Rating the Raters

−A new statistical method for evaluating

Yong Li
Institute for International Economics
Henan university
Kaifeng City, China, Tel.: +86 0378 3881821
e-mail: excellentrichard@163.com

***Abstract*—In this paper, with the actual data of a lecture contest，prevailing method’s defects to remove a maximum score and a minimum score, average of the rest to determine the ranking of the contestants are found. Comparison of three methods of evaluation that prevailing method, calculating raters’ weight once and Iterative calculation their weight, we design a quantitative method to reduce the judges’ weight who have larger deviation, increase their weights who have smaller one, so that the final result is more accurate. In addition, the computer programs of the three evaluation methods are also developed to facilitate the practical application.**

Keywords- rater; weight; iterative algorithm; computer program

I. INTRODUCTION

It is often needs subjective evaluation of things or indicators in our daily life, such as brands, benefits, investment risk and personnel assessment, etc. Things or indicators need to get an objective evaluation by subjective judgment of the raters, just, truly reflect the real situation of them, in order to strengthen management, improve planning, forecasting, and decision-making [1]. However, imcomplete information, limited knowledge of things and the subjective preferences of the raters, even the conflicts of interest between the judges, will lead to considerable evaluation deviations from the things themself. Jinwen Zhao [2] gives some typical cases, in these cases, even if a single abnormal value exists, will have a profound impact on the conclusion of the evaluation.

A very popular kind of evaluation method is to let a number of judges give subjective scores to the objects being evaluated, then remove a maximum and a minimum score, caculate the arithmetic average of the rest scores, the result determine the ranking of the objects. This algorithm at least has two drawbacks: first, the rest scores that not been removed will enter into the step for caculating arithmetic mean, it implies that all judges have the same level of evaluation, so they are assigned to the same weight. While the actual evaluation level will not be exactly the same in the subjective evaluation, the quality of the evaluation is mainly affected by the evaluator's true evaluation capacity, the physiological and mental state, the degree of outside interference. Equal weight approach does not consider the impact of the these factors on the evaluation outcome [3]. Second, removing one of the highest scores and a minimum one sometimes the score which is closer to the true value will be removed and the farther one left. All these makes the final evaluation results are inaccurate.

This paper presents a new statistical method to avoid or reduce the impact of these defects on the final evaluation results by analyzing the actual data of a lecture contest.

II. CASE AND DATA

The lecture contest case

The jury consists of six judges come from six department, plus a champion of previous lecture contest, there are total of seven judges. Seventeen contestants are elected from six departments. Seven scores of each player are removed a highest and a minimum score, the rest are calculated the arithmetic mean which determine the final ranking of the player. All player's original scores are shown in Table I.

TABLE I Player's Original Scores

Play’s code	Judge 1	Judge 2	Judge 3	Judge 4	Judge 5	Judge 6	Judge 7
A	94	92	89	92	88	90	86
B	87	90	82	91	83	86	82
C	94	84	80	90	80	82	87
D	90	87	88	87	89	***93***	***94***
E	85	86	89	88	90	95	89
F	92	88	87	90	90	87	85
G	85	90	92	89	90	92	90
H	92	94	80	87	80	80	78
I	79	80	89	80	90	90	85
J	90	88	93	90	90	***95***	***94***
K	80	79	86	80	87	90	89
L	85	90	88	89	89	***94***	***93***
M	90	***93***	93	86	88	***80***	***80***
N	94	93	90	89	90	86	90
O	90	89	91	90	88	*95*	*93*
P	89	85	94	85	89	90	88
Q	86	86	90	87	89	90	93

III. DESIGN AND COMPARISON OF EVALUATION METHODS

Observing Players' scores in Table I, we can find some abnormal phenomena. For example, the contestant M, he gets two very low score of 80 points, also gets two high score of 93 points. It indicates that the evaluation of the judges to player M has great disagreement. One possible situation is that judge 6 and judge 7 generally have more stringent requirement on all the players than other raters. But observing these two judges score the other players, we deny this assumption, because a few high scores, two of 95, three of 94, and three of 93 points are given by them. Coincidentally, Players D, J, L, and O who get these high scores are come form the same department with these two judges.

Whether to remove a maximum points and a minimum points can effectively eliminate these anomaly phenomena? Viewing from Data of the competition, this approach still does not work. For instance, judge 6 and judge 7, to several players, have appeared a high degree of consistency, even removed one score, the other still significantly impact on the final results. Player M is ranked 12, but as long as one of the judges who score M 80 points, change to score M 86 points which is the penultimate lowest score that player M gets, the final ranking of the player M will immediately rise from 12 to 7. This provoke our thinking, can we design an effective evaluation method to avoid or greatly reduce such anomalies cause a large deviation happen?

A. *Calculating the weight of raters once*

Simplicity and without loss of generality, the arithmetic mean of all the scores that judges give to certain player stands for the preliminary characterization of the real level of the contestant. The difference between a judge give this player's score with the arithmetic mean represents the deviation of this judge in the evaluation of this player, and then sum all the deviation square the judge scores for all the players, the result characterizes this judge's deviation in the entire evaluation process. All the judges are done following the above steps. We can know, one judge's sum of deviation square is smaller, more accurate he is in the whole evaluation process, so he should be given bigger weights, and vice versa. Weight calculation formula is shown in equation (1). σ_i^2 represents deviation square of judge i, s_{ji} is on behalf of the score judge i gives to contestant j. w_i stands for the weight of judge i.

$$\sigma_i^2 = \sum_{j=1}^{m}(s_{ji} - \frac{1}{n}\sum_{i=1}^{n} s_{ji})^2 \quad i = 1, 2, \cdots, n, \quad j = 1, 2, \cdots, m$$

$$w_i = \frac{(\sum_{i=1}^{n}\sigma_i^2) - \sigma_i^2}{(n-1)\sum_{i=1}^{n}\sigma_i^2} \quad i = 1, 2, \cdots, n \qquad (1)$$

The results are caculated by using Equation (1) are shown in Table II. Seeing from Table II, the weights of judge 6 and judge 7 who give out abnormal scores are not high, judge 1 and judge 2 also get low weights.

TABLE II Judges' Weights

	Judge 1	Judge 2	Judge 3	Judge 4	Judge 5	Judge 6	Judge 7
Weight	***0.1317***	***0.1372***	0.1478	0.1514	0.1540	***0.1363***	***0.1416***

B. *computing the raters stable weights*

A judge is Calculated his corresponding weight according to his deviation in the evaluation process, recalculating his composition score of the weighted average score using weight he obtained. Each judge will produce a new deviation from the weighted average score, then each judge gets a new weight, which can be obtained according to the above method, the second weight is often different from the first calculated one. For the same judge there are two different weights, and which one should be used in the final ranking process? Using iterative algorithm, we can calculate the new weighted score, then calculate the new arithmetic mean and continue to compute the next weight. So the cycle repeated until stable weights of all raters are obtained. Of course, in the process of numerical calculation, the difference between the adjacent two weights is less than the specified numerical accuracy, weights are regarded as being stable, the calculation cycle stops. Apply iterative algorithm to calculate the weights is shown in Equation (2). σ_{it}^2 represents deviation square of judge i at t round, s_{jit} is on behalf of judge i give the score to contestant j at t round, w_{it} stands for the weights of judge i at t round. For the above lecture contest, iterative calculations goes to 14th round, each of the judges's weight has changed very little, so the iterative algorithm is terminated, the finally weights of judges are shown in Table III.

TABLE III Judges' Stable Weights

	Judge 1	Judge 2	Judge 3	Judge 4	Judge 5	Judge 6	Judge 7
Weight	0.02515	0.16249	0.16249	0.16246	0.16247	0.16246	0.16248

We also found that the ranking order of all players are exactly the same by using the stable weights multiply the original scores and the stable weights multiply the adjacent weighted scores, the results do not show here.

$$\sigma_{it}^2 = \sum_{j=1}^{m}(s_{jit-1} - \sum_{i=1}^{n} w_{it-1}s_{jit-1})^2 \quad i=1,2\cdots;n;\ j=1,2\cdots;m;\ t=1,2\cdots;N$$

$$w_{it} = \frac{(\sum_{i=1}^{n}\sigma_{it}^2 - \sigma_{it}^2)}{(n-1)\sum_{i=1}^{n}\sigma_{it}^2} \quad i=1,2\cdots;n;\ t=1,2\cdots;N \qquad (2)$$

$$s_{jit} = s_{jit-1}w_{it} \quad i=1,2\cdots;n;\ j=1,2\cdots;m;\ t=1,2\cdots;N$$

It shows the final ranking results of players under the three methods together in Table IV. It can be seen, The results of the evaluation are very different obtained by usage of iterative method and remove a maximum and a minimum points method. For example, contestant N's ranking drops from three under the popular method to five under the iterative method and player G's ranking increases from five to three in reverse. Back to the raw data of the Table I, you can see the player G's lowest score is given by judge 1 and the highest score of player N is also given by judge 1. seeing from Table II and Table III, judge 1's deviation is the biggest. Intuitively, the iterative method is more accurate than the popular one.

TABLE IV Players' Final Ranking Under Three Method

Player's ranking	**Removing a maximum score and a minimum score method**	**Calculating the weight of judges once method**	**Iterative algorithm method**
1	J	J	J
2	O	O	O
3	N	N	***G***
4	A	A	***L***
5	G	G	***N***
6	L	L	***D***
7	D	D	***A***
8	E	E	E
9	F	***Q***	***Q***
10	Q	***P***	***P***
11	P	***F***	***F***
12	M	M	M
13	B	B	B
14	I	***C***	***I***
15	C	***I***	***K***
16	K	K	***C***
17	H	H	H

In order to quantitatively compare three methods, which kind of method's evaluation results is better, this paper introduces a reliability test. Reliability test in statistics weighs the stability of the measurement results, if repeated measurement results are very close, it means the high reliability of the measurement [4]. For evaluation issues, different judges rates the same contestant, can be regarded as a repeated measurement. If judges's consistency on the evaluation of a player is higher, under a common evaluation criteria, the evaluation results of the player is more accurate, i.e., The higher the degree of reliability, the evaluation results are more reliable. Next, we do reliability test using the original contest data, weighted average data by calculating weight once and weighted average data calculated by stable weight, respectively. The test results are shown in Table V. It can be seen from the reliability coefficient in Table V, the best method is the iterative method， calculating weights once method is the second.

TABLE V Reliability Coefficient Under Three Circumstances

	original scores	**weighted average data calculated by calculating weight once**	**weighted average data calculated by stable weight**
Cronbach's alpha coefficient	0.6164	0.6296	0.6484

IV. CONCULUSION

From a actual lecture contest data, we found that the popular evaluation method that remove a maximum points, removing a minimum points, the rest of the scores are calculated arithmetic average for ranking the players in the game has some drawbacks. For example, it may get rid off the score closer to the true value but retain the score farther to the true value; gives the same weight to a different evaluation level of judges; even the minority judges play a interest game makes the final evaluation results unfair, etc.

For the above reasons, a method is developed in this paper that the square of the deviation of a judge's evaluation score from the arithmetic mean of all scores is as his evaluation level characterization. Assign weights to raters according to magnitude of their deviation, the greater the deviation of a judge, the smaller weight he gets, so that the final results of the evaluation becomes more accurate. In order to maintain the logic of self-consistency in the calculation of weights, using iterative algorithm, judges' stable weights are obtained. We also borrow the idea of the reliability test to distinguish the validity of different evaluation method, reliability test results support the iterative algorithm is more accurate.

Evaluation is an important issue throughout the various disciplines, especially in recent years, the evaluation mechanism and method increasingly attracts the attention of various disciplines [5]. This paper attempts to explore the data of a small case, and reveal the information in it, then develop a more accurate evaluation method than the frequently used method. From a view of practical application, we have also developed computer programs of the three methods to quick access to the final evaluation results.

Of course, the evaluation method developed in this work also have limitations. For example, a judge is forward-looking or has change consciousness, has different view of judgment or the criteria we used to, however his useful evaluation information just play a minor role in the evaluation process. In fact, the method developed in this article, can help to find out such judges. For instance, the maximum deviation a judge has, through in-depth interviews with him, we may be able to find useful, innovative ideas.

ACKNOWLEDGMENT

The author would like to thank Professor Yougui Wang, Professor Qiang Yuan and Professor Handong Li for stimulating discussion and helpful comments. This work is supported by NSFC under Grant no.61174165 and NSSFC under Grant no.12BGL033.

REFERENCES

[1] Yushan Jiang, Konglai Zhu, "The Indicator System of Modernization Evaluation and the Methods of Comprehensive Evaluation," in Statistical Research, vol. 12, 2002, pp. 50–54.

[2] Jinwen Zhao, "The Typical Examples of Influence on the Econometric Modeling from Outliers," in Statistical Research, vol. 12, 2010, pp. 92–98.

[3] Huijun Sun, "Discussion of the Subjective Appraisal Theory," in Statistical Research, vol. 1, 2010, pp. 97–100.

[4] Wentong Zhang, The SPSS Statistical Analysis Advanced Tutorials. Higher Education Press, Beijing, 2004.

[5] S. Alonso, F.J. Cabrerizo, E. Herrera-Viedma and F.Herrera, "h-Index: A review focused in its variants, computation and standardization for different scientific fields," in Journal of Informetrics, vol. 3, 2009, pp. 273–289.