前的研究成果；通常，已经被多次使用并显示出可靠性和有效性的问题会附带一个固定的回答量表。量表比二元的回答选项（如是/否、对/错）或其他分类回答（如态度、意见、信念、情感等问题的三种选择：是/否/可能）更可取。尤其在政治方面的研究中，研究表明，量表比二元的测量更能捕捉人们对于自己信念和态度的不确定性。244 分类问题，例如一个人所属的政治党派，可以用分类来回答，但通常也会通过党派倾向的强度（您认为自己是坚定的共和党人/民主党人还是不那么坚定的？）和与哪个党派更亲近（民主党、共和党、无党派）来进行衡量。

另一方面，情感温度计的范围从0到100；例如，典型的政治学情感温度计如下：“请在下面的温度计上对您读到的两个候选人进行评分。温度计从0到100度。50分以上表示您对这个人有好感和亲切。50分以下表示您对这个人有反感和冷淡。”245

奇数的响应选择数量允许设立一个中立点。对于这是否好还是不好存在争议，这也取决于所测量的内容。例如，使用中立选项可能会遗漏真实的态度。246 有些问题不需要中立选项，或者会允许真实观点的受试者不回答问题。例如，“您对政治和公共事务有多感兴趣”这个问题在逻辑上就没有中立选项；非常感兴趣、有些感兴趣、稍微感兴趣和一点都不感兴趣这四个选择涵盖了所有可能。

避免天花板效应和地板效应也是很重要的，需要足够的响应选择。天花板效应是指变量未被高度测量，因此分数聚集在顶部，因为受试者不能再作出更高的回答，例如收入调查的最高回答是10万美元。现在，拥有10万美元或以上收入的人数明显增多，因此如果继续使用此限制，可能会导致天花板效应。

天花板效应和地板效应还可能发生在Likert量表回答和分类回答中；例如，一项研究使用了5分量表，并发现了天花板效应。247 或许选择7分量表可能更合适。

**构造效度**

如第5章所述，效度是测试一个工具测量其所声称的内容程度的能力。那一章专注于效度的理论概念，而这一节则讨论了如何测试效度。

使用之前使用过的问题的一个好处是它们很可能经过了测试，以确保它们测量了它们所声称的内容，这被称为构造效度，它在第5章首次介绍。其他类型的效度包括面对效度、收敛效度和区分效度，读者可以参阅“建议阅读”部分的其他文献了解更多内容。

构造效度表示背后的理论观念或概念，例如“人格”、“幸福”或“抑郁”。测试必须测量其声称的内容，而不是类似的内容。构造效度可以通过专家评估来衡量，看仪器测量构念的程度，或通过与测量相同或相似构念的其他工具进行比较。

构造效度可以通过统计方法来评估，以显示是否存在一个共同因素，它潜在地贯穿组成复合指数或由多个问题组合而成的情况。为形成一个指数，通常将采用相同回答集的问题求和和平均，或者除以总问题数。然而，在构建指数之前，必须确定所有问题是否"相互关联"或与彼此相关，并且代表相同的概念。一个简单的方法是看测量一个概念的多个问题是否相互关联。然而，相关系数很少用于确定这一点；相反，用于确定效度的特殊统计方法包括因子分析等。克伦巴赫的阿尔法是衡量同一构念的两个测试之间的相关性，用作内部一致性可靠性的估计。这些统计工具的优势在于它们通过识别不应该包含的项目来提高指数的强度。

**如何操作第10.5节**

**撰写仪器和测量部分**

继第9章描述刺激之后，方法部分包含了关于工具和测量的详细信息，以及执行它们的步骤。所有因变量、协变量、中介变量和调节变量的测量应该给出问卷的确切措辞。人口统计信息通常不需要如此详细的描述；有些文章会按名称指定人口统计信息（例如年龄、教育程度、性别），而其他文章可能只会说“通常的人口统计信息”。如果有很多变量需要具体操作，可以在文本中给出例子，并在附录中提供完整的问卷。248 对于反应选择也应该提供——首次详细介绍，然后其他测量可以给出更简短的版本。例如，第一次描述时应该解释反应尺度是一个5分Likert尺度，范围从完全不同意（1）到完全同意（5），但在后续引用中可以更简洁地解释为“相同的5分尺度”，或者用括号括起来的端点（例如，1  从不，5  始终如此）。

Cronbach's alpha或因子分析用于指数的测量以及形成方式的解释也必须报告。一些期刊使用符号，而其他期刊则拼写出统计测试名称（例如，Cronbach's alpha vs. 符号α）。有些期刊将其包含在方法部分，而其他期刊则将其放在结果部分；请遵循提交论文的期刊的风格。工具和测量的报告应该足够全面，以便其他研究人员可以在不联系作者的情况下复制研究。以下是作者如何描述其工具和测量的一些示例。这里只包含了一些常见的社会科学工具；其他工具，请参阅本章引用的任何研究。

示例1：问卷

这个实验使用了一个问卷，文中给出了工具项和指标的统计数据。请注意，还给出了反应选择的描述。

“对当前舆论的评估。我们用两个项目（“目前，科隆的大多数市民反对延长”[CC 1，当前舆论1]和“现在科隆的市民不希望延长快线”[CC 2，当前舆论2]）来衡量参与者对当前舆论的感知。参与者可以用五点式Likert尺度回答，范围从1（我完全不同意）到5（我完全同意）。这两个项目构成了衡量参与者对人群内部分歧感知的尺度（斯皮尔曼-布朗系数=0.70，M=3.75，SD=1.21）。”249

来自：Zerback, T., T. Koch, and B. Kramer. 2015. “Thinking of Others: Effects of Implicit and Explicit Media Cues on Climate of Opinion Perceptions.” Journalism and Mass Communication Quarterly 92 (2): 421–443.

示例2：问卷

这个实验也使用了一个问卷，包含一个情感温度计和另外两个测量，其确切措辞在附录中。它在括号中给出了指数的可靠性统计数据。

“候选人评估指数由九个变量组成：一个情感温度计、六个特质评估和两个关于麦肯纳是否适合担任职务的同意-不同意陈述。麦肯纳的评估形成了一个非常可靠的指数（α=0.924）。 ”250

来源：Thorson, Emily. 2016. “Belief Echoes: The Persistent Effects of Corrected Misinformation.”

政治传播 33 (3): 460–480。

示例3：构造效度

这个实验在文中报告了更多来自因子分析和Cronbach's alpha的细节。

“为了确保构造效度、收敛效度和鉴别效度（Perdue和Summers 1986），我们进行了一项验证性因子分析。每个因子的项目只在该因子上加载。我们估计了Cronbach（1951）的alpha值，以衡量构成一个因子的任何一组指标的可靠性。除了一个构念的值大于0.70（Nunnally 1978），支持内部一致性；我们还接受了知觉自我效能构念，其Cronbach's alpha为0.61，大于0.60（Robinson等人1991年）。最后，为了评估鉴别效度，我们比较了平均提取的方差和所有其他构念的相关系数的平方；平均提取的方差总是较高，表明具有鉴别效度（Fornell和Larcker 1981；请参见附录中的表4）。因此，因子分析显示了所有变量的令人满意的结果（Bagozzi等人1991；Cronbach等人1963；Nunnally 1978）。我们在附录中提供了构念和结果，见表3。”251

来源：Thaler, Julia, and Bernd Helmig. 2013. “Promoting Good Behavior: Does Social and Temporal Framing Make a Difference?” Voluntas: International Journal of Voluntary and Nonprofit Organizations 24 (4): 1006–1036.

示例4：眼动追踪-非商业用途（2023）

对于技术仪器，与问卷相比，对仪器和程序的描述可以更长。以下是一个眼动追踪研究的例子。

“实验在个别会话中进行。参与者口头被告知实验是关于视觉感知的。在计算机屏幕上给出了进一步的说明。说明中指出，参与者将看到几幅图片，他们的任务是对这些图片进行评分。说明还向参与者解释了实验的过程：（a）他们会看到屏幕中央的十字，（b）当他们看着十字时，十字会消失，（c）图片会呈现几秒钟，（d）然后他们会被要求给出评分。

“呈现屏幕是一台22英寸的显示器，分辨率为1680 x 1050像素，刷新率为60 Hz。参与者距离显示器约60厘米。照明保持恒定，参与者佩戴耳机以控制背景噪音。任务期间的灰色背景与图片的平均亮度相同（RGB：127,127,127）。

“实验由40个试次组成。每个试次从后续图片呈现的位置开始。在参与者注视十字1,500毫秒后，出现一个空白屏幕500毫秒。然后，在屏幕上呈现四类别（中立和积极的非购物情境，享乐和功利购物情境）中的40个图片之一，持续4,000毫秒。图片的顺序完全随机。在图片呈现期间，我们使用SMI RED 500遥感眼动追踪仪（Sensomotoric Instruments GmbH，德国Teltow）测量瞳孔直径，采样率为250 Hz。在图片呈现结束后，参与者需要对该图片进行评分。

在这个例子中，使用了潜伏性回应的研究，描述了实验过程以及仪器如何记录测量结果。

“依赖性测量：反应潜伏期。对于在词汇判断任务中的参与者，记录下的反应时间是隐含刻板印象激活的一个度量。参与者接收目标刺激，可以是一个单词（例如，聪明）或一个非单词（例如，tinpy）。所有的非单词的长度都与单词相匹配。参与者必须尽可能快速和准确地判断目标刺激是单词还是非单词，使用计算机键盘上的相应按键。该任务的第一轮是一个练习环节，参与者在此会得到反馈。在练习轮中，有24个目标刺激（单词或非单词）随机出现。每个目标刺激前面有一个感叹号，持续300毫秒。其中的12个刺激是与刻板印象无关的中性词汇（例如after、complete、more）。剩余的刺激则是等长的非单词。

“一旦练习结束，会随机出现72个目标刺激。其中16个刺激是基于预测试的刻板印象词汇（例如lazy、spiritual、poor）。其余的48个刺激包括等量的中性词汇（例如lady、delicious、pour）和与刻板印象词汇等长的非单词（例如linr、foigp-nafe、pkid）。在这些刻板印象词汇中，有些与善意的刻板印象相关（例如traditional、polite），而其他与敌意的刻板印象相关（例如lazy、uneducated）。每位参与者都有一个输出文件，记录了每个刺激的反应速度（以毫秒为单位）。”

来源：Ramasubramanian, Srividya. 2007. “Media-Based Strategies to Reduce Racial Stereotypes Activated by News Stories.” Journalism and Mass Communication Quarterly 84 (2): 249–264.

i还有其他类型的量表，包括语义差异量表、瑟斯通量表和古特曼量表。

**例子 6: 心理研究和思维列举**

和其他不引人注目的测量方法一样，心理研究或思维列举工具描述了受试者所接受的指导和遵循的程序。例如：

“然后，参与者被要求使用搜索引擎完成四个搜索任务，并遵循心理研究协议。虽然他们没有被告知使用哪个搜索引擎，但他们都选择了伊朗流行的谷歌通用搜索引擎。他们接受了心理研究技术的简要介绍（即大声思考他们的决策、行动、情感和思维）。下面是这些任务：这些任务是有意设计的，以便在搜索引擎结果中过滤出一些结果，这些结果是由政府过滤的，因此当参与者点击其中一些链接时，他们无法访问页面，并被重定向到peyvandha.ir网站，表明页面已被政府屏蔽。

搜索任务：

假设由于课程作业或个人兴趣，您需要以下项目的信息。请打开浏览器，使用搜索引擎查找有关这些主题的相关信息。在完成任务时，大声思考您的行动、决策、情感和思维：

1. 您需要有关Shojaeddin Shafa先生的传记、信仰和作品样本的信息；

2. 您需要有关新闻记者、博主和当代研究人员Mehdi Jami先生的思想和观点以及他的作品样本的信息；

3. 您需要寻找一些与20世纪30年代末纳粹主义相关的信息；和

4. 找到有关伊朗社会主义的简要历史或描述。”253

来源：Jamali, H. R., and P. Shahbaztabar. 2017

无论是创建新的问题来形成指标，还是使用以前研究中的指标，实验都应该验证研究中使用的所有指标，并在方法或结果部分报告（参见How To Do It框10.5）。在实际实验之前，进行试点研究是测试新的构念及其指标的好方法。下一章还涉及从IRB获得批准的问题，这是在从任何受试者收集数据之前必须进行的步骤。与本章，特别是关于不显眼仪器的部分相关的另一个重要主题是获得受试者的知情同意，保护他们免受伤害以及其他与实验有关的伦理考虑。

**常见错误：**

• 在更适合使用不显眼的仪器时使用自我报告

• 在没有适当的理论理解或方法论培训的情况下使用技术仪器

• 使用单个问题来代表复杂的理论构建；应该使用多个项目组合成指数

• 当已经存在经过验证的项目和指数时，使用研究人员新创建的问题

**测试你的知识**

1. 在大多数情况下，多个问题的指数\_\_\_\_\_\_\_\_。

a. 不比单个项目指标更好

b. 用于测量多层次构建的优越选择

c. 有助于缩短研究时间

d. 存在更多可靠性和有效性问题

2. 单个项目指标是\_\_\_\_\_\_\_\_。

a. 受试者用来表示其回答的设备

b. 直接提问的问题

c. 多个问题合并以测量单个项目

d. 用于测量多层次构建的劣质选择

3. 问卷调查的优点是\_\_\_\_\_\_\_\_。

a. 快速、廉价、容易进行

b. 可以保证私密性

c. 减少需求效应

d. 所有上述

4. 问卷调查的缺点是\_\_\_\_\_\_\_\_。

a. 自我报告可能不准确

b. 可能受到社会期望偏见的影响

c. 对于人们可能无法自省的事物不适用

d. 所有上述

5. 功能性磁共振成像，即fMRI，是一种非侵入性的脑部扫描仪，用于

a. 测量受试者回答问题的速度

b. 检测受试者是否在说谎

c. 识别与信任或情绪等某些心理过程相关的脑区

d. 测量受试者在网页上停留的时间

6. 当研究人员观察学生玩耍和与他人互动，并记录所见内容时，他们使用的是哪种仪器？

a. 反应时间测量

b. 问卷调查

c. 观察

d. 监视屏幕

7. 在一项研究中，教师对学生进行行为评估，包括争论、抱怨、取笑、分享、违规和不轮流等行为，是什么确保他们以相同的方式进行评估？

a. 操作检查

b. 观察者间可靠性

c. 克朗巴赫α系数

d. 因子分析

8. 在实验中，自变量在哪个层次上进行测量？

a. 类别

b. 区间

c. 比率

d. 连续

9. 在一般实验中，因变量必须在哪个层次上进行测量？

a. 类别

b. 连续

c. 名义

d. 序数

10. 协变量必须在哪个层次上进行测量？

a. 名义

b. 序数

c. 间隔

d. 以上皆是

**应用题**

1. 阅读两篇自己选择的实验性期刊文章，然后分析和评述它们的方法部分。辨识其中的内容，包括使用的仪器和程序。例如，回应选择有哪些？哪些构念使用指数，哪些使用单项指标？确切的问题措辞在文章的哪个部分报道——是在方法文本中，脚注，附录，还是在线附加材料中？还要考虑研究中可能存在的偏见和控制措施，以及研究的可信性和有效性。在使用的工具中，是否与社会期望、误解和内省无能力等问题相关？是否有足够的信息供您复制？其中一篇文章的方法部分是否比另一篇写得更好？如果是，原因是什么？此任务的目的是对方法部分进行批判性思考，以便您能够自己完成。

2. 对于您在本书中进行的研究，设计一个合适的仪器和测量方法来测试您的假设或研究问题。编写三页描述该仪器，包括合适的测量水平和回答选择。讨论您将如何验证构建的有效性。如果您的主要工具是问卷调查，请简要介绍一个无干扰的工具，也可用于您的研究。

**推荐阅读**

Question Writing

Tourangeau, R. 2004. “Experimental Design Considerations for Testing and Evaluating Questionnaires.” In

Methods for Testing and Evaluating Survey Questionnaires, edited by S. Presser, J. M. Rothgeb, M. P. Couper, J.

T. M. E. Lessler, J. Martin, and E. Singer, 209–224. Hoboken, NJ: Wiley.

Single-Item Indicators

Petrescu, Maria. 2013. “Marketing Research Using Single-Item Indicators in Structural Equation Models.”

Journal of Marketing Analytics 1 (2): 99–117.

fMRI

This paper gives guidelines on why and how to conduct fMRI studies.

Dimoka, Angelika. 2012. “How to Conduct a Functional Magnetic Resonance (fMRI) Study in Social Science

Research.” MIS Quarterly 36 (3): 811–840.

Eye Tracking

Jankowski, Jaroslaw, Jaroslaw Watrobski, Katarzyna Witkowska, and Pawel Ziemba. 2016. “Eye Tracking

Based Experimental Evaluation of the Parameters of Online Content Affecting the Web User Behavior.”

In Selected Issues in Experimental Economics, edited by K. Nermend and M. Latuszynska, 311–332. Cham,

Switzerland: Springer.

Technological Instruments

In addition to chapters on each instrument, there is a chapter on how to set up a lab for these measures.

Lang, Annie, ed. 1994. Measuring Psychological Responses to Media. Hillsdale, NJ: Erlbaum.

Interobserver Reliability

For a tutorial, see:

Hallgren, K. A. 2012. “Computing Inter-Rater Reliability for Observational Data: An Overview and Tutorial.”

Tutorials in Quantitative Methods for Psychology 8 (1): 23–34.

More on Validity

Messick, S. 1995. “Validity of Psychological Assessment: Validation of Inferences From Persons’ Responses and

Performances as Scientific Inquiry Into Score Meaning.” American Psychologist 50: 741–749.

**第十一篇 机构审查委员会与进行道德实验**

*实验方法是科学解释的最理想方法，因为它能够最大程度上排除干扰因素，提供可靠的数据，从而更好地进行科学研究。*

*—A. Lijphart*

**学习目标**

**• 设计一个实验仪器，包括适当的测量水平和回答选择，并验证其构造的有效性。**

**• 描述机构审查委员会的目的及在实验中的原则。**

**• 解释如何避免欺骗并减轻实验中的影响。**

**• 举例说明社会科学实验中的伤害和滥用权力，并减少这些影响的方法。**

**• 确定社会科学实验保护受试者隐私和保密性的方法。**

**• 创建机构审查委员会的建议书。**

**• 开展初步研究。**

尽管开篇引用声明在四十五年前，政治学家和其他人确实找到了办法克服实验的实际和伦理障碍，使其几乎成为社会科学中的增长产业。然而，即使是实验的语言——"操纵"、"控制"和"受试者"——也暗示了研究人员对参与者的权力和权威，伴随着对公平和尊重对待他们的关切。本章扩展了第2章中有关伦理的讨论，探讨这些伦理障碍和一些修正措施，以及获取机构审查委员会批准和进行初步研究的过程。

**机构审查委员会**

机构审查委员会，简称IRB，是负责审查研究提案的委员会，以确保研究人员在伦理上对待受试者，并不会给他们造成伤害。大多数发达国家的大学都设有IRB，而在美国，联邦资助的研究在大学以外进行时，也需要经过IRB的批准和监督。IRB最早是在1970年代为了回应纳粹医学实验、Tuskegee梅毒实验等引发争议的研究而设立的。在Tuskegee梅毒实验中，黑人男性受试者在四十年的时间里未接受治疗，即使已知有治疗方法，而有些人甚至从未被告知自己患有该疾病。医务人员声称他们只是在履行职责或按照命令行事，这与纳粹在大屠杀期间的说辞相似。因为这些和其他实验，为了避免此类滥用，开始着手批准、监督和规范研究。由负责阐述人体研究基本伦理原则的委员会制定的Belmont报告，是IRB用来制定研究人员必须遵守的准则的基础。其三个基本原则分别是对个体的尊重，包括不强迫或欺骗受试者并提供知情同意；善行，涉及不伤害受试者，将风险降到最低，最大化利益；和公正，包括以公平的方式选择受试者和执行程序。

IRB最初是为了回应生物医学实验的滥用而设立的，但社会科学研究也有着道德上的失误，包括第2章中描述的实验，例如斯坦利·米尔格拉姆的电击实验和菲利普·辛巴多的斯坦福监狱研究。米尔格拉姆的研究经常被引用为对实验的主要批评之一的例子：欺骗（详见第2章中的更多信息框2.3《欺骗》）。具体而言，研究人员虚假地告诉受试者实验的目的

在Stanford监狱实验中，主要的伦理问题是心理压力导致的伤害。这个实验的目的是测试惩罚是否能提高学习效果，而不是测试对权威的服从。因此，受试者在参与研究时并不真正知道他们在同意什么。此外，受试者被鼓励伤害他人，这导致了许多人，如果不是大多数人，都产生了心理困扰。

在Stanford监狱实验中，主要的伦理问题是心理压力导致的伤害。这个实验的目的是测试惩罚是否能提高学习效果，而不是测试对权威的服从。因此，受试者在参与研究时并不真正知道他们在同意什么。此外，受试者被鼓励伤害他人，这导致了许多人，如果不是大多数人，都产生了心理困扰。

这些伤害和知情同意等问题，尤其在实验中尤为突出，将在本章中在IRB的背景下讨论。由于IRB主要是为了监管生物医学研究而设立的，因此社会科学家有些不满，他们认为IRB将不适当的标准应用于他们的领域，导致对研究进行妥协的要求。因此，理解和与IRB协商的更多信息，请参见研究重点11.1。

首先，重要的是要声明，伦理研究不仅仅是IRB允许的。虽然IRB的规则和原则旨在涵盖重要的伦理关切，但有原则的实验不仅仅是遵循这些规则。从最初的想法到整个过程的每一步，设计一个符合伦理的实验应该是一个人的目标。然而，因为IRB是评估和批准伦理的实验的机构，其标准将是本章讨论的框架。申请IRB批准在每个机构都是不同的，所以本章不会尝试逐步介绍申请过程，而是在更多信息框11.2《IRB》中讨论所有IRB共有的步骤。以下是一些关于实验的常见伦理问题以及最小化或解决这些问题的建议。

**研究重点11.1**

**社会科学研究中的风险和错误**

来源：Oakes, J. Michael. 2002. “社会科学研究中的风险和错误：对IRB的评估者指南。”《评估评论》26（5）：443页。

社会科学家对最初为生物医学研究设计的IRB流程表达了不满。在本文中，作者解释了IRB的理念，帮助读者理解其规则，并给出了应对该过程的建议。

许多社会科学家认为IRB不理解社会科学调查及其方法论。作者给出了两个原因：“规章制度是为生物医学研究而制定的。对规章的解释变得更加严格。”他指出，社会科学家在处理研究伦理问题时被迫遵守与他们的研究方法不相符的规则，这可能导致不适当的审查和限制研究的深度和创新性。作者强调了在处理这些问题时的重要性，以确保IRB流程适应社会科学研究的需求。

参与制定IRB规则的社会科学家并未涉足其中。尽管缺乏此类风险的证据，但风险重点已扩大至非身体伤害。然而，他说：“社会科学研究中的风险和错误缺乏数据并不意味着受试者没有受到伤害。”他列举了本章中包括的许多伤害风险，包括侵犯隐私和污名化，并给出了具体例子，如受家庭暴力的受害者失去保密性，让施暴者能够找到她，以及患有创伤后应激障碍的受害者在回答问题时出现闪回。

IRB越来越严格的原因之一是面临罚款、法律诉讼的威胁，以及因一项研究的违规而使所有研究在某个机构中被暂停。我个人在大学附属医学院的研究人员意识到病人正在随机分配接受温暖或冷血液输注时亲身经历过这一点。他们收集了以前收集的数据，进行了分析，并发表了一项研究，所有这些都没有得到使用数据的患者的知情同意。整个大学被禁止在一段时间内获得任何新的研究批准，这甚至对正在努力完成博士学位论文的新闻学博士生造成了困扰。

除了IRB的行动令社会科学家感到沮丧外，奥克斯也责备社会科学家。他说：“简言之，许多研究人员仍然相信，或许是下意识地认为，他们有无可剥夺的研究权利。也就是说，他们有权研究其他人类，不受IRB的干预。”他将IRB比作有权要求研究作者进行更改的期刊编辑。他提出的解决方案是双方进行教育：社会科学家必须了解IRB，IRB必须了解社会科学。他继续提供关于IRB的经验主义研究，例如定义均衡性——即真正不确定一种治疗是否比另一种治疗更好——最小风险和弱势受试者。IRB制定了在更多关于IRB的框框里解释的豁免类别，适用于社会科学研究。该类别豁免了对教育实践评估、不收集任何身份数据的调查研究以及观察研究的监管。欺骗性实验并未豁免，测试干预措施的实验也未豁免。

关于受试者的道德招募也得到讨论，"公平选择"是IRB（机构审查委员会）的首要任务之一。在研究中，贫困和少数族裔人群在历史上被过度代表，例如图斯基吉实验。然后，在20世纪70年代中期至90年代中期，情况发生了变化，开始排除妇女和少数族裔。公平选择受试者确保每个人都能享受研究所产生的知识带来的好处。实证证据提供了关于奖励金额的实用建议，例如在邮寄调查中，1至5美元的奖励金额代表了在获取合作和避免强制性之间的平衡。关于更大奖励的抽奖研究表明，它们可能是强制性的，因为人们不理解概率。作者讨论了研究团体时独特的问题，例如，研究的对象是完整的教室。问题范围从一个拒绝参与的受试者到是否有权代表所有人，例如工作场所的经理。他解释了"机密性证书"如何防止敏感数据因识别信息而被传唤。

至于教育IRB，奥克斯推荐进行实证研究，了解受试者在社会科学研究中实际经历的伤害。他对社会科学研究人员未充分尝试提供有关其研究风险的数据表示遗憾，指出"实验设计的结果将最有帮助"。他列出了一个包括最佳补偿金额在内的研究主题的长列表。最后，他提供了一个有用的小贴士表，用于改善与IRB的关系。脚注中充满了有趣和有用的信息。

**更多关于……框框 11.2**

**IRB（机构审查委员会）**

虽然每个拥有IRB的研究机构都有不同的标准，但根据联邦法规，有一些共同点。其中包括三种类型的审查：免审、加速审查和全体委员会审查。"免审"这个术语可能会被误解；它并不意味着该类别的研究不需要经过IRB的批准。相反，这意味着这些研究被认为对受试者造成的风险很小，并且免于持续监督。许多社会科学研究都属于这个类别。当研究人员申请IRB批准时，他或她将被要求提出该研究应该接受哪种类型的审查。IRB将决定是否这是最合适的监督级别，或者是否需要更多的信息。通常那些不会让受试者承担比正常生活中预期风险更高的实验，属于免审类别。当然也有一些例外情况；例如，如果受试者年龄不到十八岁、是囚犯、精神病患者或者孕妇，他们被视为脆弱群体，审查级别会提升到"加速审查"。涉及学龄儿童的研究通常属于这个类别，并且需要得到儿童的同意以及父母或监护人的知情同意。"加速审查"类别可能还包括涉及视频或音频录制的研究，而有些机构也可能将这类研究视为免审。这类研究的分类可能要求研究人员每年报告项目的进展和任何不良事件，而免审类别的报告仅在研究结束时提交，需要的信息较少。"加速审查"类别的研究不一定会比被视为免审的研究得到更快的IRB批准。"全体委员会审查"正如其名——需要整个IRB委员会在定期会议上审查研究提案。免审和加速项目只受到少数委员会成员的审查。

通常，IRB的提案要求研究人员提供一个研究提案的描述性叙述，可能需要遵循标准化模板。其他文档包括刺激、工具和测量、知情同意文件和招募脚本。如果适用，还需要解释实验过程、解密文本以及欺骗的理由，以及受试者来源和研究人员是否有足够资源进行研究。从亚马逊的Mechanical Turk网站招募受试者需要遵守一系列要求，因为这些受试者实际上是"工作者"，通过他们的MTurk ID可以获得他们的身份信息，这是支付给他们报酬的必要条件。在IRB的网站上寻找关于MTurk受试者所需语言的指导或联系代表。研究人员需要详细说明研究的好处、可能的风险以及如何将风险降到最低，以及研究人员将如何处理不良事件。受试者参与研究的时间应该如实提供，因为IRB代表对此有丰富的经验。如果您不擅长判断时间，可以自己参与研究。申请IRB批准通常要求实验设计完全完成，所有材料至少处于测试阶段。有些IRB允许对刺激物的操纵检查在获得IRB批准之前进行。

研究机构通常有不同的标准，但根据联邦法规，有一些共同点。其中包括三种审查类型：免审、加速审查和全体委员会审查。"免审"并不意味着研究在这个类别下不需要IRB批准。相反，它表示这些研究被认为对受试者的风险很小，并且免于持续监督。许多社会科学研究都属于这个类别。当研究人员申请IRB批准时，他们将被要求建议该研究应该接受哪种审查类型。IRB将决定是否这是最合适的监督级别，或者是否需要更多的信息。通常，不会让受试者承担比正常生活中预期风险更高的实验属于免审类别。然而也有例外情况；例如，如果受试者年龄不到十八岁、是囚犯、精神病患者或孕妇，他们被视为脆弱群体，审查级别会提升到"加速审查"。涉及学龄儿童的研究通常属于这个类别，并涉及到儿童的同意以及父母或监护人的知情同意。"加速审查"类别可能还包括涉及视频或音频录制的研究，但有些机构也可能将这类研究视为免审。这类研究的分类可能要求研究人员每年报告项目的进展和任何不良事件，而免审类别的报告仅在研究结束时提交，需要的信息较少。 "加速审查"类别的研究不一定会比被视为免审的研究得到更快的IRB批准。

许多期刊的编辑现在要求在提交论文时提供IRB批准的证明。这可能会给那些没有审查委员会的研究人员带来问题——例如，在美国以外的机构。如果您在没有IRB的环境中进行研究，我建议至少按照联邦准则向参与者提供知情同意。此外，按照本章的其他建议，采取符合伦理的方式行事，不使用欺骗，降低风险，保护受试者的隐私，允许他们退出研究，并向他们提供解密，会有所帮助。让独立伦理专家审查您的研究提案是另一种方法。对于未接受IRB监督的论文，期刊编辑将会对研究是否符合这些伦理准则进行评估。以下是一篇发表的论文中提供的伦理声明的例子，该论文没有建立IRB：

“根据奥地利2002年的大学法案（UG2002），该法案是在研究进行时实施的，只有医学院需要为临床试验、医疗方法应用和应用医学研究委任伦理委员会。因此，这项具体研究不需要伦理批准。尽管如此，目前的研究是一个由欧洲联盟玛丽居里FP7集成奖励资助的项目的一部分，并且该项目获得了伦理批准。由于在提交项目提案时并没有机构审查委员会负责，因此该项目得到了独立专家（维也纳大学医学伦理与法律研究所）的批准。目前的研究按照赫尔辛基宣言（修订版1983年）和维也纳大学心理学系的当地准则进行。所有参与者均提供了书面知情同意，并且可以在实验期间随时撤回而不会受到进一步的影响。实验结束时，对参与者进行了详细的解密。” 来源：Serfas，Benjamin G.，Oliver B. Büttner和Arnd Florack。2014年。“Eyes Wide Shopped：购物情境引发冲动购买者的兴奋。”PLoS ONE 9（12）：1-9。

**伦理问题在实验中**

**欺骗**

如第2章所讨论的，实验中可能最主要的伦理问题是受试者可能被欺骗。没有其他方法学像实验一样容易误导受试者，这是一个引人担忧的伦理问题。

欺骗的主要方式之一是在知情同意过程中误导受试者，让他们对实验的目的产生误解。这往往被认为是为了避免需求效应或猜测假设，从而导致受试者改变回答以取悦（或不取悦）研究人员。然而，研究人员有时认为需要欺骗，其实并非如此。使用不显眼的问题或工具，不通过直接问题指出研究目的，或者使用掩饰性问题掩盖目的，都可以帮助消除欺骗受试者的必要性（有关更多信息，请参阅第10章）。例如，在道德判断研究中，受试者被要求对道德问题作出决策，这是相当明显的；告诉他们这些情景代表其他东西是愚弄不了任何人的，会损害研究人员在受试者中的信誉。然而，不必直接提问，而是让受试者对设计成反映不同道德阶段水平的不显眼陈述进行排序，可以让研究人员真实地说研究是关于道德决策的，同时不泄露假设。研究人员还可以诚实地说没有对错答案，并告诉受试者他们对于作出决策所重要的事情感兴趣。在得到知情同意时，不必向受试者透露关于回答评分的所有细节。IRB通常会批准此类研究，并要求研究结束后告知受试者使用结果的情况，尽管如果研究人员选择，可以在研究过程中进行。设计和进行不显眼的实验可能比仅仅欺骗受试者关于研究目的更为困难和耗时，然而这是有道德的研究人员应该追求的方法。

重要的是要理解，受试者需要不知道实验的具体假设或变量之间的关系以及预期结果；相反，他们需要全面了解研究的一般主题。同时，让运行实验的实验者（如研究生助理）也不了解实验的目标和假设（即第5章中描述的双盲实验），也有助于防止受试者猜测假设。

当告诉受试者实验的真实目的会使结果无效时，应尽量减少对受试者的欺骗，欺骗不应导致伤害或困扰。此外，受试者必须进行解释，并在他们参与实验后立即告知实验的真实目的。如果他们被欺骗，必须告知他们，并解释原因。研究人员需要诚实地回答受试者提出的任何问题。在解释过程中，实验者应注意不要因为受试者被愚弄而引起尴尬。某些IRB可能不会批准需要欺骗的实验，即使有解释。

有其他方式可以在实验中欺骗受试者，而不仅限于知情同意过程，例如，当使用了参与实验的同谋者，并导致受试者误以为他们是其他受试者。米尔格拉姆的电击实验使用了一个同谋者作为“学习者”，他被认为是受试者施加的电击的接收者。利昂·费斯廷格和詹姆斯·卡尔史密斯使用了一个同谋者，假装是一个刚刚完成了无聊的实验的受试者，他的任务是告诉真正的受试者该研究是有趣和重要的。在一个更现代的例子中，一项研究涉及了不同类型的谎言，使用了受过训练的同谋者通过误导和省略信息来说谎，并与不知情的受试者配对，以研究买卖财产时面对面的欺骗。受试者在事后接受了解释，报告中没有报道任何不良反应。

**伤害**

虽然欺骗可能是实验中主要的伦理关切，但IRB的主要职责是防止对受试者造成伤害。这也是研究人员的首要责任。虽然社会科学实验可能很少在身体上伤害人们，但一些实验可能会造成损害，比如泄露保密信息、侵犯隐私、尴尬、受到污名或被贴上标签。受试者不应感到过分恐惧或受到冒犯，也不应感受到自我厌恶或毫无价值，或者面临家人、朋友、同事或雇主的嘲笑或反击的风险。心理压力是社会科学实验中最可能造成的伤害。在IRB的语言中，对参与者的伤害风险不应大于日常生活中所遇到的风险。对于许多社会科学研究，这并不是一个大问题；例如，回答有关政治态度的问题可以合理地期望是在谈论政治时正常的风险。但回答有关HIV状况、性行为、药物使用或非法行为的问题将受到IRB的特别审查。为了避免心理伤害，研究人员应避免要求受试者透露可能令人尴尬、揭示创伤事件的私人信息，或被迫采取他们没有预料到的行为，就像米尔格拉姆的电击实验中那样。当提出更容易造成伤害的敏感问题时，IRB将坚持认为这些问题对于研究至关重要，并要求研究人员尽量减少伤害。然而，“日常生活”是多样的。有些人建议将伤害风险与一个非暴力的、包含言论自由的日常生活进行比较，而这可能并不适用于所有受试者。研究人员评估伤害风险时可以使用自己的日常生活作为一个标准。24

一个IRB批准的减少实验中伤害的方法是允许受试者跳过让他们感到不舒服的问题，或者在任何时候都可以自由退出研究而不会受到任何后果。例如，如果受试者因为电视节目中的暴力场面而感到不适，停止参与实验的话仍然可以获得奖励。确保受试者参与和回答研究的信息将被保密也降低了因隐私丧失而造成的伤害风险。在知情同意文件中警告受试者可能存在的伤害风险是IRB要求的，同时还要告知他们可以跳过问题或撤回参与而不会失去任何好处。不同的人对实验操作有不同的反应，有些人可能不会受到有害影响，而其他人可能在相同的处理后会感到非常沮丧。实验员如果察觉到某个受试者反应不佳，应该介入并停止该受试者的实验。最后，能够退出研究的能力还允许受试者在研究结束后将他们的回答从分析中排除掉。

如果确实发生了伤害，可以采取的纠正措施之一是为受试者提供处理意外影响的资源，例如将学生受试者转介到学校的咨询中心，或找到类似的中心并向非学生受试者提供转介。虽然许多社会科学实验造成的伤害可能性非常小，甚至可能是微乎其微的，但完全排除不可知。并不是所有的伤害在实验进行之前都是可知的。

**滥用权力**

研究人员还应当警惕与受试者之间的权力不平衡关系。经典的例子是米尔格拉姆的电击实验，研究者代表了受试者觉得必须服从的权威人物。对于大学实验员来说，亲自招募自己课堂上的学生参与实验可能导致学生参与实验，即使他们可能不真正想要参与也为了怕得到较低的成绩而参与。避免这种情况的方法包括不使用当前正在上自己课的学生，或者不是自己指导的学生。让不参与成绩评定的第三方，比如一个不为这门课程打分的研究生，进行招募是另一种方法，还可以使用参与实验池中的学生，虽然这也有一些挑战，特别是考虑到这些学生不是完全自愿参与（更多信息见Hegtvedt 2014）。参与实验获得额外学分的学生应该确保他们的名字在所有成绩完成之前都不会被告知给教师，而且额外学分只会在课程结束时授予。最好的办法是为拥有权力的学生提供礼品卡或其他奖励，而不是额外学分。此外，希望获得额外学分但不参与实验的学生必须获得一个与实验要求不相上下的替代任务。限制奖励的数量也是为了确保激励不会变得过分。

同样重要的是避免越过强迫线。为了获得IRB批准，实验的参与必须始终是自愿的，因此允许受试者在不损失额外学分或其他激励的情况下退出或选择放弃某些方面是另一种避免滥用权力的方法。

**随机分配**

另一个引发伦理问题的实验领域是随机分配的定义特征。回顾第7章，将受试者随机分配到处理组和对照组是真实实验中最重要的事情。然而，受试者往往不会欣赏这个解决实验结果偏差问题的优雅方法。对随机分配的反对意见是，如果研究人员认为某种治疗可能比其他治疗或无治疗更有效，拒绝给每个人提供治疗是不道德的，或者至少要给那些要求治疗的人提供。教育研究人员发现学校人员和家长经常有这种观点。犯罪学、商业、社会工作和其他领域也在遇到伦理困境，当有证据表明一些受试者可能比其他人更需要治疗时。斯托克称这种批评是“不当的”，解释说研究人员不知道治疗是否真的有效，这就是为什么他们要对其进行测试的原因。显然，很少有受试者相信这一点。在一项研究中，无论是医学还是非医学背景，受试者都认为随机分配是不可接受的。他们不相信医生不会对一种治疗方法持有偏好。当研究人员不确定某种特定的治疗方法是否比其他治疗或无治疗更有效时，这被称为均衡条件。研究中的受试者对于一名律师是否真的无法确定是选择和解还是上法庭也持有相同观点。此外，研究人员几乎无法说服受试者改变观点。毕竟，很少有科学家开发的治疗和干预措施不认为它们比现有措施更有效。事实上，由于受试者反对被随机分配到治疗组或对照组，医学随机临床试验的受试者招募正在下降。第7章讨论了克服受试者对随机化的反感的一些方法，包括同意允许对照组受试者在另一次实验中接受治疗。

**保护受试者**

虽然外行可能认为随机分配是不道德的，但大多数实验研究者和IRB并不这么认为（有异议的观点见Fives等人，2015）。他们更关心的是如何保护受试者免受欺骗、伤害和滥用权力。接下来

更深入地讨论了IRB要求研究人员采取的一些方法来保护受试者。这些方法中的大多数在本章前面已经预览过，但下面将更详细地介绍。

**知情同意**

确保实验受试者获得知情同意是IRB所定义的伦理研究的核心。《贝尔蒙特报告》承认，可能有侵入性或潜在危害的研究可能会被其产生的益处所抵消，但研究人员有责任在研究开始之前充分告知受试者并获得他们的许可。同意必须是自愿的，并且向受试者提供的信息必须包括研究的目的，涉及的程序和受试者必须做什么，需要多长时间，所有可以预见的风险和危害，包括心理困扰，以及任何好处，包括个人和整个社会的好处。激励措施并不算是一种好处；相反，更常见的好处描述包括新的科学知识的创造，职业最佳实践的发展，或者对于学生来说，是一种实践性的研究经验。

知情同意有各种形式。传统的书面同意书是一张实体纸张，由研究人员和受试者签字，并为每个人保留一份副本。这通常用于实验室等现场实验。随着在线实验的增加，第二种同意形式变得越来越普遍。这被称为文件免责，这意味着物理文件或纸质同意书被免除，而不是免除受试者的同意ii。这种知情同意包含与书面同意书相同的所有信息，但在在线研究的开头作为一封附信来运作。受试者阅读完同意信息后，而不是签名，点击“我同意”。IRB可能要求研究人员说明研究将使用哪种类型的同意书，并解释原因。实际上免除同意本身也是可能的；但是，对此有更高的门槛。通常，当不可能从受试者获得同意，并且风险不大于日常生活时，IRB会批准免除知情同意的请求，例如，如果研究人员正在观察人们在地铁上阅读什么，并且无法给每个人提供书面知情同意书，在公共交通工具上被观察是通常发生的事情，而且不会带来比研究人员不进行观察时更大的伤害风险。然而，在这种情况下不给予知情同意可能会损害社会科学研究的形象。

**知情同意**

确保实验对象获得知情同意是伦理研究的关键所在，这也是伦理审查委员会（IRB）所定义的。《贝尔蒙特报告》承认，入侵性或潜在有害的研究可能会被其产生的益处所抵消，但研究人员有责任在研究开始前充分告知受试者并获得他们的许可。同意必须是自愿的，并且提供给受试者的信息必须包括研究的目的、涉及的程序以及受试者需要做什么，所需时间，所有可预见的风险和危害，包括心理困扰，并且包括个体和社会的任何好处。激励措施并不算作好处；相反，更常见地描述好处是创造新的科学知识、发展职业最佳实践，或者对学生来说是实践性的研究经验。

知情同意有多种形式。传统的书面同意书是一张物理纸张，由研究人员和受试者签字，每个人都有一份副本。这通常用于实验室等面对面实验。随着在线实验的增加，第二种同意形式变得越来越普遍。这被称为免除文件记录，意思是物理文件或书面同意书被豁免，而不是受试者的同意。这种知情同意包含与书面同意书相同的所有信息，但是在在线研究开始时作为封面信。受试者在阅读知情同意信息后，点击“我同意”。IRB可能要求研究人员说明研究将使用哪种同意书，并解释原因。实际的同意豁免本身也是可能的；然而，这样的情况有更高的门槛。当无法从受试者获得同意，而且风险与日常生活中没有更大差异时，IRB通常会批准免除知情同意的请求。例如，如果研究人员正在观察地铁上的人们在读什么，并且无法给每个人提供书面知情同意书，而在公共交通工具上被观察是常见的情况，并且没有比研究人员不观察时更大的危害风险。然而，在这种情况下未能给予知情同意可能会损害社会科学研究的形象。

虽然地铁的例子并不是一个实验，但在第10章中涵盖的实验中也有类似的不显眼的工具。仅仅因为研究人员可以在受试者不知情的情况下获取数据，并不意味着可以不提供知情同意而行事。给予受试者书面或在线同意始终优于不向他们提供任何关于研究的信息。例如，在公共实验室计算机上安装的网络追踪软件可能会记录一个人的互联网行为，而受试者甚至可能对此毫无察觉。然而，在研究人员从他们的计算机收集日志文件之前，仍应通知受试者有关研究的详细信息，并允许他们选择是否参与。这也适用于其他高度不显眼的工具，如音频或视频记录和观察等。关于豁免知情同意的要求适用于公共场所的自然观察，并且没有比正常情况更大的危害风险。此外，对于普通教育场所中的正常教育实践研究、不会使受试者的就业处于风险中且保护了机密性的工作或组织效力研究、无法对受试者造成危害的档案研究，以及匿名问卷调查等都有例外情况。所有这些研究必须具有与正常生活相同的伤害风险。40

随着越来越多的实验从实验室转移到互联网，这些伦理关注点已经增加，包括对更多个体以及甚至不参与研究的人的影响的担忧。41

**隐私与机密性**

实验研究中的另一个主要问题是隐私和机密性。这不仅涉及回答敏感问题的问题。隐私涉及的是收集了哪些数据，而机密性涉及的是数据的存储和报告方式。42只有当研究人员不知道谁参与了研究，也不收集姓名或其他身份信息时，研究才是匿名的。43当实验人员知道受试者的姓名或收集可以追踪受试者的身份信息，例如用于后续跟踪的信息时，应告知受试者他们的信息是机密的。仅收集用于分析的身份信息不仅可以让受试者放心他们的身份不会被揭示，还可以缩短研究时间。此外，当受试者确信他们的身份将保持未知时，他们更有可能提供诚实的回答。在实验的报告中不使用受试者的姓名；如果使用开放性回答来说明某个观点，可以用化名或“受试者A”等标签来代替。44为了保护受试者的隐私，研究人员要么不应能够将受试者的身份与其回答联系起来，要么应去标识化数据，删除任何可以追溯回到个人的信息。

保密性可以通过各种方法维护，例如将数据存储在密码保护的计算机上，放置在锁定的办公室或文件柜中，除非去标识化，否则不与他人共享数据，使用聚合数据而不是个体数据，并销毁带有身份信息的纸质副本或文件。避免通过电子邮件发送数据。45 还重要的一点是不与受试者分享他们自己的个别结果；例如，受试者可能想知道他们在道德判断问卷上的表现如何。除了避免尴尬或失望外，不分享他们的结果可以让受试者更有信心，没有人会知道他们的个别分数。特定的法律条件和受试者对自己或他人造成伤害的威胁是例外情况；此时研究人员有责任进行干预。46对于弱势群体，如儿童、囚犯、老年人和残障人士，也有特殊情况。超出本书的范围深入讨论；然而，大多数IRB培训计划将详细解决这些问题。

在下一节中，对实验中的伦理问题的讨论超出了对受试者的关注，涉及到与研究人员自身行为相关的问题。显然，这些问题并不局限于实验，但足以引起重视，值得重复警告。

**研究人员的问题**

无需多说，研究人员应尽可能地以道德方式分析和呈现数据，但由于研究人员不当行为和被撤稿的文章屡见不鲜，这显然并不像看起来那么明显。不道德行为范围从不准确到故意不当行为、欺诈和数据伪造。这不仅仅是由于个体研究人员的诚实错误或个性缺陷，还可能包括社会层面的因素，例如学术研究中“发表或灭亡”的心态，对于发表研究的压力可能导致研究人员执行一些早前描述的不道德行为，以及在下一节中描述的行为。47 预先告知即是预先准备，以下是实验者可能违背伦理研究的一些方式。

**数据伦理**

**数据伦理问题**

众所周知的“谎言困扰，数字谎言”表明有许多方法可以操纵数据，使研究结果呈现某种特定方式或看起来更有趣或更重要。这些方法包括数据挖掘、P值操纵和HARKing，即在获得结果后进行假设。这些都是研究人员探索数据的方式，例如对数据进行处理，直到发现有趣的结果。这与科学方法的要求相反。对数据的其他不当使用还包括不报告不利的结果和失败的复制研究。对此的解药，除了研究人员抵制诱惑外，还包括报告效应大小而不仅仅依赖于p值（见第8章），并且与他人共享（去标识化的）数据供其他人进行复制研究（见第5章）。这假设研究人员没有捏造数据或进行欺诈性修改。对数据进行任何更改，例如转换以实现正态性或对缺失数据进行均值替换，都应始终进行披露。如果一名研究人员看到另一名研究人员从事不道德行为，他或她有责任报告。甚至有研究人员进行揭露行为，监管其他作者的被认为是违规的行为。如Fiske所言：“透明度是对每个人的良好建议。”

其他操纵结果的方式并不涉及修改数据，而是包括过度声称，即超出数据实际显示的范围。这在论文的讨论和结论部分中可能很普遍，其中作者可能过度推断，例如在没有数据支持的情况下得出关于受试者行为原因的结论，或者假设某些回应的原因。如果研究中没有包含中介变量，则不能对导致效应发生的因果机制提供解释。过度声称的另一个倾向出现在结果部分，当标准的p水平0.05神奇地变为0.059或0.06，或者p水平0.11被描述为几乎显著或接近显著。

超取样，即使用非常大的样本量以利用大数定律，从而更容易找到统计显著性，是另一种滥用数据的形式。从伦理上讲，这也违反了最小干预和潜在伤害的要求。报告功效分析（见第8章）可以透明地说明适当的受试者数量，并抑制这种趋势。

偶尔，实验使用次级数据，或者是先前由他人收集的现有数据。例如，美国国家选举研究（ANES）或普查社会调查（GSS）就是这样的例子。如果次级数据不是公开可用的，并且去除了所有可识别信息，那么收集数据的研究人员必须对受试者进行知情同意，以便在研究中使用它。然而，并非所有现有数据都遵循知情同意程序，因此对于已发表的科学研究来说，这是不可接受的。其中一个例子是学生对课程和教师的评估；由于学生没有被告知他们的回应将被用于研究，并被允许给予或拒绝同意，所以将这些数据用于这种方式是不道德的，并且不会得到IRB的批准。对于一些现有数据，IRB可能允许在事后获得知情同意。

**剽窃与自我剽窃**

剽窃是困扰任何类型研究的最严重的伦理问题之一，即未对不属于研究人员原创的文字和思想给予适当的引用。解决办法很简单——适当地标注出处。给予适当的引用只需简单的引用，然而这类违规行为的例子屡见不鲜。有时，这是故意为之。在某些情况下，这是由于疏忽大意的记笔记和写作习惯造成的。而另一些情况下，剽窃是因为误解需要引用什么。例如，两位教授与一位新毕业的研究生合作，在期刊编辑的指示下发现这位学生合著者几乎剽窃了他们刚刚提交的论文中的整个文献综述部分；学生不明白需要引用什么。我还有学生告诉我，原始作者的英语比他们的好，所以他们没有进行改写。因此，研究人员有责任仔细审查自己的写作以及任何合著者的写作，查找与自己风格不符的语言，并确保任何文字或思想都表明出处。

引用其他作者时，应注意将原文放在引号内，并附上页码以及典型的来源信息。没有限制可以不标明来源的文字数量；例如，常识可能认为50个字或更少的文字可以不用标明来源。但这是不正确的。即使只有一个单词或短语是从其他来源直接引用而没有适当的引用，也构成剽窃。剽窃最好通过改写来避免使用别人的陈述。改写并不意味着只改变一个或两个词；正确的改写意味着甚至句子结构都要不同。剽窃可能存在于任何类型的作品中——未发表作品、非学术作品、在线作品、演讲等等。因此，保护作品免受剽窃也是所有类型写作的必要措施，无论是为期刊、会议，甚至未发表的课堂论文。尽管剽窃是故意的还是无意的，它经常是无意识的——例如，由于记笔记时无意排除了来源。研究人员需要制定方法来避免这种情况；例如，不要从电子文件中复制粘贴，而是始终进行改写，或者在笔记中突出显示原文，并包括出处。虽然这需要更多时间，但可以消除尴尬，甚至挽救声誉。

剽窃的一个子集是自我剽窃，即从自己的写作中直接复制。这也被称为重复或文本回收。即使你是原作者，这也被视为学术不端行为，因为发表它的期刊应该得到相应的认可。避免自我剽窃有时可能涉及构图上的灵活性，特别是在多个研究依赖相同刺激和工具以寻求自我复制的实验中。已经对刺激和工具描述进行了一两次改写后，用另一种方式表达仍然简明易读会成为一种挑战。

在通信领域，相关的伦理问题日益凸显，这就是数据切片的实践，即从单一数据集中撰写多篇论文，只有微小的差异。现在，几家顶尖的通信期刊要求研究人员在投稿的附信中披露是否使用了同一数据集来撰写多篇文章。这对于调查研究尤其具有问题，但对于实验来说也有问题，如果研究人员在一项研究中经济地包含了不同的因变量（DV）、中介变量或调节变量，以便撰写两篇或更多的论文。为了避免这种伦理问题，实验研究人员应确保每篇文章都有超过“微小的差异”。

**作者身份问题**

关于作者身份的问题——谁得到了作者身份，排名是什么，何时得到——对于所有的研究方法都是棘手的。然而，每当我教授实验设计课程时，关于这个问题的课堂讨论总是最引人参与的，学生们纷纷分享着他们的“恐怖故事”：申请终身教职的教授想要单独发表一篇文章，于是将学生的名字从论文上删掉；一位作者自告奋勇将论文推进出版过程中改变了排名次序，将自己的名字排在了第一位；四位合作者对谁做的工作最多产生分歧；甚至有一位教授认为她的名字应该出现在学生的论文上，但实际上并未出现。美国心理学会等协会对作者身份问题提供了指导方针，应该咨询这些指南，因为不同学科可能存在不同规范。此外，与同事讨论以了解伦理准则也是有益的。

关于作者身份的一个主要伦理问题涉及之前提到的权力关系——这次是指教员与研究生之间的关系，而不是研究对象。例如，应该由教员还是学生获得第一作者？如果学生是以研究助理的身份得到报酬，是否还需要担任作者？如果学生虽未撰写论文，但在准备刺激和工具、收集和分析数据方面做了很多工作，是否有资格成为作者？如果学生撰写了大部分后来又被删除的论文，怎么处理？这些问题同样适用于初级和高级教员。对于这些问题和其他问题，我有两个经验法则：（1）倾向于学生或初级研究者；（2）事前制定基本规则，并在进行任何更改之前进行讨论。可以将后者视为研究人员之间的知情同意。在开始任何实验之前，应与所有涉及人员进行坦率而开放的讨论，无论是学生还是教员。谁将成为作者，谁将不是？排名将是什么？各自的职责是什么？如果有人没有履行责任，会发生什么？如果发生了任何变化，应重新进行讨论和协商，确保每个人都参与其中。最好以书面形式记录下来，通过电子邮件等方式，形成共识的摘要。研究项目可能需要很长时间，记忆会逐渐淡化；如果有书面记录，可能避免产生不愉快的情绪。

有时候，学生希望在他们的论文、学位论文或研究论文上给教授署名，而教授实际上只是提供了一些建议。学生有时会误解，指导论文或学位论文、提供建议是教师的职责，通常不应该署名。有时，学生希望向教授表示敬意，而其他时候他们认为将教授的名字放在论文上会使其更有利于发表。在这些情况下，我会婉拒，并建议学生给我致谢而不是署名。如果某人在论文中没有做出实质性的贡献，他们的名字就不应该出现在论文上。这包括给学生未经真正贡献的署名（这是对第一条经验法则过于慷慨的解释）。根据我的经验，原始论文的构思确实算作实质性贡献，但仅仅提供建议而没有做任何撰写工作则不算。我建议学生不要提出或接受这样的要求。

学位论文和学位论文应该是独立研究，并且是教员的工作职责和报酬的一部分。然而，有些教员发现在帮助学生准备论文以便发表时，付出了比预期更多的工作。在这些情况下，应参考第二条经验法则——是时候进行坦率的讨论，了解学生和教员的期望；如果教员不愿意在没有署名的情况下继续付出更多的工作，那么是时候让学生学会自己独立完成了。

作者顺序，尤其是第一作者的位置，非常重要。不仅第一作者的名字在后续作者被缩写为“等人”时会更频繁地被引用，而且第一作者通常负责将论文引导通过出版流程，包括回答审稿人的问题和进行修订工作。此外，在终身职位评审过程中，第一作者通常比其他位置被看得更有利。终身职位评审委员会也经常要求申请晋升的候选人描述其他人的贡献，并给出工作量的分配比例。在分配贡献比例之前，与所有合作者取得联系，了解他们的观点是很有帮助的，因为委员会也会联系合作者，询问同样的问题。最好不要被视为经常高估自己的贡献。

最后，作者顺序应该基于相对贡献——即具有更多能力的人应该比具有较少能力的人（例如学生）做出更大的贡献。当贡献达到学生的能力水平时，应该被视为有资格署名——例如，如果数据分析任务是共享的，但学生只能进行描述性统计，而教师则提供了差异检验。59

其他涉及作者资格的伦理问题涉及到期刊和会议。简要地说，不应该因为会议部门达到限制而将自己的名字从论文上撤下，然后如果论文被接受，再试图恢复自己的名字。这是试图绕过规则，而规则是为了给每个人一个公平的机会来展示论文。

在自我引用时，应该小心措辞和频率。过多引用自己的作品可能会被视为试图提高自己的引用率。只有在自己的作品（以及他人的作品）与当前研究有实质性的关联时，才应该引用。此外，过多地引用自己的作品可能会泄露作者身份，危害盲审过程。通常，可以在引文中使用“作者（隐藏）”这样的文字，一旦论文被接受，应该提供完整的引文。

**先导研究**

正如本书之前解释的那样，实验中有一些术语的使用存在一些变化；其中包括先导研究、前测试和操作检查。为了语义上的清晰，本书使用先导研究这个术语来指代在更大规模的实际实验之前进行的小规模实验。这些实验旨在解决研究设计中的问题，以便在进行更大规模的研究之前进行改进。先导研究可能适用于除实验以外的其他方法，但在实验中被视为最佳实践。60 先导研究有许多用途，其中一个与前面的讨论相关——发现任何未预料到的对受试者的伤害、可能冒犯他们的事物或他们可能提出的异议。除了伦理问题，先导研究还是研究人员可以发现实验中其他问题的地方，这些问题他们之前没有预料到——例如，仪器是否正常工作，测量是否具有足够的范围以避免顶限效应，实际所需的时间，指令措辞，实验者的态度以及受试者认为研究的重要性。先导研究还有利于测试新概念的内部有效性。如果涉及欺骗，先导研究可以用来确定实验的可信度。先导研究是研究人员进行“试运行”的机会，使他们能够在进行实际实验之前对其进行改进，从而节省时间和金钱。

先导研究应该按照实际研究的计划进行，但给受试者更多机会提供反馈，了解哪些方面有效，哪些方面不足。可以在研究仪器中增加开放式和封闭式问题，以收集这些反馈，同时实验者也可以就研究过程对受试者进行访谈。还应该征求实验者对研究的改进意见，例如，如果他们观察到受试者的注意力因为实验过长而分散。

受试者应该来自实际实验中将要研究的人群，但后续的实验运行中应该将他们排除在外，因为之前参与过研究可能会影响他们的行为。

关于是否应该使用先导研究来确定是否值得进行实际研究以及用于确定实际研究的样本大小存在一些争议。Westlund和Stuart认为，先导研究的结果应该是做出这些决定时的一个数据点。他们提供了不使用先导研究作为这两个目的的原因，包括因为先导研究通常具有小样本量而导致功效不足。

关于如何进行先导研究，目前没有明确的指导方针，一些人认为这也是为什么进行先导研究的不多的原因。

最后，对先导研究这个术语的另一种误用是研究人员在后期对实验进行的一种后事后的辩解，用来解释未能达到传统显著性水平或样本量较小的情况。先导研究应该与实际较大的实验一起报告，或者在需要的情况下向资助机构提交一份白皮书。

一旦获得IRB批准，就可以开始进行先导研究。在仔细考虑了先导研究中得到的教训后，需要对实际研究进行调整，然后就可以开始运行。

本书的目的是指导新的实验研究者逐步设计实验的步骤。并不是所有的步骤都会按照本书的确切顺序进行；例如，研究人员可能同时开发仪器和刺激，而且在进行过程中一定会重新考虑研究的各个方面，修改之前完成的阶段。正如前言中引用W. A. McCall的话所说，本书的目标是“获得足够和适当的数据”。在获得这些数据后，最后一步是撰写结果；关于如何撰写结果，本书中的11.3小节提供了指导和示例。

**如何操作11.3**

**撰写结果部分**

结果部分是回答假设和研究问题的地方。这是一个“就事论事”的部分，不应该推测结果出现的原因或其含义；这些应该保留在讨论和结论部分。相反，应该事实地呈现数据和统计检验，并陈述假设是否得到支持。例如：

“第一个假设预测，常规广告比广告特稿更容易激活与广告概念（H1a）和说服（H1b）相关的图式。通过独立样本t检验比较常规广告条件（M = 637.68，SD = 184.18）和未标记广告特稿条件（M = 619.70，SD = 193.47）之间的LDT均分，结果没有显著差异，t（225）= 1.07，p = ns，暗示广告特稿没有减缓与广告概念相关的图式的激活。H1a未得到支持（图1）。”

如您从这个例子中看到的，报告结果可以非常简洁。统计检验值（在这个例子中是t检验，因为有两组处理和对照组）以及括号内的自由度和p值都应该被报告。一些期刊可能要求给出精确的p值（例如p = .005），而其他期刊可能只要求在效应不显著时给出（例如p = .69），显著p值则舍入为常规水平（例如p < .05，p < .01，p < .001）。当结果不显著时，可以使用“p = ns”。

另一个例子：

“H1预测社交媒体上针对正面形象的威胁，如拒绝和批评，将比非攻击性评论引起更多的自我报告负面情感。

对此进行了方差分析（ANOVA），包括三个条件：拒绝，批评和对照。结果在一定程度上支持H1（F（2，75）= 7.37，p = .001，η2 = .16.2）。使用Scheffe事后校正，结果显示在拒绝条件（M = 2.11，SD = .70，p = .01）和批评条件（M = 2.17，SD = .61，p = .01）的参与者报告的负面情感显著高于对照条件（M = 1.59，SD = .50）。然而，拒绝和批评之间没有显著差异，回答了RQ1。对于自我报告的积极情感，没有在条件之间找到显著差异，其中较低的值表示负面情感增加，F（2，75）= .21，p = .81，η2 = .01。

这些结果部分地支持H1，提供了对积极形象遭受两种威胁（拒绝和批评）相比对照情况引起的自我报告负面情感增加的证据。”

请注意，这个例子使用方差分析（ANOVA），因为有三个组（拒绝，批评和对照）。报告组和总受试者的自由度。使用符号η2报告效应大小，表示自变量解释的方差比例。该结果在一定程度上支持H1，显示在面对拒绝和批评时，负面情感有所增加。

此外，效应大小可以使用Cohen's d进行报告。例如：

“然而，在阅读带标签的广告特稿（M = 634.09，SD = 131.69）后，对说服相关单词的反应速度比暴露在没有标签的广告特稿中（M = 672.59，SD = 144.09）的反应速度更快，t（211）= 2.06，p < .05，效应大小较小（Cohen's d = .28），支持H2b（图1）。”

在这个例子中，使用了t检验来比较两个条件（带标签和未标记的广告特稿）。效应大小（Cohen's d）被报告，以量化组间差异的大小。

请记住，结果部分应侧重于客观地呈现数据和统计检验结果，而不是解释。将讨论和结论部分留给后面，以提供对研究结果的全面分析和解释。

许多文章的结果部分以关于样本、因变量和重要协变量的描述性统计开始。有些期刊将这些内容包含在方法部分中，请查阅期刊的样式指南。以下是一个示例，展示了如何报告这些内容：

“总共有337名MTurk工作者（常规广告条件 = 124；标记广告特稿条件 = 110；未标记广告特稿条件 = 103），其中56％被归类为男性。平均年龄为31.5岁（标准差= 11.58），四分位范围为23至37岁。样本（N = 337）包括白人（n = 263，占78％）、亚裔（n = 47，占13.9％）、非洲裔美国人（n = 22，占6.5％）、西班牙裔/拉美裔（n = 16，占4.7％）、美国印第安人或阿拉斯加原住民（n = 3，占0.9％）和夏威夷或其他太平洋岛屿土著人（n = 1，占0.3％）。”

正如所示，大写的N指的是整个样本，小写的n指的是组别。

这里有另一个例子，报告因变量的汇总统计数据：

“在第二个游戏的专制治理条件中，学生选择自私的选项占总次数的83％（90/108），而成年行业专业人士选择自私的选项占总次数的61％（68/111）。”

还有一个更多人口统计学信息的最终例子：

“结果

“样本特征

“表1展示了所有残疾参与者（N = 336）的基线人口统计学和社会经济特征，根据治疗状态进行区分。平均年龄约为38岁。约80％的参与者为女性，不到一半为白人，稍微超过20％的参与者已婚。超过三分之二的参与者接受过一些大学或更高学历，平均每月家庭收入约为1400美元。四分之三的参与者报告在家庭中有孩子。样本在大多数变量上都是平衡的，除了房屋拥有权、公共住房援助和拥有银行账户。与对照组成员相比，治疗组参与者更有可能接受Section 8公共住房援助（23％对比15％；p < .05），拥有银行账户的可能性高9个百分点（88％对比79％；p < .05），以及拥有房屋的可能性低9个百分点（18％对比27％；p < .05）。总体而言，治疗组似乎在社会经济上比对照组更具劣势。与之前的ADD研究一样（例如Mills等人，2008年），塔尔萨IDA计划的样本并不代表有资格参加该计划的低收入家庭。”

索引（Cronbach's alpha，因子分析）需要进行内部一致性测试，并且有些期刊将它们放在结果部分，而其他期刊则将其包含在方法中。

当图形显示比文本描述更容易理解时，应使用表格和图形，例如在需要报告许多结果时。在这种情况下，结果部分应总结最重要的发现，其余内容则在引用表格时显示。例如：

“回到表2，我们将两个人群在信任方面进行比较，并添加性别和信念的控制变量。正如该表的第3列所示，信任与信任值的正面信念相关。然而，即使进行了这个控制，域名交易者显示的信任水平比学生高出24个百分点。性别对信任没有影响。”

结构方程模型通常以图形形式报告，其中包含统计数据和图题。例如：

“在测试拟议模型时使用了PROCESS（Hayes，2013）模型4，这是一个并行中介模型，实验操作被认为直接导致情绪和报复性攻击，同时通过情绪间接导致报复性攻击。为了测试这一点，在模型中将负面和积极情绪视为并行的中介变量，并分别检查了两种报复性攻击的测量结果。基于5000个Bootstrap样本的偏差校正Bootstrap置信区间用于测试间接效应。”

对于同一篇论文中的多个实验，可以将报告的形式分为单独报告每个研究（例如，研究1，研究2，研究3）。这在每个研究中提出不同假设时经常使用。当多个研究在不同条件下测试相同的假设时，可能更清晰地根据每个假设呈现结果，并将所有研究的结果放在每个假设下（例如，H1，研究1结果，研究2结果；H2，研究1结果，研究2结果）。在这种情况下，最后提供一个总体摘要有助于读者理清内容，例如，一项研究在展示了三项研究结果后如此概述：

“总的来说，当权衡是撒谎和羞耻时，成年专业人士撒谎的情况略微减少；而当权衡是撒谎和金钱时，成年专业人士撒谎的情况显著减少。”

当使用协方差分析（ANCOVA）进行分析时，也会报告其效应是否显著，并使用“在控制...之后”的表述。例如：

“这个假设得到了支持。在控制人口统计学协变量后，公共卫生框架（M = 4.0，SD = 1.1）相较于传统框架（M = 3.6，SD = 0.94）显著增加人们表示打算改变自己行为的可能性（F = 4.5，p < .05）。显著的协变量包括性别（F = 20.4，p < .001）和种族（F = 8.58，p < .05）。女性（M = 4.1，SD = 0.86；男性M = 3.2，SD = 1.10）和西班牙裔（M = 4.5，SD = 1.2）最有可能改变他们的行为意图。”

多元协方差分析（MANCOVA）在处理两个以上的因变量和协变量时也是类似，只是在MANCOVA和多元方差分析（MANOVA）中，都需要报告一个综合的显著性检验，考虑到所有因变量。只有在总体分析显著时，研究者才会报告各个因变量的单变量显著性检验。例如：

“将公共新闻图表因变量（问题和分析、公民观点、动员信息、与媒体联系的方式、共同点和解决方案）与作者身份协变量一起输入MANCOVA。使用Wilks’ Lambda，分析显示依赖变量考虑在内的效应显著（F = 4.87，df = 5, 641，p < .001，η2 = .037）。作者身份协变量显著（F = 50.3，df = 5, 641，p < .001，η2 = .068）。当检查依赖变量的单变量分析时，两个项目的效应显著，问题和分析以及共同点。”

最后，虽然统计软件可能生成“p = 0.000”的输出，但报告这一结果或甚至“p < .000”是不适当的。应该将其舍入为最高的常规水平，即p < .001。Lilienfeld等人（Lilienfeld et al.，2001）已经优雅地解释了这个问题，简单来说，这暗示了对真实零假设的错误拒绝的概率为零，这几乎是不可能的。

当然，报告实验结果的方式有很多种，不仅限于这里的示例。阅读您学科领域的实验结果，了解其他人是如何撰写的，是很有帮助的。

最后，来自1963年Campbell和Stanley经典专著的最好建议是：“大多数实验都会令人失望。”他们还包括“乏味”和“痛苦”。所以要对抗这个现实，坚持不懈。在小的成功中找到安慰，因为知识是在不断演进的。设计最好的实验，即使它仍然不完美。继续进行实验并解释结果，完全意识到其不足之处。尽管存在这些缺点，我发现实验是了解因果关系以及解决社会世界重要问题的最激动人心的方式。享受这个过程吧。

**常见错误**

• 为了节省时间或避免困难而使用欺骗手段。

• 没有及时向IRB提交研究项目。绝对不要在获得批准之前开始收集数据。

• 没有与IRB完结研究；报告的被研究对象数量错误；如果需要，没有就研究方案的更改提交修改申请。

• 在记录笔记和撰写论文时粗心大意，导致无意中出现抄袭。没有重新用自己的话重新表述之前论文中的内容。

**测试您的知识**

1. IRB（Institutional Review Board，伦理审查委员会）的目的是：

a. 为人类受试者研究颁发联邦资金

b. 审查研究提案，确保研究人员在处理受试者时遵循伦理原则，不对其造成伤害

c. 将生物医学研究标准应用于社会科学

d. 减缓研究过程，以便科学家改进他们的研究

2. 以下哪个选项不是《贝尔蒙特报告》中的原则之一？

a. 尊重个人原则

b. 施惠原则

c. 公正原则

d. 忠诚原则

3. 避免受试者猜测实验假设的最不受欢迎的方法是使用：

a. 不显眼的仪器

b. 欺骗

c. 分散注意力的问题

d. 操作检查

4. 当两位研究人员让一名研究助理假装成受试者，并告诉潜在受试者一项并不有趣和重要的研究时，他们使用了：

a. 不显眼的仪器

b. 预备研究

c. 同谋者

d. 分散注意力的任务

5. 当回答有关实验中的政治观点的问题类似于与他人讨论政治观点时，这是一个例子，属于：

a. 侵犯隐私

b. 泄露保密

c. 缺乏知情同意

d. 风险不比日常生活中大

6. 在社会科学实验中，减少伤害的一种方法是：

a. 允许受试者跳过问题

b. 允许受试者退出研究

c. 在知情同意过程中警告受试者关于可能的伤害风险

d. 所有以上方法

7. 当研究人员在某种程度上对受试者拥有控制权时——例如，他们的成绩——这是一个例子，属于：

a. 随机分配

b. 滥用权力

c. 丧失隐私

d. 数据操纵

8. 允许学生通过写论文而不是参加实验来获得额外学分是一个例子，属于：

a. 成绩膨胀

b. 强迫

c. 避免滥用权力的方法

d. 随机分配

9. 当一项研究被认为是匿名的时候：

a. 信息被去标识化

b. 不收集任何识别信息

c. 数据保存在受密码保护的笔记本电脑上

d. 结果以聚合形式报告

10. 当保持保密性时：

a. 信息被去标识化

b. 数据保存在受密码保护的笔记本电脑上

c. 结果以聚合形式报告

d. 所有以上方法

**Answers**

1. b

2. d

3. b

4. c

5. d

6. d

7. b

8. c

9. b

10. d

**应用练习**

1. 如果您还没有这样做，请完成您所在机构的IRB培训。您需要完成这项培训，以便提交研究提案并获得进行试点研究的批准。如果您的机构没有IRB，那么请阅读《贝尔蒙特报告》和美国卫生与公众服务部人类研究保护办公室的相关规定，网址为https://www.hhs.gov/ohrp/regulations-and-policy/regulations/45-cfr-46/index.html。《贝尔蒙特报告》和相关法规并不长，但您可以阅读Shelton, James D. (1999) 的文章《How to Interpret the Federal Policy for the Protection of Human Subjects or Common Rule (Part A)》，IRB: A review of human subjects research 21: 6–9，以获得摘要。

2. 向您所在机构的IRB提交研究提案。请咨询您的指导教师以获得具体指导，并在必要时寻求导师的支持。回答所有的修改要求。如果您的机构没有IRB，请遵循美国卫生与公众服务部网站的指南。创建知情同意文件，供被试以纸质副本或在线封面信函形式获得。如果使用学生作为研究对象，需提供替代性激励措施。遵循本章所描述的所有指南。

3. 一旦您的研究得到批准，计划并进行一个试点研究。请记住，这是您实际研究的小规模版本，旨在了解其效果，以及需要在收集真实数据之前进行哪些修改。请务必加入开放性问题，以供被试回答有关程序的问题，并在研究结束后与他们讨论对研究的印象。

**建议阅读**

American Psychological Association. 2009. Publication Manual of the American Psychological

Association, 6th edition. Washington, DC: Author.

For ethical guidelines, read Chapter 1, “Writing for the Behavioral and Social Sciences,” starting with page 11

through page 20. For more on plagiarism, read Chapter 6, “Crediting Sources.”

For another overview, see:

Hegtvedt, Karen A. 2014. “Ethics and Experiments.” In Laboratory Experiments in the Social Sciences,

edited by Murray Webster and Jane Sell, 22–51. London: Elsevier.

For codes of ethics specific to your discipline, check out your professional association’s website.

For more on the call for different IRB rules for different disciplines, see:

La Noue, George R., and Alexander Bush. 2010. “Institutional Review Board Rules: Should One Size Fit

All Disciplines?” International Journal of Interdisciplinary Social Sciences 5 (8): 239–258.

For more on P-hacking, see:

Aschwanden, Christie. 2015. “Science Isn’t Broken: It’s Just a Hell of a Lot Harder Than We Give It

Credit For.” http://fivethirtyeight.com/features/science-isnt-broken/

For a dissenting opinion on the ethics of random assignment, see:

Fives, Allyn, Daniel W. Russell, John Canavan, Rena Lyons, Patricia Eaton, Carmel Devaney, Norean

Kearns, and Aoife O’Brien. 2015. “The Ethics of Randomized Controlled Trials in Social Settings:

Can Social Trials Be Scientifically Promising and Must There Be Equipoise?” International

Journal of Research & Method in Education 38 (1): 56–71.

For more on authorship, see:

Fine, Mark A., and Lawrence A. Kurdek. 1993. “Reflections on Determining Authorship Credit and

Authorship Order on Faculty-Student Collaborations.” American Psychologist 48 (11): 1141–1147.

For more on pilot studies, see:

Westlund, Erik, and Elizabeth A. Stuart. 2017. “The Nonuse, Misuse, and Proper Use of Pilot Studies

in Experimental Evaluation Research.” American Journal of Evaluation 38 (2): 246–261