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

iQualtrics是一个提供受访者参与研究的服务，并且拥有调查软件，可以用于管理研究，无论受试者是否来自该公司的样本池。研究人员可以使用该软件而无需使用样本池中的受试者。

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、回应率、受试者多样性和代表性方面与其他可考虑的替代方案相当甚至更好。

(Continued)

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。

ii许多研究人员自费进行研究。请与您所在的部门、学院和大学联系，了解是否有资金来源，以及外部资助机构的情况。.

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.

iii–美国心理学协会，美国心理学协会出版手册，第六版（华盛顿特区：出版商，2009年）。APSA的指南还包括报告意图分析，这在医学领域比社会科学更常见。本书无法解释这种统计分析的范围，而是将读者引用到统计学书籍、课程和文章，例如Sally Hollis和Fiona Campbell的“意图分析的含义是什么？已发表随机对照试验的调查”（BMJ 319号7211，1999年9月：670-674）。

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。

A screenshot of a computer

Description automatically generated

## MORE ABOUT . . . BOX 8.4

###### Effect Size Statistics

A screenshot of a computer

Description automatically generated

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。

A grey rectangular object with a black stripe

Description automatically generated

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。

A screenshot of a computer

Description automatically generated

Cohen186建议最小的统计功效为80%。

比较较少的组/条件需要更多的参与者来保持统计功效。187

过去，人们需要手动进行功效分析。如今，研究人员通常使用统计软件（如Stata和R）或在线工具（如G\*Power）进行功效分析，其中最受欢迎的之一是由一群德国研究人员开发的G\*Power。该软件免费提供，可在http://www.gpower.hhu.de/上获得，并附带教程。188

在进行功效分析之前，研究人员应该查阅与他们的研究类似的以前的研究，并收集基本数据，包括均值、标准差和效应大小。自己进行过的先前研究也对此很有用。例如，对于一项研究探究记者在报道主题为儿童和成年人时是否使用不同水平的道德判断的实验，使用了以前在照片和道德判断上的工作数据来计算功效并确定样本大小。

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。

.

**238**

Designing Experiments for The Social Sciences

**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

Chapter 8 ■ Sampling and Effect Sizes **239**

**测试你的知识**

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

**240** Designing Experiments for the Social Sciences

**Answers**

1. d

2. b

3. c

4. a

5. d

6. d

7. a

8. c

9. b

10. b

Screens screenshot of a computer

Description automatically generated

Chapter 8 ■ Sampling and Effect Sizes **241**

**注释**

1. Michael D. Basil，“学生样本在传播研究中的使用”，《广播和电子媒体杂志》40卷3期（1996年）：439。

2. William D. Wells，“面向发现的消费者研究”，《消费者研究杂志》19卷（1993年）：492。

3. Elizabeth Tipton，“你的实验有多具有普适性？一种用于比较实验样本和总体的指标”，《教育和行为统计杂志》39卷6期（2014年）：478–501。

4. Basil，“学生样本的使用”；John A. Courtright，“理性思考非概率性的重要性”，《广播与电子媒体杂志》40卷3期（1996年）：414–421；Annie Lang，“逻辑从非概率样本的实验数据中使用推断统计学：受到Cooper、Dupagne、Potter和Sparks启发”，《广播与电子媒体杂志》40卷3期（1996年夏季）：422–430；William R. Shadish、Thomas D. Cook和Donald T. Campbell，《实验和准实验设计用于广义因果推断》（加利福尼亚州贝尔蒙特：Wadsworth Cengage Learning，2002年）。

5. Shadish、Cook和Campbell，《实验和准实验设计》，373页。

6. Basil，“学生样本的使用”。

7. Robert A. Peterson，“关于在社会科学研究中使用大学生的见解：基于第二阶段元分析的研究”，《消费者研究杂志》28卷（2001年）：450–461；I. Simonson等，“消费者研究：寻找身份”，收录于《心理学年度评论》，S. T. Fiske、D. L. Schacter和C. Zahn-Waxler主编（加利福尼亚州帕洛阿尔托：年度评论，2001年）：249–275。

8. Lisa Marriott，“在实验研究中使用学生对象：对使用学生作为纳税人代理的挑战”，《国际社会研究方法学杂志》17卷5期（2014年）：503–525。

9. W. J. Potter、R. Cooper和M. Dupagne，“主流传媒研究的三种范式”，《传播理论》3卷（1993年）：317–335。

10. Joanne M. Miller，“审视议程设置的中介因素：一种新的实验范式揭示情绪的作用”，《政治心理学》28卷6期（2007年）：696。

11. Elizabeth A. Bennion和David W. Nickerson，“方便的代价：一项实验证明电子邮件的推广降低了选民登记”，《政治研究季刊》64卷4期（2011年）：861。

12. Renita Coleman，“视觉对伦理推理的影响：对于做出道德决策的新闻工作者来说，照片有多重要？”《新闻与传播季刊》83卷4期（2006年冬季）：840。

13. Emily Thorson，“信念回声：纠正误传的持久影响”，《政治沟通》33卷3期（2016年）：466–467。

14. M. Kim和G. G. Van Ryzin，“政府资助对艺术组织捐款的影响：一项调查实验”，《非营利与志愿部门季刊》43卷5期（2014年）：916。

15. Hoon Lee和Nojin Kwak，“政治讽刺的情感效应：讽刺幽默、消极情绪和政治参与”，《大众传播与社会》17卷3期（2014年）：315。

16. Renita Coleman，“新闻工作者对儿童的道德判断：行动胜于言辞？”《新闻实践》5卷3期（2011年夏季）：257–271。

17. D. O. Sears，“实验室中的大学二年级学生：狭窄数据库对社会心理学对人性的观点的影响”，《个性与社会心理学杂志》51卷（1986年）：515–530。

18. 同上。

19. Basil，“学生样本的使用”，434页。

20. 同上。

21. 同上。

22. 同上。

23. 同上；Sears，“实验室中的大学二年级学生”。

24. B. H. Newberry，“有关实验信息的实验对象的说真话：谁被欺骗了？”《个性与社会心理学杂志》25卷（1973年）：369–374。

25. R. Hertwig和A. Ortmann，“实验中的欺骗：重新审视其辩护论据”，《伦理与行为》18卷（2008年）：59–92。

26. C. D. Kam，“隐含态度，明确选择：当潜意识启动预测候选人偏好”，《政治行为》29卷（2007年）：343–367。

27. Basil，“学生样本的使用”。

**242** Designing Experiments for the Social Sciences

28. Deborah Nazarian and Joshua M. Smyth，“上下文调节表达性写作干预的效果：两项随机复制和扩展研究”，《社会与临床心理学杂志》29卷8期（2010年）：903–929。

29. Basil，“学生样本的使用”。

30. 同上；Sears，“实验室中的大学二年级学生”。

31. Basil，“学生样本的使用”。

32. 同上。

33. 同上。

34. Renita Coleman和H. Denis Wu，《选民决策中的形象和情感：情感议程》（纽约：列克星敦书籍，2015），第9章。

35. W. A. Cunningham、T. Anderson和J. Murphy，“学生是真实的人吗？”《商业学杂志》47卷（1974年）：399–409。

36. Basil，“学生样本的使用”。

37. 同上。

38. Yair Levy、Timothy J. Ellis和Eli Cohen，“信息系统研究中实验和准实验研究的新手研究员指南”，《综合信息、知识与管理杂志》6卷（2011年）：151–161。

39. J. W. Lucas，“理论测试、泛化和外部效度问题”，《社会学理论》21卷（2003年）：236–253；J. G. Lynch Jr.，“关于消费者研究中实验的外部效度”，《消费者研究杂志》9卷（1982年）：225–239；John G. Lynch Jr.，“外部效度在理论研究中的作用”，《消费者研究杂志》10卷（1983年）：109–111。

40. Peterson，“论大学生的使用”。

41. Paul H. P. Hanel和Katia C. Vione，“学生样本是否能准确估计一般公众？”《公共科学图书馆综合期刊》11卷12期（2016年）：1–10。

42. Coleman和Wu，《选民决策中的形象和情感》。

43. C. D. Kam、J. R. Wilking和E. J. Zechmeister，“超越“狭窄数据基础”：另一种用于实验研究的便利样本”，《政治行为》29卷（2007年）：415–440；P. J. Henry，“重复实验室中的大学二年级学生：狭窄数据库对社会心理学对偏见本质观点的影响”，《心理探究》19卷2期（2008年）：49–71。

44. M. H. Birnbaum，编，《互联网心理实验》（圣地亚哥，加利福尼亚州：学术出版社，2000年）；M. H. Birnbaum，“通过互联网进行人类研究和数据收集”，《心理学年度评论》55卷1期（2004年）：803–832；Henry，“重复实验室中的大学二年级学生”；U. D. Reips，“网络实验方法：优点、缺点和解决方案”，收录于《互联网心理实验》，编者M. H. Birnbaum（圣地亚哥，加利福尼亚州：学术出版社，2000年）；M. P. Wattenberg，《年轻人投票了吗？》（上斯特里姆河岸，新泽西州：皮尔逊，2011年）。

45. Erin C. Cassese等，“社交媒体促进的互联网调查：在线实验中招募参与者”，《政治科学与政治》46卷4期（2013年）：1–10。

46. Marriott，“在实验研究中使用学生受试者”。

47. Cassese等，“社交媒体促进的互联网调查”。

48. 同上。

49. E. Diener、M. Diener和C. Diener，“预测国家主观幸福感的因素”，《个性与社会心理学杂志》69卷（1995年）：851–864；R. Fischer和S. Schwartz，“价值优先级的来源：个体、文化还是人为因素？”《跨文化心理学杂志》42卷（2011年）：

50. F. R. Kardes，“捍卫实验消费心理学”，《消费者心理学杂志》5卷（1996年）：279–296；Lucas，“理论测试、泛化”。

51. Basil，“学生样本的使用”；Lang，“使用推论统计学的逻辑”。

52. J. Alm和M. McKee，“将税务遵从实验的教训扩展到管理和决策经济学”，《管理与决策经济学》19卷（1998年）：259–275。

53. D. Bello等，“编辑部：国际商业研究中的学生样本”，《国际商务研究期刊》40卷（2009年）：361–364；D. G. Mook，“捍卫外部无效性”，《美国心理学家》38卷（1983年）：379–387；R. E. Pernice等，“学生样本在量表构建中的应用”，《心理报告》102卷（2008年）：459–464。

54. Peterson，“论大学生的使用”；Robert A. Peterson和Dwight R. Merunka，“大学生便利样本与研究再现性”，《商业研究杂志》67卷5期（2014年）：1035–1041。

.

Chapter 8 ■ Sampling and Effect Sizes **243**

55. A. Bardi等人，"自主选择的生活转折中的价值稳定性和变化：自我选择与社会化效应的比较研究"，《个性与社会心理学杂志》106（2014年）：131-147。

56. M. Host, B. Regnell和C. Wohlin，"使用学生作为研究对象——学生和专业人员在提前时间影响评估方面的比较研究"，《实证软件工程》5（2000年）：201-214；G. Liyanarachchi，"会计实验中使用学生作为研究对象的可行性：综述"，《太平洋会计评论》19，第1期（2007年）：47-67；Marriott，"在实验研究中使用学生作为研究对象"。

57. Marriott，"在实验研究中使用学生作为研究对象"。

58. C. M. Megehee，"广告时间扩展、压缩和认知处理对消费者信息和品牌接受的影响"，《商业研究杂志》62，第4期（2009年）：420-431。

59. Basil，"使用学生样本"。

60. 例如，参见：S. Hazari，C. Brown和R. Rutledge，"调查营销学生对博客中的积极学习和社会协作的看法"，《商业教育杂志》88（2013年）：101-108；Host，Regnell和Wohlin，"使用学生作为研究对象"；Jin Huang等人，"个人发展账户与低收入成年残疾人的购房：来自随机实验的证据"，《应用社会科学杂志》10，第1期（2016年）：55-66；Yanna Krupnikov和Adam Seth Levine，"样本间比较和外部效度"，《实验政治科学杂志》1，第1期（2014年）：59-80；Liyanarachchi，"使用学生作为研究对象的可行性"；G. Liyanarachchi和M. Milne，"比较会计从业者和学生的投资决策：关于学生替代品充分性的实证研究"，《会计论坛》29（2005年）：121-135；D. W. Schanzenbach，"研究项目星计划中的研究结果"，收录于《布鲁金斯教育政策论文选辑》（2007年）：205-228；Valmi D. Sousa，Jaclene A. Zauszniewski和Carol M. Musil，"如何确定方便样本是否代表总体"，《应用护理研究》17，第2期（2004年）：130-133；T. Waller，C. Lampman和G. Lupfer-Johnson，"评估护理和心理学学生对超重个体的偏见：一项隐性联想测试"，《临床护理杂志》21（2012年）：3504-3512；James N. Druckman和C. D. Kam，"学生作为实验参与者：对“狭窄数据库”的辩护"，收录于《剑桥实验政治科学手册》（由James N. Druckman等人编，纽约：CUP，2011年），41-57。

61. A. Abdel-Khalik，"会计研究中主题替代的效率"，《会计评论》49（1974年）：743-750；D. Bean和J. D'Aquila，"在研究道德困境时，会计学生作为会计专业人员的替代品：一则警示"，《教授商业伦理》7（2003年）：187-204；R. Copeland，A. Francia和R. Strawser，"行为商业研究中的学生作为实验对象"，《会计评论》48（1973年）：365-372；M. T. Gordon，A. Slade和N. Schmitt，"再访“大二学生科学”的设计：从推测到经验主义"，《管理学评论》11（1986年）：191-207。

62. Sousa，Zauszniewski和Musil，"如何确定方便样本是否代表总体"。

63. W. G. Cochran，抽样技术，第3版（纽约：Wiley，1977年）。

64. Renita Coleman，Esther Thorson和Lee Wilkins，"测试公共卫生框架和丰富资源在健康新闻报道中的影响"，《健康传播杂志》16，第9期（2011年）：941-954；Rebecca S. McEntee，Renita Coleman和Carolyn Yaschur，"比较生动文字和照片对公共关系中道德判断的影响"，《新闻与大众传播季刊》94，第4期（2017年）：1011-1030；Aimee Meader等人，"数字时代的伦理：移动图像和照片对道德判断的影响比较"，《传媒伦理学杂志》30，第4期（2015年）：234-251。

65. Nazarian和Smyth，"上下文调节效应"。

66. Krupnikov和Levine，"样本间比较"。

67. Peterson，"使用大学生"；Peterson和Merunka，"大学生方便样本"。

68. R. M. Lindsay和A. S. C. Ehrenberg，"复制研究的设计"，《美国统计学家》47（1993年）：236。

69. Peterson和Merunka，"大学生方便样本"。

**244** Designing Experiments for the Social Sciences

70. Basil，“学生样本的使用。”

71. Hanel和Vione，“学生样本是否能够准确估计总体公众？”

72. Christine E. Meltzer，Thorsten Naab和Gregor Daschmann，“所有学生样本都不同：关于传播科学中参与者选择的问题，”《传播方法与测量》6 (2012)：251-262。

73. Coleman，“视觉效果的影响。”

74. Claire A. Dunlop和Claudio M. Radaelli，“教授监管谦逊：与学生从业者的实验，”《政治学》36，第1卷(2016)：79-94。

75. Basil，“学生样本的使用。”

76. T. D. Cook和D. T. Campbell，准实验：场景中的设计与分析问题(芝加哥：Rand-McNally，1979)。

77. Peterson，“关于使用大学生的研究。”

78. Marriott，“在实验研究中使用学生受试者。”

79. Peterson，“关于使用大学生的研究”；Peterson和Merunka，“大学生方便样本。”

80. Marriott，“在实验研究中使用学生受试者。”

81. Sousa，Zauszniewski和Musil，“如何确定方便样本是否代表总体1。”

82. Winter Mason和Siddharth Suri，“在Amazon的Mechanical Turk上进行行为研究，”《行为研究》44 (2012)：1-23。

83. Srinivas Rao和Amanda Michel，“ProPublica的Mechanical Turk指南，”《ProPublica》(2010年10月15日)。

84. Bassam Tariq，“以Mechanical Turk为生计，”《纽约客》(2015年3月20日)。

85. Mason和Suri，“在Amazon的Mechanical Turk上进行行为研究。”

86. Christoph Bartneck等人，“比较从Amazon的Mechanical Turk中获得的回应与在线和直接招募的研究的相似性，”《PLoS ONE》10，第4期(2015)：1-23。

87. http://en.wikipedia.org/wiki/The\_Turk。 2017年6月24日访问。

88. http://en.wikipedia.org/wiki/The\_Turk。 2017年6月24日访问。

89. http://en.wikipedia.org/wiki/The\_Turk；Bartneck等人，“比较回应的相似性。”

90. J. Ross等人，“谁是众包工作者？Mechanical Turk中的人口统计学转变”(于2010年ACM报告的论文)。

91. Adam J. Berinsky，Gregory A. Huber和Gabriel S. Lenz，“评估在线劳工市场进行实验研究：亚马逊的Mechanical Turk，”《政治分析》20，第3卷(2012)：351-368；Gabriele Paolacci，Jesse Chandler和Panagiotis Ipeirotis，“在Amazon Mechanical Turk上运行实验，”《判断与决策制定》5，第5卷(2010)；Ross等人，“谁是众包工作者？”

92. Berinsky，Huber和Lenz，“评估在线劳工市场。”

93. Michael Buhrmester，Tracy Kwang和Sam Gosling，“亚马逊的Mechanical Turk：一种新的廉价但高质量的数据来源？”《心理科学观点》6，第1卷(2011)：3-5；R. Kosara和C. Ziemkiewicz，“Mechanical Turk梦见方形饼图吗？” BEyond Time and Errors: Novel EvaLuation Methods for Information Visualization (BELIV) 10 (2010)：373-382。

94. Connor Huff和Dustin Tingley，“评估MTurk调查受访者的人口学特征和政治倾向，”《研究与政治》(2015年7月至9月)：1-12。

95. Krupnikov和Levine，“跨样本比较。”

96. Thomas Bernauer，Robert Gampfer和Aya Kachi，“欧洲单边主义和全球气候政治中的非自愿负担分担：来自另一方公众观点的证据，”《欧盟政治》15，第1卷(2014)：132-151；Robert Gampfer，Thomas Bernauer和Aya Kachi，“在捐赠国进行共同实验中获取对南北气候资金的公众支持：”《全球环境变化》29(2014)：118-126。

97. Huff和Tingley，“评估MTurk调查受访者的人口学特征和政治倾向。”

98. Bartneck等人，“比较回应的相似性。”

99. Huff和Tingley，“评估MTurk调查受访者的人口学特征和政治倾向。”

100. Kevin E. Levay，Jeremy Freeze和James N. Druckman，“Mechanical Turk样本的人口学和政治组成，”《SAGE开放》1-17(2016年1-3月)。

101. Ibid.，第10页。

Chapter 8 ■ Sampling and Effect Sizes **245**

102. Berinsky，Huber和Lenz，“评估在线劳工市场。”

103. 同上；Buhrmester，Kwang和Gosling，“亚马逊的Mechanical Turk”；Huff和Tingley，“谁是这些人？”；Panagiotis G. Ipeirotis，“Mechanical Turk的人口统计学，”CeDER工作论文第10-10号纽约大学，检索自http://www.ipeirotis.com/wp-content/uploads/2012/02/CeDER-10-01.pdf（2010）；Krupnikov和Levine，“跨样本比较”；Kevin J. Mullinix等人，“调查实验的泛化性，”第2页（2015）：109-138。

104. 同上；Krupnikov和Levine，“跨样本比较”；Mullinix等人，“调查实验的泛化性”。

105. Birnbaum，互联网心理实验；Paolacci，Chandler和Ipeirotis，“在亚马逊的Mechanical Turk上运行实验”。

106. Buhrmester，Kwang和Gosling，“亚马逊的Mechanical Turk”。

107. Bartneck等人，“比较回应的相似性”，5/23。

108. Matthew J. C. Crump，John V. McDonnell和Todd M. Gureckis，“评估亚马逊的Mechanical Turk作为实验行为研究工具，”PloS One 8，第3卷（2013年）：1-18。

109. Krupnikov和Levine，“跨样本比较”。

110. 同上；Levay，Freeze和Druckman，“Mechanical Turk样本的人口学和政治组成”。

111. Krupnikov和Levine，“跨样本比较”。

112. 同上。

113. Emily Thorson，“信念回响：修正错误信息的持久影响，”《政治传播》33，第3卷（2016年）：460-480。

114. Huff和Tingley，“谁是这些人？”

115. Ellen Cushing，“亚马逊Mechanical Turk：数字血汗工厂，”Utne Reader January-February，http://www.utne.com/science-and-technology/amazon-mechanical-turk-zm0z13jfzlin（2013年）；John Horton和Lydia Chilton，“有薪众包的劳动经济学，”《第11届ACM电子商务会议论文集》（2010年）。

116. Bartneck等人，“比较回应的相似性”。

117. Buhrmester，Kwang和Gosling，“亚马逊的Mechanical Turk”。

118. Travis Simcox和Julie Fiez，“使用亚马逊的Mechanical Turk和Adobe Flash收集响应时间，”《行为研究》46（2014年）：95-111。

119. Scott Clifford和Jennifer Jerit，“方便的成本是否存在？实验室和在线研究数据质量的实验比较，”《实验政治科学杂志》1，第2卷（2014年）：120-131。

120. Berinsky，Huber和Lenz，“评估在线劳工市场”。

121. Bartneck等人，“比较回应的相似性”；Berinsky，Huber和Lenz，“评估在线劳工市场”；A. J. Berinsky，M. F. Margolia和M. W. Sances，“从游手好闲者中分辨工人？确保受访者在自行管理的调查中引起注意，”《美国政治学杂志》58，第3卷（2014年）：739-753；Huff和Tingley，“谁是这些人？”

122. D. Rand，“Mechanical Turk的承诺：在线劳动市场如何帮助理论学家进行行为实验，”《理论生物学杂志》299（2012年）：172-179。

123. J. Heer和M. Bostock，“众包图形感知：使用Mechanical Turk评估可视化设计”（论文发表在第11届ACM人类计算系统人因因素会议上，纽约，2010）。

124. Buhrmester，Kwang和Gosling，“亚马逊的Mechanical Turk”；Mason和Suri，“在亚马逊的Mechanical Turk上进行行为研究”。

125. Paolacci，Chandler和Ipeirotis，“在亚马逊的Mechanical Turk上运行实验”。

126. Buhrmester，Kwang和Gosling，“亚马逊的Mechanical Turk”；Heer和Bostock，“众包图形感知”；D. Kelly，D. Harper和B. Landau，“交互式信息检索实验中的问卷模式效应，”《信息处理与管理》44（2008年）：122-141。

127. Bartneck等人，“比较回应的相似性”；Simcox和Fiez，“收集响应时间”。

128. Paolacci，Chandler和Ipeirotis，“在亚马逊的Mechanical Turk上运行实验”。

129. M. Smith和B. Leigh，“虚拟受试者：使用互联网作为替代受试者和研究环境，”《行为研究方法》29（1997年）：496-505。

**246** Designing Experiments for the Social Sciences

130. Marriott，“在实验研究中使用学生样本”。

131. Gregg R. Murray等人，“方便但不是方便样本：陪审团作为实验对象池，”《社会科学研究》42，第1卷（2013年）：246-253。

132. Alan Gerber等人，“实验研究的报告指南：实验研究部门标准委员会的报告，”《实验政治科学杂志》1，第1卷（2014年）：81-98。

133. Alan S. Gerber等人，“报告平衡表，响应率和操作检查在实验研究中：来自准备报告指南的委员会的回复，”《实验政治科学杂志》2，第2卷（2016年）：216-229。

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

135. Bartneck等人，“比较回应的相似性”。

136. Cassese等人，“社交媒体调查”。

137. 同上。

138. 同上，777页。

139. Kam，“隐含态度，明确选择”。

140. Murray等人，“方便但不是方便样本”。

141. 有关将实验条件纳入英国选举研究2012年的研究示例，请参见Rosie Campbell和Philip Cowley，“选民想要什么：对调查实验中候选人特征的反应，”《政治研究》62，第4卷（2013年）：745-765；Murray等人，“方便但不是方便样本”。

142. David Niven，“乐观的另一面：高期望和拒绝现状政治，”《政治行为》22，第1卷（2000年）：71-88。

143. Lee Sigelman，Carol K. Sigelman和Barbara J. Walkosz，“公众与领导力的悖论：实验分析，”《美国政治科学》36（1992年）：366-385。

144. Nayda Terkildsen，“当白人选民评估黑人候选人：候选人肤色，偏见和自我监控的处理影响，”《美国政治科学杂志》37（1993年）：1032-1053。

145. Christopher J. Lewis，Dona-Gene Mitchell和Cynthia R. Rugeley，“迎合公众意见：利用陪审团池进行实验研究，”见于政治方法学会议（塔拉哈西，佛罗里达州2005年）。

146. 同上。

147. Murray等人，“方便但不是方便样本”，252页。

148. McEntee，Coleman和Yaschur，“比较生动的写作效果”。

149. Mitchell Hoffman和John Morgan，“谁淘气？谁是好人？对善意的工作者是否从残酷的商业环境中被挑选出来的实验，”《经济行为与组织》109（2015年）：173-187。

150. Kim和Ryzin，“政府资金的影响”；Murray等人，“方便但不是方便样本”。

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

152. 同上。

153. Willem Wetzels等人，“预付激励在面对面调查中的影响：用邮票进行大规模实验，”《国际舆论研究》20，第4卷（2008年）：507-516。

154. Dunlop和Radaelli，“教授监管谦卑”。

155. S. Mayr等人，“Gpower的简短教程，”《心理学量化方法教程》3，第2卷（2007年）：51-59。

156. 同上。

157. 同上。

158. 同上。

159. 例如，参见J. Cohen，“我所学到的事情（到目前为止），”《美国心理学家》45（1990年）：1304-1312；Jacob Cohen，“地球是圆的（P < .05），”《美国心理学家》49，第12卷（1994年）：997-1003。

160. L. J. Cronbach等人，行为测量的可靠性：得分和轮廓的普遍性理论（纽约：Wiley，1972年）；G. A. Marcoulides，“在预算限制下进行泛化性研究中的最大化功率，”《教育统计杂志》18，第2卷（1993年）：197-206。

161. Wilson Van Voorhis和Morgan，“理解权力和经验法则”。

162. Barbara G. Tabachnick和Linda S. Fidell，使用多变量统计（纽约：HarperCollins，1996年）；Esther Thorson，Robert H. Wicks和Glenn Leshner，“新闻学和大众传播研究中的实验方法，”《新闻学与大众传播季刊》89，第1卷（2012年）：112-124。

Chapter 8 ■ Sampling and Effect Sizes **247**

163. J. Descôteaux，“统计力量：历史介绍，”《心理学量化方法教程》3，第2卷（2007年）：28-34；Thorson, Wicks和Leshner，“新闻学中的实验方法”。

164. Rebecca B. Morton和Joshua A. Tucker，“实验，期刊和伦理，”《实验政治科学杂志》1，第2卷（2015年）：99-103。

165. Chia-Hu Chang等人，“网球视频的虚拟焦点广告，”《视觉传播与图像表示杂志》21，第7卷（2010年）：595-612；Justin W. Owens，Evan M. Palmer和Barbara S. Chaparro，“文本广告失明的普遍性，”《可用性研究杂志》9，第2卷（2014年）：51-69。

166. James R. Rest，判断道德问题的发展（明尼阿波利斯，MN：明尼苏达大学出版社，1979年）。

167. Thorson，Wicks和Leshner，“新闻学中的实验方法”。

168. Leland Wilkinson和统计推断工作组，美国心理学协会科学指导委员会，“心理学期刊中的统计方法：指南和解释，”《美国心理学家》54，第8卷（1999年）：594-604。

169. Thorson，Wicks和Leshner，“新闻学中的实验方法”。

170. D. R. Atkinson，M. J. Furlong和B. E. Wampold，“统计显着性，评审人评估和科学过程：是否有（统计上）显着关系？”《咨询心理学杂志》29（1982年）：189-194。

171. Graham N. Dixon，“典型效应的负面情感机制，”《传播研究》43，第6卷（2015年）：768。

172. H. Denis Wu和Renita Coleman，“政治判断上的情感效应：候选人属性和问题一致性的比较影响，”《新闻学与大众传播季刊》91，第3卷（2014年），530-543。

173. Renita Coleman等人，“公共生活和互联网：如果你建设一个更好的网站，公民会参与吗？”《新媒体与社会》10，第2卷（2008年）：179-201。

174. J. Cohen，行为科学的统计力量分析，第2版（Hillsdale，NJ：Erlbaum，1988年）。

175. Russell V. Lenth，“功率和样本大小的Java小程序，”爱荷华大学文理学院数学科学系，https://homepage.divms.uiowa.edu/~rlenth/Power/。

176. Erik Westlund和Elizabeth A. Stuart，“实验评估研究中未使用，误用和适当使用先导研究，”《美国评估杂志》38，第2卷（2017年）：246-261。

177. Descôteaux，“统计力量”。

178. Wilkinson和统计推断工作组，美国心理学协会科学指导委员会，“心理学期刊中的统计方法”。

179. S. B. Green，“进行回归分析需要多少受试者？”《多变量行为研究》26（1991年）：499-510。

180. R. J. Harris，“多元统计的基础，”第2版（纽约：学术出版社，1985年）。

181. J. Cohen和P. Cohen，应用多元回归/相关分析于行为科学（Hillsdale，NJ：Erlbaum，1975年）。

182. D. C. Howell，心理学统计学方法，第4版（贝尔蒙特，加利福尼亚州：Wadsworth，1997年）。

183. Tabachnick和Fidell，使用多变量统计。

184. E. J. Pedhazur和L. P. Schmelkin，测量，设计和分析：一种综合方法（Hillsdale，NJ：Erlbaum，1991年）。

185. A. L. Comrey和H. B. Lee，因子分析初阶，第2版（Hillsdale，NJ：Erlbaum，1992年）。

186. Cohen，行为科学的统计力量分析，第2版。

187. A. Aron和E. N. Aron，心理学统计学，第2版（上塞德尔河，新泽西州：普林斯顿大厅，1999年）。

188. 另请参阅：F. Faul等人，“G\*Power 3：社会，行为和生物医学科学的灵活统计力量分析程序，”《行为研究方法》39，第2卷（2007年）：175-191；Mayr等人，“Gpower的简短教程”。

189. Coleman，“记者对儿童的道德判断”。

190. Coleman，“视觉效果的影响”。

191. Cohen，行为科学的统计力量分析，第2版。

192. Descôteaux，“统计力量”。

193. Sarah V. Bentley，Katharine H. Greenaway和S. Alexander Haslam，“探索自我参照效应的在线范例，”《PloS ONE》12，第5卷（2017年）：1-21。

**248** Designing Experiments for the Social Sciences

194. Erica A. Hornstein和Naomi I. Eisenberger，“揭示社交支持人物的缓冲效应：社交支持减弱了恐惧习得，”《PloS ONE》12，第5卷（2017年）：1-9。

195. Satoris S. Culbertson，William S. Weyhrauch和Christopher J. Waples，“行为线索作为“结构化就业面试中欺骗指标，”《选择与评估国际期刊》24，第2卷（2016年）：119-131。

196. John M. Hoenig和Dennis M. Heisey，“滥用权力，”《美国统计学家》55，第1卷（2001年）：19-24；Daniel J. O’Keefe，“事后功效，观察功效，先验功效，回顾性功效，预期功效，实现功效：整理统计力量分析的适当用途，”《传播方法与措施》1，第4卷（2007年）：291-299；J. Descôteaux，“编辑手记：不受损害的统计力量，”《心理学量化方法教程》3，第2卷（2007年）：26-27。

197. Hoenig和Heisey，“滥用权力”。

198. 同上；O’Keefe，“事后功效，观察功效”。

199. Hoenig和Heisey，“滥用权力”；O’Keefe，“事后功效，观察功效”。

200. Hoenig和Heisey，“滥用权力”。

201. O’Keefe，“事后功效，观察功效”。

202. 同上。

203. Cohen，行为科学的统计力量分析，第2版；Hoenig和Heisey，“滥用权力”；O’Keefe，“事后功效，观察功效”。

9

# 刺激和操作检查

寻找一个大的物体而不是一个小的。 —R. Barker BausellLook for a large object rather than a small one.1

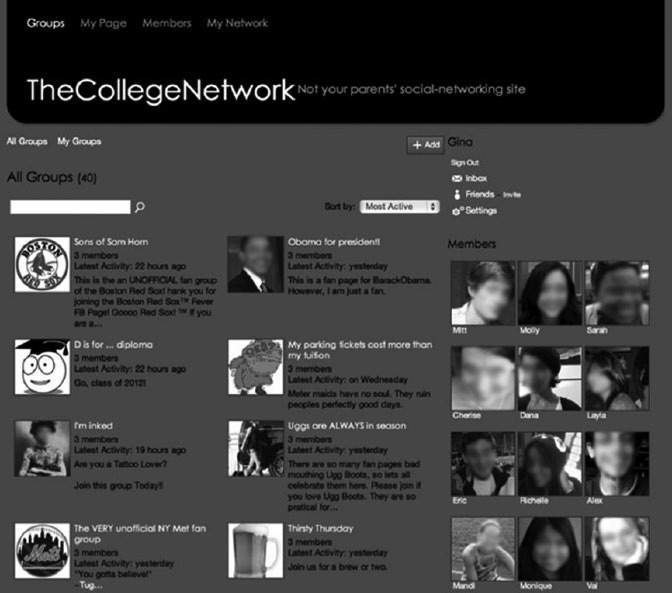
**—R. Barker Bausell**

|  |
| --- |
|  |
| 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。

T

**249**



**FIGURE 9.1**

## 研究重点9.1

## 创建逼真的刺激

SAGE Journal Article: study.sagepub.com/

coleman

Gina Masullo Chen

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。



**FIGURE 9.2A FAST CONDITION**

**Lee, Angela Min-Chia. 2014. How Fast Is Too Fast? Examining the Impact of Speed-Driven Journalism on News Production and Audience Reception (unpublished dissertation, The University of Texas at Austin).**

Dozens Dead in San Antonio Flooding, Landslides Live Text As It Happened

By NICHOLAS KULISH

**10:02 a.m**.: Torrential rains began on Sunday night and continued into Monday morning, causing power failures and cutting off roadways near The Far Northwest.

**9:40 a.m**.: Mayor Martinez de Vera said burials for the dead and treatment for the injured are being organized.

**9:28 a.m**.: “Many houses in the suburbs were destroyed by the rain,” Greenberg said. “Many of the victims are children, and the houses collapsed on them.”

For Review-No Commercial Use(2023)

**9:23 a.m**.: Police spokesman, Jon Hermes, said that hundreds more had been displaced because of damage to their homes. Willy Greenberg, spokesman for Mayor Art Martinez de Vera, said the damage is extensive.

**9:17 a.m**.: *CORRECTION—*At least 60, not 100, people were killed.

**9:12 a.m**.: At least 100 people were killed and dozens injured in San Antonio, after heavy rains caused flooding and landslides in Bexar County, officials said.



**FIGURE 9.2B SLOW CONDITION**

尼古拉斯·库利什报道：圣安东尼奥洪水和山体滑坡造成数十人死亡

据官方消息，由于暴雨引发了贝克萨尔县的洪水和山体滑坡，圣安东尼奥至少有60人死亡，数十人受伤。

警方发言人乔恩·赫尔姆斯表示，由于家园受损，还有数百人被迫离开。市长阿特·马丁内斯·德·维拉的发言人威利·格林伯格表示，损失十分严重。

格林伯格说：“郊区的许多房屋被雨水摧毁了。许多受害者是儿童，房屋倒塌在他们身上。”

马丁内斯·德·维拉表示，目前正在组织死者的葬礼和伤者的治疗。暴雨从星期天晚上开始，持续到星期一早上，导致电力故障并切断了位于远北西地区附近的道路。

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

###### Creating Realistic Social Media Stimuli

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。

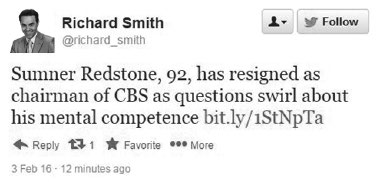
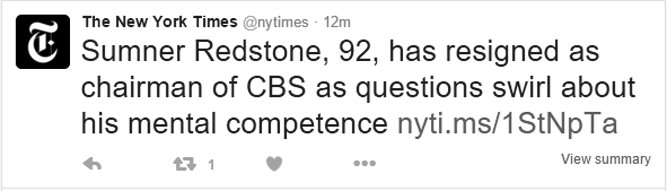
可以在以下网址找到这些工具：

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

CloneZone: clonezone.link

Prank Me Not: www.prankmenot.com

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



（续）

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](http://www.Simitator.com/)

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

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。



**FIGURE 9.4 ADVERTISEMENTS**



来源：金珊尼和杰弗里·T·汉考克，2016年。《广告特稿如何使广告模式失活》。《传播研究》，DOI：10.1177/0093650216644017。

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

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......"。

i感谢匿名审稿人提供这个明确简明的例子。

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”。

ii对于研究人员所谓的操纵检查是否实际上是中介心理状态的测量存在一些争议。操纵检查测量的是刺激的结构或内容属性，而不是受试者对内容的态度或感知；这些是参与者的属性，而不是刺激的属性。对于这一问题的详细讨论超出了本书的范围，但有兴趣的读者可以在O’Keefe的文章“Message Properties, Mediating States, and Manipulation Checks”中了解更多信息。

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

iii在这本书中，讨论中介变量的分析已超出了范围，但可以使用Baron和Kenny（1986）描述的回归分析来进行测试，也可以使用Efron和Tibshirani（1993）描述的Bootstrap分布测试，通过名为PROCESS的统计软件的免费宏来进行测试（Hayes，2013）。您可以在http://www.processmacro.org/index.html上获得该软件。R. M. Baron和D. A. Kenny，“社会心理学研究中的调节和中介变量区别：概念、策略和统计考虑”，《人格与社会心理学杂志》第51卷，第6期（1986年12月）：1173–1182；Andrew F. Hayes，《中介、调节和条件过程分析导论，第二版：基于回归的方法》（纽约：吉尔福德出版社，2013年）。

## 关于...的更多内容

## 中介变量和调节变量

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：

o 如果是虚构的，请描述如何让受试者相信它们是真实的，并

o 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。

**如何操作 9.4**

写作刺激描述

以下是一些研究人员描述刺激的例子。描述应该足够详细，让读者能够想象刺激的外观，并能够为他们自己的研究重现它们。

示例1

在这项研究中，刺激的描述与研究设计的细节相结合，涵盖在第6章中。请注意，这也涉及到使用虚构刺激和关于欺骗的被试反馈，以及如何讨论潜在的混淆因素以及它们是如何处理的：

“方法设计和刺激

研究1采用了单因素设计，以探究降低和提高声音音调对组织发言人认知的影响。每位参与者首先在线收到一小段文字，告知他们虚构公司正在发生的危机事件，然后听取组织发言人的声音录音。使用虚构组织使我们能够避免潜在的混淆效应，因为被试没有听说过这个组织或事件。然而，为了避免被试因为他们对该组织或事件没有了解而产生怀疑，危机发生地点被设定在国内的不同地区。实验结束时，被试被告知这次危机是虚构的。

简介中的文字解释了一个公司正在遇到食品污染危机。食用了这家公司生产的沙拉的消费者出现了胃部问题，并且似乎是由于该公司使用了当地政府禁止的农药而导致的。这次危机对公众构成威胁，增加了对适当组织反应的需求。

通过从虚构公司在危机中举行的新闻发布会中提取的简短音频片段来操纵声音音调。组织发言人确认了简介中提供的信息，告知潜在的受害者，并简要地阐述了所采取的纠正措施，并为此次危机道歉。我们通过录音了一名男性广播新闻学专业的学生在朗读关于危机事件的声明。之前的研究没有发现性别对声音音调与能力和支配的关系产生影响（DeGroot＆Motowidlo，1999; Jones，Feinberg，DeBruine，Little和Vukovic，2010; McHenry等，2012）。与以前的研究类似，我们通过计算机软件（例如Audacity）操纵录音中的声音音调（Imhof，2010; Jones等，2010; O'Connor等，2014）。

（接下来续）

注意：上面的示例演示了如何提供关于刺激、研究设计和程序的详细描述，同时讨论了潜在的混淆因素和被试反馈程序。它包括了虚构刺激的使用方式以及如何控制顺序效应的解释。此外，该示例讨论了使用操纵检查确认声音音调操纵效果的方法。这种详细程度有助于读者理解研究的方法，并允许将来的研究进行复制。

（续）

2011年；Puts等，2006年，2007年）。该软件允许我们操纵声音的一个特定方面，即音高，同时保持其他所有方面（如速度，强调，停顿，口音和发音）恒定。这种方法还允许我们控制内容（Imhof，2010年）。

我们制作了原始声音录音的两个版本，一个声调较高，一个声调较低（参见Jones等人，2010年；Puts等，2006年；Tigue等人，2012年）。基线水平的音高调高和调低了一个半音（参见Puts等，2006年，2007年）。”

来源：Claeys，An-Sofie，和Verolien Cauberghe。2014年。“保持控制：危机时刻组织代言人的非语言权力表达的重要性。”《传播学杂志》64：1162–1163。

例2

接下来的两个例子特别关注使刺激物尽可能真实。在第一个刺激物描述中，注意研究人员报告如何适应实际信息进行研究：

“刺激物材料

我们从媒体来源收集了一系列广告性文章和广告，以生成研究2的刺激物材料。每篇广告性文章都包含一个以健康饮食或睡眠为重点的文本信息部分，以及展示带有广告字幕的视觉部分。提供的健康饮食和睡眠信息与研究1中使用的信息相似。使用视觉编辑程序，我们将信息内容与广告内容分开并分别放置在信息优先和广告优先的广告性文章刺激物中。除了在结构上进行的实验性操作外，信息和传达的信息是相同的。仅包含广告内容的部分被用作普通广告（请参阅附录A以获取示例刺激物材料）。”

来源：Kim，Sunny Jung，和Jeffrey T. Hancock。2016年。“如何使广告合法化：基于MTurk的实验研究健康广告中的说服策略和结果。”《传播研究》44（7）：1019–1045。10.1177/0093650216644017。

例3

在这个例子中，研究人员请专家帮助他们创建真实的刺激物。同时注意这项研究从2003年开始进行，其中的“操作检查”不太可能通过对操作检查结果的充分报告进行评估：

“首先，我们向当地的政治顾问展示了我们的广告副本，以查看它们是否真实。他表示，这些广告是可能在选举期间听到的广告的合理表现，这使我们对我们的结果的普遍性有了更大的信心。其次，我们请一名作者所在大学的学生评估了模拟广告的负面程度和正面程度。我们希望确保我们的广告准确捕捉了我们感兴趣的变量。本科生对广告的评分与我们的判断非常吻合。换句话说，学生们认为我们的“攻击”广告是攻击性的，认为我们的“积极”广告是积极的。这个验证增加了我们对操作的信心。”

来源：Geer，John G.，和James H. Geer。2003年。“回忆攻击性广告：一项广播实验调查。”《政治行为》25（1）：69–95。

**A screenshot of a computer

Description automatically generated**

*A screenshot of a cell phone

Description automatically generated*

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。

Screens screenshot of a computer screen

Description automatically generated

**282** Designing Experiments for the Social Sciences

c. 控制混杂变量

d. 测量并在统计上控制混杂因素

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了解是很重要的。

a. 真

b. 假

4. 统计控制是 。

a. 不能在刺激中控制的因素

b. 被测量并控制的因素

c. A和B都是

d. 像ANOVA或卡方这样的统计技术

5. 统计控制应该在 。

a. 给予刺激之前测量

b. 给予刺激之后测量

c. 在实验进行中测量

d. 研究两周后测量

6. 为了避免受试者的先有了解污染，通常在实验中使用虚构的人物、地点和事件。这些被称为 。

a. 统计控制

b. 实验控制

c. 对照组

d. 处理组

7. 选择与特定政党没有明确联系的政治问题是一种 。

a. 统计控制

b. 实验控制

c. 对照组

d. 处理组

8. 协变量应该 。

a. 总是包括人口统计学信息

b. 包括在操作检查中

c. 随机分配

d. 与结果或因变量相关联

Chapter 9 ■ Stimuli and Manipulation Checks **283**

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。

**284** Designing Experiments for the Social Sciences

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？

创建研究刺激材料的建议阅读：

1. Graesser, A. C. 1981. Prose Comprehension Beyond the Word. 纽约，纽约：斯普林格-费尔拉格出版社。

2. Reeves, B. 和 S. Geiger. 1994. “设计评估心理反应的实验。”收录在《测量媒体心理反应》（A. Lang 编），第165-180页。希尔斯代尔，新泽西州：LEA出版社。

3. Slater, M. D. 1991. “在大众传播实验中使用信息刺激：方法论评估和讨论。”《新闻学季刊》68 (3): 412–421。

操纵检查的建议阅读：

1. O’Keefe, D. J. 2003. “信息属性，中介状态和操纵检查：实验说服信息效应研究中的声明，证据和数据分析。”《传播理论》13 (3): 251–274。

2. Tao, C. 和 E. P. Bucy. 2007. “概念化实验研究中的媒体刺激：心理学与归因为基础的定义。”《人类传播研究》33: 397–426。

固定因素与随机因素的建议阅读：

1. Chang, Yu-Hsuan A. 和 David M. Lane. 2016. “不仅限于受试者的泛化：混合效应模型的非数学教程。”《心理学定量方法》12 (3): 201–219。

2. Clark, H. 1973. “语言作为固定效应的谬误：对心理学研究中的语言统计的批判。”《口头学习与口头行为》12: 335–359. 10.1016/S0022-5371(73)80014-3.

3. Raaijmakers, J. G. W. 2003. “进一步探讨‘语言作为固定效应的谬误’。”《加拿大行为科学杂志》57: 141–151. 10.1037/h0087421.

4. Raaijmakers, J. G. W., J. M. C. Schrijnemakers 和 F. Gremen. 1999. “如何应对‘语言作为固定效应的谬误’：常见误解和替代解决方案。”《记忆与语言》41: 416–426. 10.1006/jmla.1999.2650.

Chapter 9 ■ Stimuli and Manipulation Checks **285**

**注释：**

1. R. Barker Bausell，《Conducting Meaningful Experiments: 40 Steps to Becoming a Scientist》（千橡市，加利福尼亚州：Sage出版社，1994），第95页。

2. James J. Gibson，《The Concept of the Stimulus in Psychology》，“Psychologist”，第15卷，第11期（1960年）：第694-703页。

3. 同上。

4. Michael D. Slater，《Messages as Experimental Stimuli: Design, Analysis, and Inference》。在第72届年度美国新闻与大众传播协会（华盛顿特区，华盛顿特区，1989年8月10日至13日）会议上发表的论文。

5. Esther Thorson、Robert H. Wicks和Glenn Leshner，《Experimental Methodology in Journalism and Mass Communication Research》，“Journalism and Mass Communication Quarterly”，第89卷，第1期（2012年）：第112-124页。

6. S. Shyam Sundar，《Theorizing Interactivity's Effects》，《Information Society》第20卷（2004年）。

7. Shana Kushner Gadarian和Bethany Albertson，《Anxiety, Immigration, and the Search for Information》，“Political Psychology”，第35卷，第2期（2014年）：第133-164页。

8. Elizabeth A. Bennion和David W. Nickerson，《The Cost of Convenience: An Experiment Showing E-Mail Outreach Decreases Voter Registration》，“Political Research Quarterly”，第64卷，第4期（2011年）：第858-869页。

9. Debby Vos，《How Ordinary MPs Can Make It Into the News: A Factorial Survey Experiment with Political Journalists to Explain the Newsworthiness of MPs》，“Mass Communication and Society”，第19卷，第6期（2016年）：第738-757页。

10. Theresa Beltramo等，《The Effect of Marketing Messages and Payment Over Time on Willingness to Pay for Fuel-Efficient Cookstoves》，《Journal of Economic Behavior and Organization》，第118卷（2015年）：第333-345页。

11. An-Sofie Claeys和Verolien Cauberghe，《Keeping Control: The Importance of Nonverbal Expressions of Power by Organizational Spokespersons in Time of Crisis》，《Journal of Communication》，第64卷（2014年）：第1160-1180页。

12. Horacio Alvarez-Marinelli等，《Computer Assisted English Language Learning in Costa Rican Elementary Schools: An Experimental Study》，《Computer Assisted Language Learning》，第29卷，第1期（2016年）：第103-126页。

13. Vos，《How Ordinary MPs Can Make It Into the News》。

14. Beltramo等，《The Effect of Marketing Messages》。

15. Claeys和Cauberghe，《Keeping Control》。

16. Alvarez-Marinelli等，《Computer Assisted English Language Learning》。

17. Angela Min-Chia Lee，《How Fast Is Too Fast? Examining the Impact of Speed-Driven Journalism on News Production and Audience Reception》（未发表的博士论文，德克萨斯大学，2014年）；Angela M. Lee，《The Faster the Better? Examining the Effect of Live-Blogging on Audience Reception》，《Journal of Applied Journalism and Media Studies》第8卷，第1期（待发表）。

18. Nick Lin-Hi、Jacob Horisch和Igor Blumberg，《Does CSR Matter for Nonprofit Organizations? Testing the Link between CSR Performance and Trustworthiness in the Nonprofit Versus For-Profit Domain》，《Voluntas》，第26卷（2015年）：第1944-1974页。

19. André Blais、Cengiz Erisen和Ludovic Rheault，《Strategic Voting and Coordination Problems in Proportional Systems: An Experimental Study》，《Political Research Quarterly》，第67卷，第2期（2014年）：第386-397页。

20. Pazit Ben-Nun Bloom，《Disgust, Harm and Morality in Politics》，《Political Psychology》，第35卷，第4期（2014年）：第495-513页。

21. Nicole LaFave，《Fake Tweets & More-Simitator》，取自http://edtechpicks.org/2016/05/fake-tweets-moresimitator/。

22. Canterbury College，《Simitator》，取自https://www.canterburycollege.ac.uk/teaching-learning-hub/technology/simitator/。

23. Trent R. Boulter，《Following the Familiar: Effect of Exposure and Gender on Credibility of Journalists on Twitter》，（德克萨斯大学，2017年），取自https://repositories.lib.utexas.edu/bitstream/handle/2152/62635/BOULTER-DISSERTATION-2017.pdf?sequence=1&isAllowed=y）。

24. Agne Kajackaite，《If I Close My Eyes, Nobody Will Get Hurt: The Effect of Ignorance on Performance in a Real-Effort Experiment》，《Journal of Economic Behavior and Organization》，第116卷（2015年）：第518-524页。

##### Notes

**286** Designing Experiments for the Social Sciences

25. Mina Tsay-Vogel，《我与他们：Facebook用户之间的第三人效应》，《新媒体与社会》，第18卷，第9期（2016年）：第1956-1972页。

26. John C. Tedesco，《2008年美国总统选举中的政治信息效能和互联网影响》，《美国行为科学家》，第55卷，第6期（2011年）：第696-713页。

27. Renita Coleman、Esther Thorson和Lee Wilkins，《测试公共卫生框架和丰富信息源在健康新闻报道中的影响》，《健康传播学杂志》，第16卷，第9期（2011年）：第941-954页。

28. Sophie Lecheler和Claes H. de Vreese，《一天之内有什么不同？随着时间的推移，重复和竞争性新闻框架的影响》，《传播研究》，第40卷，第2期（2013年）：第147-175页。

29. Claeys和Cauberghe，《保持控制》。

30. 例如，参见H. Denis Wu和Renita Coleman，《政治判断中的情感效应：候选人属性和议题一致性的影响比较》，《新闻学与大众传播季刊》，第91卷，第3期（2014年）：第530-543页。

31. Vos，《普通国会议员如何入选新闻》。

32. Claeys和Cauberghe，《保持控制》。

33. Vos，《普通国会议员如何入选新闻》。

34. 同上，第752页。

35. Bausell，《Conducting Meaningful Experiments》，第76页。

36. 同上，第95页。

37. C. Sansone、C. C. Morf和A. T. Panter，《社会心理学方法的Sage手册》（千橡市，加利福尼亚州：Sage出版社，2008年）。

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

39. Renita Coleman，《图像对伦理推理的影响：对新闻从业者做出道德决策的照片价值是多少？》，《新闻学与大众传播季刊》，第83卷，第4期（2006年）：第835-850页。

40. Stanley Milgram，《顺从权威：一个实验的观点》（纽约：哈珀和罗，1974年）。

41. 同上。

42. Bausell，《Conducting Meaningful Experiments》；Sansone、Morf和Panter，《社会心理学方法的Sage手册》。

43. Mutz和Pemantle，《实验研究的标准》。

44. Thorson、Wicks和Leshner，《新闻学的实验方法》。

45. 同上。

46. 同上。

47. Tsay-Vogel，《我与他们》。

48. Daniel J. O'Keefe，《劝说：理论与研究》（千橡市，加利福尼亚州：Sage，2002年）。

49. 同上。

50. 同上。

51. S. Jackson和S. Jacobs，《对消息的泛化：设计和实验分析建议，以及詹姆斯·布拉达克和迪恩·休斯的回应》，《人类传播研究》，第9卷，第2期（1983年）：第169-181页。

52. Coleman、Thorson和Wilkins，《测试公共卫生框架》。

53. Charles M. Judd, Jacob Westfall, 和 David A. Kenny，《将刺激视为社会心理学中的随机因素：一个全面且广泛被忽视问题的新解决方案》，《人格与社会心理学杂志》，第103卷，第1期（2012年）：54-69页。

54. Thorson, Wicks, 和 Leshner，《新闻学中的实验方法》，第119-120页。

55. Glenn Leshner, Paul D. Bolls, 和 Erika Thomas，《吓唬他们或恶心他们：图形健康促进信息的效果》，《健康传播学》，第24期（2009年）：447-458页。

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

57. D. J. O'Keefe，《消息属性、中介状态和操纵检查：实验性说服消息效应研究中的主张、证据和数据分析》，《传播理论》，第13卷，第3期（2003年）：251-274页。

58. Judd, Westfall, 和 Kenny，《将刺激视为随机因素的处理方法》。

59. 同上。

60. 同上。

61. Yu-Hsuan A. Chang 和 David M. Lane，《在刺激和受试者之间进行泛化：关于混合效应模型的非数学教程》，《心理学定量方法》，第12卷，第3期（2016年）：201-219页。

Chapter 9 ■ Stimuli and Manipulation Checks **287**

62. Judd, Westfall, and Kenny，《将刺激视为随机因素的处理》。

63. 同前文。

64. Chang 和 Lane，《跨刺激泛化》；Judd, Westfall, 和 Kenny，《将刺激视为随机因素的处理》。

65. Chang 和 Lane，《跨刺激泛化》。

66. 同前文。

67. 同前文。

68. 同前文。

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

70. Annie Franco, Neil Malhotra, 和 Gabor Simonovits，《社会科学中的出版偏见：打开文件抽屉》，《科学》345卷，第6203期（2014年）：1502–1505。

71. M. Foschi，《假设、操作化和操作性检查》，载于《社会科学中的实验室实验》，编者：Murray Webster 和 Jane Sell（纽约：Elsevier，2007）；Mutz 和 Pemantle，《实验研究的标准》；Thorson, Wicks, 和 Leshner，《新闻学中的实验方法学》。

72. Sansone, Morf, 和 Panter，《社会心理学方法手册》。

73. D. J. O’Keefe，《信息属性、中介状态和操作性检查》。

74. Sansone, Morf, 和 Panter，《社会心理学方法手册》。

75. Mutz 和 Pemantle，《实验研究的标准》。

76. Rebecca S. McEntee, Renita Coleman, 和 Carolyn Yaschur，《比较公共关系中生动写作和照片对道德判断的影响》，《新闻与传播学季刊》94卷，第4期（2017年）：1011–1130。

77. P. C. Cozby，《行为研究方法》，第10版（纽约：McGraw-Hill，2009）；B. C. Perdue 和 J. O. Summers，《检查营销实验中操作的成功性》，《市场营销研究杂志》23卷，第4期（1986年）：317–326。

78. Mutz 和 Pemantle，《实验研究的标准》，194页。

79. Rosie Campbell 和 Philip Cowley，《选民的需求：对调查实验中候选人特征的反应》，《政治研究》62卷，第4期（2013年）：745–765。

80. O’Keefe，《信息属性、中介状态和操作性检查》。

81. McEntee, Coleman, 和 Yaschur，《比较生动写作的影响》。

82. O’Keefe，《信息属性、中介状态和操作性检查》。

83. C. Tao 和 E. P. Bucy，《实验研究中媒体刺激的概念化：心理学与属性化定义的比较》，《人类传播研究》33卷（2007年）：397–426。

84. C. Tao 和 E. P. Bucy，《实验研究中媒体刺激的概念化》。

85. 同前文。

86. Amery D. Wu 和 Bruno D. Zumbo，《理解和使用中介变量和调节变量》，《社会指标研究：国际跨学科生活质量期刊》3期（2008年）。

87. 同前文。

88. James Hall 和 Pamela Sammons，《中介、调节和互动》，载于《教育研究定量方法手册》，编者：Timothy Teo（荷兰鹿特丹：Sense，2013年）。

89. Erica Scharrer，《超男性、侵略和电视暴力：一个实验》，《媒体心理学》7卷，第4期（2005年）：353–376。

90. Tao 和 Bucy，《实验研究中媒体刺激的概念化》。

91. 同前文，第413–414页。

92. R. M. Baron 和 D. A. Kenny，《社会心理学研究中的调节-中介变量区分：概念、策略和统计考虑》，《个性与社会心理学杂志》51卷，第6期（1986年）：1173–1182；M. D. Sobel，《结构方程模型中间接效应的渐近置信区间》，载于《社会学方法论》，编者：S. Leinhardt（旧金山：Jossey-Bass，1982年），212–290；Andrew F. Hayes，《中介、调节和条件过程分析入门》，第二版：基于回归的方法（纽约：Guilford Press，2013年）。

93. Mutz 和 Pemantle，《实验研究的标准》；B. Efron 和 R. J. Tibshirani，《引导入门》（纽约：Chapman & Hall，1993年）。

94. Mutz 和 Pemantle，《实验研究的标准》，197页。

95. 同前文。

96. 同前文。"

.

**288** Designing Experiments for the Social Sciences

97. 同前文。

98. 同前文。

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

100. McEntee, Coleman, 和 Yaschur，《比较生动写作的影响》。

101. Mutz 和 Pemantle，《实验研究的标准》。

102. 同前文。

103. Beltramo 等，《营销信息的效果》，336页。

104. 例如，参见 Vos，《普通议员如何进入新闻》。

105. 同前文，747页。

106. Lin-Hi, Horisch, 和 Blumberg，《非营利组织是否关心企业社会责任？》

107. 同前文，1955页。

108. Jueman (Mandy) Zhang, Di Zhang, 和 T. Makana Chock，《HIV/AIDS公益广告对态度和行为的影响：知觉威胁和自我效能的相互作用》，《社会行为与个性》42卷，第5期（2014年）。

10

# 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

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

|  |
| --- |
|  |
| 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。 |

**289**

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。

## MORE ABOUT . . . BOX 10.1

###### Writing Questionnaires

• 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。

A screenshot of a cell phone

Description automatically generated

A screenshot of a computer

Description automatically generated

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

###### Successful Measurements

• 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。

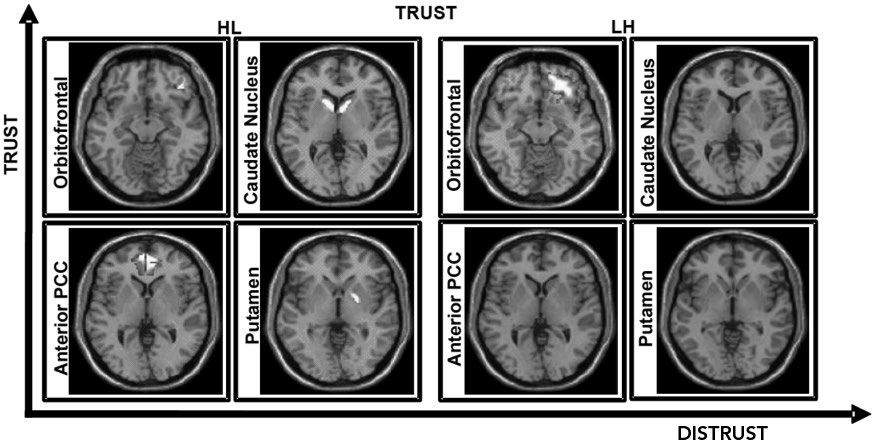
这些以及未在此列出的其他仪器，因其种类繁多，代表了隐式测试，即受试者并不一定有意识地意识到自己的反应。其中许多记录了心理或生理反应，例如心率。在社会科学中，有些仪器被更频繁地使用，而有些则更多地用于社会科学的特定学科领域。接下来，我们将重点介绍社会科学中一些较为流行的高科技仪器。由于篇幅限制，无法对每一种仪器的优势和劣势进行全面讨论，本文仅作简要介绍，感兴趣的读者应寻求更全面的文献来了解这些问题以及如何解决它们。

**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。

Dimoka (2012)

[fMRI的完整中文名称为功能性磁共振成像，是一种神经成像技术，可通过检测大脑血液流量的变化来探测大脑活动区域。它被广泛用于神经科学和心理学研究，用于研究不同心理过程和认知活动在大脑中的活动模式。fMRI是一种非侵入性的方法，对被测者没有任何伤害。



topic.54 Marketing studies that use fMRI have included research on cognitive processing when people see their favorite brand label,55 or choose between brands of beer and cof- fee,56 to understand willingness to pay,57 to see the effect of expert statements,58 and to

Dimoka (2012)

forecaFstosarlesRcheanvgiees bwas-edNoon pCoinot-omf-samle aedrs.c59 iIan lpoUlitiscael s(c2ien0ce2, 3fM)RI has been used to predict changes in attitudes toward candidates60 and to examine the relationship

between emotion and voting,61 among other things. In health communication, fMRI has been used to examine memory for health messages62 and to test the effectiveness of public service announcements (PSAs) about risky alcohol use.63 In economics, it has been used to examine reputation and trust in exchanges.64 This is not an exhaustive list of all fMRI studies in social science; nearly every discipline has at least some, including one that used fMRI to see the differences in cognitive processing between artistic and popular music.65

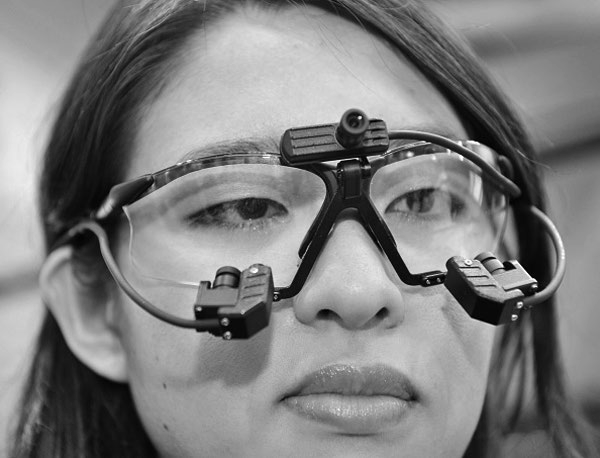
Using fMRI as an instrument is not as simple as using an instrument like a questionnaire. It takes a considerable amount of training,66 time, and effort to conduct such a study.67 It may even require a trained lab technician and possibly even a radiologist to look for brain abnor- malities if required by an Institutional Review Board.68 The equipment is expensive and large, consisting of a full-body tube where subjects lay down flat inside.69 Subjects can view stimuli through goggles or on a screen, and respond verbally or on a keyboard.70 Their heads must stay perfectly still, and they cannot have any metal piercings, implants, or medical issues.71 The scanner is noisy, and experimental tasks need to be limited to forty-five to sixty seconds.72 Because of the high cost and time it takes to run subjects one at a time, sample sizes are small, usually between ten and twenty subjects per study.73 Currently, fMRI in the social sciences is used mainly to complement other instruments, allowing for triangulation with self-reports or interviews, for example, or to measure concepts subject to the biases discussed earlier.74

**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

Image courtesy of Cedrus, [http://cedrus.com](http://cedrus.com/)

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



For Review-No Commercial Use(2023)

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/>）。

A screenshot of a computer

Description automatically generated

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。

## STUDY SPOTLIGHT 10.4

Unobtrusive Instruments—List, Choice, and Real Effort Task Experiments

Experimentalists do not necessarily have to avoid using questionnaires in order for an instrument to be unobtrusive. Several methods have been developed that still use questionnaires but avoid asking ques- tions directly, increasing truthful answers and avoiding nonresponses.225 Among them are three popu- lar with political scientists but adaptable to any social science field. In fact, the two choice-experiment studies described here come from marketing and food policy. This spotlight focuses on three types of unobtrusive instruments—the list experiment, choice experiment, and the real effort task—but there are others, such as the endorsement experiment. It should also be noted that these instruments can be used for surveys as well as experiments.

List Experiment

Kramon226 demonstrated how a list experiment can be more externally valid in that it is more reflec- tive of reality when the topic was the sensitive issue of vote buying. The strength of the list experi- ment is that it avoids the social desirability bias that can encourage participants to give socially acceptable answers rather than truthful ones. It does this by giving subjects a list of activities that are both sensitive and not sensitive, such as taking a bribe in exchange for their vote, or having seen political signs in their neighborhoods. Subjects are specifically told *not* to identify which particular items are true for them but to simply indicate *how many* items are true for them. This assumes subjects are more truthful when asked to give a number rather than answer questions directly. The control group receives only nonsensitive items, while the treatment group receives the same items as the control group but with the addition of sensitive items. The difference between group means gives the estimate of the existence of sensitive attitudes or behaviors.227 Kramon’s results showed significantly more reports of vote buying in the experiment than on surveys, which aligned with actual instances of vote buying.

For Review-No Commercial Use(2023)

Read the study at:

**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**

Choice Experiment

In choice experiments, subjects make decisions based on trade-offs and preferences—for example, between job offers with trade-offs between salary and benefits,228 or other types of consumer decisions including willingness to pay for certain attributes. Subjects make choices based on the characteristics of the different products or services offered, and researchers then use those to infer values the subjects hold—for example, if a higher salary is more important than a pension.

Subjects are asked to choose between alternatives rather than to rank or rate them.229 This avoids the social desirability bias by not asking subjects what they prefer directly, but by observing the attributes of the things they choose over those they do not. For example, in a study of recycling, respondents may

*(Continued)*

(续）

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。

**HOW TO DO IT 10.5**

Writing Up the Instrument and Measurement Sections

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。

（继续）

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. “The Effects of Internet Filtering on Users’ Information-Seeking Behaviour and Emotions.” Aslib Journal of Information Management 69 (4): 408–425.

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。.

**322** Designing Experiments for the Social Sciences

* Using self-reports when unobtrusive instruments are better suited
* Using technological instruments without proper theoretical understanding or methodological training
* Using single questions to represent complex theoretical constructs; multiple items combined in an index should be used instead.
* Using new questions created by the researcher when validated items and indexes already exist

**Common Mistakes**

##### Test Your Knowledge

1. In most cases, indexes of several questions .
   1. Are no better than single-item indicators
   2. Are superior for measuring multilevel constructs
   3. Help shorten a study
   4. Have more issues with reliability and validity

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

1. Single-item indicators are .
   1. Devices that subjects use to indicate their response
   2. Questions that are worded directly
   3. Multiple questions combined to measure a single item
   4. Inferior for measuring multilevel constructs
2. Advantages of questionnaires are that they .
   1. Are fast, cheap, and easy
   2. Can be private
   3. Reduce demand effects
   4. All of these
3. Disadvantages of questionnaires are .
   1. Self-reports can be inaccurate
   2. They can be subject to the social desirability bias
   3. They are not good for things people may not be able to introspect about
   4. All of these

Chapter 10 ■ Instruments and Measures **323**

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

**324** Designing Experiments for the Social Sciences

c. 间隔

d. 以上皆是

**答案**

1. b

2. d

3. d

4. d

5. c

6. c

7. b

8. a

9. b

10. d

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。

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。

2.

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。

**应用题**

问题撰写

图朗日，R. 2004。在《测试和评估调查问卷的方法》中，由S. Presser，J. M. Rothgeb，M. P. Couper，J. T. M. E. Lessler，J. Martin和E. Singer编辑，第209-224页。Hoboken, NJ: Wiley。

**推荐阅读**

Chapter 10 ■ Instruments and Measures **325**

单项指标

Petrescu，Maria。2013。“在结构方程模型中使用单项指标进行市场研究。”

《营销分析学杂志》1（2）：99-117。

功能性磁共振成像（fMRI）

该论文提供了关于为何以及如何进行fMRI研究的指导。

Dimoka，Angelika。2012。“如何在社会科学研究中进行功能性磁共振成像（fMRI）研究。”

《管理信息系统季刊》36（3）：811-840。

眼动追踪

Jankowski，Jaroslaw，Jaroslaw Watrobski，Katarzyna Witkowska和Pawel Ziemba。2016。“基于眼动追踪的在线内容参数对网页用户行为影响的实验评估。”

收录于《实验经济学的选定问题》（由K. Nermend和M. Latuszynska编辑），311-332页。瑞士，洛桑：斯普林格。

技术仪器

除了对每种仪器的章节外，还有一章讲述如何建立这些测量的实验室。Lang，Annieo，编。1994。《测量心理认知反应对媒体和心理学研究的意义》（第20卷第23期）：123。

观察者间一致性

有关教程，请参见：

Hallgren，K. A. 2012。“计算观测数据的观察者间可靠性：概述和教程。”

《心理学定量方法教程》8（1）：23-34。

更多关于效度的内容

Messick，S. 1995。“心理评估的效度：从个体行为的验证推论。” etrescu, Maria. 2013. “Marketing Research Using Single-Item Indicators in Structural Equation Models.”

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

|  |
| --- |
| **注释** |

**326** Designing Experiments for the Social Sciences

1. Christoph Bartneck等人，《比较亚马逊Mechanical Turk研究和在线与直接招募研究的相似性信息》，《政治心理学》第35卷，第2期（2014年）：133-164。

2. Shana Kushner Gadarian和Bethany Albertson，《焦虑，移民和对公益组织的信任寻求》，《PLoS ONE》第10卷，第4期（2015年）：1-23。

3. Nick Lin-Hi、Jacob Horisch和Igor Blumberg，《公益组织的企业社会责任是否重要？测试企业社会责任绩效与非营利组织信誉之间的联系》，1. Christoph Bartneck等，《从亚马逊Mechanical Turk研究与在线研究和直接招募研究中收到的回应相似性的比较》，《政治心理学》第35卷，第2期（2014年）：133-164。

2. Shana Kushner Gadarian和Bethany Albertson，《焦虑，移民和对公益组织的信任寻求》，《Voluntas：国际志愿与非营利组织期刊》第24卷，第4期（2013年）：1006-1036。

3. Nick Lin-Hi，Jacob Horisch和Igor Blumberg，“CSR是否对非营利组织重要？在非营利和营利领域测试CSR绩效与信誉之间的关联”，《Voluntas志愿与非营利组织国际期刊》第26卷（2015年）：1944-1974。

4. Julia Thaler和Bernd Helmig，“促进良好行为：社会和时间框架是否有影响？”《Voluntas：国际志愿与非营利组织期刊》第24卷，第4期（2013年）：1006–1036。

5. D. A. Dillman和C. D. Redline，“测试纸质自填问卷：认知访谈和现场测试比较”，收录于《测试和评估调查问卷方法》，S. Presser等人（Hoboken，NJ：Wiley，2004），299–317；R. Tourangeau，“测试和评估调查问卷的实验设计考虑”，收录于《测试和评估调查问卷方法》，S. Presser等人（Hoboken，NJ：Wiley，2004），209–224。

6. J. P. Wanous和M. J. Hudy，“单项可靠性：复制和扩展”，《组织研究方法》第4卷，第4期（2001年）：361–375。

7. J. P. Wanous和A. E. Reichers，“估计单项测量的可靠性”，《心理学报告》第78卷，第2期（1996年）：631–634。

8. C. Fuchs和A. Diamantopoulos，“在管理研究中使用单项测量：概念问题和应用指南”，《Die Betriebswirtschaft》第69卷，第2期（2009年）：195–210；Wanous和Hudy，“单项可靠性”。

9. Renita Coleman和Lee Wilkins，“新闻从业者的道德发展：与其他职业的比较和高质量道德推理的预测模型”，《新闻与传媒季刊》第81卷，第3期（2004年秋季）：511–527。

10. 同上。

11. A. Bandura，《自我效能：控制的实践》（纽约：W. H. Freeman and Company，1997）；Albert Bandura，“走向人类主体心理学”，《心理科学视角》第1卷，第2期（2006年）：164–180。

12. M. Kayhan Kurtuldu和Damla Bulut，“钢琴课程的自我效能量表的发展”，《教育科学：理论与实践》第17卷，第3期（2017年）：835–857。

13. Keiko Nanishi等，“确定母乳喂养自我效能量表-简式版的得分截点：日本干预研究的次要数据分析”，《PLoS ONE》第10卷，第6期（2015年）：1–12；R. C. Knibb，C. Barnes和C. Stalker，“父母管理食物过敏的信心：家长食物过敏自我效能量表（FASE-P）的发展和验证”，《临床与实验性过敏》第45卷，第11期（2015年）：1681–1689；Yanping Zhao等，“避孕自我效能量表（CSES）的翻译和验证中文版”，《艾滋病教育与预防》第28卷，第6期（2016年）：499–510；Franklin N. Glozah等，“评估本科生戒酒自我效能：酒精戒断自我效能量表的心理测量评估”，《健康与生活质量成果》第13卷（2015年）：1–6。

14. Shane W. Kraus等，“色情使用回避自我效能量表的发展和初步评估”，《行为成瘾杂志》第6卷，第3期（2017年）：354–363。

15. Carole Rodon和Aline Chevalier，“朝着更全面的中国互联网用户研究：中文普通话版8项网络信息检索自我效能量表（CH-IROWSE）的翻译和验证”，《国际人机交互杂志》第33卷，第10期（2017年）：846–855。

16. A. Bandura，“构建自我效能量表指南”，收录于《青少年的自我效能信念》，Tim Urdan和Frank Pajares编（夏洛特，NC：Information Age，2006）。

17. 同上。

18. Syeda Shahida Batool，Sumaira Khursheed和Hira Jahangir，“学术拖延作为低自尊的产物：学术自我效能的中介作用”，《巴基斯坦心理研究杂志》第32卷，第1期（2017年夏季）：195–211。

19. P. R. Pintrich和E. V. De Groot，“课堂学业表现的动机和自我调节学习成分”，《教育心理学杂志》第82卷（1990年）。

20. 同上，200。

21. Albert Bandura，“围绕感知的自我效能的错误概念的大肆报道”，《社会与临床心理学杂志》第26卷，第6期（2007年）：641–658。

Chapter 10 ■ Instruments and Measures **327**

22. Bandura，“构建自我效能量表指南”，308页。

23. Martin Fishbein和Icek Ajzen，“预测和改变行为”（纽约，NY：Taylor & Francis，2010年）。

24. Angela M. Lee，Renita Coleman和Logan Molyneux，“从思考到行动：不同社会规范对新闻伦理行为的影响”，《媒体伦理学杂志》第31卷，第2期（2016年）：72–85。

25. Chang-Dae Ham，Michelle R. Nelson和Susmita Das，“如何衡量说服知识”，《国际广告学杂志》第34卷，第1期（2015年）：17–53。

26. Renita Coleman，Esther Thorson和Lee Wilkins，“测试公共卫生框架和丰富新闻报道中的源头对健康新闻的影响”，《健康传播杂志》第16卷，第9期（2011年）：941–954。

27. Murray Webster和Jane Sell，“社会科学实验室实验”（阿姆斯特丹：Elsevier，2007年）。

28. Timothy Wilson，“知道何时提问：内省和适应性无意识”，《意识研究杂志》第10卷，第9/10期（2003年）：131页。

29. Rob Hoskin，“自我报告的危险”，《科学脑电波》。http://www.sciencebrainwaves.com/thedangers-of-self-report/

30. John B. McConahay，Betty B. Hardee和Valerie Batts，“美国的种族主义是否减少？这取决于谁在问什么”，《冲突解决杂志》第25卷，第4期（1981年）：563–579。

31. Cengiz Erisen，Elif Erisen和Binnur Ozkececi-Taner，“政治心理学中的研究方法”，《土耳其研究》第13卷，第1期（2013年）：13–33。

32. A. I. Jack和A. Roepstorff，编辑，《信任主题：认知科学中使用的内省证据》，2卷（埃克塞特，英国：Imprint Academic，2003年）；R. E. Nisbett和T. D. Wilson，“说得比我们知道的多：关于心理过程的口头报告”，《心理学评论》第84卷（1977年）：231–259。

33. Wilson，“知道何时提问”。

34. Sean P. Wojcik等，“保守派报告，但自由派显示更大的幸福”，《科学》第347卷，第6227期（2015年）：1243–1246。

35. Hoskin，“自我报告的危险”。

36. C. K. Madsen，“实际应用音乐练习与估计练习的30年后跟踪研究”，《音乐教育研究杂志》第52卷（2004年）：77–88。

37. C. C. Wang和D. W. Sogin，“小学音乐一般课堂活动的自我报告与观察活动”，《音乐教育研究杂志》第45卷（1997年）：444–456。

38. Renita Coleman，“视觉对道德推理的影响：新闻工作者对道德决策的照片价值是多少？”《新闻与传媒季刊》第83卷，第4期（2006年）：835–850。

39. Werner Bönte，Sandro Lombardo和Diemo Urbig，“经济学与心理学：个人竞争力的实验和自我报告测量”，《个性与个体差异》第116卷（2017年）：179–185。

40. Ralph Hertwig和Andreas Ortmann，“经济学中的实验实践：心理学家的方法论挑战？”《行为与脑科学》第24卷，第3期（2001年）：383页。

41. Emily Thorson，“信仰回响：修正错误信息的持久影响”，《政治传播》第33卷，第3期（2016年）：460–480。

42. 同上。

43. Rebecca B. Morton和Kenneth C. Williams，《实验政治学与因果研究：从自然到实验室》（纽约：剑桥大学出版社，2010年）。

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

45. 同上。

46. Eugene J. Webb等，《社会科学中的非反应性研究》（芝加哥：Rand McNally，1966年）；Eugene J. Webb等，《非反应性研究：修订版》（千橡市，加利福尼亚州：Sage，2000年）。

47. Immo Fritsche和Volker Linneweber，“非反应性（不显眼）方法”，收录于《心理测量手册——多方法视角》，M. Eid和E. Diener编辑（华盛顿特区：美国心理学会，2004年）。

48. Annie Lang编辑，《测量媒体对心理反应》（希尔斯代尔，新泽西州：Erlbaum，1994年）。

49. B. F. Skinner，“鸽子中的迷信”，《实验心理学杂志》第38卷（1948年）：168–172。

**328** Designing Experiments for the Social Sciences

50. Annie Lang，《测量心理对媒体的反应》中的“心脏能告诉我们什么？”（希尔斯代尔，新泽西州：Erlbaum，1994年）。

51. 同上。

52. Angelika Dimoka，“在社会科学研究中进行功能性磁共振（fMRI）研究的方法”，《管理信息系统季刊》第36卷，第3期（2012年）：811–840。

53. 同上。

54. Simone Kü̈hn，Enrique Strelow和Jürgen Gallinat，“大脑中的多个“购买按钮”：基于功能性脑激活使用fMRI预测销售巧克力的情况”，《神经影像》第136卷（2016年）：122–128。

55. S. Kuhn和J. Gallinat，“主观愉悦的神经基础”，《神经影像》第61卷（2012年）：289–294。

56. M. Deppe等，“在额叶前部皮层内的非线性反应揭示了特定隐含信息何时影响经济决策”，《神经影像学杂志》第15卷（2005年）：171–182。

57. H. Plassmann，J. P. O'Doherty和A. Rangel，“决策时在前额叶中编码愿意支付的积极和消极目标价值”，《神经科学杂志》第30卷（2010年）：10799–10808；Joshua B. Plavnick和Summer J. Ferreri，“教育研究中的单案例实验设计：教学和学习中因果分析的方法论”，《教育心理学评论》第25卷（2013年）：549–569；H. Plassmann，J. O'Doherty和A. Rangel，“在决策时编码每天经济交易中的支付意愿”，《神经科学杂志》第27卷（2007年）：9984–9988。

58. V. Klucharev，A. Smidts和G. Fernandez，“说服的大脑机制：‘专家权力’如何调节记忆和态度”，《社会认知与情感神经科学》第3卷（2008年）：353–366。

59. Kü̈hn，Strelow和Gallinat，“大脑中的多个“购买按钮””。

60. J. Kato等，“积极和消极广告后态度变化的神经相关性”，《行为神经科学前沿》第3卷，第6期（2009年）。

61. N. O. Rule等，“不同文化下的投票行为反映在杏仁核的反应中”，《社会认知与情感神经科学》第5卷（2010年）：349–355；M. I. Spezio等，“候选人形象对选举结果的影响的神经基础”，《社会认知与情感神经科学》第3卷（2008年）：344–352。

62. D. Seelig等，“健康促进视频的低感知信息更易记忆并激活与记忆编码相关的大脑区域”，《PLoS One》第9卷，第e113256期（2014年）。

63. Martin A. Imhof等，“真实生活中的健康信息如何引起我们大脑的共鸣：有效反酒精视频的共同处理”，《社会认知与情感神经科学》第12卷，第7期（2017年）：1188–1196。

64. B. King-Casas等，“了解你：声誉和信任在双人经济交换中”，《科学》第308卷，第5718期（2005年）：78–83。

65. Ping Huang等，“美学欣赏艺术和流行音乐之间的差异：基于fMRI研究的证据”，《PLoS ONE》第11卷，第11期（2016年）：1–14。

66. Dimoka，“在社会科学研究中进行功能性磁共振（fMRI）研究的方法”。

67. J. C. Culham，《认知功能神经影像学手册》中的“功能性神经影像学：实验设计和分析”，由R. Cabeza和A. Kingstone编辑（剑桥，马萨诸塞州：MIT Press，2006年），53–82；A. W. Song、S. A. Huettel和G. McCarthy，《认知功能神经影像学手册》中的“功能性神经影像学：功能性MRI的基本原理”，由R. Cabeza和A. Kingstone编辑（剑桥，马萨诸塞州：MIT Press，2006年），21–52。

68. Dimoka，“在社会科学研究中进行功能性磁共振（fMRI）研究的方法”

69. 同上。

70. 同上。

71. 同上。

72. 同上。

73. 同上。

74. 同上。

75. Lang，《测量心理对媒体的反应》中的“心脏能告诉我们什么？”

76. 同上。

77. 同上。

78. 同上。

79. A. Kadir ÖZer、Cemile Ekin Eremsoy 和 Emel Kromer，《测试焦虑中心理想象过程的生理测量》，社会科学电子期刊第15卷，第58期（2016年夏季）：903-916。

80. Haruka Shoda、Mayumi Adachi 和 Tomohiro Umeda，《现场表演如何动人心弦》，PLoS ONE 11卷，第4期（2016年）：1-11。

81. René A. de Wijk 等人，《自主神经系统反应和面部表情在商业早餐饮品口感中的区别》，PLoS ONE 9卷，第4期（2014年）：1-9。

82. Zachary Hohman 等人，《反毒品公益广告处理的生物心理学模型：发展有效的说服信息》，预防科学第18卷，第8期（2017年）：1006-1016。

83. Jie Xu，《信息感和认知价值：竞争或整合因素？》，健康传播第30卷，第6期（2015年）：589-597。

84. Himalaya Patel 等人，《对坏的接收开放：抖动运动使说服信息更有效》，人类行为计算机第32卷（2014年）：32-39。

85. Mareile Opwis 等人，《奇怪或联网的名人：名人代言人在能源商业广告中对心理生理反应模式的影响》，神经心理经济学会议论文集（2012年）：54。

86. Kevin Wise、Paul D. Bolls 和 Samantha R. Schaefer，《选择和阅读在线新闻：可用选择如何影响认知加工》，广播与电子媒体杂志第52卷，第1期（2008年）：69-85。

87. Zheng Wang、Alyssa C. Morey 和 Jatin Srivastava，《政治广告处理中的积极选择性关注：情感广告内容与候选人评估之间的动态相互作用》，传播研究第41卷，第1期（2014年）：119-156。

88. Erik Bucy 和 Samuel D. Bradley，《总统表情与观众情感：对电视领导人展示的反共情感回应》，社会科学信息第43卷，第1期（2004年）：59-94。

89. Sandra Ononogbu 等人，《信息和通信技术使用与学龄儿童一周学校期间睡眠质量的关联》，睡眠障碍（2014年）：1-6。

90. Robert Hopkins 和 James E. Fletcher，《电皮测量：对销售吸引力信息影响进行预测的特别有效方法》，在《测量媒体对心理反应的影响》一书中，Annie Lang 编著，（新泽西州希尔斯代尔：Erlbaum，1994年），113-132；Marieke G. N. Bos 等人，《对情感电影刺激的心理生理反应模式》，PLoS ONE 8卷，第4期（2013年）：1-8；Jonathan Renshon、Jooa Julia Lee 和 Dustin Tingley，《生理唤醒和政治信念》，政治心理学第36卷，第5期（2015年）：569-585；Xu，《信息感和认知价值》。

Chapter 10 ■ Instruments and Measures **329**

91. Hopkins 和 Fletcher，《电皮测量》。

92. 同上。

93. Michael Bang Petersen、Ann Giessing 和 Jesper Nielsen，《生理反应和党派偏见：超越自我报告的政党认同度测量》，PLoS ONE 10卷，第5期（2015年）：1-10。

94. Renshon、Lee 和 Tingley，《生理唤醒和政治信念》。

95. Barbara Krahé 等人，《对媒体暴力的脱敏：与习惯性媒体暴力暴露、攻击性认知和攻击性行为的关联》，人格与社会心理学杂志第100卷，第4期（2011年）：630-646。

96. Andreas Gregersen 等人，《跟踪观众：通过皮肤电导测量调查电视剧情节参与度》，诗学第64卷（2017年）：1-13。

97. Gabrielle Turner-Mcgrievy、Sri Kalyanaraman 和 Marci K. Campbell，《通过播客或网络传递健康信息：对心理社会和生理反应的媒体影响》，健康传播第28卷，第2期（2013年）：101-109；Chen-Bo Zhong 和 Julian House，《霍桑效应再探：物理工作环境的组织影响》，组织行为研究第32卷（2012年）：3-22。

98. Martin Reimann 等人，《我们如何与品牌相关：有关消费者-品牌关系的心理和神经生理学洞察》，消费者心理学杂志（爱思唯尔科学出版社）第22卷，第1期（2012年）：128-142；Peter Walla、Gerhard Brenner 和 Monika Koller，《品牌态度相关的情感客观测量：量化与市场营销相关的情感方面的新方法》，PLoS ONE 6卷，第11期（2011年）：1-7。

99. Mohammed Al-Fudail 和 Harvey Mellar，《研究教师使用技术时的压力》，计算机与教育第51卷，第3期（2008年）：1103-1110。

100. Guillaume Chanel, J. Matias Kivikangas 和 Niklas

Ravaja，《社交游戏分析的生理遵从性：合作与竞争性游戏》，与计算机交互 24卷，第4期（2012年）：306-316。

101. Bos 等人，《心理生理反应模式》；Renshon、Lee 和 Tingley，《生理唤醒和政治信念》。

102. 同上。

103. Bos 等人，《心理生理反应模式》。

104. Webster 和 Sell，《社会科学中的实验室实验》。

105. Hendrika Meischke 等人，《向有限英语能力的来电者提供9-1-1心肺复苏指导：模拟实验》，移民和少数民族健康杂志第17卷，第4期（2015年）：1049-1054。

106. Manuel Alonso Dos Santos 等人，《形象正负性对慈善广告所引起的注意力影响》，非营利和公共部门营销杂志第29卷，第3期（2017年）：346-363；Janet Hernández-Méndez 和 Francisco Muñoz-Leiva，《针对旅游业2.0的最有效在线广告类型：基于游客特征的注视跟踪研究》，人类行为与计算机第50卷（2015年）：618-625；C. S. Longman、A. Lavric 和 S. Monsell，《更多对注意力的关注？任务切换过程中感知属性选择的注视跟踪调查》，实验心理学杂志：学习、记忆和认知第39卷，第4期（2013年）：1142-1151。

107. Zijian Harrison Gong 和 Erik P. Bucy，《当风格掩盖实质：2012年总统辩论中显示适应性的视觉注意力》，传播学文集（2016年）：1-24；Hernández-Méndez 和 Muñoz-Leiva，《针对旅游业2.0的最有效在线广告》。

108. Yi-Chun Lin、Tzu-Chien Liu 和 John Sweller，《改善计算机模拟学习框架设计：确定孤立元素或瞬态信息效应的优先性》，计算机与教育第88卷（2015年）：280-291。

109. 同上。

110. 同上。

111. Benjamin G. Serfas、Oliver B. Büttner 和 Arnd

Florack，《睁大眼睛购物：购买冲动者的购物情境引发唤醒反应》，PLoS ONE 9卷，第12期（2014年）：1-9。

112. Yunying Dong 等人，《测试焦虑学生对威胁图片注意偏向的眼动证据》，认知与情绪31卷，第4期（2017年）：781-790；Serfas、Büttner 和 Florack，《睁大眼睛购物》。

113. Guillaume Hervet 等人，《标语瞎眼？眼动追踪互联网文本广告》，应用认知心理学第25卷，第5期（2011年）：708-716；Yoon Min Hwang 和 Kun Chang Lee，《根据购物动机使用眼动追踪研究消费者的视觉行为在移动环境中》，网络心理学、行为和社交网络学第20卷，第7期（2017年）：442-447；Nicole S. Dahmen，《干细胞辩论中的摄影构图：整合眼动数据进行媒体效应研究的新维度》，美国行为科学家第56卷，第2期（2012年）：189-203；Nora A. McIntyre、Robert M. Klassen 和 M. Tim Mainhard，《你是否希望成为教师？专家教师注视的文化、时间和动态见解》，学习与指导第49卷（2017年）：41-53；Cornelia Eva Neuert 和 Timo Lenzner，《将眼动追踪纳入认知访谈以预测试调查问题》，国际社会研究方法学第19卷，第5期（2016年）：501-519。

114. Marc Resnick 和 William Albert，《广告位置和用户任务对横幅广告视觉失明产生的影响：一项眼动跟踪研究》，国际人机交互第30卷，第3期（2014年）：206-219。

115. Gong 和 Bucy，《当风格掩盖实质》。

116. Alonso Dos Santos 等人，《形象正负性对慈善广告的影响》。

117. 同上。

118. Olena Kaminska 和 Tom Foulsham，《面对面和网络模式中的现实世界眼动跟踪》，调查统计学与方法学杂志第2卷，第3期（2014年）：343-359。

119. McIntyre、Klassen 和 Mainhard，《你是否希望成为教师？》

120. Hervet 等人，《标语瞎眼是否真实？》

121. Hwang 和 Lee，《使用眼动追踪探索》；Bartosz W. Wojdynski 和 Hyejin Bang，《在线新闻处理中的上下文广告干扰效应：眼动追踪研究》，行为与信息技术第35卷，第8期（2016年）：654-664。

**330** Designing Experiments for the Social Sciences

122. Jaroslaw Jankowski 等人，《基于眼动跟踪的在线内容参数对网络用户行为的实验评估》，收录于《实验经济学中的选定问题》，K. Nermend 和 M. Latuszynska 编，第311-332页，Springer出版社，2016年。

123. Hwang 和 Lee，《使用眼动追踪探索》。

124. Alonso Dos Santos 等人，《图像情感的影响》；Gong 和 Bucy，《风格遮盖实质》；Wojdynski 和 Bang，《上下文广告的干扰效应》。

125. Jankowski 等人，《基于眼动跟踪的实验评估》。

126. Dong 等人，《眼动证据》。

127. McIntyre, Klassen, 和 Mainhard，《你是否希望成为教师？》

128. Hernández-Méndez 和 Muñoz-Leiva，《在线广告的类型》；Hwang 和 Lee，《使用眼动追踪探索》。

129. Anastasia G. Peshkovskaya 等人，《囚徒困境游戏中决策的社会化效应：一项基于眼动追踪的研究》。

130. Dahmen，《干细胞辩论中的摄影构图》。

131. Gong 和 Bucy，《风格遮盖实质》。

132. Serfas, Büttner, 和 Florack，《睁大眼睛购物》。

133. Hwang 和 Lee，《使用眼动追踪探索》。

134. Alonso Dos Santos 等人，《图像情感的影响》；Hervet 等人，《标语瞎眼是否真实？》；Resnick 和 Albert，《广告位置的影响》；Wojdynski 和 Bang，《上下文广告的干扰效应》。

135. Sabrina Heike Kessler 和 Lars Guenther，《盯住框架》，互联网研究杂志，第27卷，第2期（2017年）：303-320。

136. Jankowski 等人，《基于眼动跟踪的实验评估》。

137. Lin, Liu, 和 Sweller，《改进框架设计》。

138. McIntyre, Klassen, 和 Mainhard，《你是否希望成为教师？》

139. Hernández-Méndez 和 Muñoz-Leiva，《在线广告的类型》。

140. Alonso Dos Santos 等人，《图像情感的影响》。

141. Neuert 和 Lenzner，《将眼动追踪融入认知访谈》；Hervet 等人，《标语瞎眼是否真实？》。

142. Kaminska 和 Foulsham，《现实世界眼动跟踪》。

143. Esther Thorson，《将眼睛用于电视测量》，收录于《测量心理反应于媒体》, Annie Lang 编（Hillsdale, NJ: Erlbaum出版社, 1994），65-84页。

144. 同上。

145. 同上。

146. D. Tewksbury，《美国人真正想知道什么？跟踪互联网上的新闻读者行为》，传播学杂志，第53卷，第4期（2003年）：694-710；Martijn Kleppe 和 Marco Otte，《通过多模态研究设计跟踪在线用户行为分析和理解新闻消费模式》，数字人文学，第32期（2017年）：158-170。

147. Renita Coleman 等人，《公众生活与互联网：如果你构建一个更好的网站，公民会参与其中吗？》新媒体与社会，第10卷，第2期（2008年）：179-201。

148. E. Menchen-Trevino 和 C. Karr，《使用Roxy采集经过知情同意的观察性网络数据》，信息技术与政治学杂志，第9卷，第3期（2012年）：254-268；S. A. Munson、S. Y. Lee 和 P. Resnick，《使用浏览器小工具鼓励阅读多样的政治观点》，收录于第七届国际博客与社交媒体大会论文集（Cambridge, MA: AAI Press, 2013），419-428，http://www.aaai.org/ocs/index.php/ICWSM/ICWSM13/paper/view/6119，访问日期为2015年9月25日。

149. Kleppe 和 Otte，《分析和理解新闻消费模式》。

150. P. Batista 和 M. Silva，《挖掘在线报纸的Web访问日志》，第2届国际自适应超媒体和自适应Web系统会议论文集，http://xldb.di.fc.ul.pt/xldb/publications/rpec02.pdf.（2002）。

151. Kleppe 和 Otte，《分析和理解新闻消费模式》。

152. 同上；Menchen-Trevino 和 Karr，《研究现实世界网络使用》。

153. O. Findahl，《瑞典人和互联网2009年》（Gavle：World Internet Institute，2009年）。

Chapter 10 ■ Instruments and Measures **331**

154. Coleman 等人，《公众生活与互联网》。

155. Batista 和 Silva，《挖掘Web访问日志》。

156. S. Ebersole，《学生对Web的使用和满足》,《计算机中介交流》 6，第1期 (2000)。

157. Ham, Nelson, 和 Das，《如何测量说服知识》。

158. Glen T. Cameron 和 David A. Frieske，《回答所需的时间：记忆响应延迟的测量》，收录于《测量心理反应于媒体》, A. Lang 编（Hillsdale, NJ: Erlbaum出版社, 1994），149-164页。

159. Marco Meyer 和 Harald Schoen，《响应延迟和直接民主环境中的态度-行为一致性：来自德国地方公投的证据》，《政治心理学》第35卷，第3期（2014年）：431-440。

160. Dermot Barnes-Holmes 等人，《隐含关联评估程序：探索私人和公共环境以及反应延迟标准对爱尔兰白人个体的亲白和反黑刻板印象的影响》，《心理记录》第60卷，第1期（2010年冬季）：57-79。

161. A. G. Greenwald, D. E. McGhee, 和 J. L. K. Schwartz，《测量隐性认知中的个体差异：隐性联想测试》，《个性与社会心理学杂志》第74卷，第6期（1998年）：1464-1480。

162. A. G. Greenwald 等人，《理解和使用隐性联想测试：III. 预测效度的元分析》，《个性与社会心理学杂志》第97卷（2009年）：17-41。

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

164. 同上。

165. Cameron 和 Frieske，《回答所需的时间》。

166. W. Jeffrey Burroughs 和 Richard A. Feinberg，《使用反应延迟测量发言人效果》，《消费者研究杂志》第14卷，第2期（1987年）：295-299。

167. Cheryl Bracken, Gary Pettey, 和 Mu Wu，《重新探讨在电视存在感研究中使用二次任务反应时间措施：探索沉浸和注意力的作用》，《AI与社会》第29卷，第4期（2014年）：533-538；Justin Robert Keene 和 Annie Lang，《公益宣传中情感轨迹的动态激励处理》，《传播学专著》第83卷，第4期（2016年）：468-485；Glenn Leshner 等人，《当恐惧言论不仅仅是恐惧言论：图形反烟草信息的效果》，《广播与电子媒体杂志》第54卷，第3期（2010年）：485-507。

168. Thierry Olive 等人，《儿童写作的认知努力和流利性：体裁和书写自动化的影响》，《学习和教学》第19卷，第4期（2009年）：299-308。

169. Cornelia Schoor，Maria Bannert，和 Roland Brünken，《多媒体学习中测量认知负荷的双重任务设计的角色》，《教育技术研究与发展》第60卷，第5期（2012年）：753-768。

170. Meyer 和 Schoen，《响应延迟和态度-行为一致性》。

171. 同上。

172. 同前引文。

173. Coleman，《视觉对道德推理的影响》。

174. Saul Fine 和 Merav Pirak，《快速和慢速伪装：用于人员测试中测量伪装的个体反应时间延迟》，《商业与心理学杂志》第31卷，第1期（2016年）：51-64。

175. Noam Tractinsky 等人，《评估Web页面即时审美感知的一致性》，《人机交互国际研究杂志》第64卷，第11期（2006年）：1071-1083。

176. Ham, Nelson, 和 Das，《如何测量说服知识》。

177. 同上。

178. Meyer 和 Schoen，《响应延迟和态度-行为一致性》。

179. Cameron 和 Frieske，《回答所需的时间》。

180. Fine 和 Pirak，《快速和慢速伪装》。

181. Glenn Leshner，Paul D. Bolls 和 Erika Thomas，《吓唬或让人厌恶：图形健康促进信息的影响》，《健康沟通》第24卷（2009年）。

182. Cameron 和 Frieske，《回答所需的时间》。

183. 同上。

184. M. D. Basil，《次级反应时间测量》，收录于《测量心理反应于媒体》, A. Lang 编（Hillsdale, NJ: Erlbaum出版社, 1994），85-98页。

185. 同上，85页。

**332** Designing Experiments for the Social Sciences

186. 同上。

187. Bracken, Pettey, 和 Wu，《重访使用次级任务反应》。

188. Leshner, Bolls, 和 Thomas，《吓唬或让人厌恶》。

189. Bracken, Pettey, 和 Wu，《重访使用次级任务反应》。

190. Keene 和 Lang，《动态动机过程的情感轨迹》。

191. 同上；Schoor, Bannert, 和 Brücken，《双重任务设计的角色》。

192. Bracken, Pettey, 和 Wu，《重访使用次级任务反应》。

193. Gong 和 Bucy，《当风格掩盖实质时》；Frank Biocca, Prabu David 和 Mark West，《连续反应测量（CRM）：一种用于研究沟通信息认知加工的计算工具》，收录于《测量心理反应于媒体》, A. Lang 编（Hillsdale, NJ: Erlbaum出版社, 1994），15–64页。

194. 同上。

195. Keene 和 Lang，《动态动机过程的情感轨迹》。

196. Gong 和 Bucy，《当风格掩盖实质时》。

197. Jeremy Saks 等，《电视政治辩论中的连续反应测量及其对观众的影响》，《播音与电子媒体杂志》第60卷，第2期（2016年）：231-247。

198. Jason M. Silveira，《音乐欣赏课中节奏感知及其与教师效果和教师强度的关系》，《音乐教育研究杂志》第62卷，第3期（2014年）：302-318。

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

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

201. M. A. Shapiro，《大声朗读和思考清单程序在调查心理过程中的应用》，收录于《测量心理反应于媒体信息》，Annie Lang 编（Hillsdale, NJ: Lawrence Erlbaum Associates，1994年），1-14页。

202. Coleman 等，《公众生活与互联网》。

203. Hamid R. Jamali 和 Pria Shahbaztabar，《互联网过滤对用户信息寻求行为和情绪的影响》，《Aslib信息管理杂志》第69卷，第4期（2017年）：408-425；Pengyi Zhang 和 Dagobert Soergel，《信息寻求与意义建构过程中的过程模式和概念变化：一项定性用户研究》，《信息科学杂志》第42卷，第1期（2016年）：59-78。

204. Heather Desurvire 和 Magy Seif El-Nasr，《游戏用户研究方法：研究玩家行为以提升游戏设计》，《IEEE计算机图形与应用》第33卷，第4期（2013年）：82-87。

205. Coleman，《视觉对道德推理的影响》；Carter Gibson 等，《学术界伦理情境中权力差异的定性分析》，《伦理与行为》第24卷，第4期（2014年）：311-325。

206. Ham, Nelson, 和 Das，《如何测量说服知识》；Robert C. Sinclair, Tanya K. Lovsin 和 Sean E. Moore，《情绪状态、问题参与度和论据强度对对说服呼吁的反应》，《心理报告》第101卷，第3期（2007年）：739-753。

207. Jeffrey L. Bernstein，《大声朗读和思考清单协议可以教会我们有关人们如何处理政治信息的内容》，2008年美国政治学会教学与学习会议，加利福尼亚州圣何塞，2008年2月。

208. Shapiro，《大声朗读和思考清单程序》。

209. 例如，参见Coleman等，《公众生活与互联网》。

210. Shapiro，《大声朗读和思考清单程序》。

211. Debra Malmberg, Marjorie Charlop, 和 Sara Gershfeld，《两个实验治疗比较研究：对自闭症谱系障碍儿童进行社交技能教学》，《发展和生理残疾研究杂志》第27卷，第3期（2015年）：375-392

212. Sandra L. Gibbons, Vicki Ebbeck, 和 Maureen R. Weiss，《儿童公平竞赛：对体育课道德发展的影响》，《运动研究季刊》第66卷，第3期（1995年）：247-255。

213. Albert Bandura，《模型的强化条件对模仿反应习得的影响》，《人格与社会心理学杂志》第1卷，第6期（1965年）：589-595。

Chapter 10 ■ Instruments and Measures **333**

214. 同上。

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

216. Wojcik等，《保守派报告》。

217. Bönte，Lombardo和Urbig，《经济学与心理学》。

218. Meischke等，《提供急救CPR指令》。

219. K. A. Hallgren，《计算观察数据的一致性：概述和教程》，《心理学定量方法教程》第8卷，第1期（2012年）：23-34。

220. Shadish，Cook和Campbell，《推广因果推断的实验和准实验设计》。

221. Kimberly A. Neuendorf，《内容分析指南》（千橡，加利福尼亚州：Sage，2002年），145页。

222. Webster和Sell，《社会科学实验室实验》。

223. Webb等，《无干扰测量》。

224. Stanley Milgram，Leon Mann和Susan Harter，《丢失的信件技术：社会科学研究的一种工具》，《公共舆论季刊》第29卷，第3期（1965年）：437-438。

225. Bryn Rosenfeld，Kosuke Imai和Jacob N. Shapiro，《敏感问题流行调查方法的实证验证研究》，《美国政治科学杂志》第60卷，第3期（2016年）：783-802。

226. Eric Kramon，《购票在肯尼亚的有效性在何处？来自肯尼亚的列表实验的证据》，《选举研究》第44卷（2016年）：397-408。

227. Rosenfeld，Imai和Shapiro，《实证验证研究》。

228. Jing Chen和Pallavi Chitturi，《用于估计主效应和相互作用的选择实验》，《统计规划与推理杂志》第142卷，第2期（2012年）：390-396。

229. Kimberly Klaiman，David L. Ortega和Cloé Garnache，《美国消费者对食品包装回收的知觉障碍：来自选择实验的证据》，《食品控制》第73卷（2017年）：291-299。

230. 同上。

231. M. L. Loureiro和W. J. Umberger，《牛肉的选择实验模型：美国消费者对食品安全、原产地标签和可追溯性的相对偏好》，《食品政策》第32卷，第4期（2007年）：496-514。

232. Chen和Chitturi，《用于估计主效应和相互作用的选择实验》。

233. Loureiro和Umberger，《牛肉的选择实验模型》。

234. Chen和Chitturi，《用于估计主效应和相互作用的选择实验》。

235. Klaiman，Ortega和Garnache，《美国消费者对食品包装回收的知觉障碍》。

236. Loureiro和Umberger，《牛肉的选择实验模型》。

237. Axel Sonntag和Daniel John Zizzo，《提醒效应、辍学和占优势：来自在线实验的证据》，《PLoS ONE》第10卷，第8期（2015年8月7日）：1-17。

238. Rosenfeld，Imai和Shapiro，《实证验证研究》。

239. James R. Rest等，《后常规道德思维：新柯尔伯格心理学方法》（马瓦，新泽西州：Erlbaum，1999年）。

240. 劳伦斯·柯尔伯格，《道德发展的哲学：道德阶段与正义观念》（剑桥，马萨诸塞州：Harper & Row，1981年）；劳伦斯·柯尔伯格，《道德发展的心理学：道德阶段的本质与有效性》（旧金山：Harper & Row，1984年）。

241. James R. Rest、Lynne Edwards和Stephen J. Thoma，《设计和验证道德判断的测量方法：阶段偏好和阶段一致性方法》，《教育心理学杂志》第89卷，第1期（1997年3月）：5-28。

242. Coleman，《图像对道德推理的影响》。

243. R. M. Alvarez和C. H. Franklin，《不确定性和政治观念》，《政治学杂志》第56卷，第3期（1994年）：671-688。

244. Emily Thorson，《信念回响》。

245. Erisen、Erisen和Ozkececi-Taner，《政治心理学研究方法》。

246. H. Cho和F. J. Boster，《增益对损失框架的反毒广告对青少年的影响》，《传播学杂志》第58卷，第3期（2008年）：428-446。

247. Alan Gerber等，《实验研究的报告指南：实验研究部门标准委员会的报告》，《实验政治学杂志》第1卷，第1期（2014年）：81-98。

**334** Designing Experiments for the Social Sciences

248. Thomas Zerback、Thomas Koch和Benjamin Kramer，《考虑他人：隐性和显性媒体线索对舆论氛围感知的影响》，《新闻与传播学季刊》第92卷，第2期（2015年）：421-443。

249. Thorson，《信念回响》，467页。

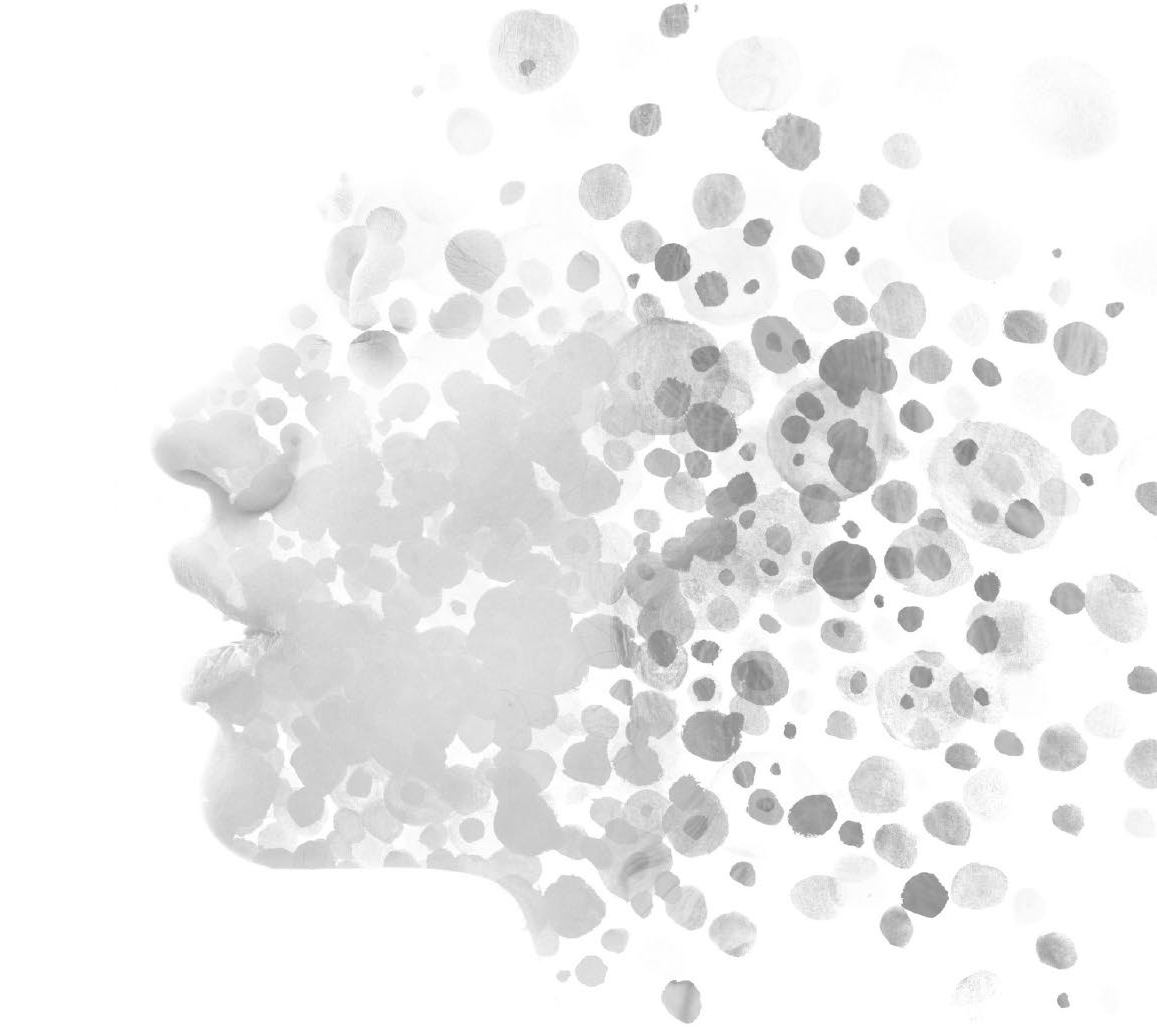
250. Thaler和Helmig，《促进良好行为》，1018页。

251. Srividya Ramasubramanian，《通过媒体手段减少新闻报道激活的种族刻板印象》，《新闻与传播学季刊》第84卷，第2期（2007年夏季）：249-264。

252. Jamali和Shahbaztabar，《互联网过滤的影响》，412页。

Chapter 10 ■ Instruments and Measures **335**

For Review-No Commercial Use(2023)



For Review-No Commercial Use(2023)