世界范围内的大学科研评价方法(下)

五、来自大学方面的评估

从整个社会的观点来看,加强对科研产出的评估可以成为一种提高科研产出的途径。而从高校的观点来看,评估也有其优势。

某一领域或某所高校内部的评估已被广泛接受。其目的是确立研究工作在内部受众心目中的合理性,比如,保证研究质量的提高。而为外部受众建立研究的合理性也越发重要——质量保证——这会是十分困难的,因为外部受众往往掌握较少的专业知识。外部合理性的要求源于科研人员对资源不断增长的需求,而为这些研究分配资源就会涉及有关机会成本的问题。另外,这样公开证实研究质量将会吸引高质量的研究人员和学生。

一个很重要的挑战是如何保证通过外部"交帐"程序能增强内部"交帐"。例如,在美国,一个独立教育委员会对教学质量的认可程序会加强内部提高教学标准的责任感。

1. 内部——质量提升

高校内部"交帐"的第一阶段是评估研究人员是如何执行其"任务"的。这种"交帐"义务是法定并有经费资助的:资源是如何使用的,是否用于其初衷?

第二种"交帐"主要是要解释那些资源之所用:例如,研究质量是否有所提升。这第二种"交帐"实现起来明显地较第一种要更困难,因其与结果有关。同行评议希望能够通过向同行"交帐"来提升研究质量。

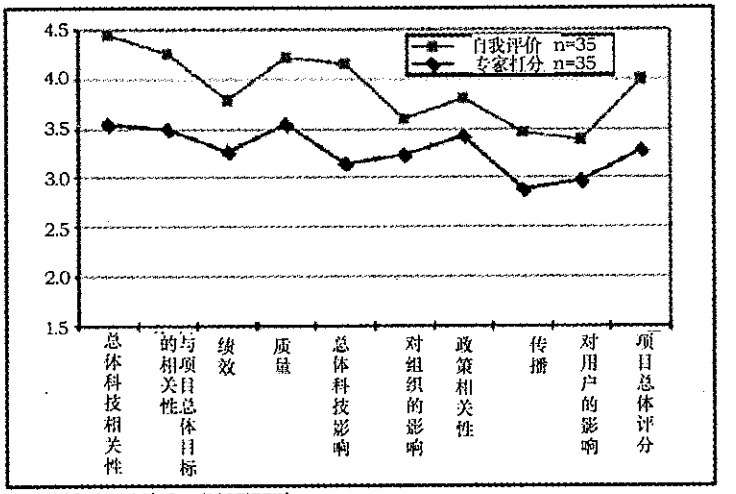
有人认为社会规范存在于每一个科学共同体中,由能够保证研究产出质量的同行评议和仲裁体制维持。该封闭质量评价系统无论在过去还是现在,对于建立科学的威信都是至关重要的。因此,科学的根本和理想化的特征一直都是自主的。

2. 外部——质量保证

尽管科学研究从很多方面看来都势在必行,对科研工作投入资源的需要必须缘于对研究的价值以一种具体的形式予以证实,因为其他更有形的客观目标也需要公共投入。从事科学并投入资金值得吗?

合理的研究对于拉动外部资源来说是很重要的。资源能够使研究者制造产出,这可以通过科学出版物来揭示。出版物对于科研人员获得未来资源是一种可靠的途径。这种可靠性来自于出版物的评议过程。但是出版物是一种学科内部评价的行为。只有该学科领域内的专家——同行才能够评价这些文章。

图10显示的是怀疑论关于自我评估的看法。该图比较了瑞典的一个由项目内部和外部进行的评估。进行内部评估的人都比外部的评估者要乐观。



来源: Arnold与Guy(1997)77页

图10 瑞士能源与环境项目绩效中专家打分与自我评价的比较

因此,研究机构感兴趣的是使他们的资助者相信其成果的重要性。高校同其所在环境以及"支撑社会"之间通过以下3种方式发生关联:

(1) 义务

义务是指向他人报告的责任,尤其是对政府部门。

(2) 信任

当高校凭着所提供的服务(包括研究和教学)获得资源而不用详细说明经费的使用时,信任就很必要了:例如,当高校获得财政拨款的时候。

(3) 市场

高校可以通过收取学费和教学收入或承担合同项目等产生市场交易行为。

义务、市场交易以及信任相互替代或补偿:"信任并核实"是里根总统过去常说的一句话。在世界范围内盛行的这种高校与社会之间的相互作用是有所差异的。乍一看,似乎市场是美国高校与社会之间的主要纽带。但是,高校为了加强外界对高校自身的信任,也积极地同家长尤其是校友会团体进行着沟通。由于美国市场力量十分强大,高校十分热衷于向选民说明他们正在进行的研究以建立信任,并体现出使命感。

从政治科学家那里我们知道,英国高校系统运行在这样一个世界中,即政府撤除了对高校的信任,但同时也限制了市场的角色(至少在学费方面)。从而,正式的义务被取代了。

如今加拿大的科技企业大量地依赖彼此之间的信任,然而科学与社会间

联系最弱的恰是信任。任何一个人的能力都不能完全保障外界能够相信科学产出并建立信任,如果这种信任可以建立,也只能是脆弱的。正如在导言中说到的,欧洲目前对转基因食品极端的怀疑可能来自于过去失败的科研活动,如近来英国对疯牛病的恐惧。由于信任十分脆弱,社会学家相信全球会产生一个从信任关系转向一个协作保障的系统转变。不断增加的义务会加强彼此间的信任。

在欧洲大陆,科研"交帐"主要落实在对资金的控制。然而,欧洲的作法在证实研究的合理性方面会引起一些麻烦。正如坎贝尔和费德尔曾经提到的:政策制定者和广大专家学者中,有一些人对德国现行的高校体系结构和研究绩效并不满意。而且这种不满在很大程度上解释了公众反对增加高校活动(如高校科研)的公共基金的原因。部分德国研究型高校对系统评估施加的阻力产生了负面效应——如一些专家所说——在公众和高校之间存在的"信任鸿沟"是难以逾越的,我们可以引用威廉·克鲁尔的断言:某些分析家希望把西德大量的高校看作德国国家研发系统中唯一致命的弱点。

评估在信任的建立过程中是一股重要的力量。

六、评价方法

这一节讨论三个问题:

1. 对什么样的产出进行评估?

2. 对谁进行评估?

3. 如何进行评估?

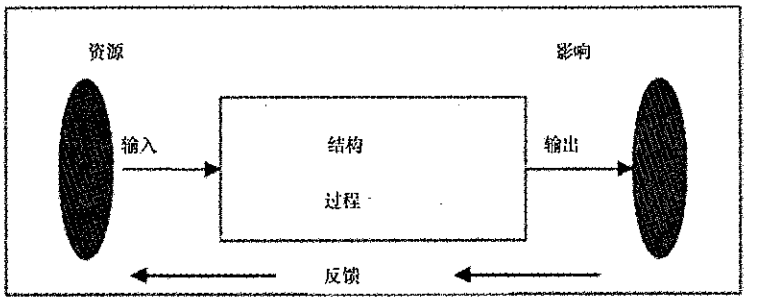
1. 对什么进行评估?

评估的首要目的是要解决在学术界中的组织问题。具体而言,要评估优化制度的组织结构;强化责任义务的原则;强调决策以及政策制定的外在和理性标准的应用。

评估的目的并非执行一项简单的无过失研究,而是"一个连续的研究,在共同迈向规范的评价方法过程中各种研究能够相互批评与改进。"所有的方法都有其优缺点。由于各有利弊,很多方法就能互相补充。例如,量化指标的一些潜在缺陷可以通过同行评议的方式得以修正。

在深入探讨需要评价什么之前,建立一个关于如何把研究过程作为一个系统呈现出来的标准模型将会是有益的。在图11当中,资源——美元和时间——是对科研生产机构的投入,而其产出,如学术论文,反过来会对社会有影响,比如节能技术。研究机构将投入转化为产出的能力有赖于其结构和运行程序,但整个转化过程也依赖于该国经济的运行方式。

评估过程的核心在于判断目标是否实现。显然,为了达到这个目标,需要



来源: Arnold和Guy(1997)pp.74

图 11 一个简单的系统模型

对预期目标进行明确的设定。在本阶段,我们仅观察政府要求对社会产生有益的影响方面。

图 11 也可用于揭示一些概念和认识该系统在运行中遇到的困难。检测一个研究系统的传统方法是量化研究机构的投入,如研究人员数或投入的资金。很明显,这些方法并没有说明该研究系统是如何运转的或者它是否产生了有益的影响。

图 11 同样也说明了有关时间方面的问题。一般而言,指定款项会在研究开始之前到位。因此,……然而,这些调配到高校的资金所发挥的效用只能进行外部评价。从某种意义上来说,外部评价是用以评估研究资源分配方法的效果的。

在该系统中,各种测度方法是与不同的组成部分和参与机构相关的。对不同类型信息或数据的选择与观测者的兴趣相关。数据的分类大概由以下 3 部分组成:

· 管理统计:某一机构不同特征的描述数据,如科研人员数、在校学生人数;

· 管理信息:统计数据可符合用于内部和外部的报告和监测用途,但不能用于管理决策;

· 绩效指标:能够提供关于某系统或组织在功能和效果方面的定量或定性的战略信息。

显然,政府作为一个社会的代表,对于绩效指标的重视可以监控系统整体功能的健康运行。对于评估者而言,面临的挑战在于寻求一种可用来报告既定的战略目标和社会目标是否得到了实现的方法。

另一个特殊的评估特点是看该系统体现的是一种累积性的还是格式化的功能。累积性评估是在与相似对象相比较的累积例证的基础上,对一个对象单元(部门,高校等)的绩效进行评价。格式化评估则用于帮助对象获取其改善绩效所需的信息以帮助其实现设定目标。就某种程度而言,累积性评估可

以作为一种连续进行的过程来监控系统运转而格式化评估显得更具有适应性。20世纪80年代期间,像美国、瑞典这些多元化的国家曾对技术投资的合理性表示过质疑,由此导致了累积性评估的增加,即追究我们从投入的经费中到底得到了什么?

对于政府而言,对高校研究的评价可以包含对以下元素的测定:

- 质量
- 效率
- 相关性
- 可维持性
- 效果

原则上可对这些内容的部分或全部进行评价。英国的研究评价活动(RAE)明确地仅评价对象的质量而荷兰要评估前四项。表6详细地给出了评价需要达到的目标列表。

表6 典型的评估内容

主题	问题
适当性	这样做是正确的吗?
经济性	实际费用比我们预期的要少吗?
效果	实现预期效果了吗?
效率	投资回报如何?
效力	同预期的投资回报相比效果如何?
过程效率	实现过程顺利吗?
质量	产出质量好吗?
影响	其结果如何?
附加值	得到了什么超出预期的效果?
替代	什么事情本应发生而没有发生?
过程改进	我们怎么能做得更好?
战略	下一步我们需要做什么?

来源:Arnold 和 Guy(1997)pp.72

2. 谁应该接受评价?

评价高校科研产出的根本性困难在于:(1)学科之间异质性(如出版物的成本和时间消耗);(2)高校之间有不同的专业分工。结合以上两个因素意味

着高校研究产出的不同不仅因为存在质和量的差异,而且还由于构成的不同而不同。

表7展现的是不同学科之间引文率差异:例如,分子生物学和遗传学方面的论文引文率要远远大于数学论文。不同的引文率可能是由于一些领域业内研究人员数目较少,实现突破难度更大,对同一领域内不同研究项目的补助比例差异,该学科已是较成熟的学科领域或大量其他因素造成的。当把高校作为一个整体对其研究产出进行评估的时候,这些困难就凸现出来了。

因此将大学作为整体来进行评价效果常常适得其反。这样一个指标对那些消耗较大、发表文章时滞较长的学科不利。高校整体的评价指标并不能提供有价值的信息,因为这些指标并不能给那些发展不太好的部门以好的机会。每个学科都应该进行独立的评估。但需强调的是,在对每个学科进行评估的时候,不应该只对该学科的学术表现进行评价。

表7 不同国家在1993~2001年间发表论文的平均引文率

领域	平均引文率	领域	平均引文率
所有领域	8.31	数学	2.47
农学	4.31	微生物学	13.38
生物学与生物化学	14.86	分子生物学与遗传学	23.99
化学	7.46	交叉学科	3.48
临床医学	9.82	神经科学与行为学	15.7
计算机科学	2.27	药物学与毒物学	8.82
经济学与商务	3.82	物理学	6.81
工程学	2.84	植物与动物科学	5.72
环境/生态学	6.98	神经病学/心理学	7.61
地理科学	7.35	社会科学、综合学科	3.14
免疫学	18.92	空间科学	10.88
材料科学	3.77		

来源:ISI引文数据库

对科学研究的评估给予了高校各院系、高校、资助部门以及政府有益的启示。这些机构在资金数量和如何分配上的选择对于结果十分重要。其目的是要评价那些获得资助的机构所做工作的成果。因此,现在资助部门的做法较为合理,即对得到了财政拨款的研究人员所作的研究成果进行评估。如果一

所高校作为一个整体单位获得资助,如财政拨款,那么对该所大学整体上的经费使用进行评价则较为合适。通过考查高校各学科的绩效,我们能够考虑这样一个问题:该高校是否最有效地分配和使用了得到的资金。资助部门的资助选择是否有助于被资助对象提高其研究的水平?这些问题只有通过对独立学科的考查才能够获得解答。

3. 如何评估科研成果?

研究如何评估很大程度上取决于评估者所希望评估的内容。图 12 展现的是适当的科研评估的类型。在水平轴上,评估者很可能会对投入、产出以及其影响产生兴趣,而政府应该对后两者更感兴趣。

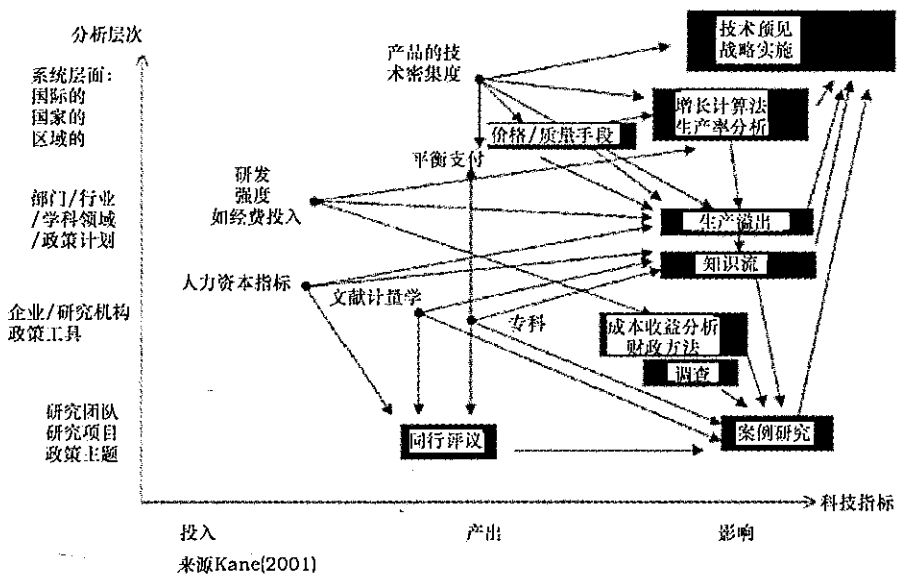


图 12 研究的评价指标和评价方法

纵轴上包括从个体研究人员的绩效到一国研究体系绩效各个值具体的受关注程度。由此,如果评估者对某一研究人员的成果感兴趣,那么最好做一个同行评议,但是如果评估某研究人员研究成果的社会影响,那么最好进行案例研究。强调一下,这些方法之间不是互斥的。

一国政府,原则上来说应会看重某个研究系统整体的影响力,例如,图 12 右上角的部分。然而,不同的评估有相当大的替换效应。每种方法的很多细节和成本都有所不同。重复进行增长算法并不实用,但专利和文献计量数据却相对容易获得。因此,即使政府部门只对图 12 中右上角的部分感兴趣,该图剩余部分给出的由这些方法得到的信息能够让社会相信该研究系统正在良性运转中。希望在于,鼓励研究产出的政策能够对社会产生积极的影响。

基本概念

要想获得令人满意的方法来测度科研活动对经济的影响力似乎成本会很高而且通常有较大的不确定性。所以,这个部分将从考查研究产出着手展开论述。评估产生的总体方法将在深入进行相关子部分之前进行讨论。

对于高校研究成果的评价有两种比较理想的方法:同行评议以及指标法。并不存在显然正确的方法而且这些方法通常在改进的同行评议方法中被结合使用,下文将会进行阐述。传统的同行评议是指某一学科领域内的专家组——有建树的专家(一个或多个人)进行科学评估工作。实际上,指标法也是一种间接进行的同行评议过程:这些指标是用同行评议法度量成功与否的结果(例如出版物数量)或者是那些论文对科学共同体所产生的影响的一种衡量(如引文数)。

很明显,其中一个重要的区别就在于在一个同行评议过程中的同行数要大大小于引用那些重要的论文的研究者人数。这样一来,引文数就可能比仅在一个小国范围进行的同行评议过程要有效,这也是为什么一些北欧国家(如瑞典、挪威等)要使用国际同行专家组进行同行评议的原因。同行评议和指标法的优缺点可以相互弥补,见表8。

表8 高校研究评价方法的优缺点

	数量指标	改进的同行评议法
优点	它们相对客观而且实现成本相当低。很多国家都进行这些类型的统计,如澳大利亚现在就实行这种方式	它可以对复杂的事物进行整合
缺点	它们可能看起来比较肤浅并因而不能够完全地展示研究的全过程。因此,学术界可能会认为它们缺乏合理性	成本高且有潜在的主观性

由于同行评议的参评者对评估领域知识有所了解,可以通过一些方法(如参评项目的难度)来整合研究过程的复杂程度。然而,消极的一方面在于,即使是有经验的评委也会在评价过程中加入个人的主观观点,例如,宁可维持现有的研究对象也不愿利用未尝试过的方法进行试验。

指标,从本质上看,更加简单且不能被复制,例如,科研项目的难度。然而,指标确实可以更精确、更客观地测量人们想要测量的内容,即使这种方法的结论是有争议的。

由于本节中我们将会提到一些有争论的问题,传统的同行评议方法并不常用于集合层面的系统评估当中,而多用于评估个人或者小规模的研究小组。英国和香港使用了一种改进的同行评议方法来帮助他们进行经费的分配。改

进的同行评议是指:(1)不仅限于同行(2)除了学术论文的质量以外还考查其他内容。荷兰也使用一种类似的方法,尽管这种方法在现行的资助决策中没有什么实际的意义。其他一些国家使用数量指标来分配资金。

表9 依据评估方法对大学评估进行分类

修正同行评议		数量指标	
应用于经费分配	应用于信息反馈和改进目的	应用于经费分配	应用于信息反馈和改进目的
英国 香港	荷兰	澳大利亚 波兰	芬兰 瑞士 丹麦 日本

在更广泛的研究领域中,还可以使用其他评估方法,如用户满意度和经济影响力研究。举例来说,芬兰对研究项目进行评价最常采用的手段是用户评价。经济影响研究是为了获得度量研究项目的净现值、成本收益比以及社会回报率的方法。

接下来会对不同类别的评估进行更详细的讨论。只有那些作者认为对所有高校的院系都相当简单易用的技巧才会被考虑进来。因此,用户调查和更加复杂的收益成本分析将不会予以考虑。

指标

如上所述,指标的作用在于提供数据告知用户项目的战略目标是否实现。这里的困难是如何保证数据能够提供我们所需要的信息。关于指标如何使用户误入歧途的一个例子是把班级的大小作为教学质量的替代标准,这是由于班级大小和教学质量之间的相关性是非常弱的。

下一节将着重讨论出版物成果。显然,仅有出版物指标并不充分,但是,出版物数量的增长是学术研究必要目标之一。表示科研产出其他方面的指标也是可以使用的。最容易想到的是硕士和博士生的数目以及他们完成学业所用的时间。其它易于测量的指标主要是商业化的指标,比如专利和许可收入。

应该指出,尽管某指标体系相对另一个指标体系而言会存在一些问题,不时地考查一下该指标的表现将会是有益的。

出版物

只对出版物数量本身进行考查显然过于简化:在"斯诺经济学"杂志当中的一篇文章不可能有"计量经济学家"中的一篇文章那么重要。因此,一种方法是根据文章所在的期刊来赋予权重。每年公布在《科学引文索引期刊引文报告》中的期刊中所有文章的平均引文率指标可派上用场:

单纯基于出版物方法所存在的原则性问题有：

·剔除了其他的度量方法，如专利数；

·有可能受到出版过程当中的人为偏见的影响；

·可能在合作作者问题的处理上有困难；

·出版率会由于所属院系的大小不同而不同。

引文分析

这些指标多使用于ISI的《科学引文索引》。

这种方法中存在的问题可能有：

·有利于英文出版物；

·无法区别引文评价是正面评价还是负面的；

·数据库的限制。

从积极方面来看：

(1)文献计量学可能不适于对个人的研究产出进行评价，因为将质量考虑进去十分困难。但是如果出版物集中在某一学科领域内，而且假设质量能够达到一般水平，则可以把文献计量方法在聚类的层面上作为一个有用的信息源加以挖掘。

(2)由于同行评议的出版物构成了分析的基础，文献计量学多多少少是一种量化的质量体系，或者说一种间接的同行评议方法。

(3)从某种程度上来说，引用某一篇重要文章的论文数远大于参与某一审批委员会进行评估的同行人数。而且，全球范围内都可以引用这样的文章。原则上，一篇论文在更宽范围内的引文数可能得到比某小组的两个同行专家认可更能成为一种在学术上受到尊重的标志。

定标比超

在美国，国家科学院科学、工程与公共政策委员会(COSEPUP)希望了解美国是否在各学科领域处于领先地位。其目标是要回答联邦政府怎样才能“从整体或局部来衡量国家研究事业的运行情况，并研究国家投资是否充足以及是否支持国家的研究目标”。

为此COSEPUP建立了同行评议专家组。该专家组由业内科研人员、同该领域密切相关的其它学科领域的研究人员以及研究成果的使用者组成，还包括外籍科学家。这些专家组会就目前或将来美国的研究状况被询问对相关问题的看法。在评价过程中，专家组用到以下方法：

·虚拟代表大会

所有参与者都需要根据自己的判断挑选出每一领域最优秀的专家，或通过征求该领域其他学科带头人的意见来进行评价。这种方式是为了挑选出能

够参加仅由全球顶尖的5~20个领域专家组成的会议人选。

- 引文分析
- 期刊出版物分析
- 定量数据分析(如经费)
- 获奖分析
- 国际会议发言人

瑞士人使用一种整体类比法,把高校冠军联盟作为他们比照的标准。瑞士人意识到尽管他们在国与国的比较中很出众(例如平均水平的对比),但他们同样需要将其机构同世界上最好的机构相比较(如整体分布情况)。

"传统"同行评议

要注意的是同行评议可以利用出版物以及引文方面的定量信息,该过程通常被称为知情同行评议,我们将在下一节中进行讨论。

表10中给出了来自于科学共同体对同行评议的褒贬意见。争论主要针对评价过程中的主观程度。这个小小的同行专家小组是否进行了独立且公正的观察判断。瑞典的经验告诉我们这种主观性可以通过邀请国际专家进入评判委员会而得以降低。然而,瑞典的经验也暴露出一个问题,即对所有的项目几乎都得到的是4分(5分制)。

需要强调的是评议过程中常遇到的一个困难,即对有关同行评议过程变得越来越保守的抱怨。评委通常在业内有所建树,且倾向于对研究成果比较丰硕的领域持有偏好。由此有人抱怨同行评议过程对交叉学科的研究持有偏见,而在现今的研究环境中学科交叉的趋势正越来越明显。

表10 对同行评议的正负面评价

对同行评议的正面评价	对同行评议的负面评价
有效的资源分配机制	评估者的偏好导致非技术因素影响评估结果
高效的资源分配者	保持已有领地的关系网
科学责任的推动者	更有可能资助那些已经有知名度的科学家/部门/机构的所谓光环效应
帮助决策者把握科学导向的机制	评估者在评价和解读的标准上存在差别
一个理性的过程	该过程假定了对什么是优秀的研究以及有前途的机会已达成共识
一个公平的过程	
一个有效且可靠的科学绩效评测方法	

来源:Kostoff(2002)

同行评议中存在的一个不为学术界察觉的难题是同行评议是一种内部的程序。因此,同行评议可能对决定谁获得研究项目很有用,因为知情评价要求识别出潜在的多产课题。内部系统的负面效应在于这种定性评价对于外部的观众显得不够有说服力。确实,在其他国家,对评价的要求主要来自于研究共同体外部。

第二个难题是同行评议的成本。进行一次彻底的评估所需要的时间很可能是相当长的。一个部门聘请外界同行来评估其运作也可能十分耗时。

了解了正负两方面的评价,让我们来看一些文献当中广泛认同的同行评议所适用的领域:

(1)该领域存在一个受到明确界定的科研共同体,这些科研人员在共同所用期刊和奖励体系方面相对稳定;

(2)该领域由科学共同体本身选择研究问题;

(3)该领域可以通过增加资助得到扩展。

有人质疑单纯的同行评议对于拥有更广泛目标的过程是否有效,如对研究系统的评估或者不同部门间分拨经费。同行评议不能够决定基础研究与应用研究之间的优先级,也不能对分属不同领域的研究课题进行排序。

Kostoff (2002)得出结论:要想进行一种"完美的"同行评议是非常耗费的。这个过程的组织所带来的社会影响虽然是未知的但是由于同行评议过程内在的保守性,其结果可能并非有效。

"改进"或者"知情"的同行评议

越来越多的国家引入了一种改进的同行评议。考虑同行评议在以下方面有所改进:(1)召集一个不同的专家小组;(2)对研究系统的多个不同方面进行评价。因此这种新的评议方法能够在更宽泛的环境中使用并且能够带来更多的信息。举个例子,同行可能会在更广的社会目标环境中对研究项目进行评价。他们还可能使用科学计量学方法来得到其结论。

改进的同行评议可以实现以下目标,例如:

·不仅仅吸收科研人员参与评估。可以吸收工业界或那些研究成果应用方参与评估,即请"用户"来评价。该方法已被瑞典、荷兰和加拿大使用过。同时还可以吸收其他学科的代表参与评价。

·英国的RAE从根本上来说是一种改进的同行评议活动,其目标是对科研质量进行评价。这些同行同样可以使用科学计量学方法来评价研究质量。

换句话说,同行或者任务的定义已经被做重大改进。这种方法在20世纪80年代在北欧国家有着广泛的应用。如下一节我们所要谈到的,当丹麦要求评估的压力变得越来越大的时候,改进的同行评议在英国以RAE这样一种折

中的形式得到了发展。丹麦的模式看上去同英国的比较类似,但其评估者和具体的评估程序还在协商阶段。

不同评估方法的对比研究

从某种意义上来说,对两种不同评估方法即指标和同行评议系统所存在的抱怨是相同的。他们都对使用英语进行的研究有偏好,而对年轻的学者以及有创新性但同样有风险的研究抱有偏见。Roberto Perotti 注意到,基于产出的评估是建立在质量噪声信号之上的,但这至少要好过一点信号都没有的评估。他举出了英国 RAE 的例子:通过反复尝试和纠正错误,最终取得几乎一致意见而结果也令人满意。

英国研究人员比较了基于同行评议的 RAE 方法以及文献计量研究方法所得的结果。他们发现:

·对于自然科学,RAE 评价等级和文献计量影响指标之间有较强的相关性;

·在对艺术类和人文类的评价上二者几乎没有相关性;

·在分析取样的 1988~1996 年之间,英国的自然科学论文以及引文率都在增加;

·社会科学领域发表的论文在增加,但同一时期内其引文率维持在原有水平。

采用文献计量方法可能要比用"改进"同行评议法成本低很多,但是,由于目前对于文献计量学是否能够作为评价质量的有效信号存在着一种普遍的质疑,因此,像丹麦的评价系统那样,由专业小组来操作文献计量学分析似乎是有必要的,因为他们可以把相关重要观点和信息在一定程度上应用到评价过程中,以保证所用量化指标是有效的。

资源的分配是否应该以对产出的评价为依据?

这个部分我们将讨论资源的分配是否应该依据对研究产出进行评价的结果。这种分配方式可以对被评价者产生"强力"激励来提高那些被评测的产出。这种激励机制给予我们两点启示:

(1)被评价的产出同潜在的我们期望的产出之间在一定程度上存在差异。测度只是一种不完美的替代手段,例如拿出版物来衡量科研活动。由此可得,强力激励只会提高测度标准之下的科研产出而不一定会提高人们的期望产出。因此,强力激励是否合意取决于它在何种程度上表征了我们期望得到的产出。

(2)以分配资源为目标来衡量产出对那些不被评测的研究产出或研究工作可能是一种伤害。比方说,如果只测量出版物数量而不测量引文数,那么很

可能导致研究人员只注重提高出版物数量而不重视研究质量。

如果目前的这种强力激励只是针对活动而非研究那么则有可能是积极的改变。比如,如果有很多提高本科生教育质量的激励(如目前加拿大的一些省所做的那样),那么,研究活动可能会因为没有足够的诱因而遭受损失。引入较强的激励会对此进行补救。

只依赖量化指标来评价科研产出并以此为依据来证明引入强力激励的合理性似乎有些牵强。显然,如果仅仅是额外的那么有限的一点经费,那么这些资源可以依据对研究产出的测度来进行分配。但英国的体制在对大量的资源进行分配时是以研究产出作为依据的,并因此建立了强力激励机制。英国的做法也可说明其合理性,因为他们同样建立了一套完整且昂贵的研究质量评价系统,该系统为保证测研究质量同实际研究质量相符提供了更强大的保障。

这一结论支持了对研究活动(如研究前沿性)的评价,但还有其他科研"产出"的形式。科研活动最重要的产出是传播最新科学进展的研究人才(硕士和博士)。如果政府能够有充分的理由确信加强研究生教育是有益的,那么强力激励机制就应该在这个方面得到加强。例如,芬兰、丹麦和澳大利亚等国都在以研究生人数为依据来配置资源。

七、英国的研究评估实践(RAE)带来了什么样的影响?

英国在20世纪80年代中期引入了一个评价研究产出的体系。引入RAE曾遭到来自外界的一些质疑。进一步说,似乎最初的评估尝试在提高研究质量方面并不是特别成功。然而,当前,尽管英国没有其他很多国家投入的经费多,但其学术研究的质量是相当高的,见表11。而且,从下文中可以看到,RAE似乎并没有危害到跨学科研究。

表11 英国科学研究的地位如何?

英国在17个进行研发经费投入对比的国家中排名第13位
若考虑到人口差异因素的话,英国在博士学位授予的数量上与其他国家相当
在出版物排名中原列第二位,目前落后于日本列第三位
英国的引文数在全球排名第二,约占到全球总引文数的11%左右
除了数学(列第三)和物质科学与工程(列第四),英国在几乎所有领域的排名均在第二
除了社会科学和工程,英国占全球引文数的份额都在增加
除了医药行业,英国公司的研发水平同国际上相比一般认为还显不足

来源:OST(2003)

对RAE的评价指出,制度上对于RAE的回应包括:

(1)进行结构改革,包括有:①给研究赋予更高的优先权;②建立评价的内部程序;③有选择性地分配研究资源和④安排一些高级经理人来监督和管理相关工作;

(2)分配由 RAE 而产生的研究资金以及其他经费以在后继的评估中获得最高的评价;

(3)进行一些博弈。

这些调查指出,对于个人来说,排名更高的学术机构的个人主要受到来自于发表高质量期刊论文的压力。而评分等级较低的机构员工则应该发表更多文章,不管这些文章是发表在何种期刊上。调查指出,学术机构在研究上花了更多的时间,其研究成果的数量和质量也都有所提高。

再看来自经济部门的研究产出情况。Moore 等人在 2002 年指出,名列前茅的研究机构中的经济学家,在 1992~1996 年之间发表在高质量期刊上的论文比 1980~1989 年间大大增加了。排名较低的机构也增加了其出版物数量。增长的出版物是科研人员付出更大努力的结果,而不是源于新兴的高质量研究机构的进入。大多数行为上的改变都发生进行评估之前发表论文数低于平均水平的个人身上。而对于已发表过高质量论文的个人则没有太大的改变。1992 和 1996 年的评估对研究人员的个人职业生涯带来的影响是累计产出上的惊人变化:高质量期刊中的累计研究产出要高于假定不评估时所会发生的情况。1992 年的 RAE 使得研究人员个人在该评估之后 4 年间在高质量期刊上的平均发文量比以往增加了约一篇。

人们讨论过 RAE 是否对跨学科研究不利。正如在绪论中提到的,随着跨学科研究的不断增加,这是一个非常重要的问题。RAE 的运作机构很关注这个问题并撰写了一篇独立报告。该报告的主要结论如下:

·跨学科的工作确实很重要:差不多有 80% 的科学研究都是这种类型;

·有人认为 RAE 抑制了这种研究;

·然而从 1996 年的 RAE 的情况看,并没有证据支持这一主张;

·把研究评估交给特定学科领域的专家组在程序上存在一些困难。

但是这些问题在 2001 年通过以下方式得到了纠正:

·改变评议专家组的构成,纳入更广泛领域的专家;

·改变评估方法;

·增加对跨学科研究的报道。

从摘要中可以看出,在一个优化设计的系统中,跨学科研究不应该受到伤害。同单纯同行评议相比较,由更广泛范围的同行进行的改进同行评议应该能减轻对专门学科的偏见。其次,如果学科之间能相互补充,那么可以预期绩

效评估会随之增加。将外溢的成果内在化是高校义不容辞的责任。

2001 年进行的 RAE 中,自评估的科研质量呈现出了显著的提高。接受评价的中有 40% 的研究部门列在前两个评价等级中。下议院委员会对这一结果进行了评估以确定其真实性。该委员会得出以下结论:为了提高排名,评价中有人使用了"不光彩"的方法,然而 RAE 果真反映出了研究质量确实有所提高;主要通过改进研究管理并瞄准有研究价值的领域,且已经从中获益。下议院委员会指出为提高绩效向大学研究提供额外的资源是必要的。

在英国一份独立报告中我们看到以下结论:英国高等教育研究的表现在国际上是"十分有竞争力的",在过去 15 年中绩效表现日渐且明显提高;英国的研究无论以产出数量还是以质量来衡量都获得了确实值得的收益,研究绩效有显著改善;在案例研究中,研究人员发现,高校员工自 1986 年以来一直在谈论研究文化的变革,因为新的制度结构和程序被引入了科研管理。

所付出的代价是:把一些资源从资本项目中转移到提高研究绩效的需要中去,以满足研究经费的不足;增加工作负荷。

还有一些人指出了评估给接受评估的高校增加的管理负担。

英国一家学术机构对 RAE 的利弊作了一个概括,详见表 12。上述的一些证据似乎表明 RAE 的优点是明显提高了研究绩效。进而分析其消极的一面可知,表 12 中列出的一些 RAE 的缺点可能同设计有关而与评估的根本理念无关。例如,如果在两次评估之间有充分的时间,那么就不会有对稳妥研究的偏好。而且,并没有令人信服的证据证明教学与研究间的互补性,尽管人们希望存在这种可能。

表 12 分配研究经费的 RAE 系统的优缺点

优 点	缺 点
奖励好的研究	会引发游戏的态度
鼓励组建大的研究团队	会导致某些研究的过度集中,使得研究缺乏多样化
为提高个体研究绩效提供了激励	可能会导致进行无风险的研究
竞争系统——鼓励完成研究,去粗取精,更好的科研管理	鼓励优秀的研究人员流向评价得分更高的研究机构,但是这会减弱教学与研究之间的协同关系
可以把国家的优先政策和研究投资渠道联系起来	会导致政府的过度干涉
	奖励过去的研究绩效
	可能会阻拦新的研究
	高成本

来源:Hare(2002)

总之,从有关英国体系的多种讨论中可以得知,RAE 毫无疑问改善了研究产出。然而这种收益是建立在一定代价基础上的。直接的财政花费似乎并不太高,大约占研究总经费的1%。

然而,RAE 所带来的各研究机构为管理工作耗费时间所导致的机会成本可能会更大。

在《科学》上的一篇社论文章中,Jonathan Adams (2002)指出:英国的研究水平相对于世界标杆有所提高,扭转了 20 世纪 80 年代的回落态势。在 1986~2000 年间,根据出版物数量和引文数的数据,其平均研究绩效有了很大进步。RAE 为把自然的研究竞争引向一种广泛追求卓越的驱动力提供了激励和动力。它通过证明做出好的研究的人员并没什么可怕的反而可以从这种结构完善的审查制度中获益良多,从而克服了共同体内部最初的反对情绪。

英国通过全面评估和以此为依据来分配的大量资源,从而产生强力激励并因而提高研究水平。

起初这种激励结构并不一定在任何场合都适用。因此,其它仿效英国的国家采用的方法要么激励不足,要么只是把目标放在减小成本。由于改进同行评议系统的成本,澳大利亚对 RAE 的翻版最初主要以出版物数量为依据。尽管如此,其它国家还是在发展评估系统上做了大量工作,对此我们将会在下两节中进行讨论。

八、评估系统的演化

澳大利亚

澳大利亚在 20 世纪 90 年代早期引入了研究产出评估系统。初期出版物的数量被当作指标来使用。这一政策的结果可从以下文献计量数据中看到:在 20 世纪 90 年代,澳大利亚在世界主要期刊发表的论文份额从 2.2% 增加到了 2.8%。澳大利亚占世界 GDP 的 1.2% 左右。然而,不出所料的是,由于使用的是简单的出版物数据统计方法,用引文表征的出版物质量似乎下降了。

这一研究系统开始得到评估。尽管很多人指出了引入类似英国 RAE 的系统是需要成本的,但(例如)澳大利亚国立大学还是倡议引入改进的同行评议而宁愿废除指标法,因为这一评价系统:

·在研究经费的分配中提供了能够真实地反映出研究质量的机制而不是依赖数量这种替代品。RAE 系统中的学科委员会正是为了这一目标而建立的。

·政府高等教育研究投资更可能流向那些承担最高质量研究的大学。

·运行这种系统的高校,研究绩效指标不断增加,将会为高校研究的内部管理提供更清晰的研究管理信息,这同样也是向外界报告和交流的依据。

·这些指标所代表的绩效也会为高校提供战略营销资料;

·小型高校中的优秀研究中心,更有可能因其研究绩效而得到财政支持;

·为相关领域的专家进行研究绩效评价提供了一种新的机制。

丹麦

评估不是一种强加且无限期保持原状的静态过程。在丹麦,最初引入评价的动机来自于20世纪80年代早期的政治家。引入评估的第一阶段遵循一种自上而下的政府评价路径。丹麦政府参照瑞典的模式,于1983年引入了改进同行评议,纳入了更多的参与对象而不仅限于科学家。经费通过改进同行评议在研究中心之间进行分配。如今,丹麦高校邀请了新的外部用户参加研究和教育咨询委员会。

绩效指标于1981年引入丹麦,但直到1989年教育部采纳这种评价机制以前一直是不为教育部认可的。到了90年代中期,各研究实体自身也开始了评估。当时的评估是基于学科的,比如,社会科学在1997年进行评估。自从新兴的研究所引入了各种评估方法之后,评估成为一种强有力的合法手段。更多高校也启动了评估工作。现在每个院系都要进行年度的绩效考评。

2000年,出现了一种向"发展合同"的转变。"发展合同"即每所高校和荷兰研究部签署协议。这些合同主要针对研究产出的成功标准和测度。现在,一定比例的研究经费按照外部经费数额分配给高校。

然而,高校和教育主管部门并没有妥协,1995年,引入英国评估体系的建议被遭到否决。然而该体系目前似乎也在开放地面对变革:"随着高校以更加开放的姿态面向世界,参与更新的合作并同外部团体有更多的接触,我们必须看到,科研质量和生产力的评价将会出现更加崭新的形式,以更深刻地映射出外部合理的需求以及研究的需要……一个向公众和股东提高开放程度和透明度的例子是,高校建立公共研究数据库,让公民能够搜索本校的活动信息,从而可以从学校院系中找到相关合作伙伴或获得更详细的研究成果。"

一个专门针对产出(毕业生数量)的计量器被开发并已用于教育基金的分配。

荷兰

1988年荷兰建立了教育评估体系在1993年进行研究评估。荷兰的研究机构每3年进行一次自评估。其目的一是为了准备外部评价,二是作为一种内部中期评估手段(荷兰每所高校每6年要接受一次外部评估)。外部评估由国际同行委员会进行,同时也会使用文献计量学工具。

该评价体系在研究和研究管理方面希望实现以下三个目标:

·以国际上评价研究质量和相关性的标准为依据,通过评估工作提高研

究的质量；

·提高研究领导能力和管理水平；

·向更高层的研究组织、投资机构、政府及社会“交帐”。

荷兰评价工作的效果：

·加强了参与同一研究项目的研究人员之间的合作；

·增加了论文发表数量，尤其是在高影响因子的国际期刊上发表的论文；

·高校管理者有更多的权力（以牺牲院、系领导的权力为代价）；

·评估作为一种工具为管理者进行质量控制提供了可靠的基础；

·研究政策的重要性得到提升；

·得到较高评价的研究者会获得更高的声誉，研究人员的威望更高了；

·公开发布的报告“使得较差和低产的研究团体不可能再安然存在而不为人发觉”。

但是可能会导致：

·更加正统的研究；

·更注重用英语进行研究；

·忽略教学和研究的并重。

荷兰的评价一直主要着眼于研究计划的管理，较少用于“交帐”和经费分配目的。

瑞士

一份于1999年通过的联邦法案指出“对组织结构更明确的界定将趋于以产出为导向的资助金和有效的评价系统”。科学技术研究中心（CEST）对受到联邦政府资助的研究项目进行了评估，包括所有政府部门自己的研究，特别是在科学领域。为了保证其独立性，CEST由联邦内务部和联邦经济部资助。CEST的目标包括：

·对瑞士科学研究和技术发展的强势弱势、潜力以及发展前景作经验诊断性的比较分析和评估；

·对包括联邦投资在内的整个瑞士科学体系的效率和效果进行评估；

·在瑞士科学体制中发展评估文化。

九、结论

加拿大的研究似乎还不错，但目前距离“卓越”尚有差距。然而，这种解读可能源自贫乏的数据。通过可获得的引文数和论文数的数据可以看出，加拿大的平均水平相对于其他国家来说是很不错的。但从质量分布高端的排名来看，加拿大的情况似乎并不是很理想。

研究产出是资源分配和宏观经济激励机制的一种功能。仅仅增加资源的

投入并不一定会增加研究的产出。本文所讨论的是世界范围内使用的一种重要工具,即研究产出评估方法的使用和演化。很多在评价研究产出方面表现较好的国家都有自己评价研究产出的方法工具。引入这些评估系统的国家都通过更多的激励来巩固他们的体系。这些国家的政策制定者必定相信,通过评估研究产出,提高对高质量研究产出的激励,一定会得到切实的收益。

评估作为激励机制是很重要的,但从一个更基本的层面看,评估提供一国研究活动的相关数据。如果没有透明而且客观的方法来考察科研活动,就难以确定该研究系统是否运行良好,应该在何处以及如何改进该研究系统。

(彭颢舒 译)

马普学会新建6个国际研究院

对德国未来的研究与创新而言,培养科学后备力量具有决定性意义。所以,在本世纪初,马普学会与德国大学校长会议(代表全德国的大学和高等院校)就拟定了培养科学后备力量的新计划——即关于创办"马普国际研究院"的计划。实践表明,这一计划是非常成功的。2006年,又将有6个新的马普国际研究院开始运营。

通过马普国际研究院、马普研究所和德国的以及部分国外的大学和其他公共研究机构建立起一种更为紧密的合作伙伴关系。依托高等院校杰出的科学中心,马普国际研究院设立了诸如分子生物学、神经科学、信息学、人口统计学、等离子体物理学或者聚合研究等一批崭新的教学与研究领域,具备了国际最高水平的教研条件和环境,得以为德国乃至其他国家的后备科学家提供唯一的定向教育与科研实践培训的机会。学业结束后,所有在马普国际研究院就读的德国及外国的博士生,都能到马普研究所、邻近的大学和高等院校以及科研机构参加科研实践并准备其博士学位论文。

由于马普研究所没有授予博士学位的法定权利,所以,马普国际研究院的博士生可选择参加德国或者国外的某一所大学(一般是马普国际研究院的协办大学)的毕业考试。在世界各地的高校进行马普国际研究院的招标时,马普研究所会根据他们对当前研究领域的技能和能力挑选博士研究生。入选的后备科学家——即马普国际研究院的博士生——将得到一份特殊的教学计划,内容包括大学授课、研究班和科学工作会议等,并与马普研究所正在进行的科研计划捆绑在一起。同时,作为独立的授予博士学位计划,还必须及时反映与当前意义重大的科学项目有关的各种理论与实验方面的问题,尤其是学科交叉研究项目或者需要特别配备研究仪器或材料项目以及与之相关的科学论