随机实验

随机方法是经济学家们工具箱中不可或缺的一部分。与早期在美国进行的“社会实验”不同——美国有庞大的预算、庞大的团队和复杂的实施——近年来在发展中国家进行的许多随机实验（randomized evaluations）的预算比较小，这使得发展中国家的经济学家能够负担得起。与当地合作伙伴的小规模合作也给了研究人员更大的灵活性，他们通常可以影响项目设计。因此，随机实验已成为一种强有力的研究工具。

这篇文章是一个实用指南和“工具箱”。如果你有兴趣将随机试验作为自己研究设计的一部分，我们希望这篇文章对你有用。

本文的大纲如下。在第2节中，我们使用标准的“潜在结果”框架来讨论随机评估如何克服回顾性评估所特有的诸多问题。我们关注选择偏差的问题，当个人或群体根据可能影响其结果的特征被选择接受处理时，就会出现选择偏差，这使得从诸多驱动选择的因素中分离出处置效应变得困难。这一问题由于对支持先前信念并呈现统计显著结果的回顾性研究的发表偏倚而复杂化。我们将讨论如何精心构建随机评估来解决这些问题。

在第3节中，我们讨论了如何在该领域中引入随机化。与哪些合作伙伴合作?如何使用试验计划?在道德和政治上可接受的方式下，随机化有哪些不同的引入方式?

在第4节中，我们将讨论研究人员如何影响设计的最终效果，或者有机会得出统计上有意义的结论。如何选择样本大小?随机化的程度、控制变量的可用性和分层的可能性如何影响设计的最终效果?

在第5节中，我们讨论了研究人员在进行随机评估时将面临的实际设计选择:在什么水平上进行随机?析因设计的优点和缺点是什么?什么时候收集什么数据?

在第6节中，我们将讨论如何在偏离最简单的基本框架时从随机评估中分析数据。我们回顾了如何处理不同群体的不同选择概率，不完善的依从性和外部性。

在第7节中，我们讨论了当数据被分组和考虑多个结果或亚组时，如何准确估计估计治疗效果的精度。最后，在第8节中，我们讨论了从随机评估中得出一般结论所涉及的一些问题，包括在设计评估和解释结果时必须使用理论作为指导。

第9节中，我们附上了三个实证案例，和一篇社会偏好方面的实证综述。

为什么要做随机实验?

因果推断中存在的问题

任何试图得出因果推理问题的尝试，如“教育对生育率的因果影响是什么?”或“班级规模对学习有什么因果影响?”要求回答的，都是反事实的问题，这些问题的困难程度是显而易见的。

在某一时刻，一个人或者参加了某项目，或者并没有。因此，我们无法估计该项目对特定个人的影响。然而，我们可以通过将受到某项目、政策或变量(后面我们将其称为处理)影响的群体与另一个没有接受该处置的类似群体进行比较，从而能得到处理对于一组个体的平均影响。

为此，我们需要一个对照组。理想的对照组，在不接受处置的情况下，与实验组不接受处理的结果是相似的。然而，在现实中，接受处理的人与没有接受处理的人通常是不同的。由于自选择、习俗、法规等各种原因，与接受处理的人相比，未接受处理的人往往是较差的对照组。两组之间的任何差异都可以归因于项目的影响或预先存在的差异(即“选择偏差”)。如果没有一种可靠的方法来估计这种选择偏差的大小，就无法将总体差异分解为处理效应和偏差项。

引入鲁宾(Rubin, 1974)提出的潜在结果概念。考虑教科书对学习的影响。如果学校i有教科书，令Y_i^T为等于学校里学生的平均测试分数;如果学校i没有教科书，则令Y_i^C为同一所学校里学生的平均测试分数。将Y_i定义为实际观测到的结果。我们感兴趣的是差异$Y_i^T - Y_i^C$，这就是教科书对学校的影响。正如我们上面所解释的，虽然每个学校都有两种潜在的结果，但我们只能观测到一种。我们感兴趣的是教科书对学习的平均影响，即$E[Y_i^T - Y_i^C]$。

假设我们有某地区大量学校的数据。有些学校有教科书，有些没有。一种方法是取两组学生的平均值，并检查有教科书的学校和没有教科书的学校之间，学生平均的考试分数之间的差异。在样本量足够大的情况下，这种差异会收敛到:

$$D = E[Y_i^T|\text{学校有教材}] - E[Y_i^C|\text{学校没有教材}] = E[Y_i^T|T] - E[Y_i^C|C]$$

减和加$[Y_i^T|T]$，即实验组中未接受处理的样本的预期结果(这个量无法观察，但在逻辑上能很好地定义)，我们得到，

$$\begin{aligned} D &= E[Y_i^T|T] - E[Y_i^C|T] - E[Y_i^C|C] + E[Y_i^C|T] \\ &= E[Y_i^T - Y_i^C|T] + E[Y_i^C|T] - E[Y_i^C|C] \end{aligned} \tag{2.1}$$

第一项$E[Y_i^T - Y_i^C|T]$就是我们试图求出来的处理效应(即处置对被处置样本的影响)。在教科书的例子中，它回答了这样一个问题:平均而言，在实验组的学校中，教科书有什么作用?

第二项$E[Y_i^C|T] - E[Y_i^C|C$为选择偏差。它捕获了实验组和对照组之间不接受处理时的潜在差异;即若没有接受处理，学校的平均考试分数也可能不一样。因为除了教科书的可能影响之外，实验组和对照组之间可能存在系统性差异。

由于观测不到$E[Y_i^C|T]$，一般没办法评估选择偏差的大小(甚至不知道它是正是负)，而选择偏差的大小在一定程度上解释了实验组和对照组之间的结果差异。所以，许多实证工作的一个基本目标是确定我们能否假设选择偏差不存在，或找到纠正选择偏差的方法。

随机化解决了选择偏差

一种完全消除选择偏差的方法是将样本随机分配到实验组和对照组。在随机评估中，从感兴趣的总体中选择一个非个体样本。注意，“总体”可能不是整个总体的随机样本，而是根据可观测到的数据选择的;因此，我们将了解处理对抽取样本的特定子总体的影响。我们将回到这个问题上。然后将实验样本随机分为两组:实验组(N_T个体)和对照组(或控制组(N_C个体))。

然后，实验组暴露于“处置”(处置状态是T)，而对照组(处置状态C)未接受处理。观察并比较实验组和对照组的结果。例如，在100所学校中，随机选择50所学校接受教科书，50所学校不接受教科书。平均处置效应可以被估计为两组间Y的经验平均之间的差异:

$$\hat{D} = \hat{E}[Y_i|T] - \hat{E}[Y_i|C].$$

其中$\hat{E}$表示样本平均值。随着样本量的增加，这种差异收敛为:

$$D = E[Y_i^T|T] - E[Y_i^C|C].$$

由于是否被处理是随机分配的，被分配到实验组和对照组的个体在预期上的差异只有他们是否接受了处理。如果这两个人都没有接受处理，他们的预期结果是一样的。这意味着选择偏差$E[Y_i^C|T] - E[Y_i^C|C]$等于零。此外，如果一个个体的潜在结果与其他个体的处理状态无关(Angrist, Imbens, and Rubin(1996)，“稳定的单位处理价值假设”(SUTVA))，我们有

$$E[Y_i|T] - E[Y_i|C] = E[Y_i^T - Y_i^C|T] = E[Y_i^T - Y_i^C].$$

这正是我们感兴趣的对处置T的因果参数。

那么，如何求回归对应的$\hat{D}$呢？考虑下式:

$$Y_i = \alpha + \beta T + \epsilon_i, \tag{2.2}$$

其中T是分配到实验组的“虚拟人”。(2.2)式可以用普通最小二乘法估计，进而能很容易地得到
$\hat{\beta}_{OLS} = \hat{E}(Y_i|T) - \hat{E}(Y_i|C)$.

这个结果告诉我们，当一个随机评估被正确地设计和执行时，它提供了对所研究样本中处理效应的一个无偏估计——这个估计在内部是有效的。在实操中进行随机实验时，这种简单的假设在许多方面会遭遇失败。本文描述了如何正确地实施随机评估，以减少此类失败，以及如何正确地分析和解释此类评估的结果，其中也包括背离这一基本设置的情况。

在继续之前，记住表达式(2.1)的意思是很重要的。评估的是特定项目对结果(如考试分数)的整体影响，允许其他输入根据项目的变化而变化。它可能不同于教科书对考试成绩的影响，使其他一切保持不变。

为了理解这一点，假设产出的生产函数Y具有形式$Y = f(I)$,其中I是投入向量，其中一些可以直接使用政策工具改变，另一些依赖于家庭或企业的反应。这种关系是结构性的;无论受政策变化影响的个人或机构会如何行为，它都是有效的。这里的I就是一个结构参数，是内置的输入向量。

考虑对向量I中的一个元素t进行更改。一个感兴趣的估计是，当所有其他解释变量保持不变时，t的改变如何影响到Y，即Y对t的偏微分。第二个感兴趣的估计是Y对t的全微分，它包括了因t的变化其他参数的改变。一般来说，如果其他投入是对t的补充或替代，那么I的外生变化将导致其他投入j的变化。

总的来说，偏微分和全微分可能有很大的不同，而且两者都可能是政策制定者感兴趣的。全微分是值得关注的，因为它揭示了在外生性输入和代理重新优化后，结果会有怎样的呈现。实际上，它能告诉我们政策对利益结果的"真实"影响。但全微分可能无法提供总体福利效应的衡量标准。重新考虑向学生提供教科书的政策，家长可能会减少家庭购买教科书，转而购买一些不具有教育生产功能的消费品。考试分数或其他教育结果变量的总导数并不能捕捉到这种重新优化的好处。然而，在适当的假设下，偏微分也可以为投入福利的影响提供适当的指导。

随机评估(以及其他内部有效的程序评估)的结果提供了处置效应的简化形式的估计，而这些简化形式参数是全微分。如果研究人员指定模型，将各种输入与感兴趣的结果联系起来，并收集这些中间输入的数据，就可以获得偏微分。这强调了要估计一项政策的福利影响，随机化需要与理论相结合，这是我们在第8节中谈到的主题。

控制选择偏差的其他方法

除了随机化，还有一些方法可以用来解决选择偏差问题。这些方法都是为了在某一识别假设下，创建出一套有效的组间比较。这些识别假设不可直接验证，并且任何特定研究的有效性都取决于这些假设看上去的说服力。本章的目的不是为了详细回顾这些方法，本节我们将简要地讨论它们与随机评估之间的关系。

通过控制可观测对象来控制选择偏差

第一种可能性是，在控制住一组可观测变量X的条件下，处理效应可以被认为与随机分配一样好。也就是说，存在这样一个向量X使得

$$E[Y_i^C|X,T] - E[Y_i^C|X,C] = 0.$$

当处置被随机分配给可观测变量X时，这显然是正确的。换句话说，对实验组与对照组的观察分配并非无条件随机的，但是在X集合中变量相互作用定义的每一层中，分配是随机进行的。在这种情况下，在对X施加条件作用后，选择偏差消失了。我们将在6.1节中讨论如何分析这种设置产生的数据。然而，在大多数观察环境中，在任何一点上都没有明确的随机化，人们必须假设适当地控制可观察变量就足以消除选择偏差。

有不同的方法来控制变量X的集合。第一种方法是完全非参数匹配，即当X的维度不太大时，先计算由X的各种可能值形成的每个单元内实验结果和对照组之间的差异，然后处置效应就是这些单元之间差异的加权平均值(参见Angrist (1998)，用于该方法对服兵役影响的应用)。

第二，如果X有许多变量或包含连续变量，则上述方法(完全非参数匹配)不实用。在这种情况下，可以根据变量X分配给实验组的概率，或者设计一些其他的方法来实现基于"倾向得分"的匹配。

第三种方法是在回归框架中使用参数或非参数化方法来控制变量X。然而，使用参数或非参数方法进行回归，需要满足基本假设:以可控的可观察变量为条件，实验组和对照组的个体间的潜在结果没有差异。要做到这一点，变量集合必须包含实验组和对照组之间的所有相关差异。这一假设是不可检验的，每个案例的可信度都不一样。多数情况下，控制变量仅仅只是那些恰好在数据集中可用的变量。无论如何灵活地引入控制变量，选择偏差(或"忽略变量")仍然是一个问题。

断点回归估计

控制可观测变量的一个非常有趣的特殊情况发生在这样的情况下:分配给处理组的概率是一个或多个可观察变量的不连续函数。例如，一个小额信贷组织可能会限制家庭土地不足一英亩的妇女获得贷款的资格;成绩在50%以上的学生可通过考试;班级人数不得超过25人。如果任何不可观察的变量与用于分配治疗的变量相关的影响是平滑的，以下假设是合理的一个小ϵ

$$E[Y_i^C|T, X < \bar{X} + \epsilon, X > \bar{X} + \epsilon] = E[Y_i^C|C, X < \bar{X} + \epsilon, X > \bar{X} + \epsilon],$$

这里X是潜在变量，$\bar{X}$是赋值阈值。这个假设意味着在$\bar{X}$的ϵ邻域内，选择偏差为零，这个假设是"断点回归估计"的基础(Campbell 1969)。这个方法使用恰好低于阈值的个体作为对照组，使用恰好高于阈值的个体作为实验组来估计处置效应。

断点回归估计在从事项目评估的研究人员中非常流行，许多人认为，当分配规则被严格地执行时，它消除了选择偏差。但是在发展中国家，断点回归估计普遍会遭遇两种障碍，导致这些地方的经济学家们很少采用这种方法。首先，分配规则的执行并不严格。比如乡村银行向客户放贷的例子（Morduch，1998；Pitt，Khandker，1998)，默多克表明，尽管官方规定不向拥有一英亩以上土地的家庭放贷，但信贷官员仍在行使自己的决定权。在一英亩门槛上借款的概率没有间断。第二个问题是，执行程序的官员可能能够操纵决定资格的潜在变量的水平，这使得个人的地位高于或低于阈值是内生的。在这种情况下，我们不能说分界线两边的个体具有相似的潜在结果，方程(2.4)就不成立了。

DID和固定效应

当数据在处理前后都存在时，DID(difference-in-difference)估计使用实验组和对照组之间结果的差异来控制两组之间已存在的差异。用$Y_1^T(Y_1^C)$表示处理后第1阶段"如果处理"("如果不处理")的潜在结果，$Y_0^T(Y_0^C)$表示处理前第0阶段"如果处理"("如果不处理")的潜在结果。个体属于T组或者C组。T组在第1期接受处理，并在第0期未接受处理。C组从未接受过处理。

DID的估计量为:

$$\hat{DD} = \hat{E}[Y_1^T|T] - \hat{E}[Y_0^C|T] - \hat{E}[Y_1^C|C] - \hat{E}[Y_0^C|C],$$

并在$\hat{E}[Y_1^C|T] - \hat{E}[Y_0^C|C] = \hat{E}[Y_1^C|T] - \hat{E}[Y_0^C|C]$的假设下，给出了一个处理效应的无偏估计，即在不进行处理的情况下，两组的结果将相互平行。

当存在多个时间段或多个实验组时，DID估计量扩展为固定效应。在控制了时间和组内因素后，通过对控制变量的结果进行回归得到了固定效应估计值。DID和固定效应估计在应用中都很常见。它们是否令人信服取决于在不处理情况下结果的平行假设是否令人信服。特别要注意的是，如果两组样本的结果在处置前差别很大，那么随时间变化的处理结果所选择的功能形式将对最后的结果产生重要影响。

实验与非实验估计的比较

由于随机试验具有的优势，越来越多的文献通过同时使用随机实验和非实验方法进行某项目、并比较两者结果的偏差，来估计该项目的影响。$LaLonde$的开创性地研究发现，在项目评估中不少常用的计量经济学分析办法没有产生精确估计，甚至得到与实证大相径庭的结果($LaLonde$1986)。许多后续的研究都侧重于分析倾向评分法的匹配表现，其中一些研究发现非实验方法可以很好地复制实验结果，但另一些的结果却更加消极。Glazerman、Levy和Myers(2003)对美国福利、职业培训和就业服务项目研究中的实验性非实验性方法进行了更全面的评述。综合12项实验与非实验的比较研究结果，他们发现回顾性评估结果往往与随机评估结果大相径庭。他们找不到一个，既能持续消除偏差，又能明确界定处理效应与偏差项的策略。

Cook, Shadish, and Wong(2006)的研究大部分是与教育有关的领域，他比较了随机和非随机研究，并得出了更加微妙的结论。文章发现，当非实验技术接近断点回归、或“中断的时间序列”(即往期数据较长的DID数据)时，实验结果和非实验结果是相似的，但匹配等其他方法没有类似的结果。他总结道，类似于实施良好的随机评估，设计良好的准实验(尤其是断点回归实验)可能有同样令人信服的结果，但“你不能通过统计来纠正你在设计上犯的错误。”库克的发现很有意思，但他所说的准实验标准(例如，严格遵守阈值规则)太严格了一些。

在发展中国家已经有了一些类似的比较研究。一些人认为遗漏变量的偏差是一个重要问题;其他人发现非实验性方法在某些情况下可能表现良好。budlemeyer和Skofias(2003)以及Diaz和Handa(2006)都关注PROGRESA，这是一个在20世纪90年代末在墨西哥采用随机设计的扶贫项目。budlemeyer和Skofias(2003)使用随机评价结果作为基准来检验回归不连续设计的性能。他们发现这种设计的性能很好，这表明如果政策的不连续性被严格执行，回归不连续设计框架可能是有用的。Diaz和Handa(2006)再次使用PROGRESA数据，将随机实验估计和倾向得分法估计进行了比较。他们的结果表明，当有大量的控制变量可用时，倾向得分法表现很好。

在肯尼亚的几项研究发现，遗漏变量可能导致严重的偏差，使得前瞻性随机评估与回顾性评估的结果显著不同。Glewwe、Kremer、Moulin和Zitzewitz(2004)研究了一个非政府组织项目，他们发现，使用DID方法可以一定程度上减轻该问题，只是没有完全解决问题。

Miguel和Kremer(2003)和Duflo、Kremer和Robinson(2006)分别比较了在使用驱虫药和使用肥料的情况下，同伴效应的实验值和非实验值。两项研究都发现，个人的决定与他们所接触到的其他人的决定有关。然而，正如Manski(1993)所指出的，这可能是由于同伴效应以外的许多因素，特别是这些个体共享相同的环境这一事实。在这两种情况下，随机化提供了外生变异的机会，使特定网络的某些成员采用了创新(分别是除虫或施肥)。这两项研究都发现了与非实验结果明显不同的结果:Duflo, Kremer, and Robinson(2006)发现没有学习效应，而Miguel和Kremer(2003)发现了消极的同伴效应。

比较研究可以用来评估回顾性估计中偏差的大小和普遍程度，并据此判断方法的可靠性。然而，正如下一小节所讨论的，这种比较不同方法准确性的研究必须谨慎进行。如果这些比较研究的回顾性部分是在了解实验结果的情况下进行的，那么为了与实验估计相匹配，就会自然而然地倾向于从貌似合理的比较组和方法中进行选择。为了解决这些问题，未来的研究人员应该在随机评估结果公布之前进行回顾性评估，或者在不了解随机评估和其他回顾性研究结果的情况下进行盲回顾性评估。

发表偏倚

发表偏倚（publication bias）又称为出版偏倚，是指在同类研究中，具有统计学显著性研究意义的结果相较于无显著性意义和无效的结果而言更容易被接受和发表的现象。研究者、审稿人或编辑依赖研究结果的方向和强度进行决策，从而产生了偏差，这使得出版的过程不是随机事件。简言之，阳性的结果（例如，发现A与B相关）比阴性的结果（例如，未发现A和B之间存在关联）更有可能得到发表，而这可能会混淆我们最终所见到的结论，掩盖结论的真实性。

非实验研究中的发表偏倚

发表偏倚的存在加剧了非实验性研究结果的不确定性。发表偏倚的原因甚多，归纳起来有如下几个方面：

1. 研究者是产生发表性偏倚的主要因素。考虑到参考文献的声望和权威性以及自身研究的可被接受性，研究者通常重点关注和报告那些具有统计学显著性研究意义的结果，而不显著的研究结果往往会被搁置（即所谓的“文件抽屉”现象）。例如，在进行回归分析时，控制变量和工具变量的选择是非常重要的，这些变量的选取往往会根植于一些现有的文献。如果基于这些变量进行回归所得到的结果在统计意义上是显著的，并且符合研究者期望得到的结果，那么研究者很可能会认为这些设定或者变量的选取是合适的，而不会去寻找其他的替代方案。然而，如果回归结果在统计上不显著或与研究者的预期相违背，那么研究人员就更有可能考虑选择其他可能的方案。以上这种行为将导致对普遍持有的观点的拒绝不足，可能会掩盖结论的真实性。

2. 期刊编辑和审稿人的决策。例如，编辑或审稿人认为不显著的结果没有价值，建议退稿。
3. 政策制定者倾向于依据显著的研究结果制定政策，阳性研究的被引频率过高也会导致发表偏倚的产生。

对于发表偏倚，可以通过多种方法进行检验，例如漏斗图法、Begg检验、Egger检验等。然而漏斗图存在较大的主观性，而Egger检验及Begg检验则给出了统计描述，$p > 0.05$表示没有明显的发表偏倚，$p \leq 0.05$则表明存在一定的发表偏倚。

研究表明，在经济学领域的研究文献中，存在着许多发表偏倚现象。例如，Ashenfelter, Harmon和Oosterbeek（1999）检验了研究教育回报的文献的发表偏倚，并得出结论：教育回报的IV估计值大于OLS估计值这一常被引用的结果可能只是一种发表偏倚的产物。Card和Krueger（1995）对最低工资文献进行了发表偏倚的检验，同样发现了导致显著结果过度报告的发表偏倚现象。

随机化与发表偏倚

部分发表偏倚问题可以通过随机评估的方法解决。如果科学地实施了随机评估，那么从理论上讲我们可以得到处置效应的无偏估计。这意味即便出现了违背预期的结果，它们也不会被认为是错误的。此外，发表偏倚实际上是一种选择性偏差，而随机化的分组则在一定程度上限制了研究人员对实验组和对照组的自主选择，从而减少选择性偏差。

在实践中进行随机评估

在现实世界中，我们对某项政策或者干预措施进行评估时，为了消除选择性偏差，通常需要进行随机化处理。下面本文将从随机评估的合作者以及随机化的方式等方面进行阐述。

合作者

不同于经济学家可以独立完成的实验室实验，在实地活动中进行随机评估往往需要一些合作者，这些合作者通常是对政策效果感兴趣的政策制定者，包括政府，非政府组织以及盈利性公司。

考虑到预算约束以及实施的便利性，一些政府项目并不是在整体的范围内实施的，而是分时分批试点实施。例如，墨西哥政府实行的PROGRESA项目。当项目启动时，由于预算资金的限制，墨西哥政府随机选择了506个试点社区进行政策评估，并随机将这些试点社区平均分为两组，其中一组作为实施政策的实验组，另一组作为未进行政策干预的对照组。相关部门公开了对应的数据，而政策评估的任务则交给了相应的学术研究者。研究表明，PROGRESA项目在改善健康和教育方面是有效的 （详见 Gertler和Boyce (2001)，Schultz（2004））。这类政府试点项目在发展中国家同样很常见。

许多以发展为重点的非政府组织经常探索一些创新型的项目，并有意愿与研究人员合作，以评估现有项目或者新型项目的有效性。近年来，这些非政府组织也与研究组织或基金会展开合作，进行了大量的随机评估。

此外，盈利性公司同样也要开展随机评估，通常是为了更好地了解他们的业务运作情况，从而更好地为客户服务，增加企业利润。例如，许多微型金融机构现阶段都加强了与研究人员的合作，进行金融产品评估，以便更好服务于投资者。

试点项目：从项目评估到田野实验

政策制定者倾向于在试点阶段开展随机评估以检验政策的有效性，并对政策进行完善。这些试点项目的实施促进了标准的“项目评估”转化为“田野实验”，即政策制定者和研究人员共同进行实验，以找到问题的最佳解决方案（Duflo 2006）。随机田野实验在发展经济学领域得到了广泛的应用。值得指出的是，单一的项目评估与田野实验之间的区别是非常大的，前者仅涉及某个项目的某一个实验组和某一个对照组之间最直接的比较，而后者则涉及大量小组的政策评估。

随机化的方式

随机化分组的方式有四种——超额认购法、分阶段法、组内随机化、鼓励设计。

超额认购法

这种方法通常适用于资源或者能力限制使得某一项目或者服务无法面向所有民众，即供不应求的情景。例如，Kremer通过与一家银行合作，评估了扩大消费信贷在南非的影响。这家银行扩大消费信贷并进行随机分组的具体操作是：随机批准了一些通常会被拒绝的边缘贷款申请。这样的评估之所以可行，是因为该项实验设计对银行的正常业务活动造成了最小的干扰。

分阶段法

财政和行政方面的限制常常导致非政府组织随着时间的推移逐步实施项目，而随机选择往往是确定逐步实施顺序的最公平的方式。小学驱虫项目就是这种随机分阶段试验的一个例子(Miguel和 Kremer, 2004)。1998-2002年期间，该项目向肯尼亚Busia农村地区的75所小学的儿童提供肠蠕虫和血吸虫病的治疗以及预防蠕虫的健康教育课程。该计划将学校随机分为三组，每组25所小学。第一组于1998年开始实施项目，第二组于1999年开始，第三组于2000年开始。通过将1998年第一组的数据与其他两组进行对比，将1999年第一、二组的数据与第三组对比，研究人员发现，驱虫项目能显著改善健康状况，并且提高学校参与度。

组内随机化

印度非政府教育组织Pratham 对 balsakhi项目的评估就是组内随机化的一个典型例子。该项目旨在为印度城市贫困学校提供补救性教育援助。Pratham培训了一批导师，这些导师被称为balsakhi，他们为孩子们提供数学和阅读理解方面的辅导。为确保校方的配合，每一所被试学校每年都会收到一批balsakhi。然而，基于随机分配原则，一些学校被要求在三年级使用balsakhi，另一些在四年级使用（Banerjee, Duflo, Cole和 Linden, 2007）。

激励设计

激励设计适用于由于伦理或者现实原因而难以随机分配项目的情景。研究人员并没有针对项目本身对受试者进行随机分配，而是随机地给受试者分配鼓励。早期的激励设计之一是关于GRE学习是否能提高考试成绩的研究(Holland,1988)。在这项研究中，显然每个人都可以学习，但研究人员通过给随机选择的一组GRE考生邮寄免费材料，增加了为学习而学习的学生人数，从而实现了随机化。值得注意的是，激励设计只是增加了接受处置的概率，而不会将其从0变为1，因此激励设计加大了对研究结果分析的难度（详见6.2节）。

样本量、实验设计以及实验功效

实验设计的功效是指在给定的效应大小和给定的统计显著性水平下我们能够拒绝零效应假设的概率。样本量以及其他实验设计的相关选择都会影响实验的功效。

在本节中，我们首先回顾计算实验功效计算的基本原理，然后讨论实验设计因素（如多个处置组，组别随机化，部分依从性，控制变量和分层问题）的影响。最后讨论实验功效计算所涉及的实际步骤以及在实际评估时它们所扮演的角色。

功效计算的基本原理

功效计算的基本原理可以用一个简单的回归框架来说明。正如我们上面讨论的，两组样本均值的差异(我们对平均处置效应的估计)是回归中的OLS回归系数β 。

$$Y_i = \alpha + \beta T + \varepsilon_i$$

假设只有一种可能的处置，并且只处置样本的一定比例P。现在假设每个个体都是从一个相同的总体中随机抽样的，这样观察结果就可以被假定为i.i.d.，方差为σ^2.

那么回归系数β的OLS估计量$\widehat{\beta}$的方差由以下公式给出：

$$\frac{1}{P(1-P)}\frac{\sigma^2}{N}$$

处置组和对照组之间划分样本的最优分配规则应是$\frac{P}{1-P} = \sqrt{\frac{c_c}{c_t}}$，即处置组的受试者与比较组的受试者的比例应该与对他们进行数据收集的单位成本的平方根的倒数成正比。

分组误差

我们上面讨论的许多实验设计都是组别随机化而不是个人随机化。在这种情况下，研究人员往往也可以获得个人数据。

当在组别随机化的项目中分析个体数据时，重要的是要考虑到误差项可能不是独立于个体的。同一组别的人可能受到某种共同的冲击，这意味着他们的结果可能是相关的。由于处置状态在组内也是一致的，因此结果的相关性可能被错误地解释为项目的影响。

考虑以下组别随机化方程（沿用（Bloom 2005）的方法）：

$$Y_{ij} = \alpha + \beta T + v_j + w_{ij}$$

其中j表示组别，ij表示个体。这时OLS估计量$\widehat{\beta}$的标准差可表示为：

$$\sqrt{\frac{1}{P(1-P)}}\sqrt{\frac{n\tau^2+\sigma^2}{nJ}}$$

而如果随机化是在个体层面进行，那么估计量$\widehat{\beta}$的标准差可表示为：

$$\sqrt{\frac{1}{P(1-P)}}\sqrt{\frac{\tau^2+\sigma^2}{nJ}}$$

由此可见，在每组成员数量固定的情况下，组别层面随机化的标准差与体层面随机化的标准差之比为设计效果：

$$D = \sqrt{1 + (n-1)\rho}$$

对于给定的样本量，增加每个集群的抽样个体数量所增加的精度远远小于增加随机化集群所增加的精度。

不完全依从

在确定最佳样本量时，我们还应考虑依从性可能不完美的可能性。

实验设计的功效来自那些最初被分配到处置组和那些没有被分配到处置组的人的结果之间的差异，不管他们是否被处置。

依从性在实验中的重要性体现在对于样本容量的影响上。如果一个研究人员在两个实验设计之间进行选择，这两个实验设计有着不同的依从水平，那么选择一个有较高依从水平的实验将会对最佳样本量有重要影响。这也将有助于考虑在何种水平上引入随机化。

控制变量

在一个简单的随机实验中，控制可能影响或预测结果的协变量的基线值并不影响β估计量的期望值，但它可以减少其方差。注意，控制受处置影响的协变量会通过反映处置的部分影响使得处置效果德估计产生偏差。因此，应在基线调查中收集有关协变量的信息。

一般来说，控制对结果有很大影响的协变量可以帮助减少估计的标准误差，从而减少所需的样本量。

分层

由于要使用的协变量必须提前选择，以避免数据挖掘，因此它们可以用来将样本分层(或块)，以提高估计的精度。这种技术(由Fisher(1926)首先提出)。

这种方法是将样本分成其某些可观察特征相同或相似值的组。随机化确保了处置组和对照组的期望是相似的。但分层是用为了确保在某些重要的可观测维度，在实际的样本中，处置组和对照组的期望是相似的这一点也能得到满足。

实际设计与执行问题

本节讨论进行随机评估时所面临的各种设计和执行问题。我们从随机化水平的选择开始：应该对个人还是更大的群体进行随机选择？然后我们讨论在同一样本内同时检验多个处置的交叉设计问题。最后讨论一些数据收集问题。

随机化水平

一个重要的实际设计选择是在哪个层面（个人、家庭、村庄、地区等层面）进行随机化。美国早期的社会实验是在个人层面上进行随机化的，而在发展中国家，许多评估是对群体进行随机化的。

对于许多干预或处置措施来说，可以选择在个人水平或群体水平上进行随机化。但哪种水平上的随机化更科学，却往往并不非常明显。

当随机化水平有一定的灵活性时，需要考虑以下几个因素。

首先，正如第4.2节所讨论的，随机分组的组规模越大，实现给定功效的总样本量就越大。因此，随机化程度对评估处置效应的预算和行政负担有潜在的巨大影响。这使得个人层面的随机化在可能的情况下显得更具吸引力。

其次，从处置组到控制组的溢出效应可能使得处置效应的估计产生偏差。在这种情况下，随机化应该在能更准确地捕捉到这些处置效应的层面上进行。例如，Miguel和Kremer(2004)发现驱虫药物的效果在对学校水平进行随机化时比之前在个体水平上进行随机化时要大得多。他们认为，由于蠕虫感染很容易在儿童之间传播，个体随机分组的对照组可能也从治疗中受益，这减少了处置组儿童和对照组儿童之间的结果差异。虽然在更大的随机水平上可能也存在这种溢出效应(例如，在Miguel和Kremer的样本中确实显示了学校之间也存在溢出效应)，但它们通常要小得多。另外，处置组中的个体很可能因为即将接受的处置而改变他们的行为，这也会使得处置效应的评估产生偏差。因此，随机化应该在能更准确地捕捉到这些处置效应的层面上进行。

第三，从执行的角度看，在群体层面上的随机化即使需要更大的样本量，有时也仍然更容易。这有很多原因。比如，在一些有较强的固定成本因素的处置措施中，让尽可能多的人能够利用处置措施是有成本效率的。再者，个体层面的随机化可能更容易导致个体对于处置执行机构的抵触与反感。因此对于处置执行机构来说，群体层面的随机化将会比个体层面的随机化更容易。

交叉设计

交叉设计概念

导致随机评估数量大幅增加的制度创新之一是更多地使用交叉设计。在交叉设计中，多个不同的处理被同时测试，同时进行随机化，使得处理彼此正交。

交叉设计的理解

在某些实验研究中，研究者要考察一个具有两水平的实验因素(设其两个水平分别为A和B)，但根据具体的专业需要，希望该实验因素的两个水平要先后施加于每一个受试者，以便节省受试对象的个数。显然，我们首先就会想到采用“自身配对设计”。但是，如果所有的受试者都是先接受A处理后接受B处理，那就人为地引入了“顺序误差”。因为我们很难保证A处理在每个受试者身上所产生的效应都能被完全清除之后，才使用B处理。也就是说，B处理所产生的实验效应中可能混杂进了A处理的实验效应，这样，对B处理的实验效应的评价就会出现严重的“偏性(bias)”。纠正这种错误的方法，就是采用交叉设计，即让全部受试者中的一半接受处理的顺序为“先A后B”，另一半接受处理的顺序为“先B后A”，从而使“A、B两种处理”在两组受试者中施加的先后顺序是“交叉”开的，故得名“成组交叉设计”。交叉试验设计的优点： 可以增加被试的利用率（组内设计）；排除被试人口统计学特征无关变量的影响，顺序效应等。

案例举例

Kremer (2003)描述了在肯尼亚西部的教育中进行的许多这样的实验。有两种方法可以考虑交叉设计。首先，它们可用于测试相对于比较组和相对于彼此的各种干预和干预组合。他们还可以确定治疗是否有重要的交互作用。决策者往往有兴趣使用各种战略来改变一种结果。例如，我们上面讨论的PROGRESA方案是几个方案的结合:现金转移、向妇女重新分配资源和奖励部分。从政策角度来看，对" PROGRESA全面方案"的评估可能足以让墨西哥政府决定是否继续实施PROGRESA。但为了了解行为，并出于政策目的，为了了解PROGRESA的哪些组成部分应该扩大规模，人们可能想知道该计划的激励部分是否必要，将资金分配给女性而不是男性是否重要等。

可能面临的问题

如果一个研究者正在交叉干预A和B，其中每一个都有一个比较组，她获得四个组:没有干预(纯控制)；仅a；仅b；以及A和B加在一起(充分干预)。如果研究人员想测试B与A联合使用时是否会产生与单独使用时不同的效果，那么样本量必须足以让她从统计上区分A与A和B，以及B与A和B。正如我们在第4节中讨论的，可以考虑将完全干预组和纯控制组的规模设置为大于仅使用A和仅使用B的组。当这种交叉设计太昂贵或需要大量的样本量时，会面临问题是：评估组合程序(A和B)还是单独评估两个组成部分。政策制定者可能倾向于评估A和B组合，只要它有扩大规模的潜力，因为A和B在一起的整体组合更很可能分别产生比A或B更大影响。

从经济学家的角度来看，评估一揽子干预措施的缺点在于，它很难理解是什么促使人们做出了反应，也就是说很难找到其中的机制。优点是，能更密集的干预更有可能产生影响，从而表明结果确实会受到影响。如果其中任何一个因素都可能对研究结果产生重大影响，这一事实存在很大的不确定性，那么有必要首先对综合方案进行评估，然后再进行后续研究，以便理清各种潜在的工作机制。在初始研究中，可以使用可能受一种干预而不是另一种干预影响的中间变量来阐明哪部分干预是有效的。

案例：在驱虫试验中(Miguel和Kremer，2004年)，两个方案得到了结合:分发了驱虫丸，并向儿童提供了关于预防行为的建议(穿鞋、洗手等)。).研究人员收集的行为变量表明，在治疗学校没有行为改变。这有力地表明，干预措施中产生影响的部分是提供驱虫丸。

数据收集

我们在这里不讨论具体的勘测设计问题，因为它们已经被大量的文献所覆盖(例如Deaton (1997))。我们这里的主要重点是选择收集哪种类型的数据、基线调查的价值以及行政数据的使用。

基线调查

基线调查的目的是为了解研究对象的基础状态或研究开始阶段的情况而进行的调查，其目的是为以后活动方案的制定和展开提供基础资料。原则上，在随机化前提下不需要基线调查，但存在以下两个个原因使得基线调查非常有必要。

（1）基线调查所得到的控制变量会减少最终结果的可变性，从而减少样本量要求。就评价成本而言，进行基线调查与不进行基线调查之间的权衡可归结为比较干预的成本、数据收集的成本以及基线调查中可收集数据的变量可能对最终结果的影响。当干预措施昂贵而数据收集相对便宜时，进行基线调查将节省资金。当干预成本较低但数据收集成本较高时，在不进行基线的情况下运行较大的实验可能更具成本效益。

（2）它们使得检查初始条件和程序影响之间的相互作用成为可能。在许多情况下，这对于评估外部有效性将是相当重要的。其次，基线调查提供了检查随机化是否适当进行的机会。第三，收集基准数据为测试和完善数据收集程序提供了机会。在调查后回顾性地收集“干预前数据”的替代策略通常是不可接受的，因为即使计划不影响那些变量。它很可能影响对这些变量的回忆。有时，充足的行政数据已经可用，可以替代基线来衡量随机化的有效性，并为观

察干预措施的重要性提供控制变量。

使用管理数据

使用与之相关的行政数据(执行组织收集的数据，作为其正常运作的一部分)可以大大降低数据收集的成本，减少损耗。使用行政数据在发达国家更为普遍，但即使在发展中国家，研究人员也可以获得这种数据。例如，Angrist，Bettinger和Kremer (2006)研究了哥伦比亚代金券计划的中期影响，将代金券彩票的数据与哥伦比亚完成学业/大学入学考试的注册数据联系起来。

然而，在这种情况下，重要的是要确保治疗组和比较组之间的数据具有可比性。例如，感兴趣的结果变量可以作为节目的一部分而仅在节目区域中被收集。仅仅通过在比较区域进行新的调查，并依靠项目数据来获得治疗区域的结果变量，来降低数据收集成本可能很有诱惑力。然而，这可能引入偏差，因为治疗和比较区域之间的测量结果差异可能反映不同的数据收集方法。

但要注意的另一个问题是，程序对感兴趣的底层变量的测量的影响可能大于变量本身。考虑这样一种评估，对于这种评估，感兴趣的结果是一些潜在的潜在变量(例如学习)，而这些变量由一些代理(例如测试分数)不完全地测量。在许多情况下，潜在变量和代理之间的关系似乎不受程序的影响。然而，如果程序本身创建了绑定到代理的激励，那么将期望测量

使用另一个代理变量的干预，该代理变量也与潜在变量高度相关，但不与计划的激励相关。例如，Glewwe和Kremer (2003)在根据地区测试分数评估教师奖励方案时，不仅收集了地区测试分数(奖励所依据的)的数据，而且还收集了非政府组织管理的“低风险”测试的数据，该测试提供了独立的学习衡量标准。

偏离完全随机化的分析

现实情况下，随机试验设计肯定面临一些潜在的问题，这些问题主要是偏离了随机化假设，因此，这一部分需要讨论主要影响化非随机化因素，以及如何去消除这些因素。这些因素包括不完全依从性的随机评估、外部性和自然减员。

选择性偏误

完全随机化的第一个偏离是当随机化以可观察变量为条件时，根据可观察变量的值选择不同的概率。在前文4.5节中，我们讨论了使用分层设计来减少估计治疗效果的方差的设计。在所有区块中，分配给治疗组和比较组的观察结果相同。然而，选择的概率在不同的阶层也可能不同。以已经讨论过的哥伦比亚代金券计划为例。随机分组在每个城市进行，每个城市预先确定获奖者的数量。因此，每个城市中彩票中奖者与总申请人的比例是不同的。这意味着彩票状态在整个样本中不是随机的(例如，如果波哥大对于给定数量的位置有更多的申请者，则波哥大的输家可能比卡利的更多)。然而，它在每个城市内部仍然是随机的。换句话说，治疗状态是一组可观察变量(阶层:在这种情况下，是城市)的随机条件。

部分合规

在某些情况下，评估旨在覆盖分配到实验组的所有个人，并非常注意确保合规性接近完美。在第3节介绍中讨论的印度尼西亚铁补充试验就是这种情况，其依从率超过92% (Thomas，Frankenberg，Friedman，Habicht和Al 2003)。然而，在其他许多情况下，人们并不期望法规遵从性是完美的。有时，只有一小部分接受治疗的人会接受这种疗法。相反，比较组的一些成员可以接受治疗。这被称为“部分(或不完全)合规”

外部效应

实验性干预可以产生溢出效应，也就是外部效应，使得未经治疗的个体受到治疗的影响。溢出效应可能是有形的——例如，在对肯尼亚一个小学驱虫方案的评估中发现了大量减少疾病的外部效应(Miguel和Kremer，2004年)。它们也可能是价格变化的结果——Vermeersch和Kremer (2004)发现，在肯尼亚的一些学校为学龄前儿童提供学校膳食，使得附近的学校降低了学费。溢出效应也可以以学习和模仿效应的形式出现(见Duflo和Saez (2003)、Miguel和Kremer (2004))。

外部效应解决方案

在溢出可能很重要的地方，可以专门设计实验来估计其程度和程度。以下有三种方式：

（1）第一种技术是有目的地改变一个群体内的治疗暴露水平。例如，在他们对信息和401(k)参与的研究中，Duflo和Saez (2003)将获得参加信息会议的激励的提议随机化为两个级别。首先随机选择一组大学部门进行治疗，然后随机选择治疗部门中的一组个人进行奖励。这使得作者能够探索获得激励对出勤率和计划注册的直接影响，以及在其他人获得激励的部门的溢出效应。

（2）第二种技术是利用随机化自然产生的各组之间的暴露差异。例如，Duflo，Kremer和Robinson (2006)在随机选择的农民样本中进行了现场农业试验。

（3）第三种估计溢出效应的技术是将个体随机分配到不同的对等群体。例如，美国的“转移到机会”实验(Liebman，Katz和Kling 2004)提供了随机选择的个人代金券，以转移到较低贫困的社区。将收到优惠券的人和没有收到优惠券的人进行比较，可以估算出邻居效应的重要性。

数据损失

随机的数据损失只会降低研究的统计能力，然而与被评估的治疗相关的损耗可能会使估计产生偏差。例如，如果那些从一个项目中获益最少的人倾向于退出样本，忽略这个事实将导致我们高估一个项目的效果。虽然随机化可确保初始治疗组和比较组潜在结果的独立性，但在非随机摩擦后不成立。这个问题发生在2006年的第一次大规模随机评估中美国的负所得税实验，并产生了丰富的计量经济学文献的方法来解决这个问题(豪斯曼和怀斯1979，赫克曼1979)。

解决方法如下。第一步必须始终是报告处理组和比较组中的数据损耗水平，并使用基线数据(如果可用)比较损耗与非损耗，以查看它们是否系统地不同，至少在可观察的维度上不同。如果磨损仍然是一个问题，统计技术可用于识别和调整的偏见。这些技术可以是参数化的(参见Hausman and Wise (1979)、woodridge(2002)或Grasdal (2001))或非参数化的。我们在这里将重点关注非参数技术，因为参数方法更广为人知。此外，非参数样本校正方法对于随机评估是有趣的，因为它们不需要参数方法特有的函数形式和分布假设。讨论非参数界的重要研究包括Manski (1989)和Lee (2002)非参数Manski-Lee界的思想是使用关于潜在结果和损耗的单调性以及潜在结果的分布，以导出可

从可用数据估计的治疗效果的界限。普通治疗效果估计将提供真实效果的上限或下限，具体取决于摩擦偏差的方向。当摩擦偏差为负且治疗效果为正时，普通估计提供真实效果的下限，并且使用Manski-Lee方法估计上限。需要注意的是，界限间距越小，说明数据损耗越低，这也表面需要进一步尽可能地限制数据损失方向。

结论推广问题

这部分讨论了与从随机评估中进行有效推断相关的一些关键问题。我们首先回到分组数据的问题，解决如何计算导致分组结构的标准错误。然后，当研究人员有兴趣评估一个项目对几个(可能相关的)结果变量的影响时，我们考虑这种情况。我们接下来转到评估在人群子群中的异质性治疗效果，最后讨论在估计中控制协变量。

数据分组

当随机化发生在群体水平时，标准误差需要考虑相同群体成员之间结果变量的可能相关性。可以通过广义最小二乘更有效地估计。如果希望避免公共协方差结构的假设，用分组数据计算标准误差的一种方法是使用群集相关Huber-White协方差矩阵估计器。当随机分组的数量足够大时，更适合使用这种方法。然而，Donald和Lang (2001)和woodridge(2004)已经指出，该估计量的渐近证明假设了大量的总单位。当集群的数目小时(小于50)表现较差，导致过度拒绝无效的零假设。当集群的数量很小时，也可以使用ran-domination推断生成假设检验(Rosenbaum 2002)。

随机化推断的优点在于，其对于任何样本大小都是有效的，并且因此即使当样本的数目非常小时也可以使用。Bloom，Bhushan，Clingingsmith，Hung，King，Kremer，Loevinsohn和Schwartz (2006)在他们对柬埔寨公共医疗保健中心分包管理的影响的研究中使用该方法计算标准误差。随机分组在地区层面进行，只有12个地区参与研究。因此，聚集的标准误差可能受到相当强的偏差的影响。然而，要注意的是，当真正的影响较大时，尽管无偏随机推断相对于更多的参数方法而言具有较低的威力，因为它甚至没有给误差项设置最小的结构(参见Bloom，Bhushan，Clingingsmith，Hung，King，Kremer，loevinshnson和Schwartz (2006)中的讨论)。

多重结果

政策评估往往影响许多不同的结果，实验者随后衡量。测试关于多重结果的假设需要特殊的技术。标准的假设测试假设实验者对每个结果分别感兴趣。但当测试多个结果时，对于至少一个结果拒绝真实零假设的概率大于用于每个测试的显著性水平(Kling和Liebman 2004)；在5%水平上测试十个独立假设的研究者将以大约40%的概率拒绝其中的至少一个。解决办法是，无论其是否重要，都应报告所有情况。在存在大量变量的情况下，提前列出哪些变量属于哪些族以进行族测试对于研究者也是有用的。

子样本分析

干预措施往往对其所影响的人口产生不同的影响。例如，我们可能期望补救教育计划对考试分数低的学生比对考试分数高的学生有更大的影响。如果我们在包含这两种类型的学生的组(例如教室)之间随机分配干预，那么治疗的效果将是对得分低和得分高的儿童的平均效果。然而，研究人员和决策者可能有兴趣对高分和低分的孩子分别测试这种效果。

报告事后分组的结果以及最初计划的结果是非常必要的，因为它们可以对第一组结果提供额外的说明。然而，如果报告了按事后分组的结果，研究人员的报告必须非常清楚地说明哪些分组是事前定义的，哪些分组是事后定义的。

协变量问题

分析实验时的另一个选择因素是选择控制什么。正如我们上面所讨论的，控制变量可以减少结果的方差，导致更精确的估计。但同样，这些变量应该事先指定。通常的做法是报告"原始"差异以及经过回归调整的结果。

外部有效性和随机评估的推广

到目前为止，我们主要关注的是内部有效性的问题，即我们是否可以得出结论，测量的影响确实是由样本中的处理措施引起的。在本节中，我们讨论外部有效性，即我们测量的影响是否会延续到其他样本或人群中。换句话说，结果是否具有普遍性和可复制性。虽然内部有效性是外部有效性的必要条件，但它并不充分。在围绕使用随机评估的讨论中，这个问题已经得到了很多的关注。Bardhan, Basu, Mookherjee和Banerjee在"新发展经济学"研讨会上的论文(Banerjee 2005, Basu 2005, Mookherjee 2005, Bardhan 2005, Banerjee, Bardhan, Basu, Kanbur和Mookherjee 2005)对这场辩论进行了一个非常有趣的概述。在本节中，我们将讨论为什么人们会担心随机评估的外部有效性，以及有哪些方法可以改善这些担忧。

部分均衡和一般均衡效应

因为随机评估比较的是特定区域内处理人群和比较人群之间的差异，所以它们无法得出一般均衡效应(Heckman, Lochner, and Taber, 1998)。这种影响对于评估扩大项目对福利的影响可能特别重要。例如，在对哥伦比亚的教育券计划进行评估时，研究人员比较了那些接受教育券进入私立学校的学生和那些申请但没有接受教育券的学生的结果。这个评估能够确定在有一个教育券的情况下赢得教育券的影响——换句话说，它衡量了项目对接受者的局部或局部影响。但是，它不能说明引进教育券制度对哥伦比亚教育制度的总体影响。因为日益激烈的竞争可以提高公立学校的教育质量(有更多的比较儿童在公立学校接受教育)，从而减少处理和比较结果之间的差异。即使教育券对系统有积极的影响，但也会减少教育券的测量效果。公立学校由于失去了最忠诚的学生和家长而导致的成绩下降，会表现为处理和比较的学生成绩之间的更大差距。换句话说，教育券的负面影响越大，它的结果就越好。这种一般均衡效应可以被认为是另一种外部性：如果观测的单位足够大，就有可能发现某些一般均衡效应——尽管这并不总是实际的。例如，可以通过在社区层面随机分配教育券来分析教育券对学校之间竞争、对学校分类以及对留在公立学校的儿童的影响(假设一个社区足够大，可以容纳几所学校，一些公立学校和一些私立学校)。如果大多数孩子留在这个社区上学，将(随机地)引入教育券的社区与不引入教育券的社区进行比较，我们就能了解这些一般均衡效应。在其他情况下，一般均衡效应可能在国家乃至世界层面发挥作用(例如，如果它们影响工资或价格)。因为这涉及到国家甚至国际层面的随机评估，所以很难对这些数据进行随机评估。

霍桑和约翰-亨利的影响

前瞻性评估的另一个局限性是评估本身可能会导致实验组或对照组改变其行为。实验组的行为变化被称为霍桑效应，而对照组的行为变化被称为约翰·亨利效应。实验组可能会感激接受处理，并意识到被观察，这可能会诱使他们在实验期间改变自己的行为(例如，更加努力地工作以取得成功)。对照组可能会因为自己是对照组而感到被冒犯，也会改变自己的行为(例如，评价对照组的人可能会与处理组的人“竞争”，或者相反，决定偷懒)。解决这种问题的一种方法是收集长期数据。例如，Duflo和Hanna(2006)在官方“实验”结束后的一年多时间里继续监测项目的影响(但非政府组织决定继续将该项目作为一个永久性项目实施)。事实上，当项目不再被正式评估时和在评估期开始时的结果是相似的，这表明最初的结果在存在的地方不是由于霍桑效应。

超越特定程序和样本的概括性分析

更为普遍的是，在随机评估中经常出现的一个问题是，评估结果在多大程度上可以复制或推广到其他环境中。有三个主要因素会影响随机评估结果的普遍性：项目的实施方式（项目的实施方式是否特别谨慎，使得复制它们非常困难），评价是在特定的样本中进行的，以及特定项目的实施情况（稍微不同的项目会有相同的结果吗？）。

田野实验和理论模型

虽然有必要在不同的环境下重复研究，但要严格检验研究结果在所有可能情况下的成立程度，是永远不可行的。然而，当实验与经济理论或模型相结合时，它们可以提供更多的一般性经验。将理论和结构与随机评估相结合有两种主要方式。首先，经济模型可以与来自随机评估的变化相结合，以估计更丰富的参数集。其代价是一组额外的假设，但好处是有一组更丰富的参数，可以用来预测项目的变化会如何影响行为。在发展经济学中，理论和随机化的一个更有野心的应用是建立实验，明确地检验有关经济行为的特定理论。Karlan and Zinman (2005c, 2005a)和Bertrand, Karlan, Mullainathan, Shafir, and Zinman(2005)是三个相关项目，提供了利用现场实验检验理论的优秀例子。

应用实例

实例1 Exposing Corrupt Politicians: The Effects of Brazil's Publicly Released Audits on Electoral Outcomes

主要内容

本文使用公开发布的审计报告来研究披露腐败行为信息对选举问责制的影响。

2003 年，作为反腐败计划的一部分，巴西联邦政府开始随机选择市镇来审计其联邦转移资金的支出。这些审计的结果随后被公开并传播给媒体。

使用根据审计报告构建的腐败数据集，我们比较了 2004 年选举之前和之后审计的市政当局的选举结果，审计前后的腐败程度相同。

结果表明，审计结果的发布对现任者的选举表现产生了重大影响，并且这些影响在有地方电台存在因而可以向大众透露信息的城市更为明显。

研究结果强调了知情程度更高的选民的价值以及当地媒体在规范政治选举方面所发挥的作用。

主要方法

它通过使用一种实验设计克服了以前的数据限制和识别问题：该实验设计在将政治家的腐败信息向公众披露的相关选择上产生外生变化。

我们的研究设计利用了两点： 审计的随机时间选择和审计的公众传播。

该分析比较了 2004 年市政选举前后经过审计的市政当局之间有资格连任的市长的选举结果。

1.实验设计的两个潜在威胁

威胁之一：

即使待审计的市政当局是随机选择的，如果实际的审计过程在选举前和选举后有系统性的不同，设计也会受到影响。

然而，我们没有发现任何证据表明审计员腐败或在选举前接受审计的市政当局受到差别对待。

我们还表明，拥有更多政治权力的市长、隶属于更高级别政府的市长以及获得较大竞选捐款的市长没有获得优先审计。

威胁之二：

尽管审计时间的变化是外生的，但市政当局的腐败程度或当地媒体的可用性并非如此。

因此，我们对于腐败程度和当地媒体可用性的衡量可能会受到市政当局一些其他特征的影响。

我们提供证据表明情况并非如此。即使允许审计的效应随着腐败和当地广播电台作用方式的不同而不同（例如政治竞争、教育、人口规模、城市化和其他媒体来源），估计结果也保持不变。

2.数据与主要模型方法

截至 2005 年 7 月，在前 13 次抽签中随机选择的 669 个市镇的审计报告已经公布。

为了估计审计对连任机会的影响，我们必须将样本限制在第一任市长（有资格连任）的集合中。 这将我们的估计样本减少到只有 373 个城市。

其中，在 2004 年 10 月市政选举之前，审计并发布了关于随机选择的 205 个城市的腐败程度的信息。在市政选举之后，审计并发布 168 个市政当局的审计报告 。

模型的主要方法是在挑选城市对其市长进行随机审计的条件下，研究对该市市长的审计是否在选举前进行会对该市长的连任结果造成何种程度的影响。

实验仍存在的问题

尽管我们在确定信息发布对腐败的影响时，是基于对市政当局的随机审计。但遗憾的是，审计实验并未根据当地媒体的可用性进行随机化。因此，我们对媒体可用性的衡量可以视为市政当局其他特征的代理变量。其作用机制是这些市政特征会导致审计报告对政治家的连任结果产生不同的影响。

一种可能性是，对电台可用性的衡量只是捕捉了具有不同教育水平的城市之间的审计效果。如果受教育程度较高的公民更了解政治家的腐败活动，那么在公民受教育程度较高的城市中，审计的影响可能较小。或者，受教育程度更高的公民也可能更多地参与政治并愿意对腐败的政客采取行动，在这种情况下，公民受教育程度更高的城市的影响可能更明显（Glaeser 和 Saks 2006）。

另一个可能性是人口规模。如果有关政治家违规行为的信息在大城市流动得更好，那么审计的影响可能会更小。

最后，随着选民在经济上变得更加多样化，选举选择可能基于重新分配而不是政治家的诚实（Alesina、Baqir 和 Easterly 2002；Glaeser 和 Saks 2006）。因此，在收入不平等程度较高的城市，审计的影响可能较小。

为了测试这些潜在的混淆，我们在后续估计中包含了一系列交互项，以允许选举前审计指标和腐败违规数量随着指征其他市政特征的代理变量在城市之间变化。

这些指征其他市政特征的代理变量包括：人口密度、识字率、城市人口比例、人均收入和基尼系数。实证结果显示，对腐败程度和当地电台可用性对审计产生影响的估计结果仍然是显著的，而且审计效应的幅度仍然相似。

结论

公开发布的审计报告为选民提供了有关市长腐败行为的新信息，选民使用这些信息来更新他们的先验信息并惩罚被发现比平均水平更腐败的政客。

在当地媒体可以更广泛地传播这些调查结果的地区，审计效应更加明显。

实例2 Woman As Policy Makers: Evidence From A Randomized Policy Experiment In India

1992年，印度的第73次修正案规定，所有村委员会的三分之一委员席位以及三分之一的村长职位必须为女性保留。那么，如果女性和男性在提供哪些公共产品上有不同的偏好，为女性保留职位的制度就会对所做的决定产生实际影响。本文的研究目的就是女性政治保留政策是否会对政策的决策产生影响?

如何实施呢？如果在某些情况下，政策结果将更接近于女性想要的而不是男性想要的，就能够说明女性政治保留政策会对政策产生影响。通过对印度的西孟加拉邦（323个村级观测值）+拉贾斯坦邦（ 100个村级观测值）两个地区进行调研（由于名字太长了，我简称A地区和B地区），获得村民对一些关于不同类型公共物品（饮用水、道路、灌溉、学校等）的投诉或者是反映。另外，还调研了样本地区的公共产品的投资情况，包括饮用水、道路、灌溉、学校等。通过分析村民投诉情况发现， 在A地区和在B地区，女性最常提出的问题是饮用水问题，男性最常提出的问题是道路问题。通过分析公共产品的投资情况发现，为女性保留政治政策区域的村委会对饮用水的投资很多，说明村长的性别很有可能影响着公共产品的投资，下面通过建模分析检验：在有女性政治保留政策的村委会中，对女性更经常提到的物品是否有更多投资。

将Y_{ij}表示为第i种物品（例如，在1998年和2000年之间的饮用水投资）的标准化结果值，R_j表示为一个虚拟数，如果村委会有女性政治保留政策，则$R_j=1$。d_{il}是请求第i个物品的虚拟数，N是请求总数量，D_i是该地区来自女性和男性的关于商品i的请求比例（偏好）的平均差异。S_i是该地区的男性和女性请求商品i的平均比例（平均偏好强度）。D_{ij}是j村的女性和男性是否提到i物品的指标之差；S_{ij}是j村的女性和男性是否提到i物品的指标之和。

$$\begin{aligned} Y_{ij} =& \beta_7 + \beta_8 R_j + \beta_9 D_i R_j + \beta_{10} D_{ij} R_j + \beta_{11} S_{ij} R_j + \beta_{12} S_{ij} \\ &+ \beta_{13} D_{ij} + \sum_{l=1}^{N} \beta l d_{il} + \epsilon_{ij} \end{aligned} \tag{1}$$

衡量女性和男性对某一特定物品i的偏好差异的强度：$D_i = \frac{n_i^w}{N^w} - \frac{n_i^m}{N^m}$

衡量总人口中对该物品i的偏好强度（假设男女比例相同）：$S_i = \frac{1}{2}(\frac{n_i^w}{N^w} + \frac{n_i^m}{N^m})$

此外，$D_{ij} = \frac{n_{ij}^w}{N_j^w} - \frac{n_{ij}^m}{N_j^m}$, $D_i = \frac{1}{N}\sum_{j=1}^{N} D_{ij}$

结论：在这两个地区，公共物品的提供都更符合女性的偏好，而不是男性的偏好（D_i的系数$\beta_9 > 0$）。那自然会提出一个问题：是不是由于女性村长主观更偏向于女性投诉的公共物品呢，根据$D_{ij}R_j$的系数显示：女性对女性和男性的需求并没有显著的差别反应。在女性政治保留的村委会中，对女性更经常投诉的商品没有更多的投资（$D_{ij}R_j$的系数不显著）。

实例3 Monitoring corruption- evidence from a field experiment in Indonesia

作者：Benjamin A. Olken

期刊: Journal of Political Economy , Vol. 115, No. 2 (April 2007), pp. 200-249

腐败是发展中国家普遍存在的严重问题。在这些国家中，腐败类似一种税收的存在，不仅增加了公共服务和经营业务的成本，而且有人认为腐败可能是多个发展中国家经济低增长率的主要原因(Mauro, 1995)。对腐败的直接度量相当困难，而对如何减少腐败，有一种办法是正确的监管与处罚制度，另外还有文献认为增加基层监督也可以治理腐败。

本文在印度尼西亚608个村庄设计并进行了一项包含实验组、和对照组的随机实验，测度了由政府审计员自上而下的监控、和通过基层村庄参与监控对腐败的影响。

背景与数据

实验是在"Kecamatan分区发展项目"(简称KDP)的基础下进行的，KDP是由世界银行贷款资助的，印度尼西亚政府的一个全国性项目。KDP每年资助印度尼西亚全国约15,000个村庄的项目。村庄最常见的基础设施项目是将现有的土路铺设成由沙子、岩石和砾石构成的道路。这些道路在村庄里、或者从村庄延伸到田野，道路的长度从0.5公里到3公里不等。

为确保项目资金被合理使用，项目有若干机制。主要的机制是一系列的村级责任会议。资金分别以40%、40%和20%的比例分三批发放给实施团队。为了获得第二和第三批资金，执行小组必须向村民公开会议提交一份责任报告，说明所有资金的使用情况。只有在这次会议通过了问责报告之后，才会发放下一批资金。

本文的数据来自印度尼西亚人口最多的两个省，东爪哇省和中爪哇省的608个村庄的KDP项目，收集于2003年9月至2004年8月。

实验设计与模型

TABLE 1
NUMBER OF VILLAGES IN EACH TREATMENT CATEGORY

	Control	Invitations	Invitations Plus Comment Forms	Total
Control	114	105	106	325
Audit	93	94	96	283
Total	207	199	202	608

将608个村庄随机分成2*3共六个小组，其中，每个小区都有48%的机会被随机纳入审计处理。每个村庄有33%的机会被随机分配到邀请处理、33%的机会被随机分配到邀请和评论表格处理。审计的随机化与邀请或邀请加评论形式的随机化是独立的。

为了检验分组的随机性，文章通过$Probit$回归报告了10种村庄特征随机进入每个实验组的概率，最后发现相当部分的特征分配都是随机的；仅有两个不太显著的特征在后面也进行了专门的讨论，不影响最后的结果。

把村庄在建设项目中上报的金额叫报告金额；把根据实地调查，估算出的该村庄在建设项目中实际支出的金额叫做实际金额。定义报告金额与实际金额差异的百分比为不明支出百分比($PercentMissing$)。则不明支出的百分比就是文章测度出的腐败水平。

$$PercentMissing_{ijk} = \alpha_1 + \alpha_2 Audit_{jk} + \alpha_3 Invitations_{ijk} + \alpha_4 InvitationsandComments_{ijk} + \epsilon_{ijk}$$

这里i代表村庄，j代表街道，k代表审计的划分。由于随机化处理得比较好，模型可以直接使用OLS进行估计。整体的结果如下表所示。

TABLE 4
AUDITS: MAIN THEFT RESULTS

			No Fixed Effects		Engineer Fixed Effects		Stratum Fixed Effects	
Percent Missing[a]	Control Mean (1)	Treatment Mean: Audits (2)	Audit Effect (3)	p-Value (4)	Audit Effect (5)	p-Value (6)	Audit Effect (7)	p-Value (8)
Major items in roads (N = 477)	.277	.192	−.085*	.058	−.076**	.039	−.048	.123
	(.033)	(.029)	(.044)		(.036)		(.031)	
Major items in roads and ancillary projects (N = 538)	.291	.199	−.091**	.034	−.086**	.022	−.090***	.008
	(.030)	(.030)	(.043)		(.037)		(.034)	
Breakdown of roads:								
Materials	.240	.162	−.078	.143	−.063	.136	−.034	.372
	(.038)	(.036)	(.053)		(.042)		(.037)	
Unskilled labor	.312	.231	−.077	.477	−.090	.304	−.041	.567
	(.080)	(.072)	(.108)		(.087)		(.072)	

结果显示，审计对额外支出百分比有重大的、统计上重大的负面影响。第3栏显示，审计使得道路项目的额外支出百分比降低了8.5个百分点，将道路及附属项目的额外支出百分比降低了9.1个百分点。在施工队固定效应与区域固定效应中，也同样出现了类似的结果。

文章除了对整体效果进行建模，还对若干其他控制变量进行了检验，也得到了了一些有启发性的结果。这些检验包括检查审计师的发现、亲属的雇佣、民众监督实验、精英对会议的掌控程度与参与程度等方面。

考虑亲属雇佣方面，基于调查数据考察下列模型：

$$Worked_{hijk} = \gamma_k + \gamma_2 Audit_{jk} + \gamma_3 Family_{hijk} + \gamma_4 Audit \times Family_{ijk} + \epsilon_{hijk}$$

TABLE 8
NEPOTISM

	(1)	(2)	(3)	(4)
Audit	−.011	.004	−.017	−.038
	(.023)	(.021)	(.032)	(.032)
Village government family member	−.020	.016	.016	−.014
	(.024)	(.017)	(.017)	(.023)
Project head family member	.051	−.015	.051	−.004
	(.032)	(.047)	(.032)	(.047)
Social activities	.017***	.017***	.013*	.014**
	(.006)	(.006)	(.006)	(.006)
Audit × village government family member	.079**			.064*
	(.034)			(.034)
Audit × project head family member		.138**		.115*
		(.060)		(.061)
Audit × social activities			.010	.008
			(.008)	(.008)
Stratum fixed effects	Yes	Yes	Yes	Yes
Observations	3,386	3,386	3,386	3,386
R^2	.26	.26	.26	.27
Mean dependent variable	.30	.30	.30	.30

结果见表8：相对于未被审计的村庄，项目负责人的家庭成员在受审计村庄参与项目的可能性比在未受审计村庄高出13.8个百分点。

对于劳动力腐败方面，对比了对照组、邀请参会、邀请参会并评论三个小组，发现基层监控显著减少了14到22个百分比的劳动力支出。但是，在总体上，基层监控对控制腐败缺乏强有力的影响。一个可能的原因是材料方面的腐败符合整个村庄的利益：虽然基层监控显著减少了劳动力方面的腐败，但由于熟练劳动力支出只占总支出的20%、材料成本占道路成本的68%，所以最后的结果在整体上是不显著的。

结论

文章发现，第一，增加外部审计的可能性可大大减少项目中资金的缺失。特别是，如果把监管的抽查率从基准的4%提高到100%，不明支出比率从27.7%降低到19.2%。第二，增加基层监控仅能在有限的情况下减少腐败。结果显示，加强民众监督只减少了劳动力腐败，对材料腐败没有影响，因此，总体影响不大；只有通过村里的学校发放匿名评议表，才能避免精英阶层对会议的操控，继而使民众监督起作用。

实例4 What Do Laboratory Experiments Measuring Social Preferences Reveal About the Real World?

测算社会偏好的实验室实验能为现实世界解释什么？

作者：

Steven D. Levitt(史蒂芬·列维特)，John A. List(约翰·李斯特)

文章目的：

实验经济学真的有用吗?

经典实验经济学案例：

1.最后的通牒博弈实验（Ultimatum game）

2.独裁者博弈实验（Dictator game）

3.信任博弈实验（Trust game）

4.礼物交换博弈实验（Gift exchange game）

5.公共产品博弈实验（Public goods game）

发现的问题：

以上经典实验案例可以发现，实验结果与理论推导结果(纳什均衡点）不符。那这种实验结果可信吗，会有哪些因素影响实验结果呢?

影响实验结果的因素：

1.道德风险

2.实验是否有人监督

3.实验背景

4.自选择问题（选择性偏误）

5.赌注效应（风险厌恶）

结论：

1.实验结果推广到现实世界需要考虑很多因素

2.实验结论和自然实验一样，需要根据一定程度理论基础设计实验

展望：

1.实验室实验结果也不并不可靠，但依据实验设计可以从中提取有价值信息

2.认识到实验设计的不足，改变实验类型、环境，从而分析不同实验类型环境下结果变化（敏感性）

3.结合实验室实验和自然实验

参考文献：

Harrison, G. W., & List, J. A. (2004). Field experiments. Journal of Economic Literature, 42(4), 1009-1055.

List, J. A., & Levitt, S. D. (2005). What do laboratory experiments tell us about the real world. NBER working paper, 14-20.

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

Rosenbaum, P. R. (2002). Covariance adjustment in randomized experiments and observational studies. Statistical Science, 17(3), 286-327.

Olken, B. A. (2007). Monitoring corruption: evidence from a field experiment in Indonesia. Journal of political Economy, 115(2), 200-249.

Chattopadhyay, R., & Duflo, E. (2004). Women as policy makers: Evidence from a randomized policy experiment in India. Econometrica, 72(5), 1409-1443.

DellaVigna, S., & Kaplan, E. (2007). The Fox News effect: Media bias and voting. The Quarterly Journal of Economics, 122(3), 1187-1234.

Ferraz, C., & Finan, F. (2008). Exposing corrupt politicians: the effects of Brazil's publicly released audits on electoral outcomes. The Quarterly journal of economics, 123(2), 703-745.