# 科技评价中同行评议与指标体系关系的研究

## ——以《泰晤士报》世界大学排名为例

俞立平　潘云涛　武夷山

摘要：为了研究同行评议与指标体系评价结果之间的关系，本文基于《泰晤士报高等教育增刊》2007 年世界大学排名数据，利用多元回归分析和 Kappa 一致性检验法进行了研究。研究发现，数据丰富程度影响同行评议与指标评价的一致性；在数据不足的情况下，辅以同行评议进行综合评价是较优的选择；可用回归拟合优度对数据丰富程度进行判断；同行评议也有其适用范围，权威的同行评议可以作为指标体系权重赋值以及指标选取的重要依据；在数据较为丰富的情况下可以向同行评议专家提供原始数据。

关键词：科技评价 大学排名 同行评议 指标体系 一致性检验

中图分类号：G31

文献标识码：A

## 1 引言

同行评议是现代科学进程中的重要环节（KOSTOFF,1997）[1]，也是科研项目立项与科研经费分配、科技论文录用、科研人才选拔中的重要方法（JAYASINGHE et al.,2003）[2]，同行评议肩负着科学“看门人”的地位。由于评价的目的和对象不同，同行评议的地位与作用并不一致，在某些基金项目的立项中，同行评议起着举足轻重的作用。

然而同行评议也存在着诸多问题，因此在某些科技评价中开始引入指标体系与同行评议相结合的评价方式。H.F. Moed, W.J.M. Burger等（1985）[3]在对莱顿大学化学与生物学的科研机构的评价中，比较了文献计量学指标与同行评议结果的相关性，发现大部分文献计量学指标排序结果与同行评议结果不相关，从而得出结论说，文献计量学指标不能作为学术质量的评价标准。Lutz Bornmann, Hans-Dieter Daniel(2005)[4]利用德国勃林格殷格翰基金招收博士及博士后研究人员时收集的数据，利用多元回归法分析了同行评议的可靠性、公正性与预见性，发现在博士后录用中没有任何不公正问题，但在博士研究生录用中存在机构、性别、学科的歧视倾向。么大中、张淑芳等（2004）[5]认为在社会科学成果的评价中，应建立以同行评议为主，间接指标体系为辅的评价机制。总体上，迄今对同行评议与指标体系评价结果的关系开展研究的不多。

本文拟采用英国《泰晤士报高等教育增刊》2007年世界大学排名数据进行研究[6]，从1986年开始，《泰晤士报高等教育增刊》每年推出世界大学排名，将同行评议与指标体系结合来进行大学排名综合评价。从2007年开始，评价所利用的论文数据库已经由Elsevier的Scopus数据库取代了美国的SCI数据库，Scopus覆盖面更广，包括了许多非英语优秀科技期刊。《泰晤士报高等教育增刊》世界大学排名已经开展多年，积累了丰富的经验，数据相对客观可靠。

本文首先采用多元回归法进行指标体系总体状况的判断，然后进行同行评议与相关指标的一致性检验，最后分析同行评议结果与其他单个指标的关系，试图分析其中的深层次的原因。

## 2研究方法

### 2.1 Kappa一致性检验

对同一事物采用多种评价方法加以评价，然后再判断各种评价结果之间的关系，可以采用回归分析法，但该方法对数据要求较高，数据过少或非连续数据难以处理。卡方检验对于复杂的评价处理方法相对单调，容易丢失数据中的重要信息，因此本文采用Kappa一致性检

验法。

1960 年，Cohen[7]等提出使用 Kappa 值作为判断评价一致性程度的指标，目前得到了广泛的应用，已成为判断一致性和信度评价的一种常用的统计学重要指标。在科学计量学领域，应用 Kappa 一致性检验进行同行评议与其他指标一致性的也未见报道。Kappa 计算方法如下：

$$K_{appa} = \frac{P_A - P_e}{1 - P_e} \qquad (1)$$

其中，$P_A$ 为观测一致率，$P_e$ 为期望一致率，即两次检验结果由于偶然机会所造成的一致率。显然 Kappa 值在 0～1 之间，若 Kappa 值较大说明一致性较好；若 Kappa=1，说明检验结果完全一致。一般说来，若 Kappa>0.75，说明已取得非常满意的一致程度；若 Kappa 值在 0.4～0.75 之间，说明已取得比较满意的一致程度；若小于 0.4，说明一致程度较差。

利用 Kappa 一致性检验可以比较同行评议与指标体系综合评价结果之间的关系。

**2.2 多元回归分析**

多元回归分析是经济计量学中的常用方法，用来分析自变量对因变量的影响及其大小，其一般形式是：

$$Y = a_0 + a_1 X_1 + a_2 X_2 + \cdots\cdots + a_n X_n$$

Y 为因变量，$a_0$ 为常数，$X_1$……$X_n$ 为自变量，$a_1$ 按……$a_n$ 为系数。根据数据的统计特征和具体情况，有时也可以采用对数形式。

在本文中，利用多元回归来分析大学排名的同行评议得分与其他指标之间的关系。

## 3变量与数据

本文采用数据为英国《泰晤士报》高等教育增刊2007年世界大学排名数据，该大学排名所依据的指标有：同行评议、雇主评价（对毕业生）、教职员/学生比例、教职员人均论文引用次数、教职员队伍国际化情况、留学生人数，上述指标权重分别为0.4、0.1、0.2、0.2、0.05、0.05，所有单个指标的分值最高均为100。为了分析同行评议结果与其他指标间的一致性关系，本文将指标体系总得分减去同行评议分值，得到纯指标得分，然后再折算成满分100。计算公式为：

纯指标得分=（总指标得分—同行评议得分×0.4）×100/60 （2）

这样，纯指标得分的最高值也为100，从而使数据标准化。有关数据及描述统计量见表1。

表1 变量及描述统计量

| 变量名称 | 变量含义 | 权重 | 均值 | 最大值 | 最小值 | 标准差 |
|---|---|---|---|---|---|---|
| PR | 同行评议<br>Peer review | 0.4 | 72.94 | 100 | 31.00 | 18.71 |
| ER | 雇主评价<br>Employer review | 0.1 | 70.12 | 100 | 5.00 | 24.85 |
| SS | 教职员学生比<br>Staff/student | 0.2 | 62.22 | 100 | 11.00 | 25.81 |
| CS | 教职员人均论文引用次数<br>Citations/staff | 0.2 | 75.36 | 100 | 1.00 | 15.35 |
| IS | 教职员队伍国际化情况<br>International staff | 0.05 | 59.06 | 100 | 13.00 | 27.82 |
| SA | 留学生人数<br>International students | 0.05 | 61.42 | 100 | 11.00 | 26.59 |
| OS | 指标体系总得分<br>（包括同行评议） | | 70.96 | 100 | 54.80 | 11.81 |

| | | | | | | |
|---|---|---|---|---|---|---|
| PS | Overall score<br>纯指标得分<br>（不包括同行评议）<br>Pure score | | 69.55 | 100 | 42.83 | 12.04 |
| n | 大学数 | 200 | | | | |

数据来源：http://www.thes.co.uk/

# 4实证结果

## 4.1指标体系总体评价

用指标体系总得分作为因变量，用同行评议、雇主评价、教职员/学生比、教职员人均论文引用数、教职员队伍国际化情况、留学生人数作为自变量进行回归，结果如下：

$$OS=0.407PR+0.102ER+0.203SS+0.204CS+0.050IS+0.051SA$$

$$(602.009^{***})\ (202.429^{***})\ (499.031^{***})(352.357^{***})(105.739^{***})(95.432^{***})R^2=0.9998$$

从回归结果看，所有系数值与权重值基本吻合，每个变量的 t 统计量都较大，全部在 1%的水平上通过了统计检验，$R^2$ 等于 0.9998，拥有极高的拟合优度。单纯从指标体系的总体状况看，具有令人满意的结果。

## 4.2 同行评议结果与指标体系总得分的一致性检验

200 所大学的同行评议结果与指标体系总得分的关系如图 1 所示，从图中可以看出，同行评议结果与指标体系总得分之间似乎有着一定的线性关系。

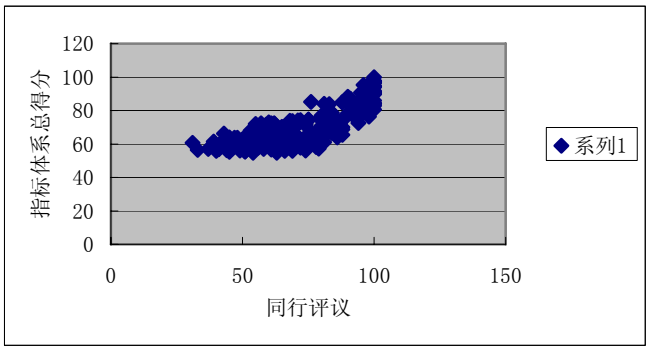


图 1 同行评议与指标体系总得分散点图

由于一些指标是人为判断的，如同行评议、雇主评价，虽然这些数据是连续的，但人的判断能力是有限的，谁也不能肯定某项指标 71 分与 72 分就一定存在多大差别，在这种情况下，采取分级数据更为精确。因此，本文将各指标分值按 10 分为一档分为 7 档（由于没有 30 分以下的低分），然后分别计算每个分值段大学的个数，由于指标体系总得分与同行评议结果不一定一致，因此用一张二维表表示，结果如表 2 所示。

表 2 同行评议与指标体系总得分比较

| | | 指标体系总得分 | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | 90- | 80- | 70- | 60- | 50- | 40- | 30- | 合计 |
| 同 | 90- | **21** | 19 | 8 | | | | | 48 |
| 行 | 80- | | **5** | 19 | 9 | | | | 33 |
| 评 | 70- | | 1 | **8** | 14 | 5 | | | 28 |
| 议 | 60- | | | 10 | **18** | 11 | | | 39 |
| | 50- | | | 3 | 13 | **10** | | | 26 |
| | 40- | | | | 10 | 12 | | | 22 |

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| 30- | | | | 2 | 2 | | 4 |
| 合计 | 21 | 25 | 48 | 66 | 40 | | 200 |

从表 2 可以看出，同行评议结果与指标体系总得分相差较大，如有 8 所大学同行评议在 90 分以上，但指标体系总得分仅在 70～80 分之间。同行评议结果与指标体系总得分完全一致的共有 21+5+8+18+10=62，占 200 所大学的 31%，比例较低。

利用 SPSS13.0 进行 Kappa 检验，Kappa 值为 0.169，概率值 P=0.000，在 1%的水平上通过了统计检验，说明同行评议结果与指标体系综合评价结果一致性极低，这和散点图分析结果基本一致。

**4.3 同行评议结果与纯指标得分的一致性检验**

同行评议结果与纯指标得分的关系如图 2 所示，从图中可以看出，同行评议结果与纯指标得分之间不相关。

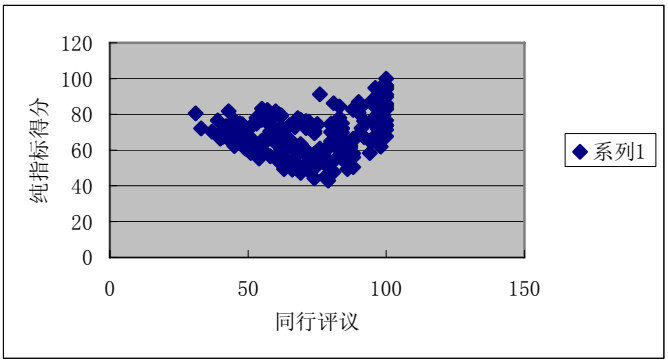


图 2 同行评议与纯指标得分散点图

同样将各指标分值按 10 分为一档分为 7 档（由于没有 30 分以下的低分），然后分别计算每个分值段大学的个数，结果如表 3 所示。

表 3 同行评议与纯指标得分的比较

| | | 纯指标得分 | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | 90- | 80- | 70- | 60- | 50- | 40- | 30- | 合计 |
| | 90- | **13** | 13 | 14 | 6 | 1 | | | 47 |
| 同 | 80- | | **4** | 6 | 12 | 7 | 2 | | 31 |
| 行 | 70- | 1 | | **7** | 4 | 13 | 4 | | 29 |
| 评 | 60- | | 3 | 10 | **12** | 11 | 3 | | 39 |
| 议 | 50- | | 3 | 10 | 10 | **5** | | | 28 |
| | 40- | | 1 | 8 | 13 | | | | 22 |
| | 30- | | 2 | 2 | | | | | 4 |
| | 合计 | 14 | 26 | 57 | 57 | 37 | 9 | | 200 |

从表 3 可以看出，同行评议结果与纯指标得分结果相差更大，如有 2 所大学同行评议在 80 分以上，但纯指标得分仅在 40～50 分之间，同行评议结果与指标体系总得分完全一致共有 13+4+7+12+5=37，占 200 所大学的 16.5%，比例相当低。

利用 SPSS13.0 进行 Kappa 检验，Kappa 值为 0.049，概率值 P=0.107，也就是说，在 5%的水平上没有通过统计检验，结论是同行评议结果与纯指标得分结果没有一致性，这和散点图结果基本一致。

**4.4 同行评议结果与其他指标的相关性分析**

为了进一步挖掘和分析同行评议结果与其他指标的关系，将同行评议得分作为因变量，雇主评价、教职员/学生比、教职员人均论文引用数、教职员队伍国际化情况、留学生人数

作为自变量进行回归，发现只有雇主评价及教职员论文人均引用数两个变量通过了统计检验（回归 1），采用变量逐步删除法进行调整，结果如表 4 的回归 2 所示。

表 4 回归结果

| 自变量 | 回归 1 | 回归 2 |
|---|---|---|
| C | 22.739*** | 20.960*** |
| | (2.901) | (3.161) |
| ER | 0.330*** | 0.321*** |
| | (6.869) | (7.018) |
| SS | -0.007 | |
| | -0.161 | -- |
| CS | 0.388*** | 0.392*** |
| | (5.146) | (5.291) |
| IS | 0.002 | |
| | (0.046) | -- |
| SA | -0.030 | |
| | (-0.543) | -- |
| $R^2$ | 0.279 | 0.277 |

同行评议结果与雇主评价及教职员论文人均引用数相关，在 1%的水平上通过了统计检验。同行评议与教职员/学生比、教职员队伍国际化情况、留学生人数无关，$R^2$ 值为 0.277，处于较低水平，这说明同行评议提供了一些指标体系所不能提供的额外信息。从另一角度分析，雇主评价主要反映了大学的总体教学水平，教职员论文人均引用数反映了大学的科研水平。雇主对新员工素质的直观认识是比较准确的，同行评议专家即使不知道某校的人均论文引用数，也能较准确判断该校的科研水平。因此这两个指标与同行评议结果是相关的。由于同行评议专家对数据掌握的不充分，因此难以掌握教职员学生比情况，对教职员队伍国际化情况和留学生人数两项国际化指标也缺乏数据，导致这三个指标与同行评议结果没有相关性。

# 5 讨论与结论

### 5.1 数据丰富程度影响同行评议与指标评价的一致性

由于科技评价的对象不同，数据获取的难度也不一样。对于一些基础工作较好，数据搜集容易，指标体系完备的评价场合，单单指标体系就能提供丰富的信息，那么，在权重设置相对合理、同行评议相对客观的情况下，同行评议与指标体系评价的结果应该是应该具有一致性的。就世界高校排名而言，由于国家众多，数据搜集不便，论文写作语言不同，在这样的情况下，纯指标得分难以和同行评议结果相一致，当然，由于评价的复杂性，在多数场合，同行评议与指标体系总得分不会完全一致。

### 5.2 在数据不足或不准确的情况下，辅以同行评议进行综合评价是较优的选择

既然数据获取困难，单纯利用指标体系进行评价必然存在缺陷，在这种情况下，将同行评议结果作为一项指标，与其他指标一起构成指标体系进行综合评价，就可以弥补数据不足或不准确的缺陷。这样的评价，能提供较为全面的信息，取得较好的评价效果。

### 5.3 可用回归拟合优度对数据丰富程度进行判断

那么，如何得知数据的丰富程度呢？在条件许可的情况下，只要进行较为权威的同行评价就可以进行判断，方法是将同行评议结果作为因变量，现有指标作为自变量进行回归，重点分析 $R^2$ 值大小，如果 $R^2$ 较小，则可以肯定数据不够丰富，但是，如果 $R^2$ 较大，说明基础数据比较丰富且准确，此时就没有必要将同行评议结果作为一项指标放到指标体系中去，这样可以大大降低评估成本，提高评估的客观性。一般情况下，$R^2>0.5$ 时就要慎重考虑是否有必要保留同行评议结果指标。

### 5.4 多元回归同样可以作为指标选取的依据之一

采用上述多元回归在某些情况下还可以进行指标选取，如果指标间相关程度较低，完全可以根据科技评价理论与统计检验结果进行多余指标的剔除。不过删除指标时一定要慎重，如果从理论上分析，该指标确实重要，那么是不能轻易删除的，比如评价大学水平如果采用获诺贝尔奖金人数作为一项指标，在一些国家如果获奖人数极少，那么该指标回归系数肯定不显著，但该指标肯定不能轻易删除。

**5.5 权威、公正的同行评议可以作为评价指标权重赋值的重要依据**

利用数据丰富、准确程度的判断方法，当 $R^2$ 较大时，多元回归中各指标系数就是指标体系权重的重要依据。在评价对象相对固定，指标体系中单个指标的数量变动相对不大的情况下，经过若干轮的实践，是可以比较客观地得到指标体系中各指标的权重的。此时，可以考虑不进行同行评议，完全利用指标体系直接进行评价。当然，即使采用较合理的指标体系进行评价，每隔一段时间也应该采用同行评议对其评价结果进行验证或修正。

迄今为止的已有几十种指标权重赋值方法，各种权重赋值方法结果与回归系数赋权很难一致。在回归拟合优度较高时，可以对同行评议与指标体系评价结果进行 Kappa 一致性检验，若结果基本相同，则以后的评价无论采取哪种方式都是可以的。若结果不同，建议以后的评价以指标体系为准。

**5.6 同行评议在某些情况下不可替代**

如上所述，在相对成熟的评价领域，如科研机构中某些学科的评价，可以不采用同行评议。但是对于国家自然科学基金申请、科技期刊论文评价等领域，同行评议可能是最主要甚至是惟一的方法。

**5.7 向同行评议专家提供有关数据作为决策参考的必要性**

那么要不要向所有的同行评议专家提供基础数据呢？首先要对指标体系进行分析，如果指标体系本身数据比较完备，能反映完全信息，此时应向同行评议专家提供相关数据，以避免专家由于数据缺失而造成的偏见。如果基础数据本来就较少，此时不提供基础数据给同行评议专家可能反而是较好的选择，让专家凭直觉进行判断即可。

**5.8 同行评议质量的判断**

如果经过理论分析和数据稽核，可以肯定指标体系的完备性和数据丰富，此时，若回归结果 $R^2$ 值较低，再结合 Kappa 一致性检验，发现同行评议与指标体系不一致，此时要从同行评议的各环节加以深入分析，如有没有明确评价目的和标准？同行评议专家的选取是否客观公正？同行评议程序是否规范等等？毕竟在同行评议中存在着各种利益相关与利益冲突问题。

**5.9 同行评议的选择问题**

在数据较为丰富的情况下，有没有必要进行同行评议呢？进一步说，如果同行评议的结果相对客观公正，那还要指标体系干吗？问题是专家也是人，他无法精确确定评价结果，比如，你让专家给个优、良、中之类的分级评价，是可以接受的。如果让专家打分，精确到97、96、93，那么专家自己也不相信这个结果就一定可信。此外，不同批次专家的评价结果可能不一致，比如，这次 30 各专家有一个结果，换另外 30 个专家可能结果不一致，可靠性低。也就是说，在数据相对丰富的情况下，同行评议的结果不如指标体系稳定。因此，采用指标体系进行评价，可以较低成本，而且具有一定的客观性。

**5.10 在原始数据较少的情况下不宜采用本文的方法**

采用回归法进行分析是有一定的前提条件的，就是数据记录的个数，必须保证一定的自由度，在评价指标较多，数据记录较少的情况下，是不能进行回归分析的。

**参考文献**

[1]Kostoff,R.N. The priciples and practices of peer review. Science and Engineering Ethics[J], 1997(3):19-34

[2]Jayasinghe,U.W., Marsh,H.W., Bond,N. A multilevel cross-classified modeling approach to peer review of grant proposals: the effects of assessor and researcher attributes on assessor ratings[J]. Journal of the royal statistical society series A-statistical in society, 2003(166):279-300

[3]H.F. Moed,W.J.M. Burger,J.G. Frankfort, A.F.G. Van Raan. A comparative study of bibliometric past performance analysis and peer judgement[J]. Scientometrics, Vol. 8. Nos 3-4(1985):149-159.

[4] Lutz Bornmann, Hans-Dieter Daniel. Selection of research fellowship recipients by committee peer review. Reliability, fairness and predictive validity of board of trustees' decisions[J]. Scientometrics, 2005, Vol.63, No. 2:297-320

[5]么大中、张淑芳等. 评价机制：同行评价制与间接指标体系的融一[J]. 黑龙江社会科学，2004（2）：117-120

[6] http://www.thes.co.uk/

[7]Cohen J.A coefficient of agreement for nominal scales.Psychological Bulletin, 1960（70）213-220.

# Study between peer review and multi-indicators evaluation in scientific and technology assessment

Yu Liping, Pan Yuntao, Wu Yishan

Institute of Scientific and Technical information of China

Beijing, 100038

Abstract: This paper analyzing peer review and multi-indicators evaluation based on multiple linear regression and kappa agreement test using the data of the 2007 times higher-QS world university rankings. The results show the data abundance affects the agreement of peer review and multi-indicators evaluation. Evaluation with peer review and multi-indicators together is a good choice while data lacking. Multiple linear regression is a good method to assess the degree of data supply and indicators choice. Excellence peer review is benchmark of indicator weight giving. Peer review can't be replaced by other methods in certain situations. Supplying original data to peer review experts is not always necessary. Multi-indicators evaluation is more stable and impersonal.

Keywords: scientific & technology assessment, university ranking, peer review, multi-indicators evaluation