关于科研指标的莱顿宣言

Diana Hicks、Paul Wouters 及其同事督促用十项原则来规范科研评估。

科学治理日益依赖于数据。建立在量化指标基础上科研评估已经取代曾经的同行评议成为主流[1]。随之而来的问题是，今日的评估已由数据而非判断主导。量化指标日益流行：通常是精心设计的，但并非总是被透彻理解，而且经常被错误地使用。主持科研评估的机构往往缺乏对于这些量化指标的透彻理解，因此这些指标虽然意在促进科学研究却经常适得其反。

2000 年之前，美国科学信息研究所（ISI）的科学引文索引（SCI）已为专家们所使用。2002 年，汤森路透公司（Thomson Reuters）整合其网络平台，使其 Web of Science 数据库更加普及。竞争者也随之而来，包括爱思唯尔（Elsevier）的 Scopus（2004 年发布）和 Google 学术搜索（Google Scholar，beta 版 2004 年发布）。相关分析工具也如雨后春笋，使比较学术机构以及个人的研究产出和影响更为容易，例如基于 Web of Science 的 InCites、基于 Scopus 的 SciVal、以及基于 Google Scholar 的 Publish or Perish 软件（2007 年发布）。

2005 年，美国加州大学圣地亚哥分校的物理学家 Jorge Hirsch 提出了 H 指数，使得被引次数更为广泛地被用于考量学者的科研。自 1995 年起，期刊影响因子也日益盛行。

最近，关于社会使用和在线评论的量化指标日渐成势，比如 F1000Prime（2000）、Mendeley（2008）、和 Altmetric.com（2011，由麦克米伦集团支持，而《自然》所属的自然出版集团亦为麦克米伦旗下公司）。

作为文献计量学者，社会科学家，以及科研管理者，我们目睹了在科研评估中量化指标被愈发广泛和严重地滥用，以下仅举数例。各国的大学日益执迷于其在各大高校排名中的位置（如上海交通的世界大学学术排名和泰晤士高等教育世界大学排名），尽管很多排名在我们看来是建立在并不精确的数据和非常武断的指标的基础之上。

一些招聘者使用 H 指数来考察候选人。一些大学依靠 H 指数以及发表在高影响因子期刊上的论文的数量来决定科研人员的晋升与否。学者们，尤其在生物医药领域，在简历中夸耀他们的 H 指数或者影响因子。教授们要求博士生在高影响因子的期刊上发表论文和申请科研经费，尽管他们还没有准备好。

在北欧和中国，一些大学根据学者个人的影响指数来分配科研经费，或者为发表在高于 15 的影响因子的期刊上的论文提供资金奖励[2]。

虽然在很多情况下研究和评估人员还是会做出相对平衡的评议，但科研指标的滥用已经到了不容忽视的地步。

因此，我们提出莱顿宣言，源于在荷兰莱顿举行的一次国际会议（参见 http://sti2014.cwts.nl）。我们所提出的十大原则对于文献计量学者而言并非前所未闻，尽管我们当中没有人可以完整地罗列出这些原则，因为我们至今没有一个系统的成文阐述。我们这一领域的启蒙者，比如 ISI 的创立者 Eugene Garfield，曾提到过这十大原则中的某些[3, 4]，但他们并未为科研评估和管理人员所知晓。同时，被评估的科学家们试图寻找相关的文献来驳斥某些评估结果，而这对于他们而言犹如大海捞针。

我们在此提出十大原则，凝练了基于指标的科研评估的规范。借此被评估者可以问责评估者，而评估者可以规范使用量化指标。

十大原则

1：量化的评估应当支持而非取代质化的专家评审。量化指标可以降低同行评议中的偏见并促进更为深入的审议。量化指标可以提高同行评议的质量，因为在没有充足信息的情况下评价别人是非常困难的。但是评估者的判断不应让位于数字。量化指标不应取代建立在充分信息基础之上的判断。评估者仍应对其评估负责。

2：科研绩效的考量应基于机构、团队、以及个人的科研使命。应当首先明确评估的目标，而所采用的指标也应切合这些目标。同时，指标的选择和应用的方式应该考虑更为广泛的社会、经济、文化环境。科学家有着各色各样的科研使命，着眼于探索未知的尖端基础研究和立足于解决社会问题的应用研究有着截然不同的任务。在某些情况下，评估者应该考虑研究的社会和经济价值而非其科学价值。世上没有一个评估方法适用于所有的情况。

3：保护卓越的本地化研究。在很多地方，研究的卓越等同于在国际期刊上发表英文论文。比如，西班牙法律明文鼓励发表于高影响力的英文期刊的论文。然而期刊影响因子所依赖的 Web of Science 数据库主要是以美国和英文期刊为主。这一数据库覆盖期刊的偏差对于社会和人文学科造成了尤为严重的后果，而在这些领域很多研究是关于本国或者当地的课题。在很多其他的领域也有偏重于本地化的题目，比如撒哈拉以南非洲的 HIV 流行病学。

这些本地化的课题往往并不为高影响因子的英文期刊所青睐。那些在 Web of Science 数据库中取得较高引用率的西班牙社会学家往往从事于抽象模型或者分析美国数据。西班牙语期刊的论文则通常关注更为相关的本地课题：本地劳动法，老年人家庭医疗，以及外来劳工等等[5]。只有基于高质量本地语言期刊的指标才能正确评价和推动卓越的本地化研究。

4：数据采集和分析过程应公开、透明、简单。数据库的建立应该遵循明确的规则，而这些规则应在评估之前就清晰阐述。这是以往数十年来相关学术单位和商业机构的惯例。而他们的数据处理的流程也发表在同行评议的文献中。这样透明的流程保证了复查的可能性。比如 2010 年荷兰莱顿大学科学技术研究中心（CWTS）所创建的一项指标引发了一场学

术争论，而这一指标随后被修改[6]。这一领域的新进机构也应遵守此标准。我们不能接受评估中的暗箱操作。

对于指标而言，简单就是美，因为简单增强透明性。但简单化的指标也可能会导致偏颇的结论（参见原则 7）。因此评估者应竭力保持平衡，采用的指标应足够简单明了但又不会曲解复杂的问题。

5：允许被评估者检验相关数据和分析。为保证数据质量，所有的被评估者应当有机会查证评估所用的数据是否准确全面地包括了他们的相关研究产出。评估者则应通过自行验证或者第三方审查来确保数据的准确性。大学可以在他们的科研信息系统中执行这一原则，并以此作为一项重要标准来选择信息系统提供商。精确和高质量的数据耗费时间和经费去搜集和处理，因此需要足够的预算。

6：考虑发表和引用的学科差异。最好能提供一套指标让不同的领域各取所需。几年前，一组欧洲的历史学家在全国的评审中得到了较差的结果，因为他们出版书籍而不是在被 Web of Science 索引的期刊中发表论文，另外他们不幸被划在了心理学系。历史学家和社会科学家往往要求学术评审考虑书籍和本国语言的论文，而计算机科学家则往往要求加入会议论文。

不同领域的引用率也有差别：数学期刊的最高的影响因子大概是 3，细胞生物学却高达 30。因而相关指标需要根据学科来标准化，最可靠的学科标准化方法是通过百分位数：每一篇论文的得分取决于其在整个学科的被引次数分布中的位置（比如说最高的 1%，10%，或者 20%）。在使用百分位数方法时，个别极其高被引的论文将略微地提高其大学的排名，但在使用被引次数均值时却可能会将其大学的排名从中等拔到顶级[7]。

7：对于学者个人的评估应基于对其整个作品辑的质化的评判。年龄越大，H 指数越高，即使是在没有新论文发表的情况下。H 指数在不同的领域也有所不同：生命科学家可高达 200，物理学家最高 100，而社会学家最多只有 20 到 30[8]。这同时也取决于数据库：有些计算机科学家在 Web of Science 中的 H 指数只有 10，但在 Google Scholar 中却有 20 到 30[9]。研读和评判一位学者的论文要远比仅仅依靠一个数字合适。即使在比较很多学者时，能够综合考虑多方面的信息更为适宜，比如个人专长、经验、活动、影响等等。

8：避免不当的具体性和虚假的精确性。科技指标不可避免会在概念上有些模糊和不确定，并且建立在一些很强但并不普适的假设的基础之上。比如说，对于被引次数到底代表了什么这一问题就存在很大的争议。因此最好能使用多个指标来提供一个更为可靠和多元的呈现。如果不确定性和潜在错误可以被量化，那么应该在发表指标结果的同时提供置信区。如若潜在错误率不可量化，那么研究人员至少不应盲目追求精确度。比如，官方发表的期刊影响因子精确到小数点后三位数，这样可以避免期刊之间打成平手。但考虑到被引次数

所存在的概念上的模糊性和随机误差，实在没有必要在相差不大的期刊之间分个伯仲。在此情形下，避免虚假的精确度意味着精确到小数点后一位就已经足够了。

9：认清科技指标对科研系统的影响。科技指标改变研究人员的动机进而改变整个科研系统，对这样的结果我们应有充分的预期。这意味着一套指标总胜于单个指标，因为单个指标更易于被操纵，也更容易取代真正的目标成为驱动研究的指挥棒。举例来说，在 90 年代，澳大利亚政府根据各高校的论文数量来分配经费，而大学可以估算出一篇论文的经济价值：在 2000 年一篇论文大约可以换来 900 澳元（折合 450 美元）的经费。可以预料的是澳大利亚的高校发表论文数据显著增加，但多发表于低被引的期刊，意味着论文质量的下降[10]。

10：定期审查指标并更新。研究的使命和评估的目标会随着时间而改变，科研体系也不停在变化演进。曾经有用的指标可能会变得不那么合适，而新的指标也会不停出现。指标体系也应随之调整。意识到不良后果后，澳大利亚政府在 2010 年推出了更为复杂的科研评估体系，而这一体系更为重视科研质量。

以后的路

遵循这十项原则，科研评估将在推动科学发展和社会进步方面发挥更为重要作用。科研指标可以提供非常有价值的信息，但我们应谨记指标只是工具，不是目标。

为作出最好的决定，我们同时需要可靠的统计数据和对研究对象的深入了解。量化和质化的证据二者不可或缺，并且这二者都是客观的。科学决策必须建立在高质量的评估过程和充分并可靠的数据的基础之上。

作者：Diana Hicks 是佐治亚理工公共政策学院教授。Paul Wouters 是荷兰莱顿大学科学技术研究中心（CWTS）的教授兼主任，Ludo Waltman 是该中心的研究员，Sarah de Rijcke 是该中心的助理教授。Ismael Rafols 是西班牙国家研究委员会和瓦伦西亚理工大学的科学政策研究员。

翻译：王健，比利时鲁汶大学研发监测中心（ECOOM）博士后。（Translator: Jian Wang, postdoctoral fellow at the Center for R&D Monitoring (ECOOM) at the University of Leuven.）

参考文献：

1. Wouters, P. in Beyond Bibliometrics: Harnessing Multidimensional Indicators of Scholarly Impact (eds Cronin, B. & Sugimoto, C.) 47–66 (MIT Press, 2014).
2. Shao, J. & Shen, H. Learned Publishing 24, 95–97 (2011).
3. Seglen, P. O. Br. Med. J. 314, 498–502 (1997).
4. Garfield, E. J. Am. Med. Assoc. 295, 90–93 (2006).
5. López Piñeiro, C. & Hicks, D. Res. Eval. 24, 78–89 (2015).
6. van Raan, A. F. J., van Leeuwen, T. N., Visser, M. S., van Eck, N. J. & Waltman, L. J. Informetrics 4, 431–435 (2010).
7. Waltman, L. et al. J. Am. Soc. Inf. Sci. Technol. 63, 2419–2432 (2012).
8. H irsch, J. E. Proc. Natl Acad. Sci. USA 102, 16569–16572 (2005).
9. B ar-Ilan, J. Scientometrics 74, 257–271 (2007).
10. Butler, L. Res. Policy 32, 143–155 (2003).