DOI:10.16262/j.cnki.1000-8217.2010.01.022

·基金纵横·

管理科学与工程学科青年评议专家评议工作质量的后评估

刘作仪

(国家自然科学基金委员会管理科学部，北京 100085)

近年来，国家自然科学基金委员会管理科学与工程学科项目申请量增长迅速。2005年申请量为835项，2008年增加到1209项，年增长率为13.13%。为了应对申请数量的迅猛增加所带来的压力，以及更好地保证项目评议质量，管理科学与工程学科在维护同行评议专家库时，重视对青年学者的培养和锻炼，把国家自然科学青年基金获得者纳入同行评议专家队伍，在每年的评议工作中，根据项目申请情况，给这类青年评议专家指派一定数量的申请项目。

青年基金获得者是否能胜任同行评议工作，青年评议专家的评议工作质量如何呢？为了回答这个问题，本文将采用统计分析方法，客观地评估这类青年评议专家的评议工作的执行情况。

1 同行评议结果的一致性测度指标

在对同行评议专家的后评估研究中，需要对不同评议人做出的评议结果的一致性进行分析。项目评议人对一个项目评价的内部一致性(inter-rater agreement)可以通过多种不同的指标来进行测度，如一致率指标、相关关系指标以及Cohen's k-coefficient指标[1,2]。

一致率指标，如Percentage agreement指标，以及利用列联表(Cross tabulation)得到的一致率指标。由于一致率指标不能测度由偶然性所造成的一致性(如一个或多个观测者通过猜测或其他偶然方式来进行评价)，因此仅利用一致率指标来测度评估结果之间的一致性是不够的。

相关关系也经常被用来测度不同评议者评价结果之间的一致性。一致性可看成是相关关系的一个特例，它更多的是关注对角线上的数值(如表3所示)，即完美一致性(perfect agreement)。值得一提的是，相关关系并没有考虑系统偏差(systematic biases)，这使得完美联合度(perfect association)并不意味着完美一致性。此外，运用相关系数(如Spearman相关系数、肯德尔秩相关系数(Kendall tau coefficient)等)指标得到的结果往往比实际的一致性程度要高。

Cohen's k-coefficient [3] 是用来评估不同评议者对同一现象评价结果之间的人际信度(inter-rater reliability)[4]的指标。其中，*Kappa*统计量是比较两个或多个评议者对同一事物，或评议者对同一事物的两次或多次观测结果是否一致，以由于偶然造成的一致性与实际评议的一致性之间的差别大小作为评价基础的统计指标，其判断结果常以C×C列联表的形式表示。当要求多个评议人根据某些标准对一份申请进行评估时，Cohen's k-coefficient指标完全可以用来测度结果之间的一致性。*Kappa*系数的计算公式为：

$$Kappa = (p_A - p_e)/(1 - p_e) \qquad (1)$$

其中p_A为实际观测到的一致率，即实际观测一致数与总检查人数的比值；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，简称期望率。式中，$p_A - p_e$为实际一致率；$1 - p_e$为非偶然一致率。

从公式(1)可以看出，*Kappa*值实际上为两个差值之比，分子为实际观测到的一致率和可能由于偶然机会造成的期望率的差值，差值越大，说明观测到的一致率远比由于偶然造成的期望一致率高；分母为(1－期望率)，表示非偶然一致率。若*Kappa*值较大，说明一致性较好。

实际上，*Kappa*值在0—1之间。若*Kappa*值等于1，说明两次判断的结果完全一致；若*Kappa*值等于0，说明两次判断的结果完全是由于偶然造

本文于2009年10月10日收到.

成的。可见，*Kappa* 值越大，表明一致程度越好。表 1 反映了 *Kappa* 值与一致性程度的对应关系。

表 1 *Kappa* 统计量与一致性程度的对应关系

Kappa	一致性程度
0.00	差的(poor)
0.00—0.20	较弱的(slight)
0.2—0.4	相当的(fair)
0.4—0.6	较好的(moderate)
0.6—0.8	实质的、显著的(substantial)
>0.8	几近完美的(almost perfect)

2 青年专家评议结果与项目综合评价结果之间的一致性分析

现以 2005—2007 年国家自然科学基金青年项目主持人对 2006—2008 年基金申请书的评议结果与对应年度项目综合评价结果作为研究对象。基本统计数据如表 2 所示。

表 2 基本统计数据

	2006 年	2007 年	2008 年	合计
上一年度青年基金项目数	32	37	40	109
参加评议青年评议人数	12	18	19	49
申请项目数	1085	1186	1350	3621
青年评议人评议项目数(占项目总数的比例)	211 (19.4%)	318 (26.8%)	227 (16.8%)	756

如表 2 所示，2005—2007 年共有 109 人获得青年项目，其中有 49 人参与到后续年度的项目评议中。评议申请书 756 份，占申请项目总数的 21%。

下面分别对 2006、2007、2008 年度青年评议人评议结果与项目综合评价的一致性进行检验与分析。

2.1 列联表分析

2006—2008 年青年评议人与项目综合评价结果之间的列联表，如表 3 所示。由表 3 可知，青年评议人与项目综合评价结果都为 A 的有 35 人，都为 B 的有 58 人，都为 C 的有 296 人。青年评议人与项目综合评价结果之间的实际一致率为 51.5%(即：(35+58+296+0)÷756=51.5%)。评议结果相差一档(即 A—B、B—A、B—C、C—B、C—D、D—C)的项目为 320 项，占总项目数的 42.3%。即青年评议人评价的项目中有 93.8%(即：51.5%+42.3%=93.8%)的比例与项目综合评价保持较高的一致性。仅有 6.2%的项目评价结果之间存在较大的差异(相差二档以上，即 A—C、B—D、C—A、D—B)。

表 3 2006—2008 年青年评议人与项目综合评价结果之间的列联表

青年评议人评议结果	项目综合评议结果* A	B	C	D	合计
A	**35**	39	38	2	114
B	20	**58**	171	0	249
C	3	24	**296**	4	327
D	0	4	62	**0**	66
合计	58	125	567	6	756

注：表中“A”代表“优”、“B”代表“良”、“C”代表“中”、“D”代表“差”。项目综合评价结果中的“E”(小额资助方式)转变为“B”。

*项目综合评议结果是项目所有反馈同行评议意见的专家对该项目所做评价的综合得分。

若将评分等级用数字表示，即 A、B、C、D 的得分分别为 4、3、2、1，则项目综合评分为 2.31 分，青年评议人的平均评分为 2.54 分，比项目综合评分高出不到 10%。

值得一提的是，用实际一致率指标衡量评议结果之间的一致性是不够的，因为它没有将由于机遇造成的一致性考虑在内。

2.2 一致性指标分析

为了消除由于机遇造成的一致性假象，得到更为有效的一致性测度，下面利用 SPSS 软件就几个关键的相关指标进行分析，计算结果如表 4 和 5 所示。

表 4 有向测度(Directional Measures)

			Value	Asymp. Std. Error[a]	Approx. T[b]	Approx. Sig.
Ordinal by Ordinal	Somers'd	Symmetric	0.422	0.044	7.969	0.000
		青年 Dependent	0.591	0.059	7.969	0.000
		综合 Dependent	0.328	0.040	7.969	0.000

a：未假定零假设；b：零假设下的渐近标准误差。

表 5 对称测度(Symmetric Measures)

		Value	Asymp. Std. Error[a]	Approx. T[b]	Approx. Sig.
Ordinal by Ordinal	Kendall's tau-b	0.406	0.029	12.187	0.000
	Kendall's tau-c	0.282	0.023	12.187	0.000
	Gamma	0.696	0.040	12.187	0.000
	Spearman Correlation	0.444	0.031	13.605	0.000[c]
Interval by Interval	Pearson's R	0.452	0.030	13.928	0.000[c]
Measure of Agreement	Kappa	0.203	0.023	9.382	0.000
N of Valid Cases		756			

a：未假定零假设；b：零假设下的渐近标准误差；c：基于正态近似。

注：指标描述：

Ordinal：表示双向有序数据；

Gamma：是依据某一水平所测得的两变量之间的联系水平。其值为 1 时，表示所有的观测量都集中在表格的左上角到右下角的对角线上，0 表示观测量相互独立；

Somers'd：萨默尔 d 值。反映了两个有序变量之间的联系水平。与 Gamma 不同的是，在计算 Gamma 时，不会对因变量与自变量作出区分，数据被认为是对称的；而 Somer'd 则是 Gamma 的非对称扩展，其区别仅在于包括与自变量不相关的成对数据。表明在那些与自变量并不相关的数据中的一致性数据(Discordant)的比例部分。

由表 4、表 5 可以看出，Gamma 值与 Somers'd 值分别为 0.696、0.422，说明青年评议人与项目综合评价结果之间具有较强的联系水平。区别在于，前者认为数据是对称的，后者则描述了那些与自变量并不相关的数据中的一致性数据的比例部分。Kendall's tau-b 值为 0.406，$p<0.001$，说明青年评议与综合评议之间存在着显著的正相关关系。*Kappa* 值为 0.203，说明了青年评议与综合评议之间具有显著一致性。

Cohen's *Kappa* 指标同时考虑了实际一致率与基于机遇的期望一致率，但它的缺陷在于没有考虑不一致程度所造成的影响（即所有的不一致情形都被考虑为等同的，无权重 *kappa* 对不一致情形的权重都赋为 0）。表 4 中的 *Kappa* 值采用的是无权重计算方法。

为了解决这个问题，可以利用加权 *Kappa* 值来进行度量。一个普遍而简单的做法是采用"absolute error weights"，如 $W_{ij}=1-\frac{(|i-j|)}{(g-1)}$（其中 g 为分类数目）作为权重，则 weighted kappa 值为 0.302。若采用"square error weights"，如 $W_{ij}=1-\frac{(|i-j|)^2}{(g-1)^2}$ 作为权重，则 weighted kappa 值为 0.41，如表 6 所示。可见，青年评议人评议与项目综合评价结果之间具有较高的一致性。

表 6　2006—2008 年主要指标数值

		2006 年	2007 年	2008 年
项目总数		211	318	227
Inter-rate agreement(评价一致的项目数)		103	170	116
Percentage agreement(实际一致率)		48.8%	53.5%	51.1%
相差一档的项目数（比例）		79 (37.4%)	127 (40%)	111 (48.9%)
相差二档及以上的项目比例		13.8%	6.5%	0.0%
Somers'd 值		0.337	0.422	0.430
Kendall's tau-b 值		0.344	0.441	0.444
Gamma 值		0.590	0.755	0.750
Spearman Correlation(r_s)		0.378	0.478	0.465
Kappa(K_w)	No-weighted	0.203	0.223	0.194
	Absolute error weighted	0.302	0.318	0.294
	square error weighted	0.41	0.405	0.41

根据表 6 的主要指标数据，可以归纳出如下结论：

（1）青年评议人与项目综合评价结果之间的实际一致率较高，基本保持在半数以上。评价结果相差一档的项目也占较大的比例。结果相差较大的项目数占很小的份额，且降幅很大，从 2006 年的 13.8%降至 2008 年的 0.0%。这反映了青年评议人具备较高的专业水平，能对申请书给予恰当的评价。

（2）从 Gamma 值和 Somers'd 值可以看出，青年评议人与项目综合评价结果之间的联系水平较高。Kendall's tau-b 值也同样反映了两者之间较高的相关性。

（3）由 Cohen's *Kappa* 值可知，青年评议人与项目综合评价结果之间具有比较显著的一致性。且一致性程度逐年增大。

3　评定等级简化对一致性分析的影响

对申请项目的评价有两个指标：等级和资助类别。等级包括 A、B、C、D，资助类别包括优先资助、可资助、不资助。大多数情况下，等级 A 对应优先资助；等级 B 对应可资助；等级 C 和 D 对应不资助。实际评议过程中，评议人对 C 和 D 的区分度并不明显，为此，对 2006、2007、2008 年申请项目的等级重新划分，将评价为 D 的项目与评价为 C 的项目合并，对项目的结果重新整理得到表 7，相应的指标计算结果如表 8 所示。

表 7　2006—2008 年数据

青年	综合								
	A			B			C		
	2006 年	2007 年	2008 年	2006 年	2007 年	2008 年	2006 年	2007 年	2008 年
A	**10**	**20**	**5**	8	18	13	12	19	9
B	7	6	7	**19**	**17**	**22**	53	82	36
C	1	2	0	11	5	12	**90**	**149**	**123**
合计	18	28	12	38	40	47	155	250	168

表 8　指标汇总

	2006 年	2007 年	2008 年
项目总数	211	318	227
Inter-rate agreement	119	186	150
Percentage agreement(%)	56.4	58.5	66.1
Somers'd	0.360	0.460	0.470
Kendall's tau-b	0.366	0.477	0.475
Gamma	0.637	0.811	0.764
Kappa	0.229	0.255	0.314

从表 8 可以看出，评议等级与资助情况之间存在如下相关性：

（1）实际一致率逐年增加，从 2006 年的 56.4%增加到 2008 年的 66.1%。

（2）Gamma 值、Somers'd 值分别从 0.637、0.360 增加到 0.764 和 0.470，说明青年评议人与项目综合评价之间的相关度在不断增加。

（3）*Kappa* 值从 2006 年的 0.229 增加到 2008 年的 0.314，说明青年评议人与项目综合评价之间的一致性在逐年增加。

（下转 46 页）

该类情况。

对专家的特殊要求进行标识，单独设置一项信息栏，专家可以表达自己的一些意愿和说明。自然科学基金委工作人员就可以根据这些信息采取措施，提高评审效率。

(3) 建立专家信誉档案

陈宜瑜主任在六届二次全委会所做工作报告中指出要推进专家库规范化管理，加强专家库建设和维护工作，逐步建立评审专家信誉档案。专家库中增加这项内容，对专家的信誉进行等级划分，自然科学基金委工作人员就可以在选取专家时参考该项信息，科学、公正的选取专家，同时专家的评审也接受到舆论的监督。建立专家信誉档案的举措必将对科学基金管理工作乃至科技界产生深远的影响。

5 结束语

专家库的维护是一项工作量大、见效慢、事务繁琐的工程，是关系到国家自然科学基金评审的一件大事，需要引起各部门的重视。要充分调动各方面积极性，对专家库实行定期系统维护，不断完善专家库的功能，建立信息准确、功能齐全、界面友好的专家库。

相信通过广大科技界人员的共同努力，国家自然科学基金专家库必定能建设成一个世界一流、功能强大的专家库，更好地服务于科学基金项目的评审和管理，更好地服务于科技界，为增强自主创新能力、建设创新型国家作出贡献。

MAINTENANCE AND CONSTRUCTION OF NSFC EXPERTS DATABASE

Ni Peigen Zhang Shouzhu

(*Division I of Physics, Department of Mathematical and Physical Sciences, National Natural Science Foundation of China, Beijing* 100085)

(上接 41 页)

4 结论

从统计结果可以看出，青年评议人评价的项目中有 93.8%的比例与项目综合评价结果保持较高的一致性，且一致性程度逐年增大；而青年评议人评议结果与项目评议结果相差较大(相差二档及以上)的项目数仅占很小的份额，为 6.2%，且该比率的年降幅很大，从 2006 年的 13.8%降至 2008 年的 0.0%。这反映了青年评议人能对申请项目给予恰当的评价。

当然，本文仅仅是从一致性方面对青年评议人的评议结果进行统计分析而做出的初步判断。如果要做出系统和精确的评价，需要综合考虑更多的因素，如评议专家对项目熟悉程度等等。但是，这种初步判断也不失为一种考察评议专家是否胜任评议工作的一种简易方法，特别是当项目管理人员在有限时间和精力情况下要做出判别的时候。正是基于这样的考虑，本文作者才做出这样的尝试。

参 考 文 献

[1] Jakobsson U, Westergren A. Statistical methods for assessing agreement for ordinal data. Scand J Caring Sci, 2005, 19: 427—431.

[2] Barnhart H X, Haber M J, Lin L I. An overview on assessing agreement with continuous measurements. Journal of Biopharmaceutical Statistics, 2007, 17(4): 529—569.

[3] Cohen J A. Weighted Kappa: Nominal scale agreement with provision for scaled disagreement or partial credit. Psychol Bull, 1968, 70: 213—220.

[4] Brousseau L, Wolfson C. The inter-rater reliability and construct validity of the Functional Independence Measure for multiple sclerosis subjects. Clin Rehabil, 1994, 8: 107—115.

TO EVALUATE THE WORKING QUALITY OF YOUNG REVIEWERS

Liu Zuoyi

(*Department of Management Sciences, National Natural Science Foundation of China, Beijing* 100085)