Behavioral Analysis of Judges in Subjective Evaluation Based on Statistics

Pei Xuesheng[1,2] Zheng Xiwen[2]

1 School of Mechatronic Engineering and Automation, Shanghai University, Shanghai 200072

2 School of Art and Design, Henan University of Science and Technology, Luoyang 471023

Keywords: subjective evaluation, statistical analysis, evaluation criterion, judges

Abstract: In this paper, a comprehensive evaluation method of judges' subjective judgment is presented. Problems in evaluating art work were analyzed by statistical tools which were also used to analyze the scores achieved by the subjective evaluation of judges while also evaluating the fairness, level, and criterion of the judges. Advantages and disadvantages of the evaluation methods were also given here. The assessment results will provide a reference for judge selection for a review organization in the future.

1. Introduction

Subjective evaluation is indispensable in people's life and their work. Especially, in the field of art, the evaluation of artistic work is characterized by uncertainty, imprecision, and subjectivity. [1] For subjective evaluation, it is common practice for several evaluators to evaluate an objective object. This method is simple and practicable, and each judge will give an evaluation from one of the several aspects of an object in his own perspective before reaching a comprehensive score. Before collecting statistical results, two extreme scores were removed and the remaining scores were averaged to get the resultant score of the object. There was inconsistency between the scores given by different judges and this is caused by their strictness with the criterion and the consistency of the judges. [2][3]

2. Problems with subjective evaluation

2.1 Judge's strictness with the criterion

In evaluation, a judge's strictness with the criterion may be different and this will lead to varied scores and relatively greater error in statistical findings.

a. Greater numbers will dominate lesser ones.

For example, in evaluating height, three different units of "meter", "decimeter", and "centimeter" are used by the judges. If height values represented in different units are aggregated for the average value, the greater numbers will dominate the lesser ones. This will make the evaluation of the numbers in "meter" invalid. Similarly, evaluation with greater differentials will predominate those with lesser differentials.

b. Removing two extreme data.

As for judges with loose or strict adherence to criterion of the evaluation, their evaluation results may be invalidated due to greater deviation which can be misunderstood as unfairness. Thus, it's unfair to remove the two extreme scores without any treatment. [4] In actual practice, evaluation of such judges can be the most informed.

2.2 Impartiality of judges

Judges can arrive at partial evaluation due to interests, guanxi, or other factors. Subjective evaluations can be greatly influenced by the fairness of the judge and such evaluation results are often questioned.

2.3 Professionalism of the judges

Judges are usually selected from experts or leaders in a certain field. Inevitably, some unsatisfactory judges maybe selected in this process. Evaluation results given by such judges often deviate from the true picture.

2.4 Weight problem

Judges provide evaluation results on the basis of their knowledge and experience, so subjectivity of the judges should be minimized.[5] Sometimes, each judge will give their careful evaluation of all aspects of one piece of work in the hope of offering more accurate results. However, the significance of each aspect considered in the evaluation is different. As a result, evaluation by different judges of the same work may vary greatly and unfair evaluation is seen on occasion. Such evaluation results may be questioned.

3. Data processing

Reliability of this evaluation is analyzed here. Given an evaluation, there are pieces of work A_1 $A_2 \ldots A_n$, and judges B_1 $B_2 \ldots B_m$. The evaluation matrix is given as:

$$AB=\begin{bmatrix} A1B1 & \cdots & AnB1 \\ \vdots & \ddots & \vdots \\ A1Bm & \cdots & AnBm \end{bmatrix}.$$

For this matrix, the scoring result of the piece A_i is generally achieved by the average:

$$Ai = （\sum_{j=1}^{m} AiBj - maxAiBj - minAiBj）/（m-2）.$$

Possible problems of this method have been discussed in the previous section. Here, we propose a new statistical method for calculation.

3.1 To begin, the evaluation scale coefficient S_j of each judge must be determined.

$$Bj = \frac{\sum_{i=1}^{n} AiBj}{n},\ B = \frac{\sum_{j=1}^{m} Bj}{m},\ Sj = \frac{B}{Bj}.$$

Create an evaluation scale coefficient matrix

$$S=\begin{bmatrix} S1 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & Sm \end{bmatrix}$$

Next, multiply the evaluation scale coefficient matrix S and evaluation matrix AB to achieve the new evaluation matrix

AB'=S*AB

The new scoring result of the piece Ai is generally achieved by the average:

$$Ai' = （\sum_{j=1}^{m} AiBj' - maxAiBj' - minAiBj'）/（m-2）.$$

3.2 Behavioral analysis of judges

Standard deviation of work Ai is evaluation results

$$\delta i = \sqrt{\frac{1}{m}\sum_{j=1}^{m}(AiBj - Ai)\text{^}2}$$

Deviation coefficient Csij of evaluation results.

$$Csij = (AiBj - Ai)/\delta i$$

Deviation coefficient of judge Bj

$$CBj = \frac{1}{n}\sum_{i=1}^{n} Csij$$

$$B\delta j = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(Csij - CBj)\text{^}2}$$

Evaluation behavior of the judge may be achieved by the judge's scale coefficient S_j and

Deviation coefficient CBj and the standard deviation Bδj of the evaluation results.

S_j<1 or CBj>0 denotes that judge Bj is less strict, and S_j>1 or CBj<0 denotes that judge Bj is more strict.

Less Bδj indicates consistency between the judge's evaluation and the results. Greater Bδj indicates inconsistency between the judge's evaluation and the results which may be a result of the judge's less experience of professionalism. A removed evaluation may reflect extreme judgment of any judge.

The above results can be automatically analyzed by a computer program. Enter the number of pieces of work and the number of judges, as well as enter the judges' judgment of each piece of work. Afterwards, the computer will automatically analyze the evaluation of each judge and produce its views for reference.

3.3 A difference correction method can be used for the judges' evaluation:

To start, calculate the maximum difference between all of the evaluation results of the judges:

$$dj = maxAiBj - minAiBj$$

Determine the range of evaluation[a, a+b]; performa normalized process of the original data in order to determine the new evaluation matrix:

$$AiBj'' = a + b * \frac{AiBj - minAiBj}{dj}$$

$$AB''=\begin{bmatrix} A1B1'' & \cdots & AnB1'' \\ \vdots & \ddots & \vdots \\ A1Bm'' & \cdots & AnBm'' \end{bmatrix}$$

Afterwards, use this evaluation matrix to calculate the evaluation results of each piece of work before placing their ranks.

4. Analysis of cases

The following table shows the original data of the evaluated pieces of work.

Table 1 The original data

	Judge 1	Judge 2	Judge 3	Judge 4	Judge 5
Work 1	98	89	92	92	88
Work 2	93	87	92	90	93
Work 3	91	85	87	88	78
Work 4	89	83	87	86	86
Work 5	87	81	85	84	85

Table 2 Average & rank

	Direct Average score	Rank	Average score gained by removing the two limits	Rank
Work 1	91.8	1	91	2
Work 2	91	2	91.66667	1
Work 3	85.8	4	86.66667	3
Work 4	86.2	3	86.33333	4
Work 5	84.4	5	84.66667	5

Table 3 Evaluation scale factor

	Judge 1	Judge 2	Judge 3	Judge 4	Judge 5
S	0.958952	1.033412	0.991422	0.998182	1.021395

Table 4 Multiply the evaluation scale and the original data

	Judge 1	Judge 2	Judge 3	Judge 4	Judge 5
Work 1	93.97729	91.97365	91.21084	91.83273	89.88279
Work 2	89.18253	89.90682	91.21084	89.83636	94.98977
Work 3	87.26463	87.84	86.25372	87.84	79.66884
Work 4	85.34672	85.77318	86.25372	85.84364	87.84
Work 5	83.42882	83.70635	84.27088	83.84727	86.8186

Table 5 New average & rank

	Direct Average score	Rank	Average score gained by removing the two limits	Rank
Work 1	91.77546	1	91.6724	1
Work 2	91.02526	2	90.31801	2
Work 3	85.77344	4	87.11945	3
Work 4	86.21145	3	85.95685	4
Work 5	84.41439	5	83.9415	5

Table 6 Standard deviation and deviation coefficient

δi	Csi1	Csi2	Csi3	Csi4	Csi5
3.487119	1.777972	-0.80296	0.057354	0.057354	-1.08972
2.280351	0.877058	-1.75412	0.438529	-0.43853	0.877058
4.354308	1.19422	-0.18373	0.275589	0.505247	-1.79133
1.939072	1.44399	-1.65027	0.412568	-0.10314	-0.10314
1.959592	1.326807	-1.73506	0.306186	-0.20412	0.306186

Table 7 Behavioral analysis of judges

	Judge 1	Judge 2	Judge 3	Judge 4	Judge 5
CBj	1.324009	-1.22523	0.298045	-0.03664	-0.36019
Bδj	1.0997	1.165085	0.274085	0.316468	1.003639
Conclusion	Loose	Strict	General	Consistency	Inconsistency

4.1 A difference correction method

Table 8 The maximum difference of Judge

	Judge 1	Judge 2	Judge 3	Judge 4	Judge 5
dj	11	8	7	8	15

Table 9 when a=60,b=40

	Judge1	Judge2	Judge3	Judge4	Judge5	Average score gained by removing the two limits	Rank
Work1	100	100	100	100	86.67	100	1
Work2	81.82	90	100	90	100	93.333	2
Work3	74.55	80	71.43	80	60	75.325	3
Work4	67.27	70	71.43	70	81.33	70.476	4
Work 5	60	60	60	60	78.67	60	5

5. Conclusion

The actual case shows that the judges of different types judged the works at the same scale by increasing the scale factor of the judges, which makes the evaluation more objective and accurate. In addition, statistical tools which were used to analyze the scores achieved by the subjective

evaluation of judges while also evaluating the fairness, level, and criterion of the judges. The assessment results will provide a reference for judge selection for a review organization in the future.

Pei Xuesheng: (1969,12-) Shangqiu, Henan, China，Associate professor, Doctoral candidate, Research interests: Industrial design methods and applications, Product digital design etc.

Reference

[1] Antonsson E K，Cagan J． Formal engineering design synthesis[M]． New York：Cambridge University Press，2001

[2] Li Yong, Wang Yougui. Evaluation of the Judges: A New Statistic and Evaluation Method[J]. Statistics and Decision. 2014(18);

[3] LVShu-long , LIANGFei-bao, LIUWen-li. An evaluation model of rater score [J] Journal of Fuzhou University (Natural Science)Vol.38 No. 3 Jun .2010

[4] Xia Jinyang, Research of Statistics on “Invalid Judge” in Sports Teaching Competition [J], Physical Education, 2012 (11).

[5] Fan Zhongguang, Evaluation of Third-party Logistics Service Provider Based on Three-parameter Interval Number Model[J], Logistics Technology (Equipment), 2015 (1)