第46卷第2期 中国科学技术大学学报 Vol. 46, No. 2
2016年2月 JOURNAL OF UNIVERSITY OF SCIENCE AND TECHNOLOGY OF CHINA Feb. 2016

文章编号:0253-2778(2016)02-0165-08

专家群评价结果可信度分析与检验

赵亚娟

(合肥工业大学管理学院,安徽合肥 230009)

摘要:专家存在的信息结构性偏差和认知结构性偏差等问题,导致专家评价存在随机性和不一致性,专家群评价结果的可信度受到质疑.使用心理统计学的信度系数对单个专家和专家群评价结果的可信度进行度量,并运用双因素固定效应模型对评价结果的误差来源做进一步分析和检验,能同时实现分析判断评价结果的公平合理性和专家评价的可信程度.最后通过一个评标案例进行了实证分析.

关键词:评价结果;可信度;方差分析

中图分类号:C934 **文献标识码**:A doi:10.3969/j.issn.0253-2778.2016.02.00

引用格式: Zhao Yajuan. Reliability analysis of group experts and testing of evaluation outcomes[J]. Journal of University of Science and Technology of China, 2016,46(2):165-172.
赵亚娟. 专家群评价结果可信度分析与检验[J]. 中国科学技术大学学报,2016,46(2):165-172.

Reliability analysis of group experts and testing of evaluation outcomes

ZHAO Yajuan

(*School of Management, Hefei University of Technology, Hefei* 230009, *China*)

Abstract: Information structural bias and cognitive structural bias and other issues lead to randomness and inconsistency of expert evaluations, calling into question the reliability of evaluation outcomes of expert groups. To solve this problem, psychological statistics reliability was used to measure expert reliability for individual experts and expert groups; then, two-factor fixed effects model was used to further analyze and test the sources of error. This can simultaneously analyze and evaluate the fairness and reasonableness of evaluation outcome and the reliability of experts. Finally, an empirical analysis was carried out with a tender evaluation case.

Key words: evaluation outcome; reliability; analysis of variance (ANOVA)

0 引言

日常生活中的评标、体育比赛,其实质是一种多属性群决策活动,是集结有关专家群体对某一决策问题的意见的过程,其过程主要涉及3个方面:评价准备阶段(评价因素及权重的设置和专家对评价标准的熟悉)、获取评价专家偏好数据的阶段(定量或定性的评价)和专家群评价数据集结阶段.大多数文献主要针对于第一阶段评价因素及权重设置和第三阶段专家群意见集结方面进行研究,但对于第二个阶段,即专家群体数据集结之前的专家群体评价数据的一致性和稳定性分析却比较少,其实质是对专

收稿日期:2014-09-10;**修回日期**:2015-06-08

作者简介:赵亚娟,女,1982年生,博士生. 研究方向:招投标及产业经济学. E-mail: xinsheng_51@163.com

家群评价的可靠性或可信性进行再评估.

较早对专家可靠性进行分析的是吴敬业和汤理,他们对序贯群评价问题中的个体评价专家可靠性度量的准则、个体与群体的评价特征进行分析[1],李学栋和李浩志在此文的基础上,对专家可靠性基于"临界值"、"正确度"以及灰关联3种分析法进行研究[2].元继学[3]对专家主观评分比赛中群决策机制进行分析,构建了集中性指标和相似性指标对群决策过程中的一致性问题进行度量;吕书龙等[4]对评分数据采用参数和非参数结合的方法进行检验,提出了评分控制模型、偏差吻合模型以及区分度模型并进行了验证.高先务等[5-6]对评估结果的可信度评价提出了新的思路,利用方差分析和卡方分布数理统计的方法对评标结果进行了检验.Cheng等[7]利用数据库知识挖掘中的孤立点检测技术,建立了相似性系数对评标数据进行检测和风险预警;梁晶[8]基于系统偏差、非共识性偏差、同行偏差和偶然偏差4个指标建立了评标专家异常性评分的判定方法,并行实证分析.

文献[1,3]主要是对群决策可靠性的机制和特征进行分析;文献[2,3]基于决策矩阵对群决策结果的一致性程度进行分析;文献[4-8]主要使用数理统计的方法对专家群体评价结果进行检验,检验专家群体评价的结果是否存在偏差,其中文献[7]只是一种检测和风险预警机制,没有对评标结果的可信度进行分析.因此以上文献基本上是对专家群评价可信性的机制和特征进行分析以及对专家群体的可信程度进行检验,从整体上判断专家群评价的可信程度,没有具体到对单个专家的可信程度进行度量以及评价结果差异来源进行分析和检验.本文首先对专家群评价问题进行简单描述,指出专家评价阶段中存在的问题;随后借鉴心理、体育统计学等相关理论,对专家评价结果进行专家和群体专家可信度系数的度量,由于可信度系数并不能对专家可信度高低来源解释清楚,本文进一步对专家可信度进行方差分析,并对专家群评价结果不一致进一步进行检验,识别出哪位专家评价存在系统性偏差.

1 问题的描述

在专家群评价结果的可信性分析中,专家的选取是评价结果可信性或者评价公平性的一个重要因素,如果选择不当,评价结果的准确性和可信性就会受到影响,并容易遭受相关利益人的投诉和非议.

专家群决策模型可以用下面的数学模型来描述[6,9]:$E=(E_1,E_2,E_m)$是专家集合,$A=(A_1,A_2,A_n)$为评价方案集合,$F=(f_1,f_2^i,i=1,2,m,f_3)$为映射规则集,设$X$为待评价的参数,$X$在原则上具有客观的真实值,$X_i=(X_{i1},X_{i2},X_{im})$为专家$i$的评分向量,$X_{ij}$表示专家$i$对方案$j$的评分值,则专家群体评估问题可记为$P=(A,E,X,F)$,专家群体评价的过程主要包括3个阶段:

① $D=f_1(A_1,A_2,A_n)$,f_1为评价准备阶段,即数据和资料的收集和处理阶段,包括评价因素、权重及对象的确定;

② 确定专家集合E,对D进行评价,产生评价向量$X=f_2^i(X_1,X_2,X_m)$;f_2^i为D的决策映射;

③ 由评价向量X产生群体评价向量X_g,即$G=f_3(X_{i1},X_{i2},X_{im})$,$f_3$为$m$个专家数据的集结映射方法,$X_g$就为集结映射的评价结果.

在专家群评价中,f_1为数据整理的预处理阶段,f_3为专家偏好的集结阶段,在确定f_1,f_3下,D,G就已经确定,f_1,f_3具有确定性的结构.由于评价方案及评价因素本身的模糊性以及专家所掌握信息的有限性和事物认识上的局限性导致专家存在"信息结构性偏差"和"认知结构性偏差"的问题,此外专家主体有追求自身利益偏好等原因,造成评价参数X的评价结果不准确.因此f_2^i不具有明确的结构,而具有随机性和不稳定性,专家评价结果的可信度取决于专家的评价能力、素质和问题的难度[10].因此,在群专家偏好集结第三阶段f_3之前,如果没有考虑到专家的评价能力和可信性,没有依据专家的评价可信程度赋予专家适宜的评价权重而对所有专家同等程度对待,如算术平均法,其群评价结果是有失合理和公平的.所以,本文利用真分数理论的信度系数以及方差分析对专家评价结果的可信性及其差异进行深入研究.

2 专家可信度的评价

测量的可靠性或评价的可信度是指在同等条件下,对同一批受评或测量对象评价或测量时,评价或测量结果的一致性程度.可信性系数r定义为评价或测量对象真值方差与评价或实测值方差的比值,其取值范围为$0<r\leqslant 1$,其值越接近1,则可信度越高.

在上述专家群体评价问题描述中,在第二阶段专家群体评价时,专家评价具有随机不确定性,这种

随机不确定性表现为评估结果对实际对象“真值”的偏离. 本文只对单指标评分值或专家最后的综合评分值进行分析，对专家多属性评价问题不做分析. 因此，假设有 n 个评价方案，m 个专家，X_{ij} 表示第 j 个专家对第 i 个方案进行评价的得分，t_i 表示第 i 个方案的真实值，e_{ij} 表示误差，$i=1,2,\cdots,n$；$j=1,2,\cdots,m$. 根据真分数理论，评价值由被评价对象的真值和误差组成：

$$X_{ij} = t_i + e_{ij} \tag{1}$$

并假定：t_i 和 e_{ij} 相互独立；e_{ij} 和 e_{ik} 相互独立.

根据信度系数的定义可知，专家群体评价的信度系数为

$$r_j = \frac{\sigma_T^2}{\sigma_X^2} = 1 - \frac{\sigma_e^2}{\sigma_X^2} \tag{2}$$

这是专家群体评价的可信程度，黄正南[11]对信度系数的重新定义，可得到单个专家评价的可信程度，即单个专家的信度系数：

$$r_j = \frac{\sum_{k=1}^{m}(x_j - \overline{x_i})(x_k - \overline{x_i})}{\sum_{k=1}^{m}\frac{1}{2}[(x_j - \overline{x_i}) + (x_k - \overline{x_i})]} = \frac{\sum_{k=1}^{m} l_{jk}}{\sum_{k=1}^{m}\frac{(l_{jj} + l_{kk})}{2}} \tag{3}$$

式中，l_{jk} 为 x_j 和 x_k 的离均差积和，l_{jj}，l_{kk} 分别为 x_j，x_k 的离均差和.

那么，专家群体的信度系数通过斯庇尔斯-布朗公式的式子进行校正，即有

$$R = \frac{mr_j}{1 + (m-1)r_j} \tag{4}$$

式(4)即为原有信度分析中的 Cronbach's α 系数，m 为专家人数，随着 m 人数的增加，专家群体的信度就越大，随着 m 的增加，可信度趋近为 1，这就是平常评价中为了提高评价结果的信度，一般都会增加评价者人数.

3　方差方法检验

根据式(2)，可知信度系数是个相对数，表示为专家评分的误差(σ_e^2)与方案个体差异所导致的误差(σ_X^2)的比值，所以需要对误差 e_{ij} 的来源、成分及它们对可信度的影响进行分析探讨，从而能比较准确地衡量专家的可信程度. 上文中的信度系数其本质是以相关系数为基础，是组内相关系数中的一种，只能在某种程度上对评价过程中产生的误差进行估计，而不能对评价结果中所可能包含的各种来源误差进行甄别. 单纯利用信度系数对评价结果可信度进行估计是不全面的和不真实的[12]. 因此，需要对专家群体评价结果进行误差分析和检验. 方差分析能从多角度对多种误差进行分析，是检验两组以上均值之间是否有显著性差别的一种常用方法，故方差分析又被称为“F 检验”.

在某个项目评价中，m 个专家对 n 个方案关于某个指标进行一次性评价，在这个决策问题中，主要有两个因素：一个因素是方案 A，有 n 个被评价方案，就有 n 个不同水平($A_1, A_2, \cdots, A_n$)；另一个因素是专家 B，有 m 个专家，就有 m 个不同水平($B_1, B_2, \cdots, B_m$). 因素 A 和 B 作用于评价值 X，就相当于在 A，B 的每种组合水平(A_i, B_j)下做一次试验，试验结果为 X_{ij}($i=1,2,\cdots,n, j=1,2,\cdots,m$)，由于专家评价都是相互独立的，则 X_{ij} 是相互独立的. 其次，虽然专家一般是从专家库中随机抽取选择的，但每个专家都有其特有的评价特点，故评价结果可信度估计只限于样本(抽取或选取的)专家的推断. 因此，通过以上分析，这种专家群体评分模式其实质是双因素固定效应模型(two-way fixed effects model)，且不考虑方案和专家之间的交叉效应[13-15]：

$$\left.\begin{aligned} &x_{ij} = t + \alpha_i + \beta_j + \varepsilon_{ij}, \\ &\sum_{i=1}^{n}\alpha_i = 0,\ \sum_{j=1}^{m}\beta_j = 0, \\ &\varepsilon_{ij} \sim N(0,\sigma^2), i = 1,2,\cdots,n, j = 1,2,\cdots,m, \\ &\quad \text{且它们之间互相独立} \end{aligned}\right\} \tag{5}$$

那么根据以上分析，式(5)中 α_i，β_j 分别表示由被评价方案 B 因素、专家 A 因素所引起的误差，ε_{ij} 表示这两种因素之外的随机因素引起的随机误差.

假设

$$\overline{x} = \frac{1}{mn}\sum_{i=1}^{n}\sum_{j=1}^{m} x_{ij},\ \overline{x_{i\cdot}} = \frac{1}{m}\sum_{j=1}^{m} x_{ij},\ \overline{x_{\cdot j}} = \frac{1}{n}\sum_{i=1}^{n} x_{ij},$$

有

$$Ex_{i\cdot} = t + \alpha_i,\ Ex_{\cdot j} = t + \beta_j,\ E\overline{x} = t \tag{6}$$

由(6)可知，总离差分解为

$$S_A = \sum_{i=1}^{n}\sum_{j=1}^{m}(x_{i\cdot} - \overline{x})^2,$$

$$E(MS_A) = E\left(\frac{S_A}{n-1}\right) = \sigma^2 + m\sum_{i=1}^{n}\alpha_i;$$

$$s_B = \sum_{i=1}^{n}\sum_{j=1}^{m}(x_{\cdot j} - \bar{x})^2,$$

$$E(MS_B) = E\left(\frac{S_B}{m-1}\right) = \sigma^2 + n\sum_{j=1}^{m}\beta_j;$$

$$s_\varepsilon = \sum_{i=1}^{n}\sum_{j=1}^{m}(x_{ij} - x_{i\cdot} - x_{\cdot j} + \bar{x})^2,$$

$$E(MS_\varepsilon) = E\left(\frac{S_\varepsilon}{(n-1)(m-1)}\right) = \sigma^2;$$

$$s_{总} = \sum_{i=1}^{n}\sum_{j=1}^{m}(x_{ij} - \bar{x})^2,$$

$$s_{总} = s_A + s_B + s_\varepsilon.$$

因此对模型的假设检验为

$H_{01}:\alpha_1=\alpha_2=\cdots=\alpha_i=0$, $H_{11}:\alpha_i(i=1,2,\cdots,n)$ 中至少有一个不为零;

$H_{02}:\beta_1=\beta_2=\cdots=\beta_j=0$, $H_{12}:\beta_j(j=1,2,\cdots,m)$ 中至少有个不为零.

在 H_{01}, H_{02} 为真时,可得 $\frac{S_{总}}{\sigma}$ 服从 $\chi^2(mn-1)$ 的卡方分布,根据 Cochran 定理可得:$\frac{S_A}{\sigma}$, $\frac{S_B}{\sigma}$, $\frac{S_\varepsilon}{\sigma}$ 分别服从 $(n-1)$, $(m-1)$, $(n-1)(m-1)$ 的 χ^2 分布,它们相互独立且 $(n-1)(m-1)+(n-1)+(m-1)=nm-1$.

为了检验 H_{01}, H_{02} 是否为真,需要构建统计量 F:

$$F_A = \frac{S_A/(n-1)}{S_\varepsilon/[(n-1)(m-1)]} = \frac{MS_A}{MS_\varepsilon},$$

$$F_B = \frac{S_B/(m-1)}{S_\varepsilon/[(n-1)(m-1)]} = \frac{MS_B}{MS_\varepsilon}.$$

所以对给定的假设显著水平 α 查 F 分布表,得到

$$F_\alpha[(n-1),(n-1)(m-1)],$$

$$F_\alpha[(m-1),(n-1)(m-1)],$$

如果 $F_A > F_\alpha$,则说明评价方案 A 效应是显著有差异,表明了被评价方案各自的优劣性,同时说明评价标准能够很好地区分和显示被评价方案的优劣差异;如果 $F_B > F_\alpha$ 则说明专家效应 B 存在显著差异,说明专家对方案的评价或认知偏好存在不一致,需要进一步通过统计方法进行检验,找出与其他专家评价存在不一致的专家.通过方差分析可知,只有当 $F_A > F_\alpha$, $F_B \leqslant F_\alpha$,即方案效应 A 差异显著,而专家效应 B 差异不显著时,则表明此次评价效果比较好,说明此次评价即能很好地区分被评价方案的优劣性,且专家间的评价一致性程度较高,能很好地体现竞争、公平的评价原则.因此进行方差分析,是对信度系数做进一步误差分解、分析和检验,能保证专家群评价结果可信度估计的准确性.

4 实证分析

评标在招标采购中占有重要地位,现阶段主要有两种评标法:最低投标价法和综合评估法.综合评估法其实质是多属性群决策,是指在投标满足招标文件实质性要求的前提下,关于价格、质量、商务、技术等因素对投标人的方案进行综合评价,按投标人综合评价的结果由优到劣的顺序确定中标候选人的评标方法.虽然法律要求评标标准和程序刚性化,但是评标过程中,存在大量无法明确和细致规定的情况存在,评价的模糊性和主观性较强,需要评标专家自行处理和决定,从而形成评标专家的“自由裁量权”.其次,专家存在上述分析中提到的“信息结构性偏差”和“认知结构性偏差”的问题,评标专家的评价可信性是受到质疑的.最后,根据《招标投标法》中第四十二条,实际上规定了招标项目的评标权和定标权全部划归到评标委员会的评标专家手中,评标专家具有主导招标项目的生杀大权.因此,基于上述分析,有必要对专家群评标结果进行可信度分析和检验,并使用实践操作中的一个真实案例进行实证分析.

本案例是某企业的施工总承包工程项目,工程项目总投资约 3 亿,招标类别属于施工-土建,标段共 1 个标段.招标方式是公开招标,需要事前进行资格预审,资格预审采用有限数量制.通过初步审查和详细审查的投标申请人数量超过 9 家时,由招标人依法组建的资格评审委员会按资格预审评分标准进行评定,确定资格预审得分前 9 家的投标申请人为潜在投标人.具体评分标准如表 1 所示,共 11 项指标,总分 100 分,从评分标准来看,评价指标的主观性比较强,而且评分分值区间较大,如建议书是一种主观性评价指标,专家的评分区间为 15 分.

本项目共有 24 个投标人提交了资格预审投标文件,经过初步审查有 18 家投标文件通过资格预审文件的形式和符合性评审要求,评标委员会对这 18 家投标文件按照详细评审标准进行打分,取得分前 9 位的申请人通过资格审查,进入投标阶段.根据《评标委员会和评标方法暂行规定》第九条规定,评标委员会由招标人或其委托的招标代理机构熟悉相关业务的代表以及有关技术、经济等方面的专家组成,成员人数为 5 人以上单数,其中技术、经济等方

面的专家不得少于成员总数的三分之二. 本项目评标委员会成员人数为 7 人,每个专家根据各指标对各投标人的资格预审文件进行打分. 7 个专家对 18 家投标人的综合打分情况整理如表 2 所示.

4.1　专家可信度计算

根据表 2 的数据和式(4),(5),可得到单个专家的可信度系数以及专家群体平均的可信度系数,如表 3 所示. 通过表 3 可知道,除了专家评委 b 和 d 的

表 1　评标因素和分值表

Tab. 1　The table of evaluation factors and scores

评标因素	企业类似项目业绩	企业荣誉	企业质量方面取得的荣誉	安全文明方面取得的荣誉	新技术应用方面取得的荣誉	企业信用情况	企业认证情况	项目经理业绩及表彰	技术负责人表彰	项目班子配备情况	建议书
分值	10	10	10	10	5	5	5	10	5	15	15

表 2　评分明细表

Tab. 2　The list of scoring

		评委 a	评委 b	评委 c	评委 d	评委 e	评委 f	评委 g	综合平均得分	排序
1	投标人 1	85	88.00	84.00	59.50	86.00	81	84.00	81.07	7
2	投标人 2	56	61.00	56.00	58.00	60.00	55	53.00	57.00	17
3	投标人 3	81	86.00	79.00	81.00	84.00	75	77.00	80.43	8
4	投标人 4	83	84.00	78.00	79.00	80.00	75	77.00	79.43	7
5	投标人 5	82	86	79.00	84.00	84.00	78	82.00	82.14	5
6	投标人 6	81	86.00	84.00	82.00	84.00	81.5	84.00	83.21	4
7	投标人 7	58.5	63.50	57.50	59.50	61.50	54.5	55.50	58.64	14
8	投标人 8	70	72.00	66.00	71.00	70.00	63	65.00	68.14	10
9	投标人 9	96	99.00	96.00	95.00	95.00	96	98.00	96.43	1
10	投标人 10	82.5	84.50	81.50	80.50	84.50	76.5	81.00	81.57	6
11	投标人 11	90	93.00	87.00	87.00	91.00	89.5	91.00	89.79	2
12	投标人 12	72	76.00	67.00	69.00	71.00	68.5	72.00	70.79	9
13	投标人 13	77	76.00	74.00	74.00	78.00	72	75.00	75.14	8
14	投标人 14	68	71.00	62.00	64.00	69.00	64	65.00	66.14	11
15	投标人 15	51	54.00	44.00	44.00	50.00	43	49.00	47.86	15
16	投标人 16	65	70.00	61.00	63.00	67.00	61	64.00	64.43	12
17	投标人 17	86	90.00	83.00	87.00	86.00	84	89.00	86.43	3
18	投标人 18	61	66.00	58.00	62.00	66.00	57.5	60.00	61.50	13

表 3　单个专家和专家群体的信度值

Tab. 3　The reliability of singer expert and group experts

j	k	$\sum x_j x_k$	$\sum x_j \sum x_k$	l_{jk}	$\frac{(l_{jj}+l_{kk})}{2}$
1	1	103 220.50	1 809 025.00	2 719.11	—
1	2	107 674.00	1 891 070.00	2 614.56	2 637.69
1	3	99 807.50	1 744 465.00	2 892.78	2 942.78
1	4	99 616.50	1 747 155.00	2 552.33	2 883.68
1	5	104 722.00	1 838 615.00	2 576.72	2 611.28
1	6	98 160.50	1 714 875.00	2 889.67	2 945.56
1	7	101 693.25	1 777 417.50	2 947.83	3 002.62
2	2	112 380.50	1 976 836.00	2 556.28	—
2	3	104 119.00	1 823 582.00	2 808.89	2 861.36
2	4	103 960.50	1 826 394.00	2 494.17	2 802.26
2	5	109 282.50	1 922 002.00	2 504.61	2 529.86

续表

j	k	$\sum x_j x_k$	$\sum x_j \sum x_k$	l_{jk}	$\frac{(l_{jj}+l_{kk})}{2}$
2	6	102 407.50	1 792 650.00	2 815.83	2 864.14
2	7	106 087.75	1 858 029.00	2 863.92	2 921.20
3	3	96 622.50	1 682 209.00	3 166.44	—
3	4	96 351.00	1 684 803.00	2 750.83	3 107.35
3	5	101 292.00	1 772 999.00	2 792.06	2 834.94
3	6	95 001.50	1 653 675.00	3 130.67	3 169.22
3	7	98 385.75	1 713 985.50	3 164.33	3 226.28
4	4	96 792.75	1 687 401.00	3 048.25	—
4	5	101 120.50	1 775 733.00	2 468.67	2 775.85
4	6	94 783.50	1 656 225.00	2 771.00	3 110.13
4	7	98 161.75	1 716 628.50	2 793.50	3 167.19
5	5	106 319.50	1 868 689.00	2 503.44	—
5	6	99 606.00	1 742 925.00	2 776.83	2 837.72
5	7	103 173.75	1 806 490.50	2 813.17	2 894.78
6	6	93 484.50	1 625 625.00	3 172.00	—
6	7	96 808.75	1 684 912.50	3 202.50	3 229.06
7	7	100 306.25	1 746 362.25	3 286.13	—

专家	r_a	r_b	r_c	r_d	r_e	r_f	r_g
可信度	0.968	0.885	0.967	0.887	0.966	0.969	0.964
专家群体可信度平均值 r					0.944		

可信度系数低于 0.9 以下，其他专家评委的可信度系数都是 0.95 以上，专家群体评价的可信度系数为 0.944，整体来看评标结果是可信的，但是关于可信度系数取值多少是可信赖的，并没有明确的标准，有的人认为是 0.85 以上，有的则认为是 0.95 以上，且这种可信度系数的高低是由于专家群体评价的一致性造成的，即专家效应 B，还是由于投标人方案本身巨大差异造成的，即方案效应 A，需要进一步地分析.

4.2 显著性方差检验

根据本文第三部分，知道上文中求得的可信度系数只能在某种程度上对评价过程中产生的误差进行估计，而不能对评价结果中所可能包含的各种误差来源解释清楚. 因此我们要区分可信度系数的高低是因为各专家的评价标准不一致，还是因为投标人方案个体差异显著. 即根据上文中提出的方差分析法，对评标数据进行 H_{01}，H_{02} 的假设检验，得到方差分析表(表 4).

由方差分析表可知，F_A 等于 150.416，对于给定的显著水平0.01，根据 F 分布表知道，$F_{0.01}(17,102)$ 等于2.19，$F_A > F_{0.01}(17,102)$，拒绝 H_{01} 原假设，投标人方案效应显著，说明投标人方案个体差异明显，同时说明本项目设置的评标因素是合理的，较好地区分了各投标人方案的优劣性.

表 4 评标数据方差分析一览表

Tab. 4 The ANOVA list of data of tender evaluation

变异来源	离差平方和	自由度	方差	F 值
投标人方案	19 597.73	17	1 152.808	150.416(**)
评标专家	693.69	6	115.614	15.085(**)
误差	781.74	102	7.664	
总变异	21 073.16	125		

F_B 等于 15.085，对于给定的显著水平 0.01，根据 F 分布表知道，$F_{0.01}(6,102)$ 等于 2.96，$F_A > F_{0.01}(6,102)$，拒绝 H_{02} 原假设，说明评标专家个体间的评价水平差异效应显著，评标专家评价的尺度不一致，有偏高偏低评分的倾向，如评委 d 对投标人 1 的评分为 59.5，低于所有专家评分的平均分 20 分左右，其评价能力和可信度是有问题的. 虽然上文中的可信度系数比较高，但相对于各专家的评价差异，主要是由于各投标人方案差异比较大的原因造成的，专家评标结果可信度程度并不理想.

表 5　各均值之差绝对值的检验表

Tab. 5　The test table of the absolute value of difference between each mean

	评委 a=74.72	评委 b=78.11	评委 c=72.06	评委 d=72.17	评委 e=75.94	评委 f=70.83	评委 g=73.42
评委 a=74.72		**3.389**(**)	2.667	2.556	1.222	**3.889**(**)	1.306
评委 b=78.11			**6.056**(**)	**5.944**(**)	2.167	**7.278**(**)	**4.694**(**)
评委 c=72.06				0.111	**3.889**(**)	1.222	1.361
评委 d=72.17					**3.778**(**)	1.333	1.250
评委 e=75.94						**5.111**(**)	2.528
评委 f=70.83							2.583
评委 g=73.42							

因为，评标专家间存在偏差和不一致性，需要进一步检验，到底是哪位评标专家的偏差效应最明显. 由于评标专家评价的投标人方案数目都相同，本文采用 Tukey 检验法对假设 $H_{02}^{ij}\ \beta_j=\beta_k$ 进行检验. 其拒绝域为

$$W_{jk}=\left\{|\overline{x_{\cdot j}}-\overline{x_{\cdot k}}|\geqslant q_{1-\alpha}(r,f_e)\ \sqrt{\frac{MS_e}{n}}\right\},$$

其中，

$$q(r,f_e)=\max_j\frac{(\overline{x_j}-\mu)}{\hat{\sigma}/\sqrt{n}}-\min_k\frac{(\overline{x_k}-\mu)}{\hat{\sigma}/\sqrt{n}},$$

其结构类似于 t 统计量，因此称为 t 化极差统计量，其值可以通过 t 化极差统计量分位数 $q_{1-\alpha}(r,f)$ 表得到. 这里 r 为 7，f_e 为 102，MS_e 为 7.644，对于 $\alpha=0.01$，得到

$$q_{1-\alpha}(r,f_e)\ \sqrt{\frac{MS_e}{n}}=4.24\times\sqrt{\frac{7.664}{18}}=3.26,$$

且均值两两比较得到表 5.

表 5 中加粗和星号标志的，说明两评标专家均值之间存在差异，通过表 5 知道，评委专家 b 和其他专家均值之间存在较大差异，说明评委专家 b 评价偏差比较大；评委 d 与其他专家相比，也存在一定的偏差，因此对专家 b 和 d 的评标数据需要重新审核，这也与信度系数的分析是一致的，他们俩的可信度系数都低于 0.9.

5　结论

本文首先利用心理和体育学的可信度系数定义，求得单个专家和专家群体的可信程度. 但可信度系数只是对评价误差的一种估计，并不能区分可信度系数的高低是由于评价方案差异还是由于专家评价差异引起的，所以本文使用方差分析对误差来源做进一步分析和检验. 在应用分析中，由于评标在招标采购中处于重要地位，专家评价主观性较强，本文使用实践中的一个招标项目进行实证分析. 通过实证分析，本案例中专家可信度系数之所以比较高，主要是因为投标人方案差异显著造成的，但专家评价差异是显著的，即专家评价结果存在不一致性，最后通过 Tukey 检验，得出两位专家评价数据存在偏差，需要审核其评价数据.

本文不仅对单个专家和专家群体的评价可信度程度进行度量，而且还对专家评价结果的误差来源和原因进行了分析，并且识别出到底是哪位专家的评价偏差较大. 通过本文的分析可知，由于专家的评价水平存在差异，其权重应该根据评价可信度程度赋予其相应权重；其次，本文是对专家评价的精确数值进行分析，下一步将对专家评价的模糊区间值的可信度进行分析；最后，关于专家的可信度评价还可以引用心理学中的其他方法如概化理论和项目反应理论中的概化系数、多面 Rasch 模型等.

参考文献(References)

[1] 吴敬业，汤理. 评价专家的可靠性预分析[J]. 系统工程，1992，51(5)：52-59.

[2] 李学栋，李浩志. 群体评价专家可靠性分析的拓展方法[C]//管理科学与系统科学进展：全国青年管理科学与系统科学论文集(第 3 卷). 北京：中国系统工程学会，1995，376-403.

[3] Yuan Jixue. A study on group decision-making mechanism in the games of subjective scoring[J]. China Soft Science, 2009(2)：173-176.
元继学. 专家主观评分比赛中群决策机制的研究[J]. 中国软科学，2009(2)：173-176.

[4] 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.
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(自然科学版)，2010，38(3)：

358-362.

[5] 高先务,刘心报,刘林. 基于方差分析方法的群决策专家估值的一致性检验[J]. 统计研究, 2011, 28(8): 111-112.

[6] Gao Xianwu, Liu Xinbao, Liu Lin. Consistency checks for deviation of subjective evaluation in GDM based on χ^2 statistic[J]. Journal of University of Science and Technology of China, 2012,42(11):942-946.
高先务,刘心报,刘林. 基于χ^2统计量的专家估值数据偏差的一致性检验[J]. 中国科学技术大学学报,2012,42(11):942-946.

[7] Cheng T, Wang Y, Sun Y. Development and application of tender evaluation decision-making and risk early warning system for water projects based on KDD[J]. Advances in Engineering Software, 2008, 48: 58-69.

[8] 梁晶. 加强对评标专家异常性评分的判定与监管[J]. 建筑经济,2013(2):97-99.

[9] Cao Yongqiang, Qu Xiaofei. Probability model for expert group evaluation[J]. Systems Engineering: Theory Methodology Application, 1994, 3(1): 77-80.
曹永强,曲晓飞. 专家群体评价的概率模型[J]. 系统工程理论方法应用,1994,3(1):77-80.

[10] Chen Ji, Su Weihua. Some issues about group evaluation technology[J]. Statistical Research, 2008, 25(8):80-84.
陈骥,苏为华. 关于群组评价技术若干问题的探讨[J]. 统计研究,2008,25(8):80-84.

[11] Huang Zhengnan. Reliability analysis for specialists judging[J]. Chinese Journal of Health Statistics, 2000,17(3):154-156.
黄正南. 专家评判的信度分析[J]. 中国卫生统计, 2000,17(3):154-156.

[12] Zheng Kai. Analysis on the reliability of measurement and evaluation method[J]. China Sport Science, 2007(2): 90-93.
郑凯. 测量可靠性及估计方法分析[J]. 体育科学, 2007(2): 90-93.

[13] McGraw K O, Wong S P. Forming inferences about some intraclass correlation coefficients[J]. Psychological Methods, 1996, 1(1): 30-46.

[14] Zhang Jihong, Guo Shizhen. A novel approach to the decomposition of the total sum of squares[J]. Mathematics in Practice and Theory, 2008, 38(2): 150-155.
张继红,郭世贞. 方差分析平方和分解分析方法一种新形式:数理统计方差分析教学的一种新方法[J]. 数学的实践与认识,2008, 38(2): 150-155.

[15] Wei Dengyun. The analysis on objectivity of scoring in 34th World Gymnastics Championship[J]. China Sport Science and Technology, 2000,36(7):23-24.
魏登云. 第34届世界体操锦标赛裁判员评分的客观性分析[J]. 中国体育科技,2000,36(7):23-24.