

影响评估国际动议组织



工作文稿 3

基于理论的影响评估：

原理与实践

霍华德·怀特

2009年6月

关于3ie

影响评估国际动议组织（简称3ie）旨在通过支持对客观事实的主体、时间、原因及影响程度的分析和使用来改善发展中国家的人民生活水平。**3ie**是一个新的动议组织，可以满足对客观事实的更高要求，并通过制定更具针对性的政策来提升发展的有效性。**3ie**资助高质量的影响评估行动，以向发展中国家建议更好的政策和项目设计。

3ie工作文稿既涉及影响评估的有关概念性问题，又包含了个案研究和情景分析。

本篇工作文稿由**3ie**执行总裁霍华德·怀特博士创作。版权所有**3ie**，2009年。

联系方式

影响评估国际动议组织

转全球发展网络

邮编7510

瓦桑昆

印度新德里110070

电话：+91-11-2613-9494/6885

网址：www.3ieimpact.org

基于理论的影响评估：原理与实践

霍华德·怀特

执行总裁

影响评估国际动议组织，3ie

邮箱：hwhite@3ieimpact.org

摘要

对严谨的影响评估的追求一直以来都伴随着人们对于“为什么”的探求，而不仅仅限于“是什么”的问题。基于理论的影响评估方法已经被人们广泛接受，该方法通过刻画投入到结果到影响的因果关系，并对潜在假设进行验证来解释“为什么”的问题。然而基于理论的影响评估方法并未得到广泛的应用。本文提出了以下6个原则以指导实践应用：（1）分析因果关系（程序理论）；（2）环境认知；（3）预知异质性；（4）使用可靠的反事实进行严谨的影响评估；（5）严谨的事实分析；（6）使用混合法。

1. 引言

近年来越来越多的人对使用定量方法来评估发展项目的影响表现出浓厚的兴趣。例如贫困行动实验室（J-PAL）和贫困行动创新（IPA）组织等的研究计划¹，世界银行发展影响评估项目（DIME）与西班牙影响评估基金（SIEF）²资助的系列研究，以及影响评估国际动议组织（3ie）³对相关研究的资助都显示，在未来5年内，成百上千的此类研究将出炉，与当前的聊聊数十篇形成鲜明对比（以2006年的《全球发展中心》为例）。然而，多数支持开展科学的影响评估的人们，其反复强调的一点就是不但要找出哪些是有效的，还要解释为什么有效。要做到这点不是仅凭简单的报告平均处理效应就能解决的。因而影响评估网络中的网络（NONIE）指出：“基于理论的方法应用意味着一项设计合理的影响评估要涵盖过程评估与影响评估两方面。如此，这项研究在强调干预为什么成功——或为什么不成功——时，政策的相关性也能因此得到增强。”（NONIE，日期不详）。同样，影响评估国际动议组织（3ie，日期不详）在其影响评估实践指南中也提出：“研究要清楚的说明干预措施（投入）将如何影响最终结果，而且要对投入到最终结果之间的每项链接（假设）进行检验（此法有时也被称为程序理论）。评估设计应当包含从投入到影响之间整个因果关系链的分析。”

我们在此提倡的理解一个项目为什么产生或没有产生影响的方法被称之为基于理论的影响评估（TBIE）。此方法并非创新之举。基于理论的评估已是一种很成熟的方法，旨在检验反映投入与最终结果和影响间的因果关系的假设，（见1998年韦斯和2004年卡沃尔与怀特在发展中国家中的应用）。长期以来，一些实验方法或准实验方法的实践者利用程序理论来解释其研究成果（2009年布莱克曼与莱克著，第67-68页）。罗杰斯（2009）在回顾了针对各类发展措施的可能的影响评估方案后，指出基于理论的方法适用于任何案例。

尽管研究者对基于理论的影响评估已经基本达成共识，但实践当中似乎很少有研究严格照此执行。本文旨在通过列出基于理论的评估背后的步骤或是原则来缩短理论与实践的差距。本文第2部分将以孟加拉国综合营养计划（BINP）开始；并在3部分以此为例，并结合一些其他的案例阐述基于理论的影响评估的基本原则；本文第4部分将简要比较基于理论的影响评估（TBIE）与黑盒子方法的异同；第5部分为结语。

¹ 详见 www.povertyactionlab.org 与 <http://poverty-action.org>

² 详见 www.worldbank.org/dime 与 www.worldbank.org/sief

³ 详见 www.3ieimpact.org

2. 案例-孟加拉国综合营养计划

本节将简要回顾对孟加拉国综合营养计划（BINP）的评估研究，进而以此为例说明下节要讨论的基于理论的影响评估的原则。有关此计划的更多讨论见2005年世界银行，2005年怀特以及2006年怀特与马赛的文章。

孟加拉国综合营养计划效仿印度泰米尔纳德综合营养计划，是一个幼儿发育成长监测项目。婴儿每周在当地的体重监测站称重。监测站的员工是一位接受过社区营养师培训的当地妇女。根据体重、年龄两个参数编绘制体重-年龄图。生长缓慢的（发育不良）或远远低于正常标准（营养不良）的儿童可以参加该计划。该计划包含营养咨询和补食两部分。然而，项目文件里明确说明预期影响主要通过咨询实现。理由是营养不良应归咎于无知而非贫困。即便在最富裕的人群里，也存在营养不良的现象；以及妇女怀孕期间应当节食的错误信念都证明了此观点。综合营养计划同时也为孕妇提供咨询与补食。孟加拉国综合营养计划作为一个试点项目，后由国民营养计划（NNP）取代。

孟加拉国综合营养计划执行之初就成绩斐然。监测数据显示，在项目执行地区营养不良，尤其是严重营养不良的现象大大减少。为此，世界银行在未经任何评估之前，就决定在孟加拉国综合营养计划实施中期将其扩展为国民营养计划。英国儿童救助会发表报告质疑此决定，他们基于由一个简单的处理-控制事后对比设计收集的数据，发现两个地区间没有区别（儿童救助会，2003年）。

由世行业务评估部（OED，现在的独立评估集团，IEG）所做的分析使用了倾向得分匹配的方法，将从项目实施地区收集的数据与海伦凯乐国际的一项国家营养调查的数据合并构建了控制组。该分析发现，尽管该项目对那些极度营养不良的儿童有正面影响，但是对所有儿童的营养状况并没产生显著的影响。

沿着因果关系链，非常多的假说阐述了为何预期孟加拉国综合营养计划对儿童营养状况产生正面影响，图1列举了其中的一部分。

首要问题是人们究竟是否了解该计划并参与其中——很多发展计划在一开始就遭遇失败的原因就在于没有付出足够的努力向潜在受益者说明他们的干预计划，或为受益者的相对成本与收益做一个切合实际的评估。尽管一部分例外情况，但孟加拉国综合营养计划在这方面做得还不错，大概有90%的符合条件的妇女都带着小孩一起参加计划，见图1。

其次，要保证瞄准的人群正是该项目的目标人群。该计划的目标群体是年轻妈妈。然而在孩

子的健康和营养问题上，妈妈们通常做不了主，更谈不上自己一个人说了算。首先，在孟加拉国乡村地区妇女们并不去市场，购物是男人的事。对于生活在联合家庭里的妇女，就像在许多少数民族家庭那样，妈妈们与婆婆一起生活，婆婆管理家中女眷事务。事实上，在该国一些更为保守的地区参加计划的妇女比例明显低很多。

一旦妈妈带着小孩去称重，那些符合条件——发育迟缓或营养不良的小孩必须入选该项目。但资料显示，在这一过程中存在大量的瞄准不当第一类错误是符合条件的没有入选，第二类错误是入选的不符合条件。我们使用一些儿童成长图（培训时使用过的）考核社区营养师，结果发现他们大都无法根据成长图判断哪些孩子符合入选该项目的条件，从而导致瞄准不当。目标群体的瞄准对该项目的影响至关重要，就如我们发现的，最营养不良的孩子的确受益了，那么如果该项目只是关注这类儿童，其平均影响将会更大，而实际上资源都用在了不会受益的那些孩子身上。

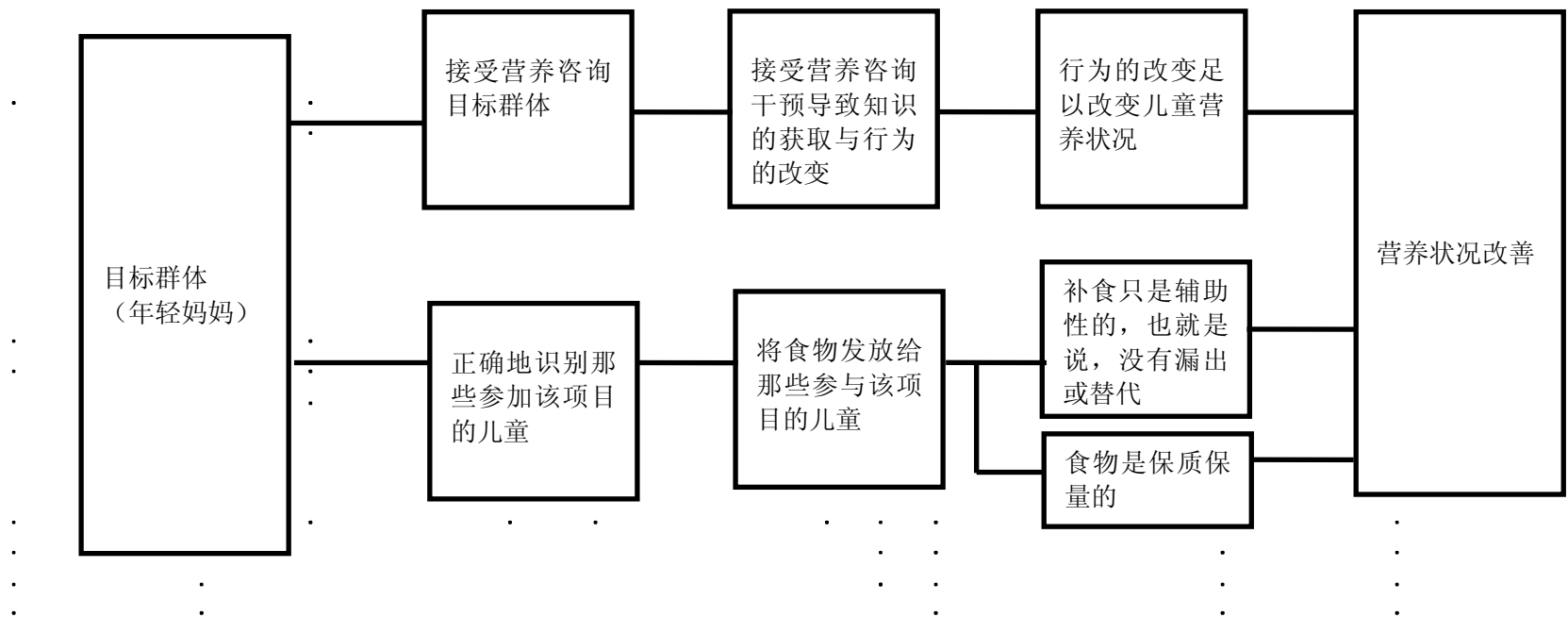
此外，补食要收到好的效果就要做到真正的补充进食，而实际上既存在漏出现象（食物给了非目标人群，提供给孕妇的食物尤其如此）也存在替代现象（目标个体接受补食后，家里减少了原本应该分配给他/她的食物）。

再来看行为的变化。行为教育达到了交流的目的，但却没有带来预期的行为改变。参加计划的妇女学到了很多好的做法，但理论与实践还存在较大差距：很多妇女没有学以致用。部分原因是资源贫乏：贫困家庭的妇女怀孕期间不大可能吃的更多，那些有田地或是与年龄较大的男性亲属一起居住的妇女在怀孕期间也不可能多休息。但其实还是婆婆的问题。一个焦点访谈小组的成员就明白的告诉研究人员说：“如果婆婆同意，我们就按你们说的做，否则还是老办法。”最后，一些旨在增加怀孕期间体重的行为变化不可能对新生儿体重偏小有多大帮助（因为这基本是由妈妈们怀孕前的体重决定的）。

总之，由于脆弱和缺失的因果关系，项目的影响大打折扣。总体上看，可以说该项目并未产生什么影响。其监测数据显示的进步实际上是全国范围内都有的，这就是为什么儿童救助会发现项目地区与其它地区并无差别。实际上，营养状况的改善是粮食增收、收入增加和粮价下降带来的，并非综合营养计划的功劳。

不过，该分析还是指出了改进计划的几条途径：（1）营养咨询要同时面向婆婆和老公；（2）进一步精确定位计划的目标群体；（3）通过加强对社区营养师的培训或是有选择的录用营养师，提高目标瞄准的准确性。然而，这么算下来该计划成本会很高，受管理和资源方面的制约，不可能大规模推广。

图1——营养计划因果关系链：营养咨询与补食



可惜以上评估方面的教训并未得到重视。世行的营养团队很热衷于印度泰米尔纳德综合营养计划与孟加拉国综合营养计划模式。在印度泰米尔纳德综合营养计划中，这种模式被认取得了成功（尽管从未按现今严格的标准考证过）；虽然有些争议，世行正在宣称孟加拉国综合营养计划也是成功的。作为对这些争论的回应，世行随后发表的一份文件毫无保留地声称印度泰米尔纳德综合营养计划是成功的（世界银行，2006年）。尽管评估结果显示可能适用印度泰米尔纳德综合营养计划的模式要在孟加拉国推广还需要做些修改，但在此思想指导下，世行决定广泛推广基于孟加拉国综合营养计划模式的国民营养计划。3年后，国民营养计划失败并提前终结，鉴于执行力度不到位，影响显然难以显现，原来计划的影响评估也就搁置了。世行总结报告最后建议该计划要参照两年前世行业务评估部提出的方案大举改革。

3. 基本原则

基于理论的影响评估的6大原则：

1. 分析因果关系（程序理论）；
2. 环境认知；
3. 预知异质性；
4. 基于可行的反事实进行严谨的影响评估；
5. 严谨的事实分析；
6. 混合使用定量和定性的方法进行检验。

分析因果关系（程序理论）

因果关系链是连接输入与输出及影响的纽带。也就是说，因果关系体现了程序理论（也称变化理论），它说明了一项干预行动将如何达成其预期影响。程序理论源于传统的日志法，不同的是日志法可能没有对潜在的假设做出明晰的阐述，而基于理论的方法其核心就是对假设的检验。

线性分析是因果法饱受争议之处，即单向或决定论的方法，有关线性一词在评估论述中的不同含义见怀特（2009）的讨论。但此种质疑是不正确的。虽然项目执行者经常设想通过一个简单的框架将投入和行动、产出、结果以及影响联系起来，而基于理论的评估检验的是此中的因果关系。其中的一个假设的联系便是观测到的结果的确是由项目活动与产出导致的，而不是反之亦然。但这种逆向或双向的因果关系是影响评估争议的核心：项目实施的选择偏误和项目参与者的自选择都意味着产出变量将影响谁会参与项目，而不是相反的情况。例如，社会资本相对丰富的社区更可能去申请社区发展项目资金。这些项目本来是意在增加社会资本的，但如果简单地比较项目处理组与控制组在项目实施前后的差异很可能只是反映了项目实施前的地区差异，并非项目的影响。

还有一种比较合理的质疑是因果法偏于静态分析，而项目通常存在变通和发展的情况。项目文件里描述的系统可能与实际执行脱节，可能是因为设计方案重新修改，也可能是一线的操作人员随意解读项目程序。前者，程序理论应当反映新的设计方案，评估要记录下形成此方案的习得过程。后者，计划实施与实际执行之间的差异便是评估的关键问题：为什么会出现这些差异，对项目会产生怎样的影响？

之前提到的社会发展资金研究便是项目习得的一个例子。关于社会基金资助的投资，还有人质疑它的可持续性，因为它没有预留运营和维护（O&M）的资金。起初，社会基金通过一个中央委员会审批所有的申请，它默认各部委在该委员会的代表在同一实施一个项目的同时将自己解决相应的运营费用。但这种体制行不通，因此社会基金改为一揽子协议，囊括各部委的全部项目，但仍然有缺陷，所以一些社会基金开始逐项签订行业协议，另一些基金则要求当地提供可持续发展计划，还有一些预留了维护资金（见世界银行，2002年）。

程序理论允许实践中的习得，应当讲它是动态的，重申了理论、数据间的联系。基于模型的统计分析法把模型作为假设，单纯验证数据如何的符合它的模型，而研究人员有各种方法保证数据是符合模型的，正如科斯所说，只要你功夫下到，数据终究会投降（里摩尔引，1983年）。然而数据分析法让数据说话，从数据中找寻模式。此方法听起来没有章法，当然统计行为都不可能撇开理论，因为理论首先就指明了去收集和分析哪些数据。而理论也要能够解释数据揭示意外情形。可能听起来有点像数据挖掘，但其实有本质区别。数据挖掘者清楚自己要找的是什么，找不到不罢休。而数据分析员通过观察数据，发现预期或未知的模式（详见默克尔吉等第一章，1998年）。

另一种对因果法可能合理的质疑是，其太多关注因果分析，会忽视某些预期外的影响。有两种方法可以弥补这一缺陷。首先，认真执行程序理论的话可以发现潜在的预期外的结果，比如环境方面的影响，可能程序设计者已经大致思考过了；其次，前期的实地调查工作，包括参与者分析，是评估设计的重要环节，可以发现预期外的结果，进而将其纳入到评估框架中。

预期外影响的问题也与采用谁的理论相关。一个好的以理论为基础的影响评估设计在评价一个项目如何发挥作用时会考虑各种竞争性的理论。项目经理有自己的观点，而项目执行人员，受益者或其他评论家可能又有不同的看法。例如社会基金项目（支持社区发展的基金），项目经理认为，通过执行过程中的示范学习（看到社会基金在做什么）和边学边干，该项目对地方和国家的机构发展将产生积极地影响。然而，反对者认为社会基金绕过了现有的政府程序，直接挤占了当地的人力，并间接干扰了有关部门对资源的优化配置。因此，评估自然就要同时考虑官方的程序理论和反对者的理论（研究全文见世界银行，2002年，总结见卡瓦罗等，2002年，所使用的基于理论的方法见卡瓦罗与怀特，2004年。）。

通常程序理论整体思路的形成从项目文件开始。如果存在一个逻辑框架，那它一定会体现程序理论。然而，尽管项目文件可能将一些假设定位为风险，但它很少能明确所有的假设。接下来是由项目经理执行所提出的程序理论，即便他们之前对那些项目文件是否有自己明确的观点，这样可以项目经理真正参与其中，并允许其以积极地方式影响评估设计。⁴ 之后是查阅与要评估的项目有关或类似项目的已有评估研究及学术文献，这样可以发现因果关系中的薄弱环节。例如，目标定位不准是经常遇到的问题，尤其是小额融资项目（莫斯利与赫尔姆，1996年）。还有，针对妇女的小额资助实际上可能被家庭的男性成员占用，进而影响到项目的执行效果，像孩子的健康、营养问题。还要综合项目操作人员及受益者的观点。评估者要经常思考的重要问题是：“普通村民如何能够体验我们的项目？通过什么渠道获取信息？为什么要参与其中？”。思考这些问题很有帮助，尽管发展人类学告诉我们项目实施地区的人们，其观点和我们预期的可能差别很大，原因是认知不同，需求不同，或是项目执行人员沟通不到位。

环境认知

环境认知对理解项目影响和评估设计至关重要。环境指的是项目执行地的社会、政治与经济条件，这些都将影响到因果关系的演绎。同样的项目在不同的环境中将会产生不同的影响，例如表面上看上去很成功的印度泰米尔纳德综合营养计划在孟加拉国效果就不是很理想。然而，同样的项目也是存在于理想当中的，在复杂的现实条件下很少有，这也是环境的重要一部分。此外，如下文所述，环境认知有助于估计一致性，也利于项目的推广。

环境认知需要在评估设计前通读项目文件，还要广泛阅读各类文献（人类学和政治经济学），详见下文混合法的使用。

环境认知还利于项目的推广。对世行援助加纳基础教育和孟加拉国妈妈与儿童健康的研究，总体来讲都是成功的案例。在加纳开展的大规模校舍重建与教科书捐赠，为提高当地入学率和学习成绩做出了巨大贡献（世界银行，2004年）。这样的成绩背后有两个重要的背景。首先，经历数年危机后，当地学校系统的确是破败不堪，基础设施严重匮乏，学校几乎停顿。校舍重建与捐赠教科书在此情况下就发挥了重要作用。若当时学校已经能够正常运转了，那此举的影响就大打折扣了。其次，该项目得到政治上的有力支持，执行起来非常顺利（该项目是教育改革的一部分）。政府承诺也是资金援助计划取得成功的关键因素，促进了孟加拉国人口转型，死亡率和出生率大幅下降（世界银行，2005年）。孟加拉国刚获独立伊始，

⁴ 项目经理通常会反映说现在进行项目评估时机还不成熟，因为项目设计刚刚改动过，他们刚刚完成自己的研究不久，或者当地政府、部长或项目经理有变动等。这些意见通常都会被婉言谢绝，任何试图影响评估的行为都会遭到拒绝。但可以从中了解到在项目经理们心中哪些评估问题最重要。

几乎没有任何设施，随后赶上饥荒被免除了债务，在10年内，建立起遍布全国的健康与计划生育体系，实现上门避孕服务。如果不是政府全程监督，此类项目可能就半途而废了。

预知异质性

环境认知有助于预知潜在的影响异质性。依据干预设计、受益人特点和社会经济条件的变化，影响（实施效果）会有不同。通过检验背后的理论有助于发现潜在的异质性，并利于评估设计预测这种差异。预测异质性其意义有二。首先，计算样本量时需要了解总体的分化水平：分化程度越高，所需要的样本量越大（对控制组和处理组都是如此）。其次，从简单概率角度看，如果对20个任意选取的样本进行测试，有5%的概率其中的某个样本影响显著。像医学随机控制试验一样，好的做法都要求在采集数据前确定好测试的样本。基于理论的方法有助于前期界定这些样本，并对不同影响的产生提供合理的解释。不过，模型与数据间的联系需要引起重视。

就儿童营养计划而言，营养不良儿童较之营养较好的儿童增加体重的可能性更大，尽管有些极度营养不良的儿童可能由于腹泻难以有效进食、增加体重。定位准确的项目将会有更大的平均影响，这种影响在粮食缺乏的季节将更加明显——孟加拉国综合营养计划就是如此。年龄小的儿童可能受益更多；婴儿时期发育迟缓的可能无法通过后天的补食达到增高的效果。同样的，良好的营养可以促进认知的发展也仅限3岁以前。因此，营养计划的影响也因受益者年龄及其自身的营养状况而有所不同，后者还受到季节性因素的影响。。社会经济条件不同，营养计划的影响也会有所差异，例如替代现象（用补食替代原有饮食）发生在贫困家庭的可能性更大。

基于以上原因，营养计划已经从以学校为基础的补食转向关注3岁以下儿童，例如前述孟加拉国的案例。但针对学校的营养计划还是有望提高学生的学业成绩。由于缺乏热量，学生们上课无精打采的，营养计划使他们上课精神更集中。考虑到饱餐后容易犯困，所以何时时间进食也很重要。孩子们上课精神集中了，学习就会成绩提高，不过也受其他条件的制约。各类干预项目持有的一个关键的假设是，即此干预正在解开问题的症结。如果老师心不在焉，学生注意力集中也没有用，同样，缺乏学习材料，那学习效果也好不到哪去。因此，营养计划在条件较好的学校比在条件较差的学校影响更显著，因为在环境较差的学校教师缺勤现象普遍存在。类似的，附条件的现金资助增加了学龄儿童上学的需求，但供应方存在制约因素，它可能并不能提高学生学习的效果，甚至不能提高入学率（拉维林，2009年）。

预测异质性的另一层意义在于发现干预间的互补性；例如，小额融资配合商业支持服务就会产生更大的影响。如果两个项目可以相互替代，那么一起实施的效果就不及分而行之。能够挖掘此种互补性的方案显然具备重要的政策意义。

尽管通常（经常是暗含的）假定影响是线性，影响也会随时间变化（*乌尔考克，2009年*）。线性影响与前面所述单向因果推理和对不断变化的项目实施采用静态方法的质疑是有区别的。即便项目设计无变化，因果关系的方向也是确定的，项目的影响也会随时间而变化，且最终结果对影响评估的时间点选择很敏感。例如，现有研究表明，那些旨在提高边缘人群参与率和赋权的项目，其影响曲线很可能呈J型，即情况开始时恶化然后才改善。基于理论的影响评估在这一领域虽然研究不深，但却做了很大努力。前面所述孟加拉国综合营养计划通过增强妇女意识，可能会在最初引发他们与丈夫和婆婆的冲突，这可能会被认为营养计划没有影响；但随着社会的变化，孟加拉国乡村地区妇女地位的提升，长期影响的评估可能会发现该计划积极的影响。

异质性的识别关乎项目是否可以推广。在肯尼亚、南非和乌干达开展的一项随机控制试验显示，行过割礼的男性感染艾滋病的几率大幅降低（见*沃尔等关于乌干达，2008年*）。其中，异质性一方面表现在年龄上。行割礼后要有一个月的禁欲期以便伤口能够愈合；在此期间性行为风险更高。也就是说，12岁的男孩做此手术后并不面临这样一个月的高危期。但年龄大些的男性经常难以整月禁欲，因而降低了手术的益处。不过研究还是发现割礼可以降低30%-50%的艾滋病传播机率。此类影响仅能推广到那些具有类似性行为的群体。对那些奉行节欲、单一性伴侣或使用安全套的群体来说，割礼就不会对艾滋病传播产生任何影响。

影响

通过恰当的反事实对影响进行严谨的评估当然是基于理论的影响评估的关键组成部分。恰当的反事实通常参照控制组而定义，要避免选择上的偏倚，即选用实验或准实验的方法。面板数据有助于强化设计，因此鼓励收集基期数据。如果没有控制组，可利用现有数据或通过回忆来构建，但要加倍的谨慎（*班姆博格，2009年*）。除了选择偏误，设计当中还应当考虑的重要问题有溢出效应（控制组受到项目的影响）以及样本污染（控制组受到其他项目的影响）。

严谨的事实分析

对影响的反事实分析要辅以各类严谨的事实分析。因果关系链的许多环节都是基于事实分析的。具体到孟加拉国综合营养计划的案例，它包括定位不准及其原因，漏出的识别，以及不能学以致用问题。

瞄准分析是事实分析最常见的形式，而事实分析又占了影响研究的全部或大部。谁将从项目中受益？其中，有多少人来自于目标群体？瞄准误差有多大？瞄准误差可以量化并进行追溯，正如孟加拉国案例中的做法一样。瞄准分析应当在不同层面上开展。社会基金的案例中

发现，贫困标示地图的使用显示，在许多国家社会基金都是关注最贫困地区，但即便是在这些最贫困地区，那些相对条件好点的社区获得项目资源的可能性还是更大些（世界银行，2002年）。反过来，以乡村通电为例，条件好的社区接通的可能性更大，但在通电的社区，贫困家庭由于无力负担连接费用，多年来依然未通电（世界银行，2008年）。

瞄准分析必须依据一套有代表性的数据来完成。由于影响评估数据允许选择偏误的存在，它通常并不能代表总体，也就无法回答诸如“最贫困的20%人口中有多少人从该项目中受益”的问题，除非在分析时使用样本权重使其具有代表性。

瞄准分析的第二项工作实际上是一个二元变量练习题，要求根据参与状况绘制出相关特征（个人、家庭和社区的特点）的图表。准实验方法要求对项目参与情况进行多变量分析，但使用这些结果进行瞄准分析通常是一个误会，瞄准分析应当依赖于描述统计。项目是否惠及最底层的20%，是基于二维表格得出的论断，而不是基于多变量回归中五分位数的统计显著性。回归分析可以发现增进参与度的因素，有助于解释二元变量瞄准分析的结果。举例来讲，对印度某项目的多变量分析发现那些最贫困地区的部落参加项目的人数明显偏低，因此解释了瞄准效率低的问题。⁵

利用事实分析一个例子是测试那些接受培训的人们是否学到了只是并将其应用于实践。孟加拉国综合营养计划研究表明妈妈们学到了新知识但很多人并没有将其应用于实践。社区营养师们会称体重，但却读不懂成长图表。类似这样的分析很少有人做，但显然其前景很好。培训过的老师们是否了解改进的教学方法并加以应用？世行在加纳开展的一项研究表明多数都不会。⁶

在孟加拉国的案例中，事实分析经常能找出因果关系中的重要漏洞，进而说明影响微弱的原因。另一项世行业务评估部的研究发现，在肯尼亚开展的培训和访问农业推广服务对提高单产并无影响。理论上讲，该项目资助研究站开展农业新技术研究，将成果传授给推广人员，最后到农户。但在实践当中，研究成果并没有传授给推广人员，而推广人员告诉农户的做法很多都是早已在使用的了（世界银行，2000年）。

然而，有时候看似是需要进行事实分析，可能实际上是需要一个反事实。例如，学校按人头补助旨在提高入学率和学习成绩。但如何做到？毫无疑问要看资金的投入。这听起来可能像

⁵ 具体来讲，当部落变量被排除在回归之外的時候，财富变量很显著，但一旦后者被加进来，财富变量便不再显著。因此，决定参与度的是部落身份而非贫困本身。鉴于解释变量间可能高度共线，这种方法并不总是可行，像前面所述因素与教育和地理位置。

⁶ 课堂观察是检验实践的最佳方法，但常因成本高昂而被排除在外。可能有人会说直接问老师会有偏误，因为即便没有使用新教学方法他们也会说用了。事实上，老师们的回答出人意料的坦率，否则就是他们就是压根儿不知道怎么回答算正确。

简单的事实分析——追踪资金使用情况：检查究竟有多少用在了学校，如何使用的，这两方面都着实是对研究非常有帮助。然而，如果学校原先就已经掌握了一些资源，就可能出现可替代性问题。在此例中，尽管处理组学校和控制组学校在学校改善和和资源获取上的对比分析可能是一个更好的设计，但关于学校支出模式的前后对照也是一个有效的反例。

混合法

混合法指的是一次评估中将定量与定性分析两种方法相结合。所有的定量研究都包含一定程度的定性分析——至少项目文件要认真阅读，因此这只是度的问题。

混合法一般由定性分析法的支持者倡导。但是在发展领域，直到不久前，定性分析法还一直主导着评估的发展历程；因此在定性研究中越来越多的使用定量分析法是向混合法发展迈出了一大步。不过，在此我关注的是在定量研究中更多的使用定性的数据，关于此问题的论述详见怀特，2008年。这里我总结三点。

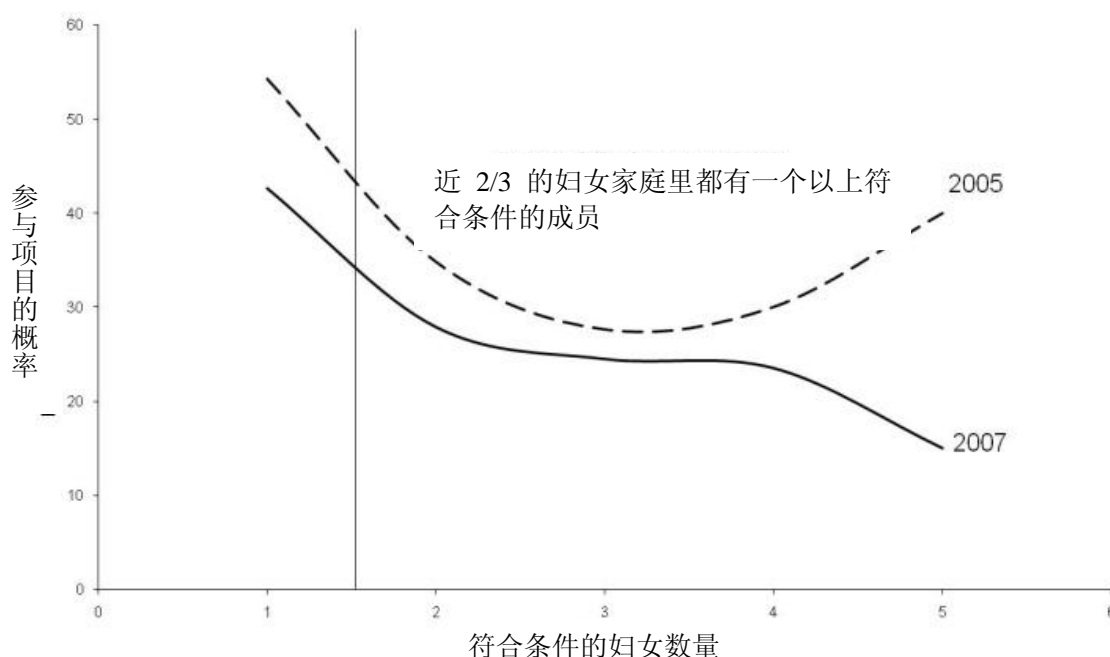
首先，使用定性数据有多种做法，而不仅仅是安排一些焦点访谈（在我看来，焦点访谈时收集定性数据的一种较弱形式，除非是做得非常的成功）。例如，阅读关于干预项目的人类学和政治方面的文献以便为评估设计提供信息。孟加拉国的案例中，婆婆影响的发现源于阅读人类学文献（怀特，1992年）。受此启发，我们打开调查问卷登记的家庭名册，找出那些跟婆婆一起生活的妇女（例如，如果户主的配偶健在，就包括户主的儿媳；如果户主的母亲健在，就包括户主的配偶和户主的嫂子与弟媳），因此按照定性分析提供的信息展开定量分析。

相关做法还包括实地考察（在项目实施地观察一两天）、利用参与式评估，以及在项目聘请一位人类学家，后者是一种正在尝试的方法，但可应用到更长期的研究中去。我的第二个观点是尽管实地考察颇受非议，它仍是基于理论的影响评估的核心部分。实地考察是不可替代的，没有这种体验就难以搞清楚如何合理地分析数据（遇到此情况的就深有感受）。花几天时间观察各种条件下项目的执行情况，最好不完全是项目人员安排的地方，将有助于研究的设计与实施。参观非项目地区同样也是有益处的。

通过考察，与项目人员、受益者和其他利益相关人士讨论很受启发，这类的例子不胜枚举。这里我仅举两个例子。第一个例子是关于一项农村生计项目的评估，该项目通过妇女自助组发放贷款。有个男人来抱怨说他的女儿未婚，22岁，无法申请到贷款，因为其妻子已经申请了一笔贷款。这表明村民们认为贷款是给家庭的而非给个人的，正因如此，那些来自家里有多个妇女符合条件的家庭得女性参与项目的比例更低（见图2）。项目的目标是让所有符合条件的妇女都参加，而只要利益分配限于家庭而非个人层面，那这个目标就是不现实的。第

二个例子是引用一句话来得出有力的结论。在赞比亚社会基金评估的实地考察中，我们惊讶地发现从经理、项目人员到村民们都异口同声的说是“社区”选择了这个项目，尽管一套筛选程序在实际运作（详见怀特与瓦加，2008年）。然而，事实是这个所谓的“社区”实际上指的范围很狭窄，就是项目委员会。正如一位正在打电话的地方项目官员对我们说的：“我得走了，我在办公室还有个会”。

图2——按家庭里符合条件妇女数量统计的安德拉普拉德妇女自助组织项目参与率



来源：影响评估调研数据

不采集新的数据，利用现成的数据进行影响评估（这种做法值得提倡，我们并没有为采集新的数据做好准备，而同时现有数据却没有充分挖掘），不过这样研究者进行影响评估时可能因没有亲身体验到项目实施而面临风险。由于缺乏对项目实际运作情况的了解，这样的研究很可能存在与政策相关度不高的问题。

最后，应当安排一部分预算开展一些行动研究型的活动，针对数据中的疑点进行额外的实地调查。孟加拉国案例中，针对知识获取后未能用于实践这一问题的焦点小组访谈，以及刚刚提到的印度安德拉普拉德自助小组的资助调查，都是此类工作的例子。我们有面板数据，调查中对家庭企业使用了标准的LSMS调查表。对这些数据的分析发现，多数企业都盈利微薄，还有不少亏损的行为。但是调查表中的数据并没有揭示这些企业是如何运营的。为此，我们委托开展了对这些拥有家庭企业的家庭进行二次访查，我称其为定量民族志研究。二次访查使用了半结构性问卷调收集了企业的日现金流以及家庭成员劳动力投入（偶尔也包括一些雇

工)。结果证实了企业的这些活动活动收入微薄（相对于日薪Rs. 50-70，每天Rs. 20-30的收入也不奇怪），面临很多高风险（牲畜死亡，特别是山羊；市场规模小）。

4. 对比基于理论的影响评估方法与黑盒子方法

基于理论的影响评估可与黑盒子方法进行对比。后者通常是简单地报告影响——关注平均处理效应的系数的统计显著性，而不去回答“为什么”显著。。本文试图说明如何回答“为什么”以及这么做的益处。然而，要提请注意以下几点：

不应过分批判只报告了平均处理效应的研究。预知异质性与理解特定影响产生的环境同样重要。大家不关注平均处理效应（通常既包括处理中的处理也包括意向性处理）的情形极少。事实上很有可能是大家最关注的参数。如果干预对样本中其中一组的平均处理效应的符号与其他组相反，那么只报告平均处理效应的显著性可能是误导。此外，平均处理效应是计算成本效益的基础。

其次，基于理论的影响评估从不同的方面分析因果关系，试图揭开因果关系链的各个环节以及干预行动中哪些元素在起作用，哪些元素没有起作用。回归分析可以达成此目标。例如，孟加拉国综合营养计划研究对理论与实践差距的影响因素进行了回归性分析。但是这些基于回归分析的方法，依赖于样本选择模型和所考察关系的参数设定。不少人对此持批判态度，他们更倾向于实验或准实验的方法，例如倾向得分匹配与断点回归设计。这些严谨的方法能够分析出哪些项目元素是有效的，不过项目设计要针对不同群体有所变化，例如有些企业得到贷款，有些得到商业支持服务，有些则兼而有之。实践当中，基于理论的影响评估会将这些严谨的方法与其他方法结合起来解密因果关系。

最后，黑盒子内涵可能如此混杂以至于最好是不要去碰？世行对乡村电气化的研究发现，通电确实大幅降低了人口生育率（世界银行，2008年）。该研究发现了通过看电视增长了避孕知识可能是原因之一，而并非是因为电气化丰富了人们的生活，减少了性行为。不过还有其他许多可能的因素，像收入的影响，其他教化的作用等等。像这种多种因素混杂在一起，难以区分，最好就采取简约模型评估影响。

5. 结束语

本文呼吁更多地使用严谨的定量研究来分析哪些因素能促进发展。不过，如果能够解释干预措施为什么能产生影响，或者为什么不能产生影响，这些研究的政策影响将会更加深远。普遍认为，基于理论的影响评估恰恰能解决上述问题。然而，很多最新的研究都没有严格遵循基于理论的方法的基本原则，只是推测产生影响的原因和影响的异质性，而没用可靠的实证分析加以解释。

我列举了一个基于理论的影响评估在实践中应用的例子，阐述了该方法是如何形成政策决策从而增强项目的影响的，这中间用到了前述的一系列的原则。程序理论要灵活理解，它能够适应实地情况的变化，也能吸纳其他竞争性的理论，接受任何意外的结果。要综合运用事实分析与反事实分析的方法，即使用多种方法。程序理论要结合干预项目的社会、政治与文化环境综合考虑，这也是发现影响预期差异的一个途径。

致谢

作者感谢玛丽·戈尔德给予本文早期版本的评论。通用的免责声明适用本文。文章的观点仅代表作者本人，不代表3ie及其任何成员或支持者。

参考文献

- 3ie受让人指南（日期不详）《3ie影响评估实践：受让人指南》，
<http://www.3ieimpact.org/page.php?pg=overview>（2009年6月1日访问）。
- 班伯格与迈克尔（2009年），《通过重建基线调查数据加强项目影响评估》，发展效率期刊，第一章第一节（37-59页）。
- 布莱克曼，伦纳德与斯蒂芬妮·赖希（2009年），《随机控制试验：靠不住的黄金法则？》，选自斯图尔特·唐纳德，克里斯蒂娜与梅尔文·马克著《在应用研究与评估实践中什么是可信的证据？》，【千橡木，加利福尼亚：赛奇】
- 卡沃尔，索尼娅，吉尔·伯金斯与霍华德·怀特（2004年），《社会基金：参与，社会资本与可持续性》，国际发展期刊，14 611-625，2002年。
- 卡沃尔，索尼娅与霍华德·怀特（2004年），《基于理论的评估：以社会基金为例》，美国评估期刊，25（2） 141-60，2004年。
- 全球发展中心（2006年），《我们何时能学会？》，华盛顿：全球发展中心。
- 影响评估国际动议组织，3ie（日期不详），《3ie影响评估实践：受让者指南》，
<http://www.3ieimpact.org/doc/3ie%20impact%20evaluation%20practice.pdf>，（2009年，6月1日访问）
- 李宙果，（1983年）《经济计量学的反对论》，美国经济评论，23（1），31-43。
- 莫斯利，保罗与大卫·赫尔姆（1996年），《扶贫融资》，伦敦：路特莱斯。
- 默克赫尔吉，让丹，马克·维斯与霍华德·怀特（1994），发展中国家经济计量学与数据分析，伦敦：路特莱斯。
- 诺尼（日期不详）《诺尼论影响评估》，<http://www.worldbank.org/ieg/nonie/members.html>，（2009年6月1日访问）

- 拉维玲与马丁（2009年），《评估三类程式化干预》，发展效率期刊，1（3）。
- 罗杰斯，帕特丽夏（2009年），《使影响评估设计与干预性质和评估目的相匹配》，发展效率期刊，1（3）。
- 救助儿童（2003年），《收效甚微——质疑世行资助孟加拉国、埃塞俄比亚与乌干达社区营养计划的效果》，伦敦：英国儿童基金会。
- 韦斯，卡罗尔（1998年），《评估：研究项目与政策的方法》，朴恩提斯·霍尔：纽约。
- 沃尔，基高兹，斯瓦达等，《男子割礼在男性艾滋病的尝试》，乌干达拉开，《对男子及其女性伙伴艾滋病影响》，第15届逆转录酶病毒与机会感染大会，2008年；波士顿，摩洛哥，2008年。
- 怀特，霍华德（2005年），《论孟加拉国综合营养计划的贡献》，健康政策与规划，20（6），408-411。
- 怀特，霍华德（2008年），《关于概率与参与度：混合法在定量影响评估中的应用》，IDS公告，2008年。
- 怀特，霍华德（2009年），《反思当前影响评估的辩论》，3ie工作文稿1，新德里：影响评估国际动议组织。
- 怀特，霍华德与埃德尔多·马赛（2006年），《孟加拉国综合营养计划：影响评估发现》，国际发展期刊，19:627-652，2006年。
- 怀特，霍华德与安朱·瓦加（2008年），《世行能否促社会资本增加？：马拉维、赞比亚社会基金的社区参与》，发展研究期刊，44（8）：1145-1168。
- 怀特，莎拉（1992年），《与鳄鱼争辩：孟加拉国性别与阶级》，伦敦：赞德。
- 乌尔库克，迈克尔（2009年），《项目评估方法多元化：理解影响路径与功效的逻辑方法》，发展效率期刊，1（1）：1-14。
- 世界银行（2000年），《农业推广：肯尼亚经验》，华盛顿：世行运营评估部。
- 世界银行（2002年），《社会基金：有效性评估》，华盛顿：世行运营评估部。
- 世界银行（2005年），《保持动力到2015？孟加拉国改善母亲与儿童的健康与营养状况项目的影响评估》，华盛顿：世行运营评估部。
- 世界银行（2006年），《将营养问题重置于发展之中心：长期大规模行动的战略》，华盛顿：世行运营评估部。
- 世界银行（2008年），《乡村电气化的福利影响：成本与收益的重新评估》，华盛顿：世行独立评估集团。