

第五章 实验室研究

研究的类型

科学研究林林总总，但是总是会涉及到理论和数据。大体上，我们根据理论和数据的关系，可以把研究归为三类：一类有数据支持但无理论指导（**data without theory**），一类有理论但无数据支持（**theory without data**），第三类既有理论又有数据（**theory with data**）。

首先来看第一类的研究。例如，你通过调查发现，中国人喜欢吃米饭，美国人喜欢吃土豆；中国人喜欢喝茶，美国人喜欢喝咖啡；中国人喜欢吃豆沙包子，美国人喜欢吃奶酪蛋糕。尽管有这些发现，但是你没有一个理论能够帮助你解释为什么中国和美国人在饮食上存在这样的差异。此外，这些数据也不能帮助你预测中国人和美国人对于其他饮食的偏好。这样的研究就属于只有数据但是没有理论指导的研究。

再看第二类的研究。假如你建立了一个模型，来描述一个人领到的奖金和他受到的激励之间的关系。你的模型有很多非常漂亮的参数，你把奖金和激励的关系完全量化了，看上去你似乎可以精确地预测多少奖金可以产生多少激励。但是问题在于，你并没有实证的数据来检验自己的模型到底对不对，也就是说，你没有办法知道你的预测在现实生活中到底成立不成立。这样的研究就属于只有理论但是没有实证数据支持的研究。需要指出的是，在这里我们所说的是实证数据，是从现实中得来的，而不是根据你的模型计算出来的数据。

令人可惜的是，很多管理学和经济学的研究往往落入此二类。第一类的文章往往有满页的表格，整段的事实，但是仅仅停留在数据层面，而没有上升为理论，例如很多管理学方面的研究。第二类的文章恰恰相反，只有理论建模没有实证，例如很多经济学的文章。这两类研究都是不可取的，可取的研究应该既有理论的指导，又有数据的支持，即第三类的研究。比方说，我们的理论认为人们在预测别人的偏好的时候，往往把自己的偏好强加于别人。由此可以导出很多预测，例如中国人因为自己喜欢吃中国菜，更容易高估喜欢吃中国菜的美国人。为了验证该理论是不是正确，可以让中国人首先回答他们自己是不是喜欢吃中国菜，然后让他们估计喜欢吃中国菜的美国人的比例；同时，让美国人回答他们是不是喜欢吃中国菜，从而得出喜欢吃中国菜的美国人的真实比例。这样可以获得一些数据并用统计方法来检验理论和数据是不是相符合，从而完成了一个既有理论指导，又有数据支持的研究。一个研究只有有了理论指导和数据支持，才可能经得起检验。

理论和假设

那么，到底什么是理论呢？理论就是解释和预测某些现象的一系列假设（Schweigert, 2006），通常用来解释已经发生的事件和预测未来的事件。我们需要用数据来支持待证的理论，或者用理论来解释现有的数据。

假设是关于自变量和因变量之间关系的陈述，用以解释某个现象。在这里现象就是因变量，而导致这个现象的原因是自变量。例如你假设使用大的电脑显示器能够提高员工的

工作积极性，那么员工的工作积极性就是因变量，而电脑显示器的大小则是自变量。这个“显示器与积极性”的例子将贯穿这一部分接下来的内容。我们将用这个例子对假设涉及到的一些概念进行解释。之后，我们还要讨论判断假设优劣的标准。

自变量 (Independent Variable)

什么是自变量呢？自变量就是在你的假设中引起某个现象的变量，也是实验中被实验者所操纵的变量。在“显示器与积极性”的例子中，显示器的大小就是自变量。实验者改变显示器的大小，来检验显示器的大小是否会影响员工的工作积极性。

显示器的大小会有不同的尺寸表示，同样的，自变量通常会拥有几个不同的取值，每一个取值就叫做自变量的一个水平(Level)，可以分为有限的或无限的，也可以分为离散的或连续的。有的自变量有有限个离散的水平。电脑屏幕的大小就是这样，我们现在在市面上能买到的电脑屏幕只有有限的几个尺寸，并且它的大小也不可能是连续变化的。而有的自变量则可以是连续的。比方说温度就是一个连续的变量。一般在实验中，我们并不能检测一个连续变量的所有可能的值。

因变量 (Dependent Variable)

因变量就是在你的假设中被预测的变量，或者实验者认为会随着自变量变化而变化的变量。在“显示器与积极性”的例子中，员工的工作积极性就是因变量。

几种简单假设的形式

从自变量个数的角度来看，最简单的假设是单一自变量假设。在单一自变量假设中，更为简单的情况是这个自变量只有两个水平。比方说，电脑屏幕的大和小。如果只想知道大屏幕和小屏幕对工作积极性的不同影响，那么一个自变量取两个水平就足够了。

但是，如果你想知道屏幕的大小和工作积极性是否存在非线性关系，你就需要多取几个值。比方说，你的假设说：电脑屏幕很小的时候，人们工作积极性很低；大一些的屏幕能够提高员工的工作积极性；但是电脑屏幕大到了一定程度，工作积极性就不再随屏幕大小变化了。为了检验这个假设，你需要最少取3个值：电脑屏幕非常小，电脑屏幕中等，电脑屏幕很大。可见，自变量的水平不是随机决定的，而是根据你的假设来确定的。

还是从自变量个数的角度来看，如果一个假设有两个或两个以上的自变量，我们称这样的假设为多自变量假设。比方说，你有一个假设说：工作年限短的员工使用大显示器比使用小显示器工作积极性高，但是工作年限长的员工使用两种显示器时工作积极性差不多。这就是一个有两个自变量的假设，一个自变量是显示器的大小，另外一个工作年限。依此类推，你也可以把自变量增加到三个，四个，甚至更多。

从因变量的角度来看，我们也可以有不止一个因变量。什么时候我们需要多个因变量呢？有时加入另外一些和主要因变量相似的因变量，只是为了从另外的角度来加强实验的有效性；有时我们的理论本身就在关注自变量对两个以上的因变量的影响。

什么是好的假设

一个好的科学研究，最首要的前提就是要有好的假设。对于一个研究者来说，假设的检验固然重要，但是首先他要有一个好的假设。在这一部分，我们将着重讨论什么样的假设才是好的假设。很多经典的研究之所以经典，都是因为他们的假设回答了一个非常重要的，并且以往的研究都没能回答好的问题。自然而然，这些研究者也成了各自领域中的佼佼者。因此可以说，提出一个好的假设是科学研究中最具魅力也最具挑战的一部分。

那么，到底什么样的假设才是好的假设呢？一个好的假设需要满足以下几个条件：

一个假设必须是能够证伪的（**Falsifiable**）。理论上，一个假设应该是有可能被数据证明是错误的。比方说，“有志者事竟成”就是一个没有办法证伪的假设。这个假设讲的是志向和成功的关系。但是如果我们不能准确定义志向和成功，就没有办法证明这个假设到底是正确还是错误的。如果一个人没有成功，我们总是可以说他的志向还不够；如果一个人成功了，我们也总是可以说他志向。所以，要想证明这个理论是否正确，我们必须对“在多大程度上有志向”算满足我们假设中的“有志”的条件有一个明确的定义。同样的道理，我们也必须对成功有一个明确的定义，否则一个人总是可以说他成功了，关键看他的这个成功是不是符合我们假设里对成功的定义。

一个假设还必须具有理论上的重要性（**Theoretically Important**）。研究者应该能够在其他人的理论基础上，对他人的理论做改进，或者提出以往理论没有研究过的新假设。所以要能提出好的假设，你还要知道别人做了些什么，能站在巨人的肩膀上想问题。

一个假设还需要具备实际意义上的重要性（**Practically Important**）。也就是说，一个假设要有实用价值，能够回答现实生活中重要的问题，对现实生活有所启迪。有一些学术研究，耗费大量的研究经费，但是研究成果仅仅在学术上有贡献，而对人们的现实生活没有指导意义。一个好的研究应该超越所在的学术小圈子，能够直接或者间接地被应用到现实的大世界中去。

在评价一个假设是不是具备实际意义上的重要性的时候，我们应该用发展的眼光看待一个研究。一个研究目前看来无法对现实生活有所贡献，但是如果它有可能在将来对我们的生活产生重要的影响的话，这样的研究也是具备实际意义上的重要性的。我们所说的有实际意义上的重要性的假设，应该要么现在就能做到解决现实生活中的实际问题，要么具备解决实际问题的潜质。牛顿的三大定律对那个时代的现实生活并没有立杆见影的影响，但是对后人生活的贡献却是不可估量的。

一个假设还应该简洁（**Simple**）。没有经验的研究者会有一个倾向，在自己的假设中加入很多自变量，来看看这些变量之间的关系。但是随着自变量的增多，这些变量之间的关系就变得越来越复杂，最后也越来越难对因变量的变化作出合理的预测。比方说，有研究者想研究天气和绩效之间的关系。但是他也意识到，性别、文化、睡眠、年龄等和绩效都有关系。如果他在他的假设里面把这几个因素都加进去，假设就会变得非常复杂，从而失去了理论和实际意义上的重要性，因为这个时候对因变量变化的描述已经受到太多自变量的影响而混杂不堪了。毋庸置疑，实际状况中影响因变量的因素一定是远远多于我们在假设里涉及到的自变量。可是，一个好的假设不是要穷尽所有的因素，而是要分离出几个主要的因素。如果你试图把太多的影响因变量的因素都包括进来，你的研究就会失去重点，也很难扩展到其他的人群和情况中去。

一个好的假设还应该具有繁衍性（**Fertile**）。也就是说，从一个假设可以推演出很多具体

的假设。打个比方说，有两个女孩子，一个叫小丽，一个叫小萍。小丽长得难看，小萍长得好看。她们现在在吵架。最为具体的假设是小丽喜欢妒忌小萍。这个假设就不是一个具备繁衍性的假设，因为你没办法把这个假设推演到其他的人和其他的情况中去。接着你提出一个假设，说长得难看的人喜欢妒忌长得好看的人。这个假设就比前一个假设的繁衍性高一些，我们可以把这个假设推演到其他的人身上。如果你继续改进你的假设，说，一个人在一个领域里面显弱了，就会喜欢在另外一个领域里面争强。这就是一个繁衍性更高的假设，我们不仅可以把这个假设推演到其他人身，而且可以推演到其他很多领域上面去（上面的例子取自March and Lave的“An introduction to models in the social sciences”一书。为适合中国读者，本章作者对原例稍做了修改）。

一个好的假设还应该是有趣的（Interesting）。也就是说，一个好的假设要给读者一个惊喜。一篇文章读下来，读者通常有三种反应：第一种反应是，不看这篇文章我也知道这个结果，之所以没做这个研究是因为我觉得不值得做。比方说“睡眠不足情况下人们的绩效比睡眠充足情况下低”之类的假设就属于这一类。第二种反应是，不读这篇文章我不会想到事情是这样的，但是读了之后我会觉得，我当时为什么没想到呢？大多数的好的文章都属于这一类。比方说我们前面提到的“人们喜欢把自己的观点强加在别人身上”就是这一类的研究。读了这样的文章人们会觉得眼前一亮，说，对呀，有道理，有新意。

最后一类反应是，事实上文章里说的东西确实是正确的，但是如果我不读这篇文章我不知道事情是这样的，而且读了之后我都不能相信文章里说的东西是正确的。比方说，哥白尼提出地球是围着太阳转的。在当时的条件下，即使人们读懂了他的文章，也都不信服，但是现在我们知道哥白尼确实是正确的。这种境界的研究确实为数不多，但这样的研究往往都成为经典之作。

心理学中Milgram的服从实验就是一个这样的例子。Stanley Milgram教授在二十世纪六十年代做了一系列实验来研究人们对权威过度服从的现象。他在纽黑文市张贴广告，招募一些男性到耶鲁大学Milgram的实验室，参加一个关于“记忆和学习研究”的实验。当每个实验参与者到达实验室时，都会发现里面已经有两个人在了，一个是穿着实验室制服的实验人员，一个是叫Wallace的中年人。实际上Wallace先生是事先安排好的，但是参加实验的人并不知情，他们以为Wallace先生是和自己一样报名参加实验的。穿着制服的实验者向参加实验的人解释，这个实验是要检验惩罚对学习的影响。每轮试验有两个人参加，一个人扮演“教师”的角色，另外一个人扮演“学生”。如果“学生”回答错误的话，“教师”会对学生实施惩罚。然后实验参与者和Wallace先生抽签决定到底谁是教师谁是学生。但是实际上这个抽签是事先做过手脚的，最后总是Wallace先生做学生，而被招募来的实验参与者总是扮演“教师”的角色。

实验者在Wallace先生身上连上电极，并让“教师”坐在一个机器面前。这个机器上有很多按钮，不同的按钮代表不同的电压。只要按下某个电钮，Wallace先生就会被对应的电压击中——这也就是惩罚。这些按钮从15伏开始，最高的高达450伏。这些按钮边上也注明有“轻微电击”，“中度电击”，一直上升到“危险：严重电击”，最后超过400伏的按钮边是大大的红叉，以示特别警告。

“学生”Wallace先生在实验中要学习一些词组，然后回答哪些词应该是归在一组的。如果答错，“教师”就给Wallace先生一次电击。第一次电击从最低的15伏开始，第二次是30伏，逐渐上升。在实验中，Wallace先生实际上是从来没受到过电击的，但是“教师”并不知道。在实验中，Wallace先生之后会不断犯错误，电击也越来越高。超过150伏之后，

Wallace先生会发出惨叫，并要求退出实验。这个时候很多“教师”就要求停止实验。他们表示很担心Wallace先生。但是，实验者总是说：“请继续，所有的责任由我来承担。”

实际上，这个实验是来检验人们会不会服从实验者并给Wallace先生更高电压的电击。实验发现，尽管实验者只是用很简单的词句，比方说“请继续”，来要求参加实验的人继续实验，大约有65%的人顺从了实验者并最终按下了高达450伏的按钮。这个实验动摇了人们一直认为的这个想法，实验结果大大出乎人们的意料。即使实验结果摆在那里，人们还是很难相信有高达65%的人给出450伏的电击。

一个假设要让读者产生第三种反应确实可遇不可求，但是，作为研究者，我们要尽量避免做第一种研究，争取做第三种研究。

实验室研究

提出了假设之后，就要来验证它是否正确。科学发展到现在，已经有了很多检验假设的方法。我们接下来先介绍一下在社会科学中常用的检验假设的三种方法，然后简要介绍一下它们之间的相对利弊，最后着重介绍一下实验室实验的研究方法。

观察性研究（Observational Study）

比方说，你有这样一个假设：同样一项工作，不付钱比付钱更能调动人们参与的积极性。怎样来检验这个假设呢？一个可能的方法是收集自然发生的数据进行分析，这就是观察性研究。比方说，在某些国家献血是无偿的，但是在另外一些国家献血是有补偿的，那么我们可以观察在这两个国家里献血的比例分别是多少。

在一项新的研究开始之初，这样的研究是非常有用处的。收集自然发生的数据可以帮助研究者对自己所要研究的问题有一个大致的了解。比方说，如果你想研究在工作中员工之间互相帮助的关系是怎样形成的，那么，首先在一些企业当中对员工之间的帮助行为进行观察会对研究者找到最关键的因素非常有帮助。

当然，观察性研究的优越性并不仅仅局限于一项研究工作的开始阶段。如果一项研究主要在实验室里进行，那么在获得了实验室数据之后，再回到现实生活中进行实地研究可以帮助我们证实在实验室里获得的结果是不是在自然环境下也会发生。比方说，在实验室的环境下，你发现女性员工比男性员工更容易获得同事的帮助，那么，在现实的工作环境下是否如此呢？实地观察性研究可以帮助我们回答这个问题。

但是这种自然发生的数据也有它的不足。首先，自然发生的数据会受到很多和我们的假设无关的因素的影响。在“献血与补偿”的例子里，一个国家有没有献血的传统，人们对献血是不是有害健康的看法等都会影响献血人口占总人口的比例。而由于这些因素的影响，我们就没有办法清楚地分辨出献血人口比例的高低到底是由于有无补偿还是由于其他的因素造成的。其次，这些自然发生的数据只能说明两个变量之间的相关关系，但是不能确认两者之间的因果关系。如果我们收集了一组关于人们的开心程度的数据，同时也收集了这些人居住的房子的大小的数据。我们通过对数据的分析发现，整体来看，住在大房子

里面的人比住在小房子里面的人开心。但是这些数据不能帮助我们确认，到底是因为住在大的房子里面，人们更加开心，还是因为人们更加开心，他们更有可能得到好的工作，所以更有可能住在大房子里面。也就是说，通过这些自然发生的数据，我们只能说“两个变量是相关的”，但是没有办法确认变量之间的因果关系。

鉴于以上的原因，研究者通常不是收集自然发生的数据，而是通过实验的方式来对假设进行检验。实验大致上可以分为两类：一类是实地实验，一类是实验室实验。

实地实验（Field Experiment）

实地实验是在自然环境下进行的有控制的实验。实验者在自然环境下操纵自变量，来检验自变量的变化在因变量上造成的影响，从而发现自变量和因变量之间的因果关系。同样是检验有没有补偿对献血积极性的影响，实验者可以选择两家医院来进行实地实验。在一个医院，实验者给参加献血的人一些金钱补偿，在另外一家医院里实验者不提供任何补偿。然后实验者可以记录在有补偿和没有补偿的两种情况下分别有多少人来参加献血。

但是这种实地实验的方法也和观察性研究一样，很容易受到很多无关因素的影响。比方说，一个医院周围住的大都是老年人，另外一个医院则在青年人聚集区，很可能年轻人比老年人更有可能参加献血。或者一个医院的工作人员态度非常好，而另外一个医院工作人员态度恶劣，这也会影响人们的献血积极性。

实验室实验（Lab Experiment）

为了保证我们的实验确实能够检测自变量和因变量之间的因果关系，更好的办法是进行实验室实验。在实验室实验中，我们需要对其他的因素加以严格控制，只改变我们希望改变的自变量，监测因变量的变化。比方说，我们把参加实验的人聚集到实验室里面，然后随机把他们分配到有补偿和没有补偿的两个实验情况中去。我们告诉有补偿组的人们，如果他们参加献血，可以得到100元的金钱补偿；我们告诉没有补偿组的人们，他们参加献血是无偿的。然后我们请这些参加实验的人回答，有多大的可能性他们会参加当前条件的献血。

通常在实验室实验中一个自变量总是取几个可能的值，也就是自变量水平，相对于这些可能值的情况就是实验组。上面的实验中涉及到两个实验组：一组是献血有补偿的情况，一组是没有补偿的情况。我们在后面的部分会经常提到实验组这个概念。

很多时候，一个假设所涉及的自变量不是研究者所能操纵的。比方说，性别、种族、年龄等。如果我们有一个假设说，男性比女性在工作中更加容易受到天气的影响。要检验这样一个假设，我们需要让一组男性和一组女性分别参加我们的实验。这个时候，一个人到底是男性还是女性是不受实验者控制的，所以我们没有办法在实验中做到对所有被试随机分配。我们把这种实验者不能直接操纵自变量、不能对被试在各个实验组之间随机分配的实验叫做准实验（Quasi-experiment）。

实地实验和实验室实验的比较

每一种研究方法都有自己的优点与缺点，不能简单地讲一种方法优于另一种方法，但是在特定的研究需求和条件下某种研究方法可能会更适合。

实地实验的局限性主要在于：实地实验由于发生在自然的环境中，和收集自然发生的数据一样，很多潜在可以影响实验结果的因素不能得到很好的控制，内部效度（Internal Validity）往往比较差。一个实验的内部效度是指在多大程度上我们能够确认因变量的变化确实是由自变量的变化引起的（Cook & Campbell, 1979）。比方前面我们说的，有可能有补偿的那家医院周围住的都是老年人，而另外一家医院周围都是年轻人，结果发现没有补偿的那家医院有更多的人献血，但是这很有可能是因为年轻人更有可能献血，而不是因为人们不想要补偿。不过，一个精心安排的实地实验通常会尽量去除那些实验者能够预见到的干扰因素。

相较而言，因为实验室实验可以对实验过程和因素进行有效的控制，研究人员可以尽量排除已知的干扰因素，所以内部效度比较高，设计灵活，另外实验费用也较低。不过实验室实验也有明显的缺点：因为在实验室实验中，研究人员营造了特殊的实验环境和条件，使被试和实验过程都处在一个“非自然态”，而且受实验室自身规模和经费等所限，测试样本难以完备，所以外部效度（External Validity）比较低。外部效度是指在多大程度上一个实验的结果能从它自身的被试和实验环境中被扩展到其他被试和实验环境中去（Cook & Campbell, 1979）。一个实验者总是希望他得到的实验结果能够代表一个普遍的现象，因此我们很关心实验的可重复性（Repeatability），也就是你的实验结果是不是在不同的被试和实验环境下仍旧能够被重复证实。如果一个实验结果只对某一个学校的学生有效，这样的研究结果必然不具备理论意义上的重要性。

这时实地实验的优点就显现出来：在内部效度高的前提下，因为实地实验的实验环境是自然态，测试样本相对比较完备等优点，具有较高的外部效度。实验室实验和实地实验的主要优劣势对比概括如表1。

表1：实验室实验和实地实验主要优劣势对比

	实地实验	实验室实验
被试是否察觉参与了实验	是	否
接近现实的程度	低	高
不相干因素	少	多
实验的可控性	好	差
实验花费	低	高

一个实验的内部效度和外部效度如果都相当高，自然是再好不过了。但是多数情况下内部效度和外部效度是一对矛盾，很难在同一次实验中做到两全。在不能做到两全其美的情况下，如果一项研究更加关注两个变量之间的因果关系，实验室实验会是一个更好的选择，因为在实验室中我们可以通过各种手段来去除其他无关因素的影响。实际上，内部效度高是外部效度高的必要非充分条件。在必要的情况下，我们可以先在实验室里对假设进行检验，以明确自变量与因变量的因果关系，然后在自然环境中用实地实验的方法再次进行实验，来检测这个假设的外部有效性。

那么我们应该如何保证实验的内部效度呢？在实验室实验中，一个实验者又需要注意哪些

基本问题呢？

实验设计的基本原则

在谈到效度之前，我们先介绍实验设计的三条基本原则：随机化（**Randomization**），复制（**Replication**）和区集（**Blocking**）。一般情况下，在被试的选择，测试等方面总会存在实验误差，所以我们需要对实验数据进行统计分析，从而得出有效的结论。这三条原则是统计分析的基础，也是实验高效度的基石。

随机化指实验材料（包括被试）的分配，被试的实验顺序等是随机产生的。如果这些因素都是随机的，那么我们称之为完全随机化（**Complete Randomization**）。其中被试被随机分到各个实验组的过程，我们称之为随机分配（**Random Assignment**）。我们可以用电脑里的各种统计软件或者简单的随机数发生器来进行随机化操作。随机化首先是统计分析的需要。统计分析中要求基础分析量，如观测值（**Observations**），误差（**Errors**）等是独立随机变量。随机化下，我们可以认为这些基础分析量是独立随机变量。随机化也可以减少甚至去除某些额外因素（**Extraneous Factors**），尤其是没有得到控制的干扰因素的影响。在研究补偿对献血影响的例子中，如果按照年龄在 30 岁以上和 30 岁以下将被试分到有补偿和没有补偿的两个实验组中去就会产生系统偏差，因为年轻人相对更有可能献血。实验结果可能显示没有补偿的实验组献血更积极，于是我们便得到了错误的结论。在样本足够大时，将被试随机分到两个实验组就可以基本消除这种情况。

复制指在相同的处理下，独立重复实验以得到重复样本。复制首先可以让实验者对实验误差有一个估计。这种估计帮助实验者了解测试结果是否有统计意义上的不同。其次由统计分析性质可知，相较于一次测试，多次的复制可以帮助我们更精确的估计样本均值（**Sample Mean**）。另外，统计分析需要一定的数据量才可以达到一定的置信度，对于复杂的实验设计来说尤其如此，而复制可以提供一定的数据量。需要特别指出的是，复制和重复测量不一样，重复测量只是从测量角度提高准确度，而复制则是重新测量整个实验从头到尾被试受到的总影响。比如研究运动与心律的关系时，被试运动后测量心律，休息一定时间进行同样的运动后再测量就是复制；被试运动后两研究员分别通过左右手动脉同时测量其心律就是重复测量。

区集是一种实验设计中的方法，用来处理无关因素（**Nuisance Factor**）。无关因素指，实验之前我们就知道一些实验个体可能含有近似的会影响测试结果的因素，但是这些因素又不是我们关心的。如果我们知道这种因素是存在而且可控的，那么可以采用区集的处理方法。我们把受相同无关因素影响的测试值放到同一个区块（**Block**），并认为同一区块的测试值受到这种因素同样程度的影响。这样在统计分析时，我们就可以去除掉这种因素对实验结果的干扰。在后面的章节我们还会提到这种做法是如何消除无关因素的影响的。比如在对比研究几种药剂对细菌的抑制作用时，因为实验条件所限，可能每天对每种药剂只能进行一次测试，但是我们需要更多的数据（复制），所以一共重复进行了 5 天的测试。由于每天的天气在变，每次测试也都把实验条件重新初始化，比如清洗培养皿等，这导致每次测试的环境不一样，对每次测试结果的影响也就不一样，这时我们需要把每天都当成一个单独的区块。

实验的效度问题

在实验中，有些实验方式或事件会影响效度，我们把这些实验方式或事件称作效度威胁因素（**Threats to Validity**）。我们往往把他们按照主要直接影响内部效度或者外部效度分为内部效度威胁因素（**Threats to Internal Validity**）和外部效度威胁因素（**Threats to External Validity**）两类。实际上，有些方式或事件稍稍改变，便会从内部效度威胁因素变为外部效度威胁因素。比如对一些因素来说，如果只影响一个实验组的人员，而不对其他实验组的人员产生影响，那么归为内部效度威胁因素，反之如果对所有的实验组都产生等效影响，那么归为外部效度威胁因素。我们将会在后面的介绍中，随时举例说明这种情况。

对内部效度产生威胁的主要是混淆变量（**Coufounding Variable**）。没有得到控制的无关变量使测试结果产生了系统性的偏差，于是我们就不能确定因变量的变化是否是由自变量的变化产生的。

比较常见的内部效度威胁因素有：

偶然事件（History）：在实验进程中，有些事件会突然发生，这些事件会影响被试在实验中的反应，影响测试结果。比如 2003 年突发的 SARS 事件，如果在此期间对比研究美国和中国的电视广告对消费者购物意愿的影响恐怕就很难得到正确结果，因为中国大多数人这时没有条件也没有心情去购物。实验环境复杂可控性差的实地实验受到影响的可能性一般会更大。大多情况下，我们意识不到的某些事件会影响被试，所以随机化和对照组（**Control Group**）对此有一定帮助。

被试选择偏差（Selection Bias）：被试被主观意愿左右，进入不同的实验组。因为这种主动性背后隐藏的某些因素可能会影响测试效果。又分为两种，即自选择（**Self selection**）与区别选择（**Differential selection**）。自选择指被试主动参与某测试时，区别选择指研究人员以自己的感觉选择被试。比如在研究工资和教育程度的相关性时，我们希望把所有样本的工资和教育程度放在一起研究。但是在现实中，当工资低于某个程度时，有些人会选择不工作。对于他们，我们可以了解他们的教育程度，却不知道如果他们工作，那么工资会是多少。特别的，在这部分人群中，受教育程度可能会比较低。那么如果在样本中只研究有工作的人群，最后得到的工资和教育程度的相关度会与真实值有差别，非常可能低估教育对工资的影响。所以我们利用志愿者做研究时，就要特别注意自选择问题，对被试进行随机分配可以帮助解决这个问题。在例子中提到的研究工资与教育程度等无法避免选择偏差的情况下，有些特别的处理方法也许会有效，例如 Heckman 在 “Shadow Prices, Market Wages, and Labor Supply” 一文提出的处理样本偏差的方法。

实验者偏差（Experimenter Bias）：由于实验者本身的行为所导致的偏差。比方说，如果实验操作者事先知道所要检验的假设，在进行实验的过程中，就可能有意或者无意地做出某些行为，影响不同实验情况下的被试的反应。另外，在对一些主观数据进行编码的时候，实验编码者也有可能由于了解所要检验的假设，使这些主观数据的编码存在某种倾向性。这些都会影响实验的最终结果。为了去除实验者偏差，我们通常要求不能让执行实验的人了解实验所要检验的假设。而且通常提出假设的研究者本人不能担当执行操作试验的角色，我们需要一个不知道所要检验的假设的人来执行实验。

测量手段（Instrument）：在实验中用到的各种手段工具会影响测试结果。一个典型例子就是实验介绍。如果某一实验组的使用一种介绍方式，介绍的清晰明白，而其他实验组使用另一种有歧义的介绍方式，那么测试结果显然会受此影响。所以我们要尽量

让所有的实验组使用相同的实验介绍。还要注意实验介绍写得是否清楚易懂，实验情景是否合情合理并且人性化，问题措辞是否可能产生歧义等等。

特别的，工具如果对所有的被试组有同样的影响，那就属于外部效度威胁因素了，比如同样的实验介绍让所有实验分组都产生了歧义。

成熟程度 (Maturation)：随着年龄的增长，我们的心理和生理会逐渐成熟，进而对实验产生影响。一般只有实验周期很长时，我们才需要考虑这种影响。当被试是儿童时，我们要特别注意这种影响。比如有些研究表明，即使没有受到任何治疗，大多数大学生也会在六个月内从心理消沉期走出来，如果有人做新药剂实验，测试结果表明服用药剂的大学生会在六个月内从心理消沉期走出来，那么显然我们不能认为药剂有疗效。我们可以采用随机化的对照实验组来解决这个问题。

偶然减员 (Mortality)：实验中，一些被试可能会退出实验，从而影响测试结果。比如被试突然被公司调去外地，不能继续参加实验。因为不知道这个被试与其它完成实验的被试有什么区别，我们很难解释这会对实验结果造成什么影响。为了避免这种情况，我们可以在正式实验前做预测试，了解被试特性。不过要注意这种预测试对实验内部效度的影响。随机选择在某些情况下会有一些帮助。

测试关联(Testing)：在测后实验 (Posttest) 之前进行相关度比较高的侧前测试 (Pretest) 可能会使被试对实验更加熟悉和敏感，从而提高测试成绩。这时应该加入对照组。所罗门四组设计也是一种理论上理想的解决方案。

统计回归(Statistical Regression)：典型情况是研究极端组时，测试值的变化会比研究一般群体时大得多。属于极端组的被试会在下一次测试中很可能会向均值靠近。比如说某一次测试中分数在 95 分以上的群体 (满分 100)，再重新接受测试时有些被试的分数就非常可能向均值靠近一些。避免研究这种极端组是一种选择，否则就应该对被试等随机分配，增加对照组。

多重因素作用 (Interaction)：两种或以上的效度威胁因素会同时起作用，我们称之为多重因素。比如实验者因为知道实验目的而自己对被试进行实验分组，那这就属于选择偏差和实验者偏差因素的多种因素作用。

比较常见的外部效度威胁因素：

样本不具代表性 (Non-representative Sample)：作为样本 (Sample) 的被试不能代表母体 (Population) 的情况。比如研究中国电视广告对消费者购物倾向的影响时，如果只研究汽车类广告对购物倾向的影响，那就不具有代表性。实际上，很多其他因素都可以使样本不具有代表性。保证样本具有代表性是保证外部效度的基石。

传递效应 (Carryover Effects)：随着实验进程，被试会随时间发生变化，进而影响测试结果。比如实验时间过长后，被试可能会感到疲惫，厌倦。这种变化也有可能是学习能力带来的，比如大学生的学习能力很强，在实验进程中对实验越来越熟悉，表现也越来越

好，而同为被试的老人的进步可能就不明显。本章后面介绍的组内设计是常用的一种处理方法，并对此因素有进一步的探讨。所罗门四组设计也是一种理论上理想的解决方案。

霍桑效应 (Hawthorne Effect)：指当研究人员存在时，由于紧张等，被试的表现会与平时不一样。如果我们不知道这种差别是否会对测试结果产生重大影响，那么应该怎么处理呢？一个取巧的方法是再安排一个对照组，对照组与实验组一样会被观察，但是不需要做接受测试，目的只是测试霍桑效应。

需求特性 (Demand Characteristics)：被试在参与实验时很自然地去猜测实验者到底想要检验什么，在试验中能引导被试作出猜测的线索被称为需求特性 (Schweigert, 2006)。一旦被试对假设作出猜测，他们在实验中的行为会或多或少受到影响。一些被试会根据他们对假设的猜测故意作出和假设一致的行为，而另外一些人也许会故意作出跟他们的猜测相反的行为。比如，在“显示器与积极性”的研究中，被试认为实验者想检验的假设是“显示器屏幕越大工作积极性越高”，那么即使事实上他们的工作积极性和显示器的大小没多大影响，他们还是努力表现得和实验者的假设一致，其实他们是希望公司管理层看到实验结果后给他们配置更大的显示器。再比如，如果人们猜测实验者要检验的假设是“惩罚越多工作表现越好”，但是因为他们不想受到惩罚，所以故意降低自己的工作表现。这就属于故意做出和假设相反的行为的例子。不管他们的行为和假设一致还是相反，实验的结果的效度都受到了影响。因此，为了减少需求特性从而降低被试猜测出一个实验的假设的可能性，一个设计缜密的实验通常会比较好地隐藏实验者的真实意图，避免被试猜测出真实的实验意图，并有意调整自己的行为。

安慰剂效应 (Placebo Effect)：安慰剂效应指被试即使没有真的接受实验或某种对待，也会给出有效果的反馈。典型例子是药剂实验，被试即使服入的不是真药剂而是安慰剂时也经常会感觉好很多，并给出药有效果的主观反馈。不让被试知道自己正在被测试自然是最好的做法。如果不可以的话，至少不能让被试知道实验的目的。

霍桑效应，需求效应，安慰剂效应等都有一个特点，那就是实验参与人员意识到正在进行实验，所以对测试的反馈不同于未参与实验时。因此有些人把有这个特点的因素都称之为副效应 (Reactivity)。

实验室实验中应该注意的问题

在这一部分，我们首先介绍一下在实验中如何把假设变成可以操作衡量的东西，然后再介绍一些实验中需要避免的问题。

首先，我们来谈一谈在实验中如何定义一个变量，以及什么叫做可操作性定义 (Operational Definition)。一般来说，一个变量通常是一个抽象的概念，你需要把它转换成具体的可以操作或者观察的形式。那么，一个实验者用来操纵或者衡量的关于这个变量的可操作的形式就是可操作性定义 (Cozby, 2001)。有了可操作性定义，其他的研究者就可以相对容易地重复某个试验 (Elmes, Kantovitz, & Roediger, 1999)。除了可操作性好，一个好的变量定义也要能准确有效地代表变量。比如，把电话客户服务人员的效率只定义为接电话的数量而忽视质量就有一定问题

变量的抽象程度不同，它所需要的可操作性定义的难易也不同。比方说，工作时间是一个相对来说具体的变量，你只需要用它的小时数来衡量。而工作积极性就是一个比较复杂而且抽象的变量，会涉及到很多因素，比方说员工愿意每个星期加班多少个小时，员工是否愿意接受困难的任務，员工是不是提前完成任务等等。一个研究者可以选择工作积极性的某一个方面来作为工作积极性的可操作性定义。关键在于，一个研究者必须先有一个方法来有效衡量或者操纵这个变量才能具体地实施一个实验。

如果你想知道情绪对工作效率的影响，那么首先你就要知道，在一个实验中，你需要怎样做来产生你需要的情绪，所谓的工作效率应该怎样来衡量。比方说，你可以分别让两组人看回忆他们过去的经历，一组人回忆快乐的经历，另外一组人回忆伤心的经历，这样你就有办法使参加实验的人产生两种不同的情绪。然后你让人们来做某种工作，比方让他们数零件，然后看他们在规定时间内可以完成多少。回忆过去的经历和数零件就是对情绪和工作效率的一个可操作性定义。一般来说，如果我们不能根据一个假设给出可操作性定义，那么这个假设就是没有办法证伪的。

在实验设计，包括形成可操作定义中，保证效度也是需要时刻考虑的问题。在我们前面的介绍中，已经介绍了应该重点考虑哪些影响效度的因素，以及如何避免它们的影响。除了需要注意前面提出的之外，还要特别注意的是，在考虑如何使你的变量可以操作的时候，要避免天花板效应(Ceiling Effect)和地板效应(Floor Effect)。在实验中有的时候会产生所有的数据都集中在可能范围的最高端的情况，这叫做天花板效应。比方说，你想证实，更多的奖金可以产生更高的工作积极性。你找了一批人，告诉他们，如果他们愿意数零件数5分钟，你付给他们每人20元钱。对另外一批人，你告诉他们，如果他们愿意数零件数5分钟，你付给他们每人40元钱。然后你让所有这些人回答，他们有多大可能性愿意来数零件。然后你发现不管是给他们20元钱还是40元钱，愿意数零件的可能性都在95%左右。这是不是意味着你的假设不成立呢？并不见得。因为很有可能你的结果受到了天花板效应的影响。也就是说，本来给20块钱大家就已经很愿意来数零件了，再多给他们钱也不可能提高他们愿意数零件的积极性。如果是这样，你需要把20元钱的奖励调低，比方说调低到5元钱。当然，也有可能是因为这个百分制的衡量方式不能体现工作积极性的区别，那么你可以换一个方法来衡量因变量，比方说，你问参加实验的人，如果我给你20元钱，你愿意数零件数多少分钟？对另外一组人，你可以问，如果我给你40元钱，你愿意数零件数多少分钟？这样就避免了天花板效应。和天花板效应相反的是地板效应，是指所有的数据都集中在可能范围的最底端的情况，它的处理方法也和天花板效应相对应。我们在实验中要尽量避免这两种情况的发生，否则就无法断定到底是因为是自变量确实对因变量没有影响，还是自变量没有设置在合适的水平，或者因变量没有得到合理的测量。

我们的初始实验设计可能并不完备，尤其在实验复杂的情况下，所以有的时候，实验者会事先请少数被试做一些“测试性实验”(Pilot Study)，小规模地测试一下实验，看看是不是有一些你意料之外的问题。为了更好地达到测试的目的，在测试性试验结束后，参加测试性实验的被试通常需要回答一些和实验的因变量无关但是和实验设计有关的问题，比方说，“你是否觉得我们的实验介绍清楚而且容易理解？”，“你在实验过程中是否有理解的困难？”等等。实验者也会征求被试的意见，从而知道哪里需要改动。有的时候，实验者还要求被试在参加实验的过程中做口头即时报告，这样被试的一些反应就可以帮助实验者对实验做出必要的改动，保证在整个实验正式开始之前能把可能出现的问题最小化。

对实验结果的理解

做完了实验，收集好了数据，我们就需要对数据加以分析。如果数据的分析结果和我们的假设不一致怎么办？是不是这就意味着我们的假设是错误的呢？先不要过早下结论，让我们来看看什么情况下我们会得不到和假设一致的结果。

当然，很有可能我们的假设是错误的。但是这并不是唯一的解释。另外一种可能性是因为我们的实验设计不妥当造成的。比方说，被试没能很好地理解你的指示；或者是被试在实验后期比较疲劳，没有认真回答你的问题等等。

此外，你这个时候要思考一下，“我的实验里有没有混淆变量？”消除混淆变量的影响是保证你得到可靠的数据的一个非常重要的方面。所以，你应该看一看，你本来应该控制不变的变量是不是得到了应有的控制？有没有其他可能的变量应该得到控制，但是你当时没有注意到？样本量的大小是不是保证随机分配消除了随机差异？真正操作实验的人是不是对每个被试都公正且没有倾向性？

除了混淆变量，你还要考虑你的可操作性定义是不是有效。比方说，你对“快乐”和“悲伤”的可操作性定义是分别让人们听一段欢快和缓慢的音乐。如果你的音乐没有达到让被试感到“快乐”或者“悲伤”的效果，那么你需要考虑修改你的可操作性定义。此外，我们在前面提到过，在考虑变量的可操作性定义的时候，我们要注意选取适当的取值范围，避免产生地板效应和天花板效应。如果你发现可能存在的地板效应和天花板效应有可能造成两个实验组没有区别，那你就需要改进你的可操作性定义，再重新进行你的实验。

实验设计

在接下来的这一部分中，我们要着重讲讲怎样设计一个实验来对假设进行检验。实验的设计在很大程度上取决于你的假设——你的假设有几个自变量以及每个自变量各有几个水平。如果你的假设只有一个自变量，那你的实验就是最简单的组间设计

（**Between-subjects Design**）或者是组内设计（**Within-subjects Design**）。如果你有两个或者两个以上的自变量，那么你的设计应该是因素设计（**Factorial Design**），当然，一个因素设计既可以是组间设计，也可以是组内设计，还可以是组间组内混合的设计。我们下面对这三种设计一一加以介绍。

组间设计（**Between-subjects Design**）

设计一个实验首先要考虑的就是如何把实验参与者分配到有不同自变量水平的实验组中。你可以有两种分配方式：1、把不同的被试分配到不同的可能值上；2、让每个被试接受所有的可能值。

所谓组间设计，是说参加不同实验组的人是不同的，即上面的第一种分配方式。假定你有这样一个假设：对于某件东西，一个人拥有之后卖出它时索要的价格要高于他拥有之

前愿意付出的价格。那么你就可以设计这样一个实验：把被试随机分成两组，其中一组的人你给他们每人一个杯子，另外一组人不发杯子。你请已经有杯子的人回答，如果要把这个杯子卖掉，买方至少要出多少钱他们才愿意卖；你也请没有杯子的人回答，如果要买这样的一个杯子，他们最多愿意出多少钱。这样的一个实验采用的就是第一种分配被试的方式，是一个典型的组间设计的实验。

再比如，你的假设是，正面的反馈比负面的反馈更能提高员工的工作绩效。那么你可以随机分配一组人，给他们提供正面的反馈，另外一组人收到负面反馈，然后你看看这两组人的工作绩效到底哪个高。和上面的例子一样，如果一组人收到了正面反馈，那他们就不会收到负面反馈；而收到负面反馈的那组人也不可能收到正面反馈。也就是说，每个人都只能参加一个实验组，这样的设计属于组间设计。

我们前面讲过的“显示器与积极性”也是一个组间设计的例子。一组员工使用大显示器，另外一组员工使用小显示器，我们分别测量他们的工作积极性。如果我们的自变量有多于两个的可能值，那么我们就有多于两个的实验组。比方说，使用大显示器可以提高工作积极性，但是显示器大到一定程度，再增大显示器就对工作积极性没有影响了。因此我们可以有三个实验组，一组人使用14寸显示器，另外一组人使用19寸显示器，还有一组人使用25寸显示器。然后我们分别检验各组人的工作积极性。很显然，不同组的人使用不同大小的显示器，这也是一个组间设计，只是有更多的实验组而已。

由于不同实验组中的被试之间存在个体差异，我们在分组时需要尽可能做到对被试进行随机分配，平衡抵消差异。

组内设计（Within-subjects Design）

另外一个减少组间差异的方法就是采用我们上面提到的第二种分配被试的实验设计方法——组内设计。所谓的组内设计，就是被试要参与某个自变量的所有可能情况。对于组内设计来说，所有的被试参加所有的实验组，被试之间的个体差异都发生在实验组之内，所以并不需要随机分配。

我们仍旧来看“显示器与积极性”这个例子。你可以给所有的人提供小的显示器，测量他们的工作积极性；过一段时间之后，你把所有人的显示器换成大一些的显示器，再测量他们的工作积极性；然后你比较这两种情况下人们的工作积极性。由于每个人都使用过两个显示器，这个实验设计就是一个组内设计。

一种比较常见的组内设计是测试前一测试后设计（Pretest-posttest Design）。比方说，你的假设是喝酒精饮料会降低人们的反应速度。你可以首先测试一下被试没有喝酒之前的反应速度，然后你让这些入喝酒精饮料，之后再让这些入做同样的测试，记录他们的反应速度。这就是一个测试前一测试后设计。同样一组人用同样的测试方法被测试了两次，一次是在自变量没有被改变之前（喝酒之前），一次是在自变量被改变之后（喝酒以后）。

组内设计和组间设计的选择

在资源充沛的情况下，很多实验者都偏向采用组间设计。组间设计是一种比较保守的设计，因为在组间设计中不会出现一个实验组污染另外一个实验组的情况。一般来说，组间设计的需求特性没有组内设计明显。很容易想象，如果一个被试回答了两个实验组下的

问题，他就可以相对容易地把这两个问题进行比较，也就更可能猜测出实验者的意图，从而调整自己的行为。这就影响了实验结果的真实性。比方说你想采用组内设计的方法来检验喝酒精饮料对反应速度的影响。由于被试在喝酒之前和喝酒之后做的测试相同，他们很容易猜测到你是想检验喝酒对他们的影响。不管他们把自己的反应速度调慢还是调快，实验的结果都存在一些偏差。如果是组间设计，需求特性的影响就相对小一些。当然，在组内设计中，我们可以通过一些实验设计的技巧来减少需求特性的影响。比方说，我们可以让被试在喝酒前和喝酒后做不同的测试，比方说，都是做数学题，但是题目不同。这样被试就很难分辨实验者的真实意图，也很难分辨哪些问题是实验者真正关心的。但是，尽管我们可以减少需求特性在组内设计中的影响，组间设计仍旧是减少需求特性更简便、更可靠的实验设计方式。

组内设计的另外一个问题就是可能产生传递效应。比如你测试正面反馈和负面反馈对工作绩效的影响。如果采用组内设计，人们先接受正面反馈，然后我们测量他们的工作绩效；然后被试再接受负面反馈，我们再次测量他们的工作绩效。由于对因变量的测量都是通过让被试参加相同的测试，因此被试在第二次参加这个测试时的成绩会提高，但是这不一定是反馈对绩效的影响，而很有可能是由于人们在第一次做测试的时候获得的一些经验可以被用在第二次测试中，从而提高了成绩。我们把这种传递效应叫做练习效应（**Practice Effect**）。但是如果被试因为重复已经做过的测试而感到无聊并逐渐对测试敷衍了事的话，成绩会降低。这也不是反馈对绩效的影响，而是另一种传递效应，叫做疲劳效应（**Fatigue Effect**）。我们在实验中应该尽量去除练习效应和疲劳效应。去除传递效应有一些常见的方法，例如让被试回答不同的测量因变量的问题。假如你想测试在不同环境下的记忆力，那么不要让被试背诵相同的东西，而是背诵类似的东西。

如果让所有的被试都以同样的顺序经历所有的实验组，就会很容易产生传递效应。为了减少这种情况的发生，我们可以用**ABBA**互相抵消的方法（**ABBA Counterbalancing**）设计实验。仍旧以反馈和绩效的关系的假设为例。你可以对每个被试都采用这样的顺序：正面反馈→负面反馈→负面反馈→正面反馈（**ABBA**）。把正面反馈放在第一和第四个位置可以在某种程度上去除练习效应。但是，如果你的自变量有三个可能值，上面的这种完全互相抵消的方法就不太可行，因为这三个可能值的顺序组合有**6种**，那么被试就要经历**3（3个可能值）×6（6种可能顺序组合）=18个实验组**，实在是太长了。

在这种情况下我们有没有其他办法呢？我们可以随机把被试分配到不同的实验顺序中去，我们把这种方法叫做抵消平衡法。如果是两个可能值的自变量，这两个可能值的顺序排列只有两种情况：**AB**和**BA**。那么你可以随机选取一半被试采用**AB**的顺序，另外一半采用**BA**的顺序。比方说，有一半的人是先接到正面反馈，另一半的人是先接到负面反馈。需要注意的是，在抵消平衡法中，顺序是一个组间变量。如果是三个可能值的自变量，你就要把所有的被试随机分成**6组**，每组采用一种排列顺序。不难看出，**ABBA**互相抵消的方法一般来说只适用于自变量有两个可能值的情况，但是抵消平衡法却可以适用于自变量有两个或者两个以上可能值的情况。

可是这样还是会有问题，随着自变量的可能值的增多，可能的顺序也在增多。比方说，**3个自变量的值有6种顺序，4个自变量的值有24种可能的顺序，5个自变量的值甚至有120种可能的顺序！**有的时候不同自变量值的排列顺序的数目甚至比参加实验的人还多，那么随机分配被试到不同的实验顺序中去的方法也不适用了。

这个时候，我们就没有办法做到完全的平衡抵消了。我们需要做的是**一个不完全的平**

衡抵消，但是我们要保证每个可能值出现的次数相同，而且这些值可能出现的位置的次数也相同。比方说，如果我们有A，B，和C三个自变量的值。那么我们要保证三个值出现在第一位，第二位，和第三位的次数相等。这种不完全平衡抵消的方法叫做拉丁方设计（Latin-square Design）。

在下面的图表中我们列出了有四个实验组（A B C D）的拉丁方设计。

		顺序			
		第一位	第二位	第三位	第四位
被试编号	1	A	B	C	D
	2	B	C	D	A
	3	C	D	A	B
	4	D	A	B	C

按照这张图表的情况，参加实验的人数需要是4的倍数。比方说，如果我们有12个被试，被试1、被试5、被试9都采用第一个实验顺序，被试2、被试6、被试10采用第二个实验顺序，依此类推。鉴于这种分配方法的复杂性，建议感兴趣的读者参考其他相关书籍，在此我们不做深入讲述。比如，Roger E. Kirk 编写的《Experimental Design: Procedures for the Behavioral Science》对拉丁方设计有详细的介绍。

此外，值得一提的是，有时由于条件限制，可能无论是互相抵消法还是拉丁方分配方法都不能使用，因此无法在实验中加以排除或控制影响实验结果的因素。在这种情况下，只有做完实验后采用协方差分析（Analysis of Covariance）或偏相关等方法，把影响结果的因素分析出来，以达到对额外变量的控制。这种事后用统计技术来达到控制额外变量的方法，称为统计控制（statistical control）。

当然，如果组间设计完全优于组内设计的话，我们就没有必要讨论组内设计了。组内设计有它自身的优点。组内设计的主要优点是，由于不存在被试的组间差异，组内设计更容易作出显著的效果。如果我们能很好地控制其他对组内设计的不利因素，组内设计也不失为一个好的选择。

因素设计（Factorial Design）

以上我们介绍了只涉及一个自变量的最基本的组内设计和组间设计。一些比较复杂的设计常常涉及多于一个自变量的情况。我们把在一个实验中同时操纵两个或两个以上自变量的实验设计叫做因素设计。

假定你想研究是否拥有相似的背景如何影响人们对他人的行为的理解。比方说你有这样一个假设：如果一个人表现出好的行为，那么和这个人有相似背景的人倾向于认为那个人的表现是出于这个人的主观意图，而没有相似背景的人则不大会这样认为；相反的，如

果一个人表现出差的行为，相对于和这个人没有相似背景的人而言，有相似背景的人更倾向于认为这个人表现出的行为不大可能是出于个人的主观意图。这个假设有两个自变量：一是拥有相似的背景与否，二是被评价的行为的好坏。这个假设的因变量是人们认为另外一个人的行为在多大程度上是出于他的主观意图。

因此，在这样一个因素设计中，我们可以同时检验多个假设，既可以看有没有相似背景如何影响人们对他人行为的理解，也可以看他人的行为是好是坏如何影响人们对这些行为的理解。在检验背景的影响的时候，我们忽略了行为好坏在这里面的影响；同样的，在检验行为好坏的影响的时候，我们也忽略了是否拥有相似背景的影响。这样的分析得出来的效应叫做主要效应（Main Effect）。

而如果我们把两个自变量同时考虑进来，看他们之间的组合对因变量的影响，这样的分析得出来的效应叫做交互效应（Interaction Effect）。之所以采用因素设计，是因为我们预测实验的结果会产生一个交互效应。当然，如果你关注的不是交互作用，就不需要采用因素设计，采用最简单的单变量实验设计就可以了。

根据上面的例子，我们预测有这样一个交互作用：在理解人们的好的行为的时候，和行为人有相似背景的人比没有相似背景的人更容易认为行为人的行为是出于他的主观意图；但是在理解人们的坏的行为的时候，有相似背景的人比没有相似背景的人更容易相信行为人的行为不是出于主观的意图。那如何来进行这个实验呢？我们可以把所有的被试随机分成四组：

1. 有相似背景的人理解他人的好的行为：我们让被试想象，他们有一个同事，被试和这个同事并不相识，但曾经和被试一同参加新员工培训。这个同事上班从来不迟到。

2. 没有相似背景的人理解他人的好的行为：我们让被试想象，他们有一个同事，被试和这个同事并不相识，而且被试和这个同事在进入公司的时候在公司的不同分部接受了新员工培训。这个同事上班从来不迟到。

3. 有相似背景的人理解他人的差的行为：我们让被试想象，他们有一个同事，被试和这个同事并不相识，但曾经和被试一同参加新员工培训。这个同事上个月上班迟到 5 次。

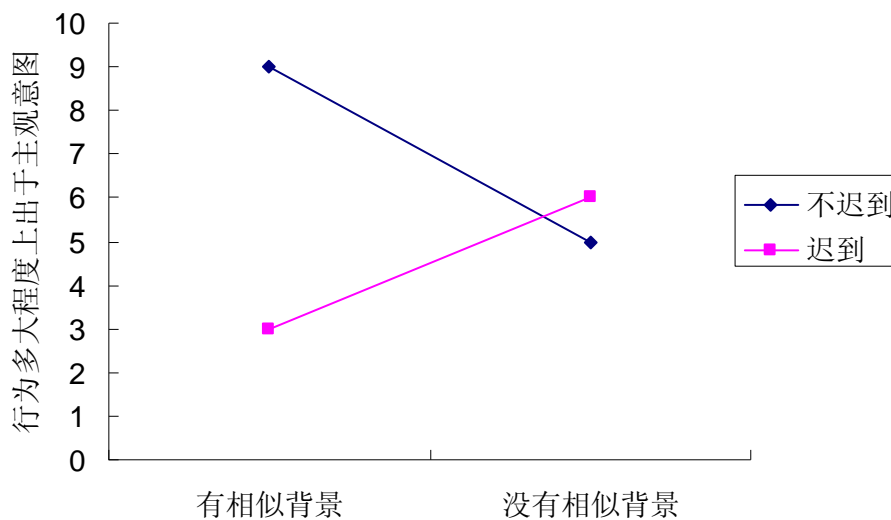
4. 没有相似背景的人理解他人的差的行为：我们让被试想象，他们有一个同事，被试和这个同事并不相识，而且被试和这个同事在进入公司的时候在公司的不同分部接受了新员工培训。这个同事上个月上班迟到 5 次。

然后我们让第一和第二组被试回答这样一个问题：“你认为你的同事上班从来不迟到在多大程度上是由于他对自己有比较高的要求？”被试在一个 1 到 11 的量表上打分，11 代表“完全由于他对自己有较高要求”，1 代表“完全不是因为对自己有较高要求”。第三和第四组被试回答的问题是：“你认为你的同事上个月上班迟到在多大程度上是由于他对自己没有比较高的要求？”类似的，被试也在一个 1 到 11 的量表上打分，11 代表“完全由于他对自己没有较高要求”，1 代表“完全不是因为他对自己没有较高要求”。

假定我们的实验得到了这样一个结果：

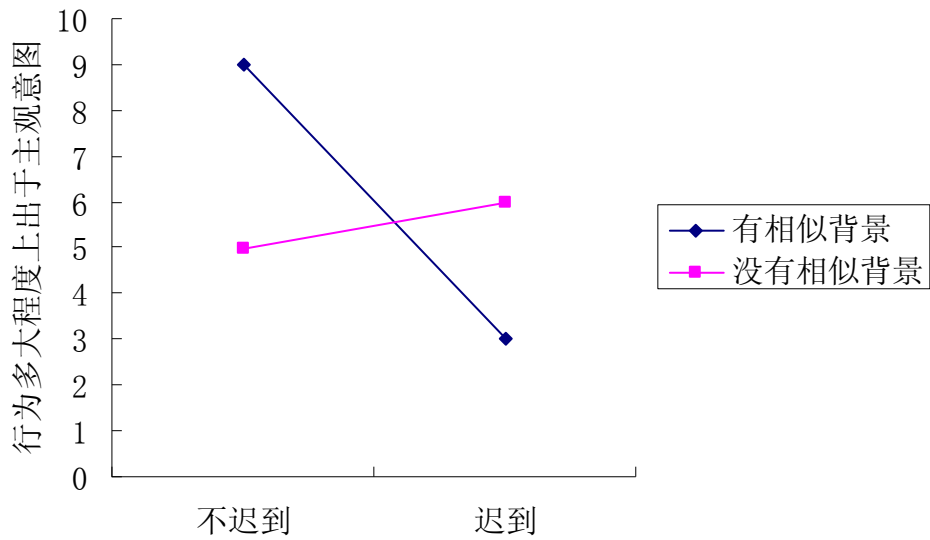
		背景		
		有相似背景	没有相似背景	边际平均值
行为	不迟到	9	5	7
	迟到	3	6	4.5
	边际平均值	6	5.5	

我们对每一行或者每一列求平均值，就是上面表格中的边际平均值(Marginal Average) 边际平均值是忽略一个自变量，仅仅对因变量在另外一个自变量的某一个可能值下求平均值。比方说，边际平均值“7”就意味着在两次对如何理解他人行为的测量中，人们认为主观意图对不迟到这个行为的影响程度是7。我们看到，对于迟到的行为，人们认为主观意图在这里的影响程度是4.5。类似的，我们还计算出，不论是否迟到，有相似背景的人认为主观意图对他人的行为的影响程度是6，没有相似背景的人认为主观意图对他人的行为的影响程度是5.5。



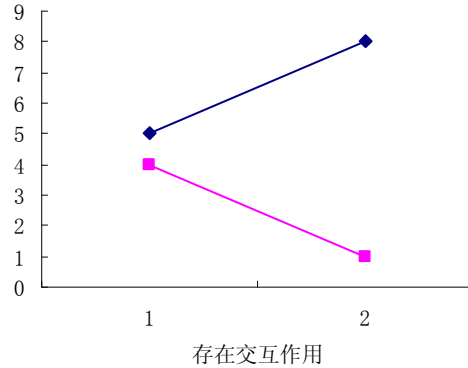
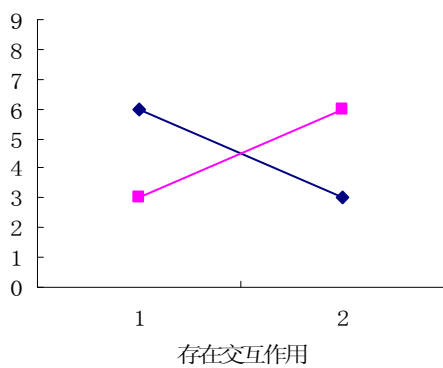
现在看一下根据上面的数据画出的图。横轴代表有无相似背景这一自变量，纵轴代表对行为意图的判断这一因变量，而另外一个自变量“行为好坏”用不同颜色的线段来代表，蓝线代表不迟到这一好行为，红线代表迟到这一坏行为。由于我们有两个自变量，但是只有一个横轴，我们必须决定用那个自变量做横轴。一般来说，这取决于假设表述的形式。在上面的例子里，我们首先是固定行为的好坏，改变背景这个自变量，所以我们就把背景这个自变量作为横轴。上图显示，对于一个人的好的行为，与他有相似背景的人比没有相似背景的人更倾向于认为那是出于他的主观意图；而对于一个人的坏的行为，与他有相似背景的人比没有相似背景的人更倾向于认为那不是出于他的主观意图。

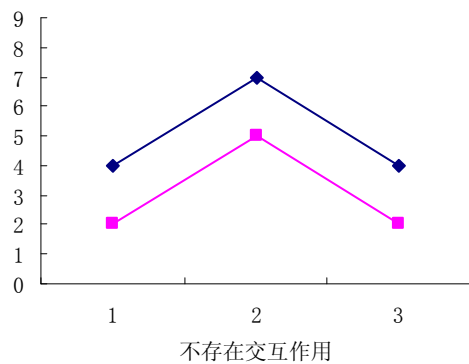
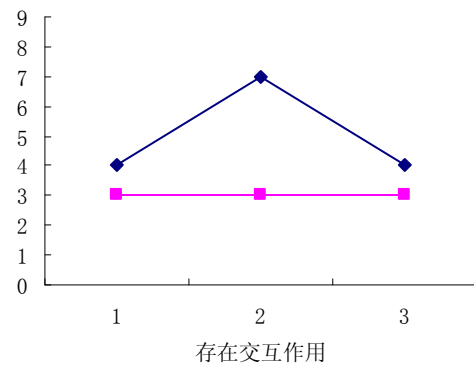
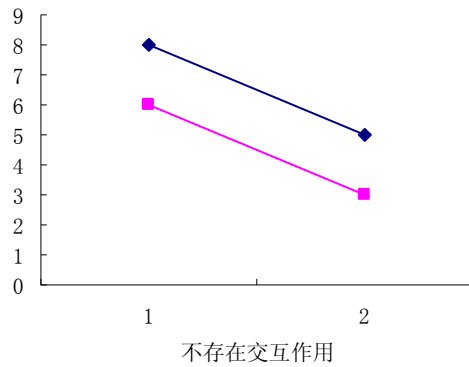
但是如果把行为的好坏作为横轴，我们就得到了这样一个图：



对于这个图，比较容易的理解方式是：对于有相似背景的人，人们容易认为他人好的行为是出于其主观意图，而差的行为则不是出于其主观意图；对于没有相似背景的人，行为的好坏对于人们对于他人主观意图的推测没有影响。

很多时候，图可以让人更直观地观察到是否存在交互效应。一般来说，如果两条线是平行的，可以推测实验结果没有交互效应；如果两条线的斜率存在较大差异，可以推测实验结果是存在交互效应的。我们在下面画出了几种可能的情况。需要指出的是，交互效应并不以主要效应的存在为前提。如图 xxx 所示，虽然这张图显示的结果并不存在主要效应，也就是说，这张图的边际平均值相同，但是由于两条线的斜率明显不同，这个结果构成了一个交互效应。





检验实验结果是否存在交互效应的常用方法是 ANOVA 分析 (Analysis of Variance)。

我们上面讲到的例子是一个典型的组间因素设计 (Between-subjects Factorial Design)。必须明确的是，因素设计和组间设计、组内设计之间不是互相排斥的，一个因素设计可以是单纯的组间因素设计，也可以是组内因素设计，甚至可以是组间组内混合的因素设计。

让我们先来说说组间组内混合的因素设计。假定我们现在有两组人，一组人先想象一个跟他同时参加新员工培训的同事，这个同事上班从来不迟到，并让被试回答他认为这个同事上班不迟到多大程度上是因为对自己有较高的要求；然后再让被试想象一个没有跟他一起参加新员工培训的同事，这个同事上班也从来不迟到，并让被试回答他认为这个同事上班不迟到多大程度上是因为对自己有较高的要求。另外一组人也回答两次问题，只不过这组被试需要想象一个同事上个月上班迟到了 5 次。这就是一个组间组内混合的因素设计。其中，是否有相似背景这个自变量是一个组间变量，而同事的行为这个变量是一个组内变量，同样的被试分别想象了两个同事，并两次回答了相同的关于因变量的问题。

如果更进一步，让被试把所有的实验组都经历一遍，那就是一个完全的组内因素设计。

到底选取组间因素设计，组内因素设计，还是混合因素设计，不是实验者任意决定的。它取决于你的假设和你的实验条件。在上面的例子中，很明显完全组内因素设计不是一个

好的选择。它不仅容易产生传递效应，而且非常容易被猜测出实验者的意图。如果我们想要保证各个实验组互不影响，减少混淆因素的影响的话，采用组间设计比较妥当。

当然，组内因素设计或者混合因素设计也有它们自身的好处。比方说，有的时候一个假设本身看的就是组内因素的变化，这个时候就应该采用组内因素设计或者混合因素设计。比如说，你想检验人们不同时间点上的心情的变化，以及是否吃早饭对心情的影响。你想知道人们是不是下午比早上心情好，而且你想研究吃不吃早饭是否和时间对人们的心情产生交互作用。由于本身就是想比较同一个自变量在同一组人身上的变化，时间变量最好作为一个组内变量。是否吃早饭当然是作为一个组间变量，因为你不可能让人同时既吃早饭又不吃早饭。

如果一个因素设计有两个自变量，相对应的交互作用就叫做两重交互作用（**Two-way Interaction**）。如果我们有多个自变量，这样的设计叫做高阶设计（**Higher-order Design**）。比方说，在背景和行为之间的交互影响的例子中，再加入时间这个自变量，分别在早上和晚上测量人们如何理解他人的行为，我们就有了一个 **2X2X2** 的高阶设计，一共有 **8** 个实验组。在一个有三个自变量的设计中，假设三个自变量分别为 **A**，**B**，**C**。那么这个实验会产生三个主要效应，分别对应 **A**，**B**，**C**。还有三个两重交互作用，分别发生在 **AB** 之间，**AC** 之间，和 **BC** 之间。还有一个三重交互作用，发生在 **ABC** 三个自变量之间。对于一个高阶设计来说，我们的假设关注的应该是多重交互作用，否则不必也不应该采用高阶的设计。

需要注意的是，到底是几重交互作用，取决于你有多少个自变量，并不取决于自变量水平的个数。比方说，我们上面的图 **xx** 和图 **xx** 就是一个两重交互的例子，因为这个图上只有两个自变量。虽然其中一个自变量有三个水平，这仍旧是一个两重交互作用，而不是一个三重交互作用。我们上面讲的都是每个自变量有两个可能值的情况。实际上很多时候自变量有多于两个的可能值。这个时候，只要我们只有两个自变量，交互作用就仍旧是两重交互作用，尽管你需要更多的实验组了。总之，实验设计中可以有多个自变量，而每个自变量又可以有多个水平，自变量既可以事组内变量，也可以事组间变量。

研究中最常见的就是两重交互作用。当自变量增多的时候，对实验结果的解释就变得困难起来。很多时候我们很难理解一个四阶的交互作用到底意味着什么。更多的自变量会混淆我们对问题的理解，而且通常不具备理论上的重要性。这时候我们可以采用一个实验设计技巧：把我们不关心的多重交互作用和区块混淆在一起（**Confound With Block**）。这种做法的指导思想是让一个区块内元素受某种多重交互作用同样的影响。这样区块的影响和多重交互作用这两种我们都不关心的，但是会影响实验结果的因素被放到一起考虑了，这样就可以把其共同作用的影响仅当作是区块的影响。具体的原理和处理方法可以参考 **Douglas C. Montgomery (2005)**。

总结

在本章中，我们首先介绍了研究的类型，以及什么样的假设才是一个好的假设。然后，我们着重讲述了如何在实验室中对假设进行检验。在实验室实验中，我们又着重讲了组间设计、组内设计和因素设计这三种最常见的实验设计方法和它们各自的优缺点。实验设计涉及到许多很多概念，有许多需要注意的问题。很多好的研究不仅有好的假设，还有让人

信服而且印象深刻的实验设计。实验设计本身是一门科学，同时也是一种艺术。

References

- Cook, T. D., & Campbell, D. T. (1979). *Quasi-experimentation: Design and analysis issues for field settings*, Chicago: Rand McNally.
- Cozby, P. C. (2001). *Methods in Behavioral Research*, Mountain View, CA: Mayfield Publishing.
- Douglas C. Montgomery (2005). *Design and analysis of experiments (Sixth edition)*, New York: John Wiley & Sons.
- Elmes, D. G., Kantowitz, B. H., & Roediger, H. L. (1999). *Research Methods in Psychology*, Pacific Grove, CA: Brooks/Cole Publishing.
- March, J. G., & Lave, C. A. (1975). *An Introduction to Models in the Social Sciences*, New York : Harper & Row.
- Schweigert, W. A. (2006). *Research Methods in Psychology*, Long Grove, IL: Waveland.

Suggested readings

Cook, T. D., & Campbell, D. T. (1979). *Quasi-experimentation: Design and analysis issues for field settings*, Chicago: Rand McNally.

Douglas C. Montgomery (2005). *Design and analysis of experiments (Sixth edition)*, New York: John Wiley & Sons.

Glossary

ABBA 互相抵消	ABBA counterbalancing
协方差分析	Analysis of covariance
ANOVA 分析	Analysis of variance
组间设计	Between-subjects design
区集	Block
传递效应	Carryover effects
天花板效应	Ceiling effect
完全随机化	Complete randomization
混淆变量	Confounding variable
对照组	Control group
需求特性	Demand characteristics
因变量	Dependent variable
区别选择	Differential selection
误差	Errors
实验者偏差	Experimenter bias
外部效度	External validity
额外因素	Extraneous factors
因素设计	Factorial design
可证伪的	Falsifiable
繁衍性	Fertility
实地实验	Field experiment
地板效应	Floor effect
霍桑效应	Hawthorne effect
高阶设计	Higher-order design
自变量	Independent variable
测量手段	Instrument
交互效应	Internal validity
实验室实验	Lab experiment
拉丁方设计	Latin-square design
主要效应	Main effect
成熟程度	Maturation
偶然减员	Mortality
样本不具代表性	Non-representative sample
无关因素	Nuisance factor
观察性研究	Observational study
观测值	Observations
可操作性定义	Operational definition
测试性实验	Pilot study
安慰剂效应	Placebo effect
实际意义上的重要性	Practical importance
母体	Population
测试前一测试后设计	Pretest-posttest design

准实验	Quasi-experiment
随机分配	Random assignment
随机化	Randomization
副效应	Reactivity
可重复性	Repeatability
复制	Replication
样本	Sample
样本均值	Sample mean
被试选择偏差	Selection bias
自选择	Self selection
统计控制	Statistical control
统计回归	Statistical regression
理论上的重要性	Theoretical importance
外部效度威胁因素	Threats to external validity
内部效度威胁因素	Threats to internal validity
两重交互作用	Two-way interaction
组内设计	Within-subjects design