DOI:10.13878/j.cnki.jnuist.2020.05.008

田林琳[1]　孙维东[1]　张弛[1]　郭明[1]　韦纳都[1]

基于互逆强化模型和数理统计方法分析专家评分偏差问题与建议

摘要

“十二五”期间,高技术研究发展计划(863计划)作为引领科技发展的主要抓手之一,为提升中国整体科技实力和创新能力发挥了重要作用.验收评审专家有着评估课题完成水平、衡量科研成果产出价值的关键作用,其评分可靠性直接关系着计划实施成效评价的合理性.鉴于此,本文结合互逆强化模型和数理统计方法,以“十二五”863计划某领域课题验收为例,系统分析了专家评分的偏差问题并按表现将专家划分为8种偏差类型并分别给出应对建议.结果表明:该领域技术验收专家评分整体合理,多数专家均能给出可靠评分;宽严尺度虽然略有差异,但处于可接受范围.本文研究将为国家科技计划相关评审工作乃至其他科研管理活动中合理开展专家评价、精细化规范评审行为以及完善领域专家库、遴选专家评委人选提供参考.

关键词

国家高技术研究发展计划(863计划);科技项目评审;科技计划管理;专家评分;偏差分析

中图分类号 F204;G311

文献标志码 A

收稿日期 2020-07-28

作者简介

田林琳,男,硕士,工程师,研究方向为项目专业化管理.tianll@139.com

1　国家遥感中心,北京,100036

0　引言

作为引领科技发展的主要抓手之一,高技术研究发展计划(863计划)为提升中国整体科技实力和创新能力发挥了重要作用.“十二五”期间,863计划重点支持了先进制造、现代农业、海洋、地球观测与导航、生物和医药等技术领域中的前沿、关键、共性技术突破与核心技术产品及系统研发.众所周知的北斗羲和系统[1]、国际大科学工程——平方公里阵列射电望远镜(SKA)[2]等均受到其资助.该计划兼顾高科技发展和产业化应用,因而其不同技术领域均在一定程度上表现出研究范畴跨度大、技术纵深链路长、项目课题类别多的特色.

从科技计划管理角度出发,如何评估数以亿计的经费投入带来的产出价值是一个关键问题.对此,技术验收专家组的整体评价往往在实践中起着主导作用.而具备类似上述特色的科技计划涉及范畴广度、深度俱足,对验收专家的综合能力提出了严格要求.特别是在部分课题属基础研究、前沿探索类的情况下,其不确定性使课题成果和潜在价值难以客观衡量,既增加了专家评分难度,又令不同类型课题间的评分难以横向比较.因此,利用评分数据分析专家偏差,帮助科研管理人员评估专家评审能力,可以为了解领域创新成果和课题实施成效的评审信度与效度,以及更好地把握技术领域发展现状提供重要参考.

评审活动受到专业因素(如评分人是否充分了解参评对象属性)、心理因素(如评分人是否对参评对象心怀同情)、外部因素(如评分人是否与参评对象有利害关系)等多方面影响,因而评分偏差分析既是科研管理人员的关注焦点,也是心理学、应用数学和信息科学等学科的研究对象.20世纪末已有学者关注到地域科研实力评价中的偏差问题[3].随后,一些科研管理人员对国家重点实验室评估偏差进行了分析,如谢焕瑛等从来源和成因上归纳了6种影响专家评分的效应[4]和4类偏差[5],张健等给出了应对潜在不公平评估的策略[6],杨晓秋梳理了实验室评估中的若干偏差问题[7],重点指出应增加专家培训力度使其更好地内化评估规则.这些研究主要是定性总结偏差成因和表现,但少有给出具体的定量分析方法.由于评委评分易受评分人经验知识、思考方式、人格特征等影响[8],心理测量学领域对评分

偏差定量分析有很大兴趣,所用理论呈现出从经典测量理论[9]、概化理论[10]到现代测量学中的项目反应理论[8]的过渡.如著名的多面 Rasch 模型可用于评估项目难度、评委宽严程度、考生能力等参数及其交互关系,在结构化面试[11]、教育教学能力测试[10]、英语听说考试[12]等方面均有应用.但上述理论过于复杂,模型需较好的先验初始值进行迭代求解,且还可能出现不收敛现象[13].忽略专业背景差异,专家评分与网购评分、书评影评评级等在形式上并无区别.随着互联网 4.0 时代到来,应用数学与信息科学学者聚焦于网络社区用户评分偏差和异常分检测,从数学和算法层面建立了评分评估模型与指标,同样可用于专家评分偏差分析.如 Lauw 等基于强化模型给出了衡量评分人偏差和参评对象争议性的两个指标[14],Dai 等利用评分人和参评对象间的正面、负面效应建立二部图以检测行为异常的评分人[15],文献[16-17]则致力于面向众包系统构建评价体系和搜索高争议性参评对象.但需注意此方面研究更多是侧重于甄别异常用户以识别恶意或虚假评价.当然,也有少数专门面向评委评分偏差的研究,如吕书龙等利用假设检验等数学思想建立评分控制和偏差吻合模型[18],而文献[19]则基于投影寻踪构建评委综合评价模型.

考虑到心理学中相关理论限制较大,本文仅以数学和信息科学中的互逆强化模型和数理统计方法为技术手段,以 863 计划某技术领域课题验收为典型案例,对"十二五"期间 863 计划的评审专家评分偏差进行初步的定量探索.此项研究是对现阶段科研管理中专家偏差分析研究的完善与延伸,可助力精细化规范评审行为和后续专家遴选.据笔者所知,这是首次面向 863 计划等国家科技计划的专家分析工作.

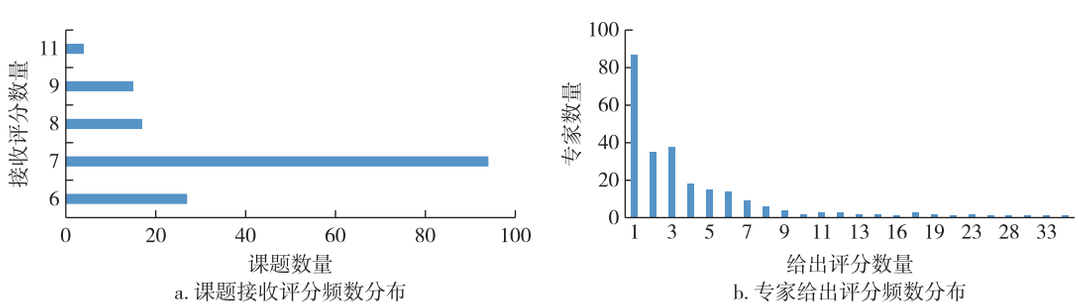
1 偏差分析数据用例

863 计划旨在面向经济社会发展需求加强技术研发和应用,同时也面向国际前沿和国家未来重大需求开展一定的前沿理论与技术探索,具有多学科交叉和兼顾研发与探索的特点.因此,其下设项目、课题的验收评审往往既要求专家组研发与集成经验丰富,又要求在领域前沿发展态势上具有敏锐的嗅觉.本文将"十二五"863 计划某技术领域的课题验收评分作为偏差分析数据,一来便于科研管理人员将本文方法迁移用至其他技术领域;二来在科技体制改革后 863 计划被延伸融入到国家重点研发计划,两个计划间专家遴选范围重叠度较高,所得经验和结论可直接用在重点研发计划相应领域的重点专项中,帮助遴选合适的评审专家开展综合绩效评价工作.所用数据包含该领域全部专家评分,但由于项目数量较少且评价采用等级制,课题数量较多且评价采用百分制,后文对项目评级情况不做讨论.数据具体由 252 位专家对 157 个课题的 1 135 次评分组成,课题平均收到 7.2 个评分,专家人均评分 4.5 次,统计信息如图 1 所示.

课题接收评分频数图中可见各课题得分数量基本能保证分数均值、方差等统计信息的有效性.但对于专家给出评分的频数有两点说明:

1)863 计划各技术领域均设有领域专家组,负责全周期跟踪项目及课题进展,从而能够较为完善地评价项目、课题完成水平,所以验收专家组一般由 1~2 位熟悉相应执行情况的领域专家组成员和同行专家共同组成.从专家给出评分频数图可知,随着评分次数增加,人数快速下降,自左到右从同行专家居多转为领域专家组专家居多.

2)对于评分次数较少的专家难以确保其评分信息有效性,下文研究仅聚焦于至少有 5 次评分的 74 位专家.虽然无法分析剩余专家评分,但这些专家给出的分数仍然有助于课题评价,在对课题情况开展分析时仍将使用全部专家评分数据.



a. 课题接收评分频数分布　　b. 专家给出评分频数分布

图 1　评分数据统计信息

Fig. 1　Statistics of score data

2 评分偏差评估模型与指标

2.1 评分偏差成因与类型

近十几年来,不同科研管理人员根据各自经验总结了科研活动中的评分偏差成因与类型.表1列出了其中比较有代表性的看法,从中不难发现:

1)系统内、外因素来自专家和课题之外,超出了本文范畴;偶然偏差较小且属于量化评分中必然出现的正常偏差[5];同行偏好偏差、非共识偏差源自同行偏好效应和非共识效应,可以一并讨论.故下文不再展开这些内容.

2)惯性思维、学术权威和个人偏好效应难以仅凭评分数据进行分析.不过遴选同行专家的回避原则显著降低了专家与课题间的关联性,一定程度上避免了其影响.此外,前两者在验收评审中未必会增加偏差:由于长期跟踪课题,惯性思维使领域专家组成员评分更可靠;专家的权威性反映了其卓越的专业素养和眼光,权威效应也可能缩小偏差.

3)同行偏好属于普遍性偏好,对绝大部分专家的作用是均衡的[5],对于课题间评分的相对影响不大.

4)各课题验收专家均为相关方向资深同行且符合回避原则,不能了解参评课题及同课题间存在好恶关系的可能较低,但部分课题的前沿性和探索性增加了量化评价难度,可见验收评审中的非共识效应由课题不确定性主导.为明确这一点,以下将因不确定性引起的非共识偏差归于课题而非专家,并称之为争议性偏差[14,17].

5)因个人习惯导致评分尺度不同,从而产生或偏高或偏低的系统性差异,所以系统偏差和严厉度偏差非常相似.另外,同情心理是形成个人评分习惯的潜在心理因素,该效应令专家倾向于高估课题分数.因此,本文将同情心理效应引起的偏差归于上述偏差,并将其统称为专家固有偏差.

6)评审活动中无法知悉体现课题完成情况的真实分数,但合理的假设是多数课题评分或其均值是较为客观和接近真实值的,所以在评价专家评分能力时实质上往往是综合参考对同一课题的其他专家评分进行判断,也即暗含了对一致性偏差的考察.

综合以上分析可知:课题争议性干扰了专家评分准确性,需在评估专家偏差时降低其影响;与其他专家评分的一致性体现在评估专家总体偏差的过程中;固有偏差代表了专家间评分松紧尺度的不同标准;除系统内、外因素和偶然偏差等不在本文范畴或可忽略的因素之外,个人偏好等因素既难以通过评分数据辨别,在课题评分中又仅对个别专家产生较大影响,本文将它们引发的极端评分不加区别,统一归于异常评分.综上,下文将结合评分一致性和课题争议性两方面建立专家总体偏差评估模型,并利用两个假设检验方法实现对异常评分和固有偏差的检测,以此开展专家评分偏差分析工作.

表1 科研评分偏差典型成因与类型

Table 1 Typical causes and types of scientific research score bias

成因/类型	出处	名称	含义
偏差成因	文献[3]	系统外因素	参评对象原始资料不正确传递的错误信息
		系统内因素	对参评对象各方面考察的权重分配不恰当
	文献[4]	惯性思维效应	专家在与参评对象的长期交互中形成“先入为主”的固有认识
		个人偏好效应	专家不以评价规则为参考标准,根据个人偏好的指标(如论文发表等)进行判断
		同行偏好效应	偏爱研究领域特别相近的参评对象
		学术权威效应	迷信或顾虑个别权威专家发表的意见
		同情心理效应	对较差的参评对象产生同情心理
		非共识效应	因不够了解参评对象、和参评对象的好恶关系、参评对象自身不确定性等导致专家间判断不一致
偏差类型	文献[5]	同行偏好偏差	由同行偏好引起的评分偏差
		非共识偏差	由非共识偏好引起的评分偏差
		系统偏差	由于专家个性或工作习惯不同产生的偏差
		偶然偏差	专家实际评分不能准确表达其意愿引起的偏差
	文献[12]	严厉度偏差	由于专家或严厉或宽松的评分习惯产生的偏差
		一致性偏差	与其他专家评分行为不一致呈现的偏差

2.2 基于互逆强化的总体偏差评估模型

设有 n 个专家参与 m 个课题的评分工作,目标是评估各专家的总体评分偏差.若已知全部评分的真实偏差,经简单聚合操作就可以得到专家总体偏差,如用均值作为第 i 个专家的总体偏差:

$$b_i = A_j \hat{d}_{ij}, \tag{1}$$

式中 b_i 表示专家 i 总体偏差,$\hat{d}_{ij}$ 为专家 i 对课题 j 的评分偏差,A_j 代表对该专家全部评分求平均.

然而,实践上存在两个问题:第一,因课题真实分数未知,真实偏差无法获取;第二,因研究方向、前沿探索性、成果形式差异导致课题定量评分的不确定性不同,即争议性程度使专家 i 对多个课题的评分偏差受到不同程度的干扰.例如,当该专家多对高争议度课题评分和多对低争议度课题评分时,根据式(1)计算前者偏差必然大于后者.对于问题一,通常认为评分均值是课题真实分数的近似估计,即是说综合多个专家评分可以较好地评价课题完成水平.那么专家 i 对课题 j 的评分与其他专家对该课题的评分差别 d_{ij} 可近似看作专家 i 在课题 j 上的偏差,这同时体现了一致性偏差.对于问题二,可通过对偏差的加权放缩来应对,从而将总体偏差近似为式(2),其中 $\bar{c}_j$ 代表课题 j 的非争议程度:

$$b_i = A_j\, d_{ij}\bar{c}_j. \tag{2}$$

此时问题转为如何衡量课题争议程度.争议度是引发专家间出现非共识和意见发散的能力,最直观的衡量方法就是对此课题接收的全部评分求偏差均值.但同样要考虑参评专家的评分能力,因此令课题争议度为

$$c_j = A_i\, d_{ij}\bar{b}_i, \tag{3}$$

式中 $\bar{b}_i$ 为专家 i 的评分能力,或者说非偏差程度.

式(2)和(3)说明了专家偏差和课题争议度的相互依赖,争议度影响着专家偏差,专家偏差又反过来影响争议度,二者联合构成了互逆强化模型[14].如果把专家和课题视作顶点,把评分视为顶点间连边的权重,上述问题将转为常用于社区网络信息挖掘的特殊二部图[20].本文定义 i 对 j 的评分偏差为 i 的评分与其他专家对 j 的评分之差的绝对值平均,有

$$d_{ij} = \frac{1}{n_j - 1}\sum_{k\neq i} \mid e_{kj} - e_{ij} \mid, \tag{4}$$

式(4)中 e_{ij} 为 i 给 j 的分数,n_j 为给课题 j 评分的专家数,在验收评审中 n_j 必然大于 1,故式中分母必为正整数.

原始评分采用百分制,为便于后续计算需将全部分数乘以 0.01 压缩至[0,1]区间,从而任一 b,c 或 d,e 取值范围均为[0,1].为保证 $\bar{b}$ 和 $\bar{c}$ 区间与其他变量一致同时保持 $\bar{c}$ 与 c,$\bar{b}$ 与 b 间的负相关,本文参照文献[14]采用线性关系令 $b + \bar{b} = 1$ 和 $c + \bar{c} = 1$.将总体偏差和争议度表示为列向量形式 $\boldsymbol{B}$ 和 $\boldsymbol{C}$,相应有如下关系:

$$\boldsymbol{B} = \boldsymbol{K}(\boldsymbol{1}_m - \boldsymbol{C}), \tag{5}$$

$$\boldsymbol{C} = \boldsymbol{L}^{\mathrm{T}}(\boldsymbol{1}_n - \boldsymbol{B}), \tag{6}$$

式(5)、(6)中的 $\boldsymbol{1}$ 分别表示长度为 m 和 n 的全 1 列向量,$\boldsymbol{K}$ 和 $\boldsymbol{L}$ 为 $n \times m$ 大小的矩阵且 i 行 j 列元素分别为 $K_{ij} = d_{ij}/m_i$ 和 $L_{ij} = d_{ij}/n_j$.m_i 类似 n_j 的定义,代表专家 i 评审的课题数.上标 T 表示矩阵转置.

互逆强化是全局性的动态过程,因为变动任何课题的争议度估计值会影响给其评分的专家的偏差估计,偏差估计值变化又会影响这些专家给予分数的课题的争议度估计,形式上相似于概率图模型[21]中的信念传播机制[22].借鉴谷歌的 PageRank 排序算法[23],Berkhin 等得到了 $\boldsymbol{B}$ 和 $\boldsymbol{C}$ 各自的自嵌套表达式,经自迭代求解出 $\boldsymbol{B}$ 和 $\boldsymbol{C}$.然而,这一求解方式需满足一定前提且在自迭代过程中要周期性规范化 $\boldsymbol{B}$ 和 $\boldsymbol{C}$.此外,笔者发现将自迭代得到的 $\boldsymbol{B}$ 代入式(6)计算出的 $\boldsymbol{C}$,与自迭代得到的 $\boldsymbol{C}$ 并不一致,反之将自迭代结果 $\boldsymbol{C}$ 代入式(5)也有相似的现象,这是与总体偏差和争议度的相互依存关系相违背的.因此,本文采用互迭代方式进行求解,即先在(0,1]区间随机初始化 $\boldsymbol{B}$ 为 $\boldsymbol{B}_0$ 并代入式(6)得到 $\boldsymbol{C}$ 为 $\boldsymbol{C}_1$,再将 $\boldsymbol{C}_1$ 代入式(5)更新 $\boldsymbol{B}$ 为 $\boldsymbol{B}_1$,如此往复直至收敛.当然,从初始化 $\boldsymbol{C}$ 开始互迭代可得到相同结果.以上方法虽然简单但非常有效,可以证明互迭代过程同样能收敛.证明如下:

不妨设任意第 k 至 $k+2$ 轮迭代中得到 $\boldsymbol{B}_k$,$\boldsymbol{B}_{k+1}$ 和 $\boldsymbol{B}_{k+2}$,则有

$$\begin{aligned}\boldsymbol{B}_{k+2} - \boldsymbol{B}_{k+1} &= \boldsymbol{K}(\boldsymbol{1}_m - \boldsymbol{C}_{k+1}) - \boldsymbol{K}(\boldsymbol{1}_m - \boldsymbol{C}_k) = \\ &\boldsymbol{K}\boldsymbol{L}^{\mathrm{T}}(\boldsymbol{1}_n - \boldsymbol{B}_k) - \boldsymbol{K}\boldsymbol{L}^{\mathrm{T}}(\boldsymbol{1}_n - \boldsymbol{B}_{k+1}) = \\ &\boldsymbol{K}\boldsymbol{L}^{\mathrm{T}}(\boldsymbol{B}_{k+1} - \boldsymbol{B}_k)\end{aligned} \tag{7}$$

收敛即要令 $\boldsymbol{B}_{k+2}$ 和 $\boldsymbol{B}_{k+1}$ 中对应元素变化不大于 $\boldsymbol{B}_{k+1}$ 和 $\boldsymbol{B}_k$ 间变化,利用向量 $\boldsymbol{l}_2$ 范数 $\|\cdot\|_2$ 可等价转换为满足 $\|\boldsymbol{B}_{k+2} - \boldsymbol{B}_{k+1}\|_2 \leqslant \|\boldsymbol{B}_{k+1} - \boldsymbol{B}_k\|_2$.引入变量 $\boldsymbol{U}_{k+1} = (\boldsymbol{B}_{k+1} - \boldsymbol{B}_k)(\boldsymbol{B}_{k+1} - \boldsymbol{B}_k)^{\mathrm{T}}$,有:

$$\begin{aligned}\|\boldsymbol{B}_{k+2} - \boldsymbol{B}_{k+1}\|_2^2 - \|\boldsymbol{B}_{k+1} - \boldsymbol{B}_k\|_2^2 &= \\ \mathrm{tr}(\boldsymbol{U}_{k+2}) - \mathrm{tr}(\boldsymbol{U}_{k+1}) &= \\ \mathrm{tr}(\boldsymbol{K}\boldsymbol{L}^{\mathrm{T}}\boldsymbol{U}_{k+1}\boldsymbol{L}\boldsymbol{K}^{\mathrm{T}}) - \mathrm{tr}(\boldsymbol{U}_{k+1}) &= \\ -\mathrm{tr}((\boldsymbol{I} - \boldsymbol{L}\boldsymbol{K}^{\mathrm{T}}\boldsymbol{K}\boldsymbol{L}^{\mathrm{T}})\boldsymbol{U}_{k+1}) &= \\ -\mathrm{tr}(\boldsymbol{R}\boldsymbol{U}_{k+1}),\end{aligned} \tag{8}$$

其中 tr(·) 为矩阵的迹,$\boldsymbol{I}$ 为单位矩阵,$\boldsymbol{R}=\boldsymbol{I}-\boldsymbol{L}\boldsymbol{K}^{\mathrm{T}}\boldsymbol{K}\boldsymbol{L}^{\mathrm{T}}$.

显然,$\boldsymbol{R}$ 和 $\boldsymbol{U}_{k+1}$ 是实对称矩阵.由于 $\boldsymbol{L}$ 和 $\boldsymbol{K}$ 中元素均为不大于1的正数 d_{ij} 除以 m_i 或 n_j 得到的,而矩阵 $\boldsymbol{L}\boldsymbol{K}^{\mathrm{T}}\boldsymbol{K}\boldsymbol{L}^{\mathrm{T}}$ 中任一元素均由 $\boldsymbol{L}$ 和 $\boldsymbol{K}$ 中元素4次相乘而来,故 $\boldsymbol{L}\boldsymbol{K}^{\mathrm{T}}\boldsymbol{K}\boldsymbol{L}^{\mathrm{T}}$ 中元素值必然极为接近0.由此知 $\boldsymbol{R}$ 接近单位矩阵 $\boldsymbol{I}$,其特征值必大于0,所以 $\boldsymbol{R}$ 为对称正定矩阵.对于任意非零实向量 $\boldsymbol{X}$,有 $\boldsymbol{X}^{\mathrm{T}}\boldsymbol{U}_{k+1}\boldsymbol{X}=\|(\boldsymbol{B}_{k+1}-\boldsymbol{B}_k)^{\mathrm{T}}\boldsymbol{X}\|_2^2\geqslant 0$,当且仅当 $\boldsymbol{B}_{k+1}-\boldsymbol{B}_k$ 为零向量时该式等于0.此时有两种情况:

1)$\boldsymbol{B}_{k+1}-\boldsymbol{B}_k$ 不为零向量:$\boldsymbol{U}_{k+1}$ 相应为对称正定矩阵.存在可逆矩阵 $\boldsymbol{P}$ 和 $\boldsymbol{Q}$,使 $\boldsymbol{R}=\boldsymbol{P}^{\mathrm{T}}\boldsymbol{P}$ 和 $\boldsymbol{U}_{k+1}=\boldsymbol{Q}^{\mathrm{T}}\boldsymbol{Q}$,则 $\boldsymbol{Q}(\boldsymbol{R}\boldsymbol{U}_{k+1})\boldsymbol{Q}^{-1}=(\boldsymbol{P}\boldsymbol{Q}^{\mathrm{T}})^{\mathrm{T}}\boldsymbol{P}\boldsymbol{Q}^{\mathrm{T}}$,即 $\boldsymbol{R}\boldsymbol{U}_{k+1}$ 与 $(\boldsymbol{P}\boldsymbol{Q}^{\mathrm{T}})^{\mathrm{T}}\boldsymbol{P}\boldsymbol{Q}^{\mathrm{T}}$ 相似,二者的迹相等.显然 $\boldsymbol{P}\boldsymbol{Q}^{\mathrm{T}}$ 可逆,从而知 $(\boldsymbol{P}\boldsymbol{Q}^{\mathrm{T}})^{\mathrm{T}}\boldsymbol{P}\boldsymbol{Q}^{\mathrm{T}}$ 是正定矩阵,其迹大于0.因此有 $\mathrm{tr}(\boldsymbol{R}\boldsymbol{U}_{k+1})>0$,故从式(8)易知 $\|\boldsymbol{B}_{k+2}-\boldsymbol{B}_{k+1}\|_2\leqslant\|\boldsymbol{B}_{k+1}-\boldsymbol{B}_k\|_2$ 成立.

2)$\boldsymbol{B}_{k+1}-\boldsymbol{B}_k$ 为零向量:此时 $\|\boldsymbol{B}_{k+2}-\boldsymbol{B}_{k+1}\|_2\leqslant\|\boldsymbol{B}_{k+1}-\boldsymbol{B}_k\|_2$ 成立.

综上得证互迭代使 $\boldsymbol{B}$ 稳定收敛,同理可证 $\boldsymbol{C}$ 的收敛性.实际上,只要 $\boldsymbol{B}$ 没有恰好初始化为收敛解,$\boldsymbol{B}_1-\boldsymbol{B}_0$ 不会是零向量,随后 $\boldsymbol{B}$ 将不断更新直至收敛;而若恰好初始化为收敛解,则无需迭代已得到了想要的结果.本文在评分数据上基于不同初始值多次求解,均经3~4次互迭代即可得到稳定且一致的结果.

2.3 基于假设检验的异常评分与固有偏差检测

异常评分反映了专家评分因某些主客观因素引起的明显偏离真实分数的现象,了解异常评分情况有助于识别问题专家.同一课题的评分数据是以真实分数为中心的随机变量,如能保证专家评分客观性,该变量将近似服从高斯分布.参照文献[18],本文以课题均分作为真实分数的近似,视均值上下2倍标准差范围为评分正常区间,以此判断专家评分是否异常并统计各专家的异常评分次数.如对于专家 i 给出的课题 j 评分,

$$N_i\leftarrow N_i+1,\quad \text{if}\quad e_{ij}\notin[\mu_j-2\sigma_j,\mu_j+2\sigma_j],\tag{9}$$

其中,N_i 为专家 i 异常评分次数,初始为0;e_{ij} 同前为专家评分;μ_j 和 σ_j 分别是课题 j 平均分和标准差,根据该课题收到的所有评分计算.依验收规范应先剔除最高分和最低分后再计算平均分,但考虑到部分课题参评专家仅6人,剔除后无法保证统计稳定性,所以本文没有剔除最值.此外,文献[18]与本文不同,其不合理地对各课题采用统一标准差,会导致低争议度课题的异常评分漏检和高争议度课题的异常评分虚警.

正如前文所述,异常评分成因难以确认,所以存在异常评分并不意味着专家一定有严重偏差,还需考虑其评分次数.对于评分较多的专家,少量异常是可以接受的,问题在于异常次数阈值如何确定.将专家 i 参与的验收评审看作独立过程,那么 N_i 近似服从二项分布 $\mathrm{bin}(m_i,p)$,从而得到零假设(专家 i 评分总体正常)和备择假设(专家 i 评分总体异常).其中 m_i 定义同前,p 代表异常评分概率,为异常评分总数与评分总数之商.给定显著性水平 α_{bin} 可利用概率 $P(N_i>T_i^{\mathrm{bin}})\leqslant\alpha_{\mathrm{bin}}$ 确定阈值 T_i^{bin},从而判断专家 i 的异常评分数量是否处于合理范围.

另一方面,我们期望评分接近真实分数,也就是二者之差接近0.若专家固有偏差较大,那么过宽或过严的评分习惯将令差值表现为普遍大于或小于0.在忽略课题争议性且评分过程足够理想的前提下,一个专家的评分数据应服从中心为0而方差未知的高斯分布.可根据 t 检验理论,利用该专家评分偏差平均值与0(即理想偏差平均值)间的离差统计量进行假设检验,判断是否存在明显的固有偏差.然而,各异的争议度使课题间评分难度和波动性不一,只有先降低争议性差异的影响才能保证 t 检验合理性.本文采用与2.2中相同的处理方式,利用 $\bar{c}_j$ 作为课题 j 的权系数对相关偏差进行放缩,以降低争议性的负面影响.例如,对于参评课题集合 $\{s_1,s_2,\cdots,s_{m_i}\}$,专家 i 产生评分向量 $[e_{i1},e_{i2},\cdots,e_{im_i}]$,则相应有加权偏差向量 $[\bar{c}_1f_{i1},\bar{c}_2f_{i2},\cdots,\bar{c}_{m_i}f_{im_i}]$.$f$ 为专家 i 对某课题的评分与该课题平均分之差.随后构造统计量

$$t_i=\frac{\mu_{f_i}-0}{\sigma_{f_i}/\sqrt{m_i}}\tag{10}$$

近似服从 t 分布 $t(m_i-1)$,式中 μ_{f_i} 和 σ_{f_i} 分别为加权偏差向量的均值和标准差.同时,得到了两个对立假设:零假设(专家 i 评分无固有偏差)和备择假设(专家 i 评分有固有偏差).给定显著性水平 α_t 后,从 t 分布表确定双侧阈值 $t_{1-\alpha/2}(m_i-1)$ 和 $t_{\alpha/2}(m_i-1)$.超出阈值即可判定该专家明显存在固有偏差:$t_i<t_{1-\alpha/2}(m_i-1)$ 说明评分过于严格,倾向于给低分,而 $t_i>t_{\alpha/2}(m_i-1)$ 说明评分过于宽松,倾向于给高分.

至此,本部分已给出评估专家评分偏差的3个指标及衡量课题争议度的指标.其中:总体偏差是对

专家偏差的整体性估计;异常评分检测极端值,是对偏差的突变性估计;固有偏差判断专家内在的评分尺度习惯,是对偏差的倾向性估计.三者间有着一定联系:

1)异常评分次数和固有偏差信息既相互影响又相互补充:固有偏差在极端情况下会引发异常评分,且异常分数将一致性地极高或极低;反过来,异常分数过多同样可能增加固有偏差.当然,因为异常评分受多种因素影响,更常见的是异常分数中同时包含高分、低分,不会引起固有偏差.这些可能的情况无法单独从异常评分或固有偏差来判断.因此,这两个指标既从不同侧面反映专家的特定偏差问题,又在特定情况下表现出一定耦合性.

2)总体偏差与异常评分、固有偏差粒度互补:总体偏差是从整体层面评估专家偏差的核心指标,涵盖了突变性、倾向性等考量.这对于从粗粒度快速锁定问题专家非常关键,但无法判断问题具体信息,如专家偏差主要受外部条件干扰,抑或评审规范不够内化,还是评分尺度异于他人? 这些细粒度信息对于采用何种处理措施很有指导性,可通过异常评分和固有偏差来判断,必要时还可继续搜索其他相关信息进一步定位问题.

综上所述,3 个指标相结合才能较完善地分析专家偏差,下文将据此完成对 863 计划某领域课题验收专家的偏差分析.

3 评分偏差评估结果与分析

3.1 评分偏差评估结果

本文所用数据涉及 157 位专家对 252 个课题的评分,课题平均分和标准差如图 2 所示.课题 82 得分最高(95.17 分),课题 96 得分最低(73.18 分),标准差在 0.7~7.55 间波动,表明这些课题无论在完成水平还是在争议性上均有很大差异,尤其后者会干扰评审评分,在评分偏差分析中将其纳入考量很有必要.因本文聚焦于专家偏差,下文对争议度不做详细讨论.

使用 2.2 和 2.3 中方法面向评分不少于 5 次的 74 位专家进行偏差评估,得到各项指标及阈值如表 2 所示(显著性水平 0.01).表 2 中序号对应总体偏差排名,序号越小意味着总体偏差越大.由于我们仅展示了 3 位小数,导致少量序号不同的专家总体偏差值看上去相同,如专家 19~21.为使估计值均匀分布在[0,1]以方便相对比较,所列总体偏差经过了最大值规范化处理.因空间有限,表 2 中只给出了 t 检验的右侧阈值,左侧阈值为其相反数.

表 2 中斜体加粗的部分为异常评分次数大于或等于阈值以及 t 检验值超限的数字.7 位专家异常评分过多,仅占专家总数 9.46%,且其中 6 位次数刚好等于阈值,可认为专家整体低异常;14 位专家出现固有偏差,占比略高,约为 18.92%.但其中多数专家检测值超限不多,造成的实际高估或低估偏差分值不大(具体见 3.3 中专家实例),说明专家整体固有偏差程度是可以接受的.另一方面,异常评分过多的专家序号均靠前,并且总体偏差最大的正是唯一超过异常次数阈值的专家,侧面证明了总体偏差指标的有效性.然而固有偏差较大的专家呈不规律分布,原因在于固有偏差表示评分会习惯性的偏高或偏低,意味着评价课题完成情况的专家给分尺度不同,大多数情况下并不会引起极端评分和高总体偏差.但大部分专家评分仅有 5~7 次,过多的异常评分引起总体偏差显著增加是很正常的.当然,过大的固有偏差仍然会对总体偏差产生不可忽略的影响,如排名第 8 位的专家.以上现象均印证了 2.3 结尾部分的推测.

3.2 总体偏差模型互迭代结果验证

原始互逆强化模型需将偏差向量 $\boldsymbol{B}$ 和争议度向

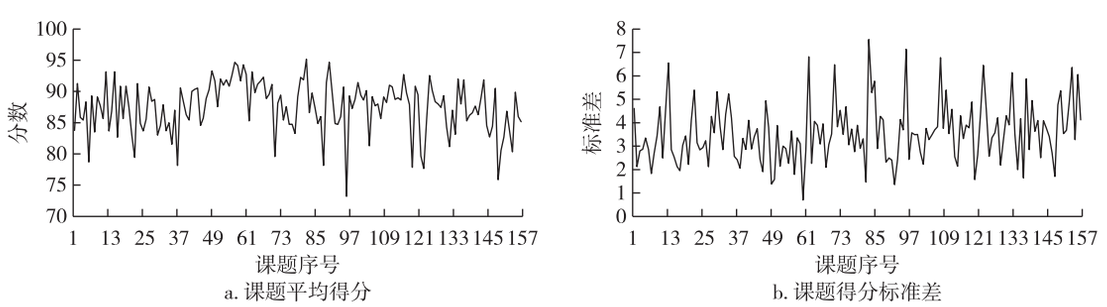


图 2 课题平均分及标准差

Fig. 2 Average scores and standard deviations of subjects

表 2　专家评分偏差指标值及相应阈值

Table 2　Index values and corresponding thresholds for expert score bias

序号	总体偏差	异常次数	异常阈值	t 检验	检验阈值	序号	总体偏差	异常次数	异常阈值	t 检验	检验阈值
1	1	3	2	**-5.422**	4.032	38	0.467	0	1	0.673	4.604
2	0.917	**1**	1	-3.290	4.604	39	0.467	0	3	-1.298	2.921
3	0.827	1	2	-2.109	3.707	40	0.463	0	2	-1.699	3.355
4	0.761	**1**	1	**-5.070**	4.604	41	0.462	0	2	-1.729	3.012
5	0.749	**1**	1	-2.530	4.604	42	0.455	0	3	**-3.805**	2.787
6	0.689	**1**	1	1.928	4.604	43	0.455	0	1	-1.319	4.604
7	0.677	0	2	-1.964	3.499	44	0.452	0	2	-0.421	3.499
8	0.665	0	2	**21.256**	3.707	45	0.447	0	3	-2.633	2.845
9	0.653	0	1	-1.847	4.604	46	0.446	0	1	0.205	4.604
10	0.651	1	2	0.286	3.250	47	0.442	0	2	-1.101	3.106
11	0.650	1	2	-2.108	4.032	48	0.441	0	2	2.310	3.012
12	0.628	0	2	-2.805	3.250	49	0.439	1	3	-1.289	2.819
13	0.604	0	1	4.298	4.604	50	0.438	0	4	-2.500	2.738
14	0.603	1	2	0.714	3.707	51	0.429	0	2	1.932	3.499
15	0.603	0	2	3.074	4.032	52	0.428	0	2	2.583	3.055
16	0.587	0	2	**5.082**	3.169	53	0.428	0	2	-0.053	3.499
17	0.580	0	2	-2.520	4.032	54	0.426	0	3	1.745	2.878
18	0.575	0	2	1.875	3.499	55	0.425	0	5	**3.385**	2.676
19	0.567	**2**	2	-2.240	3.355	56	0.421	2	4	0.854	2.744
20	0.567	0	1	**5.788**	4.604	57	0.420	0	2	0.049	3.707
21	0.567	0	2	2.456	4.032	58	0.415	0	1	**6.761**	4.604
22	0.554	**2**	2	-2.757	3.707	59	0.409	0	3	**4.667**	2.771
23	0.537	0	2	**4.654**	3.106	60	0.409	0	2	-0.569	3.355
24	0.533	0	3	0.761	2.878	61	0.393	0	3	2.647	2.921
25	0.532	2	3	-0.510	2.947	62	0.389	0	3	0.172	2.819
26	0.518	0	2	-0.491	3.169	63	0.377	0	1	-0.389	4.604
27	0.516	0	2	-0.897	4.032	64	0.376	0	1	1.988	4.604
28	0.516	0	2	-1.453	3.355	65	0.371	0	1	1.296	4.604
29	0.504	0	2	**-4.123**	3.707	66	0.359	0	2	-2.512	4.032
30	0.504	0	2	**4.245**	4.032	67	0.354	0	2	-1.665	4.032
31	0.501	1	2	1.868	4.032	68	0.353	1	2	1.104	3.055
32	0.489	0	2	-1.832	4.032	69	0.345	0	2	**3.404**	3.106
33	0.488	0	2	3.086	3.707	70	0.345	0	3	-1.015	2.921
34	0.473	0	2	**-4.826**	3.499	71	0.344	0	2	-0.611	4.032
35	0.472	0	2	-0.966	4.032	72	0.340	0	2	-1.084	3.169
36	0.471	1	2	-2.197	3.707	73	0.322	0	2	1.618	3.707
37	0.471	0	2	-0.533	4.032	74	0.281	0	1	0.826	4.604

量 $\boldsymbol{C}$ 表示为递归形式后,分别自迭代求解.自迭代过程中对 $\boldsymbol{B}$ 和 $\boldsymbol{C}$ 的规范化会导致求得的结果丢失式(5)和(6)中体现的交互关系,但模型的构建依赖于 $\boldsymbol{B}$ 和 $\boldsymbol{C}$ 的耦合性.这种矛盾并不合理,所以本文提出了互迭代策略作为替代.为验证互迭代的求解效果,本部分分别采用这两种方式得到专家总体偏差和课题争议度,结果如图 3—5 所示.

利用自迭代分别得到总体偏差 $\boldsymbol{B}$ 和争议度 $\boldsymbol{C}$,同时基于相互依赖关系,也可将自迭代结果 $\boldsymbol{B}$ 代入式(6)得到相应的 $\boldsymbol{C}$,同理式(5)又可用自迭代结果 $\boldsymbol{C}$ 得到相应的 $\boldsymbol{B}$.理论上,这两个 $\boldsymbol{B}$ 和两个 $\boldsymbol{C}$ 之间应该是一致的,但从图 3 易知实际情况并非如此.周期

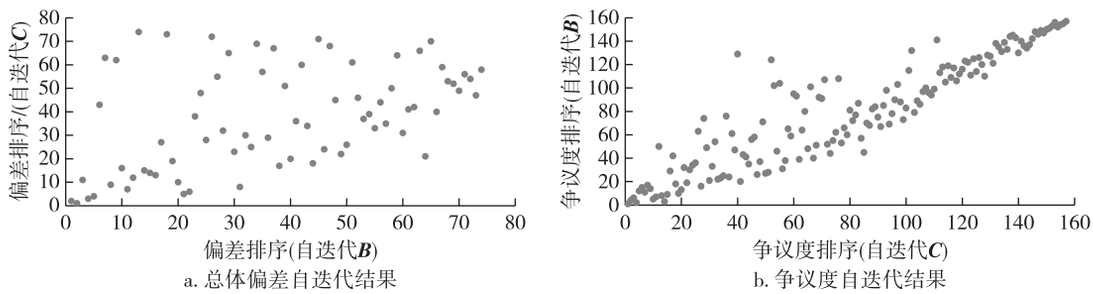


图 3　总体偏差及争议度自迭代结果

Fig. 3 Self-iteration results of overall bias (a) and controversy (b)

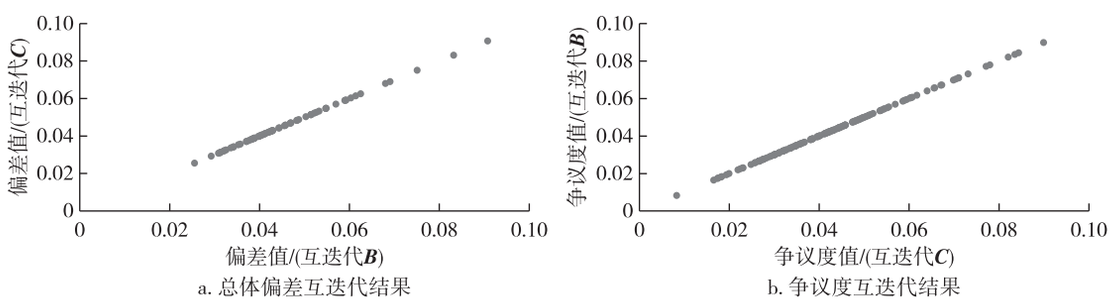


图 4　总体偏差及争议度互迭代结果

Fig. 4　Inter-iteration results of overall bias (a) and controversy (b)

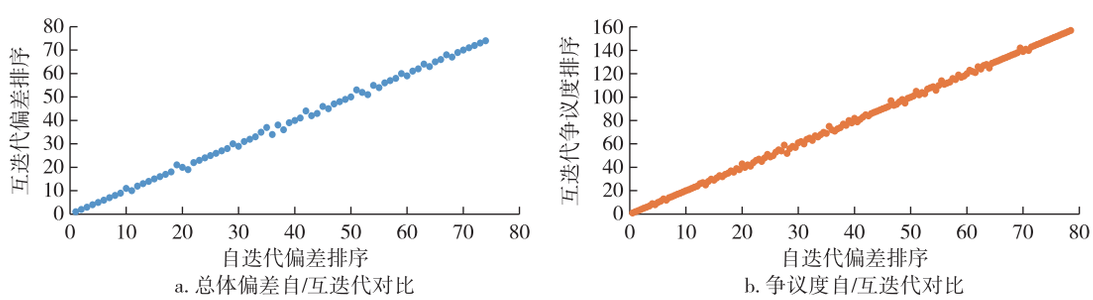


图 5　自迭代和互迭代求解结果对比

Fig. 5　Comparison of self-iteration and inter-iteration results for overall bias (a) and controversy (b)

性规范化处理导致无法定量比较,故图 3 中对比的是自迭代结果排序情况.图中横轴表示直接求解结果的排序,纵轴为将自迭代结果代入式(6)和(5)的计算结果排序,排序越一致则散点越接近对角线.可以看出偏差排序差别巨大,争议度排序略好但仍呈现出明显发散状.

利用本文的互迭代方式得到 $\boldsymbol{B}$ 和 $\boldsymbol{C}$,同样可按照上述过程再次利用两式反算出 $\boldsymbol{C}$ 和 $\boldsymbol{B}$.图 4 给出了这些结果的对比结果,互迭代结果间表现出了高度一致性.最后,图 5 展示了不同求解方式的结果排序对比,横轴为自迭代直接得到的 $\boldsymbol{B}$ 和 $\boldsymbol{C}$ 排序,纵轴为互迭代结果排序.可知,基于两种求解方式的排序基本相同,尤其是排名靠前的部分,而排序不同之处均为小幅差异,对于专家偏差分析影响非常小.但考虑到互迭代方式始终维持着总体偏差和争议度间的关联性,本文提出的求解思路明显更加合理.此外,互迭代的另一个优点是无需引入规范化处理,从而结果可定量比较,因此只有图 4 中直接展示了总体偏差值和争议度值而非其排序,更利于科研管理人员后续开展更精细的分析工作.

3.3　代表性专家实例分析

如前文所述,3 个偏差指标各有侧重,相互结合才能较好地分析专家偏差情况.本部分以几个专家实例分析一些有代表性的偏差表现,同时也验证本文所用指标的有效性.首先是表 2 中的第 1 位专家,其总体偏差最大且是唯一异常评分次数超过阈值的专家.此外,通过 t 检测认定该专家有给低分的习惯.

图 6 给出了其评分数据和相应课题平均分,其中柱状分数为专家 1 参评的课题平均分,折线为专家 1 的评分.注意图中误差棒以±2 倍标准差为上下限,以便快速确定异常评分位置(后续图 7—9 采用相同设置).图中折线一直处于平均分以下,在课题 107、108、110 处出现评分异常,其余 3 个课题中评分也逼近了下限.特别是对于评分波动性较大的课题 108,专家 1 评分仍能超出正常范围.过多的异常评分和明显的固有偏差集中体现为极大的总体偏差,这表明该专家问题严重,在后续评审活动中不建议将其继续作为技术专家.

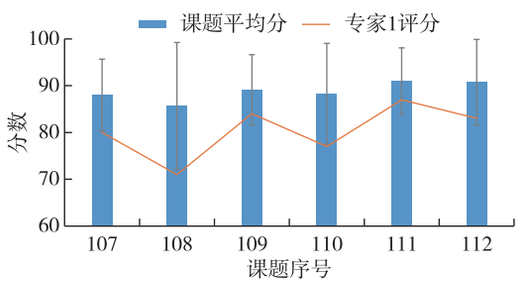
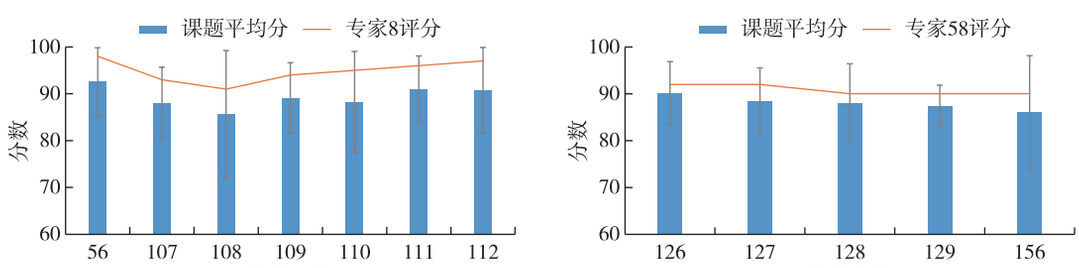


图 6 专家 1 评分及相应课题平均分

Fig. 6 Scores given by expert 1 and their corresponding biases

总体偏差同样较大是专家 8,从固有偏差检测结果知其具有很强的给高分习惯,但未有异常评分.在图 7 中专家 8 表现与分析一致,评分全部高于平均分且处于正常区间的较高位置.但对于分数波动较大的课题 108,该专家给出了较为合理的评分,这是一个比较好的现象.评分尺度过于宽松是专家 8 总体偏差较大的主导因素,证明了固有偏差过大时也会对总体偏差产生严重影响,但这种“尺子”方面的问题仅从总体偏差无法发现,说明了结合固有偏差和总体偏差的必要性.与专家 8 相反,从表 2 和图 7 中均能确定专家 58 也有给高分的倾向,但程度更低,从而总体偏差较小,仅排在第 58 位.仅以 t 检验结果而言,有明显固有偏差的专家分布在表 2 排序的各部分,表明固有偏差在整体上对专家总体偏差的影响还是可以接受的.

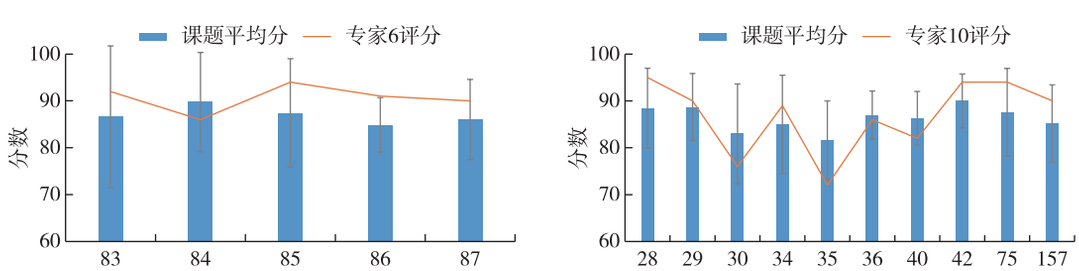
与固有偏差不同,异常次数的多少和总体偏差的大小显著相关,表 2 中异常次数达到阈值的专家均在排序前列.原因在于专家评分次数普遍较少,集中在 5~7 次,所以每个异常值的出现均会对总体偏差有不小贡献.例如专家 6 总体偏差较大而 t 检测值低,从图 8 中也可看出仅有轻微的给高分倾向,但 5 次评分中就有 1 个异常值.当然,也并非异常评分少且无明显固有偏差就意味着总体偏差小,原因有二:一是即便没有或较少出现异常评分,还可能存在较多接近但未超出正常范围的评分;二是固有偏差不



a. 专家8评分及偏差情况　　b. 专家58评分及偏差情况

图 7 专家 8 和专家 58 评分及相应课题平均分

Fig. 7 Scores given by expert 8 and expert 58 and their corresponding biases



a. 专家6评分及偏差情况　　b. 专家10评分及偏差情况

图 8 专家 6 和专家 10 评分及相应课题平均分

Fig. 8 Scores given by expert 6 and expert 8 and their corresponding biases

明显也可能是因为评分忽高忽低,如图 8 中显示的专家 10 评分情况.该专家参与了 10 次验收评审,仅 1 次评分异常(相应阈值为 2 次),t 检验值 0.286 接近于 0,表明其无过宽或过严的评分惯性.但从图 8 中可知其评分在平均分上下波动,并且过半评分接近正常范围上下限,故偏差排序靠前.对于类似表现的专家,仅凭异常评分和固有偏差检测是不够的,加入总体偏差才能正确分析其偏差情况.

在分析了 5 个存在问题及表现各不相同的实例后,图 9 给出专家 72 的评分及相关课题分数信息,作为较理想的专家示例,其总体偏差极小、无异常评分,仅评分尺度略显严格,图中也可看出该专家评分与平均分非常一致.

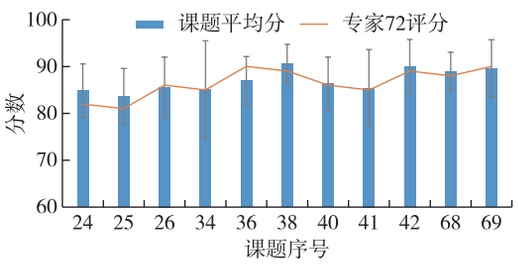


图 9 专家 72 评分及相应课题平均分

Fig. 9 Scores given by expert 72 and their corresponding biases

3.4 专家评分偏差归类

以上结合典型实例分析了 3 个偏差指标的关联性:异常评分体现突变性信息,对总体偏差影响明显;固有偏差体现一致性的评分倾向,对总体偏差有一定影响;总体偏差是综合性评价,既包含了突变性和倾向性信息,又体现了两者之外的一些因素,但不能细致区分偏差表现.分析工作应先根据总体偏差大体锁定问题专家群体,再联合异常评分、固有偏差判断专家具体问题并确定处理措施.因此,表 3 列出了以这 3 个指标划分的 8 种专家偏差类型及建议的应对措施.

异常评分次数和固有偏差均有检测阈值.为了使总体偏差保持一致,本部分简单采用大津法[24](又称最大类间差方法)寻找可将总体偏差分为差距最大的两类的阈值,这样即可利用 3 个指标的阈值将任一专家归类到特定偏差类型.根据大津法得到高总体偏差专家(排序 1~21)和低总体偏差专家(排序 22~74).各类型专家人数和占比也列于表 3.这些偏差类型不限于本文数据,在其他科技计划管理活动中同样可以应用.

对于部分专家需进一步培训和沟通,有针对性地矫正评分行为.对建议措施解释如下:

1) Ⅰ类专家严重影响评分可靠性,不建议继续参与验收评审.

2) Ⅱ类专家总体偏差大、异常多、评分忽上忽下,可以推断频繁受外在因素干扰且影响程度较大(如与课题团队间的好恶关系、不正确的刻板印象等).主要问题在于评分独立性、客观性不足,应加强此方面意识培训.此外还应观察其 t 检验值是否已接近阈值,预防Ⅱ类专家转为Ⅰ类.

3) Ⅲ类专家评分尺度问题明显,或偏高(如受同情心理效应影响)或偏低(如有高标准、严要求的评审习惯).较大的总体偏差表明该问题已明显影响到评分合理性.应多与此类专家沟通,令其加强尺度把握.

4) Ⅳ类专家偏差大但其他指标正常,说明其评分上下波动却没有过于极端.推测此类专家的主要问题在于对评分标准理解不足而非受外在因素的严重干扰,应加强培训提高验收规范内化程度.此外,也存在评分次数不多使异常评分和固有偏差检测不准确的可能,仍需跟踪观察确定其是否为潜在的Ⅰ/Ⅱ/Ⅲ类专家.

表 3 专家评分偏差类型及应对措施

Table 3 Types of experts according to their score biases and countermeasures

偏差类型	总体偏差	异常评分	固有偏差	应对措施	人数	占比/%
Ⅰ型	大	多	强	建议不再列入遴选范围	2	2.7
Ⅱ型	大	多	弱	加强评分独立性和客观性,避免外在因素干扰	4	5.4
Ⅲ型	大	少	强	加强评分宽严尺度的把握(重度)	3	4.1
Ⅳ型	大	少	弱	提高评审评分规则内化程度并后续跟进考察	12	16.2
Ⅴ型	小	多	强	无(近乎不可能类型)	0	0
Ⅵ型	小	多	弱	注意与参评课题好恶关系等因素影响	1	1.4
Ⅶ型	小	少	强	注意评分宽严尺度的把握(轻度)	9	12.2
Ⅷ型	小	少	弱	无(理想专家类型)	43	58.1

5)Ⅴ类专家仅为保证完整性而提出,基本不可能出现.原因在于异常评分多、固有偏差强均会增加总体偏差,极难同时出现低总体偏差.本文数据一定程度上证明了这一点.

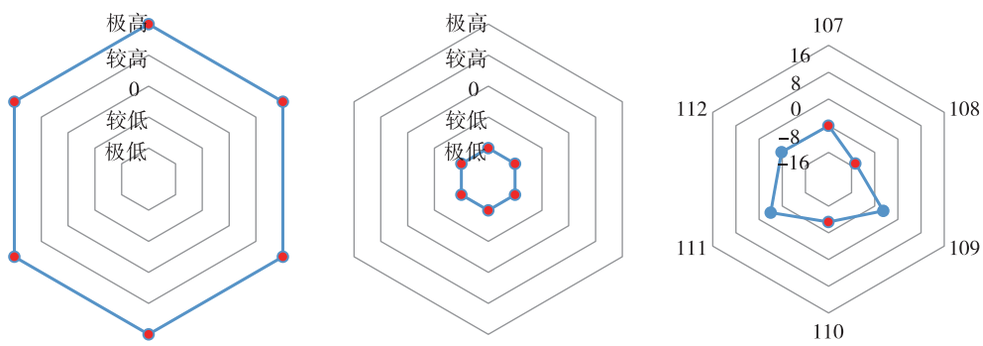
6)Ⅵ类专家与Ⅱ类成因相似但程度较轻,是在评分次数较少的专家中存在的小概率情况.因其偏差较小,不建议采用强化培训,应先进一步搜集相关信息确定外部因素来源后,提醒专家注意该因素影响.

7)Ⅶ类专家仅固有偏差偏高,提醒其稍微注意控制评分尺度即可.

8)Ⅷ类专家各项指标正常,无需任何处理措施.

需要注意的是,以硬阈值划分总体偏差只是一种粗略的分组方式.阈值附近的高、低偏差专家客观上并无太大区别,不能粗暴地认定前者一定有严重问题而后者没有.表3仅是给出了一些参考措施建议,对于接近总体偏差阈值的专家应根据情况具体讨论.雷达图因其形状的规律性和对比的便利性在分析偏差效应中非常适用[5].本文给出了部分偏差类型的理想雷达示意图和相应实例,可以看出雷达图非常形象地表达了类型间的不同特点.

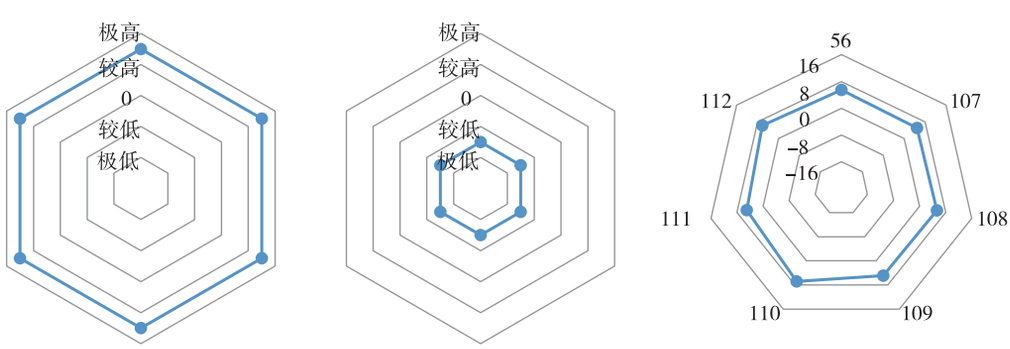
雷达图根据专家评分与课题平均分之差绘制,越外层的多边形表示高估越严重,越内层则越低估.角点上的数字代表课题序号,同一多边形的边构成了特定差值的等值线,差值列于多边形左上位置,红色点代表异常评分.对于专家实例,雷达图中显示范围统一为-16~16,便于公平比较.突变性的异常评分会造成雷达图中形状的不规律变化.对于涉及异常评分较多的类型,尤其是可能既有高异常分又有低异常分的情况(Ⅱ和Ⅵ型),并没有理想的雷达示意图可代表其多样性表现.即便异常评分少,但若总体偏差高且无明显固有偏差(Ⅳ型),评分仍然是围绕课题平均分在较大范围内上下波动,同样难以找到理想示意图.排除掉以上三类和近乎不可能的类型(Ⅴ型),图11—13展示了Ⅲ、Ⅶ、Ⅷ三种类型.另外,同时满足异常评分多和固有偏差强的条件下,异常评分或者多为极高分、或者多为极低分,不规律性显著降低,所以Ⅰ型也可找到理想雷达示意图(图10).从图10—13可知,专家1(偏差大、异常多、偏低估)、专家8(偏差大、异常少、偏高估)、



a. Ⅰ型雷达图:宽松型　b. Ⅰ型雷达图:严格型　c. 专家1偏差雷达图

图10　Ⅰ型专家偏差雷达图及实例

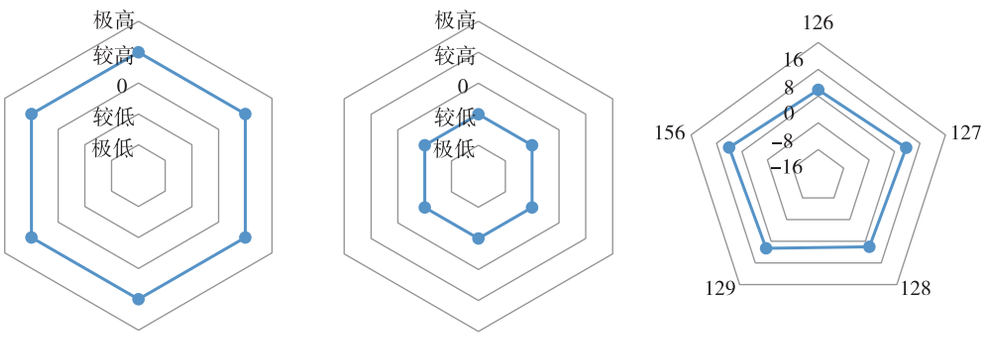
Fig. 10　Radar chart of type Ⅰ expert bias (a and b) and an example (c)



a. Ⅲ型雷达图:宽松型　b. Ⅲ型雷达图:严格型　c. 专家8偏差雷达图

图11　Ⅲ型专家偏差雷达图及实例

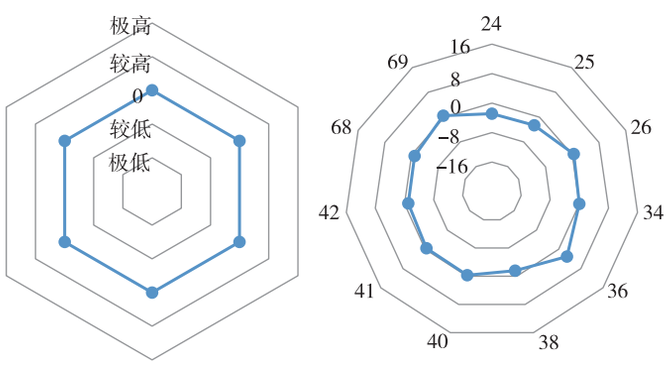
Fig. 11　Radar chart of type Ⅲ expert bias (a and b) and an example (c)



a. Ⅶ型雷达图:宽松型　b. Ⅶ型雷达图:严格型　c. 专家58偏差雷达图

图 12　Ⅶ型专家偏差雷达图及实例

Fig. 12　Radar chart of type Ⅶ expert bias (a and b) and an example (c)



a. Ⅷ型雷达图　b. 专家72偏差雷达图

图 13　Ⅷ型专家偏差雷达图及实例

Fig. 13　Radar chart of type Ⅷ expert bias (a) and an example (b)

专家 58(偏差低、异常少、偏高估)、专家 72(偏差低、异常少、固有偏差弱)与相应类型的理想雷达图非常相似,说明这 4 种偏差类型确有稳定的雷达图形状.即便不采用本文的 3 个指标,科研管理人员也可利用雷达图直接完成简单的偏差分析工作,至少能够快速找到理想专家群体(Ⅷ型)以提高评审结果可靠性,或者找到Ⅰ型专家群体减少其参评次数甚至不再作为专家人选.

4　总结

在科技管理工作中,验收评审有着评估课题完成水平、衡量科研产出价值的重要作用.开展评审专家可靠性研究对于科技评审活动是十分有指导意义的.因此,本文结合数据挖掘算法和数理统计方法给出了衡量专家评分偏差的 3 个定量指标,以对“十二五”863 计划某技术领域课题验收专家的评审行为进行初步探索.分析发现,该领域验收专家评分整体合理,仅 1 人次评分异常明显;固有偏差处于可接受范围.本文还根据偏差指标进一步归纳了 8 种偏差类型并给出应对措施建议,此项研究是对现阶段科研管理相关工作的完善与延伸.科技部近期正在开展“十三五”国家重点研发计划各重点专项首批到期项目的综合绩效评价,分析结果可用于绩效评价专家遴选和评前培训,帮助特定专家群体内化评审规范并降低评分习惯、个人偏好、外部因素等影响.此外,文中采用的评价体系和专家偏差类型同样可在其他科研管理活动中发挥评价评审过程、规范评审行为的作用.为响应“三评”(项目评审、人才评价、机构评估)改革意见[25],下一步工作将聚焦推进本文评价体系在多项国家科技计划乃至各类“三评”活动中的推广应用.一来从专家偏差性和评审对象争议性两方面综合评价评审过程、完善评审机制,同时广泛采样检验本文分析方法的泛化能力;二来基于总体偏差、固有偏差和异常评分并结合大量评分数据,既可以从不同粒度归纳总结专家潜在的共性问题和分析差异化的评审行为,又能根据所得经验和专家历史偏差评价结果辅助“三评”专家遴选工作,提升科技评审效度.正值教育部、科技部联合印发《关于

规范高等院校SCI论文相关指标使用 树立正确评价导向的若干意见》[26]之际,希望本文能够对其中的"完善学术同行评价"、"规范各类评价活动"等内容提供方法论支撑.

参考文献

References

[1] 中国科学技术部国家遥感中心.羲和系统信号获取说明[EB/OL].(2014-06-10)[2020-03-02].http://csi.gov.cn/nrscc/kjjhgl/gj863jh/863tzgg/201406/t20140610_32726.html
The National Remote Sensing Center of China(NRSCC). Description of Xihe system signal acquisition[EB/OL].(2014-06-10)[2020-03-02].http://csi.gov.cn/nrscc/kjjhgl/gj863jh/863tzgg/201406/t20140610_32726.html

[2] 中国科学技术部.中国反射面天线技术引领国际大科学工程核心设备研制[EB/OL].(2018-10-26)[2020-03-02].http://www.most.gov.cn/gnwkjdt/201810/t20181026_142436.htm
Ministry of Science and Technology of the People's Republic of China. China's reflector antenna technology leads the development of core equipment for international mega-science project[EB/OL].(2018-10-26)[2020-03-02].http://www.most.gov.cn/gnwkjdt/201810/t20181026_142436.htm

[3] 彭龙.科研活动的实力评价及评价的偏差问题[J].科学管理研究,1993,11(3):42-45
PENG Long. The deviation of strength evaluation and evaluation of scientific research activities[J]. Scientific Management Research,1993,11(3):42-45

[4] 谢焕瑛,张健.国家重点实验室评估专家的若干问题研究[J].研究与发展管理,2006,18(4):108-111
XIE Huanying,ZHANG Jian.Research on the problems of the reviewers of the state key laboratory[J].R & D Management,2006,18(4):108-111

[5] 谢焕瑛.国家重点实验室评估专家评分偏差效应分析[J].研究与发展管理,2006,18(6):134-138
XIE Huanying.Analysis on the warp of the review of the state key laboratory[J].R & D Management,2006,18(6):134-138

[6] 张健,谢焕瑛.国家重点实验室评估方法的若干问题研究[J].管理学报,2008,5(2):279-281
ZHANG Jian, XIE Huanying. Research on several problems on the evaluation of the state key laboratory[J]. Chinese Journal of Management, 2008, 5(2):279-281

[7] 杨晓秋.关于国家重点实验室评估的思考[J].实验室研究与探索,2015,34(9):141-144,148
YANG Xiaoqiu. Thoughts on the evaluation of the state key laboratory[J]. Research and Exploration in Laboratory,2015,34(9):141-144,148

[8] 孙晓敏,张厚粲.国家公务员结构化面试中评委偏差的IRT分析[J].心理学报,2006,38(4):614-625
SUN Xiaomin,ZHANG Houcan.An IRT analysis of rater bias in structured interview of national civilian candidates [J].Acta Psychologica Sinica,2006,38(4):614-625

[9] 苏永华,柴雪,丁玉洋.国家公务员录用面试初步研究[J].应用心理学,1998,4(1):15-20
SU Xuehua, CHAI Xue, DING Yuyang. Preliminary research on interview of national civil servant recruitment[J].Chinese Journal of Applied Psychology,1998,4(1):15-20

[10] 陈宛玉,戴海琦.教育教学能力测试的GT和多面Rasch模型分析[J].考试研究,2013,9(3):70-78
CHEN Wanyu, DAI Haiqi. Analysis of GT and multidimensional Rasch model for educational and teaching capability testing[J]. Examination Research,2013,9(3):70-78

[11] 孙晓敏,薛刚.多面Rasch模型在结构化面试中的应用[J].心理学报,2008,40(9):1030-1039
SUN Xiaomin, XUE Gang. A many-faceted Rasch model analysis of structured interview[J]. Acta Psychologica Sinica,2008,40(9):1030-1039

[12] 周燕,曾用强.机助英语听说考试计算机自动评分的多层面Rasch模型分析[J].外语测试与教学,2016,6(1):22-31
ZHOU Yan,ZENG Yongqiang.Multi level Rasch model analysis of computer-aided English listening and speaking test[J]. Foreign Language Testing & Teaching,2016,6(1):22-31

[13] 王佶旻,李潇.基于Rasch模型的参数估计方法比较研究[J].中国考试,2017,14(9):11-21
WANG Jimin, LI Xiao. The comparison between the method of MLE, MLE/EM and BMES under the Rasch model[J].China Examinations,2017,14(9):11-21

[14] Lauw H W, Lim E P, Wang K. Bias and controversy in evaluation systems[J]. IEEE Transactions on Knowledge and Data Engineering,2008,20(11):1490-1504

[15] Dai H B,Zhu F D,Lim E P,et al.Detecting anomalies in bipartite graphs with mutual dependency principles[C]//Proceedings of the Twelfth IEEE International Conference on Data Mining,2012:172-180

[16] Xie H,Lui J C S.Incentive mechanism and rating system design for crowdsourcing systems:analysis,tradeoffs and inference[J].IEEE Transactions on Services Computing,2018,11(1):90-102

[17] Zielinski K, Nielek R, Wierzbicki A, et al. Computing controversy: formal model and algorithms for detecting controversy on Wikipedia and in search queries[J]. Information Processing & Management,2018,54(1):14-36

[18] 吕书龙,梁飞豹,刘文丽.关于评委评分的评价模型[J].福州大学学报(自然科学报),2010,38(3):358-362
LÜ Shulong, LIANG Feibao, LIU Wenli. An evaluation model of rater score[J]. Journal of Fuzhou University (Natural Science Edition),2010,38(3):358-362

[19] 梁薇.基于投影寻踪模型的网评评委综合素质评价[J].统计与决策,2017,33(23):60-63
LIANG Wei. Evaluation of comprehensive quality of web evaluation judges based on projection pursuit model[J]. Statistics and Decision,2017,33(23):60-63

[20] Gao M, Chen L, Li B, et al. Projection-based link prediction in a bipartite network[J]. Information

Sciences,2017,376:158-171
[21] Liu W, Wu S, Wu X, et al. Mixed probability inverse depth estimation based on probabilistic graph model[J]. IEEE Access,2019,7:72591-72603
[22] Soldi G,Meyer F,Braca P,et al.Self-tuning algorithms for multisensor-multitarget tracking using belief propagation [J].IEEE Transactions on Signal Processing,2019,67 (15):3922-3937
[23] Berkhin P.A survey on PageRank computing[J].Internet Mathematics,2005,2(1):73-120
[24] Otsu N.A threshold selection method from gray-level histograms[J].IEEE Transactions on Systems,Man,and Cybernetics,1979,9(1):62-66
[25] 中共中央办公厅 国务院办公厅印发《关于深化项目评审、人才评价、机构评估改革的意见》[EB/OL].(2018-07-03)[2020-03-02]. http://www.gov.cn/zhengce/2018-07/03/content_5303251.htm
[26] 教育部 科技部印发《关于规范高等学校 SCI 论文相关指标使用 树立正确评价导向的若干意见》的通知[EB/OL].(2020-02-18)[2020-03-02]. http://www.gov.cn/zhengce/zhengceku/2020-03/03/content_5486229.htm

Analyze the problems and suggestions of expert score bias based on inverse reinforcement model and mathematical statistics methods

TIAN Linlin[1] SUN Weidong[1] ZHANG Chi[1] GUO Ming[1] WEI Nadu[1]
1 National Remote Sensing Center of China,Beijing 100036

Abstract During the 12th Five Year Plan period,National High Technology Research and Development Program of China (863 Program),as one of the main drivers of science and technology development,has provided important support for improving China's scientific and technological strength and innovation ability.Acceptance experts play a key role in assessing the level of subject completion and measuring the value of scientific research achievements.The reliability of their scores is directly related to the rationality of 863 program implementation evaluation.Therefore, taking the subject acceptance in a certain field as an example,this paper combines an inverse reinforcement-based model and mathematical statistics methods to systematically analyze the rating bias of technical acceptance experts. Finally,these experts are divided into 8 categories according to their rating performances and corresponding suggestions are given respectively.The results show that the scores are reasonable as a whole,and most experts can give reliable scores;although there are some differences in the rating scales,they are basically in an acceptable range.This study will provide a reference for the review work related to national science and technology plan,and other scientific research management activities,in order to reasonably carry out expert evaluation,refine and standardize review behaviors,improve field expert databases,and select acceptance expert candidates.

Key words National High-tech R&D Program of China;science and technology project review;science and technology program management;expert rating;bias analysis