A Study of Judging Scheme for Innovative Category Competitions Based on Stochastic Simulation and T-Score Methods

Jiawei Dai *

School of Artificial Intelligence and Software, LiaoNing Petrochemical University, Fushun, China

* Corresponding Author Email: welltai@foxmail.com

Abstract. This study proposes a set of fair and accurate judging optimization scheme for the judging problem in science and technology innovation competitions. First, by simulating random distribution, a work allocation model is designed to achieve balanced judging workload and minimize work intersection. Second, the T-score method is introduced to optimize the calculation of standard scores in order to reduce the subjectivity and polarity of scoring. Again, a segmentation management strategy is proposed to effectively screen and highlight innovative works by setting score thresholds and dividing works into different grades. This scheme improves the scientificity and fairness of evaluation through reasonable allocation, scoring standardization and segmented screening, which is of guiding significance to the evaluation of science and technology competitions.

Keywords: Stochastic simulation; T Score method; Competition Judges.

1. Introduction

In today's rapidly developing scientific and technological fields, innovation competitions have become an important platform for stimulating innovative thinking and discovering outstanding talents. However, with the increase in the number of entries and the complexity of innovation content, how to establish a fair and efficient judging system has become a major challenge for competition organizers. A scientific and reasonable judging program is not only related to the fairness of the competition results, but also affects the confidence of the participants and their motivation to innovate in the future [1]. Therefore, this study focuses on the key aspects of judging and aims to construct a comprehensive and efficient judging mechanism by optimizing the distribution of entries, improving the scoring calculation model, implementing segmented management, and enhancing the scientificity of the scoring model, so as to ensure the fairness and accuracy of the judging results. This study not only provides a feasible solution for the evaluation of science and technology innovation competitions, but also provides new ideas and methods for the evaluation of similar competitions.

2. Distribution model

In this paper, assuming 3000 entries, 125 judges, and the requirement that each entry must be scored by 5 experts, we now need to determine the optimal "cross-distribution" of entries to increase the total number of cross-reviews among the different experts.

First, we define the constants M, N and K, which denote the number of entries, the number of judges and the number of times each entry has to be judged, respectively, so that for this problem, their values are as shown in equation (1).

$$\begin{cases} M = 3000 \\ N = 125 \\ K = 5 \end{cases} \tag{1}$$

Next, define a 0-1 decision variable x_{ij}, which is defined in equation (2) as follows.

$$x_{ij} = \begin{cases} 1 & \text{if work } i \text{ is judged by expert } j, \\ 0 & \text{if entry } i \text{ not reviewed by expert } j, \end{cases} \quad i \in \{1,2,\dots,M\}, j \in \{1,2,\dots,N\} \tag{2}$$

Now we choose one entry and two different judges to illustrate the significance of the decision variable x_{ij}, as shown in Table 1.

Table 1. x_{ij} Significance of Decision Variables.

x_{ij_1}	x_{ij_2}	$x_{ij_1} \times x_{ij_2}$	significance	cross-evaluation
0	0	0	Neither expert j_1 nor j_2 reviewed entry i	non-existent
0	1	0	Only expert j_2 has reviewed entry i	non-existent
1	0	0	Only expert j_1 has reviewed entry i	non-existent
1	1	1	Experts j_1 and j_2 both judged entry i	remain

From Table 1, we can know that, for the same entry, if there is cross-evaluation between two experts, the value of $x_{ij_1} \times x_{ij_2}$ is 1, and vice versa is 0. The size of the intersection of any two different experts can be found by Equation (3) [2]. The size of the intersection of any two different experts can be found by the formula (3):

$$U_{j_1 j_2} = \sum_{i=1}^{M} x_{ij_1} x_{ij_2}, \ \ j_1 \in \{1,2,\dots,N\}, j_1 \in \{1,2,\dots,N\}, j_1 \neq j_2 \tag{3}$$

According to the nature of the exchange law of addition and multiplication, it is easy to get $U_{j_1 j_2} = U_{j_2 j_1}$, so in order to calculate the size of the intersection of the works of any two different experts, Equation (3) can also be simplified into Equation (4).

$$U_{j_2 j_1} = U_{j_1 j_2} = \sum_{i=1}^{M} x_{ij_1} x_{ij_2}, \ \ 1 \leq j_1 < j_2 \leq N, \ \ j_1, j_2 \in \mathbb{Z} \tag{4}$$

Combined with the above analysis, the 0-1 decision variable x_{ij} is introduced as the main meaningful variable. The objective function can be defined by equation (5) as follows.

$$\min Z = \max_{\substack{1 \leq j_1 < j_2 \leq N \\ j_1, j_2 \in \mathbb{Z}}} \sum_{i=1}^{M} x_{ij_1} x_{ij_2} - \min_{\substack{1 \leq j_1 < j_2 \leq N \\ j_1, j_2 \in \mathbb{Z}}} \sum_{i=1}^{M} x_{ij_1} x_{ij_2} \tag{5}$$

The constraints on this objective function are.

$$\text{s.t.} \begin{cases} \sum_{j=1}^{N} x_{ij} = K, \ \ i \in \{1,2,\dots,M\} \\ \frac{M \times K}{N} - \varepsilon \leq \sum_{i=1}^{M} x_{ij} \leq \frac{M \times K}{N} + \varepsilon, \ \ j \in \{1,2,\dots,N\}, \varepsilon \in \mathbb{N}, \varepsilon \ll \frac{M \times K}{N} \\ x_{ij} \in \{0,1\}, \ \ i \in \{1,2,\dots,M\}, j \in \{1,2,\dots,N\} \end{cases} \tag{6}$$

In the objective function, $\max_{1 \leq j_1 < j_2 \leq N} \sum_{i=1}^{M} x_{ij_1} x_{ij_2}$ and $\max_{1 \leq j_1 < j_2 \leq N} \sum_{i=1}^{M} x_{ij_1} x_{ij_2}$ indicate that to find the maximal and minimal value of the intersection of the works of any two different experts, the difference of the two can represent the extreme difference between the intersection size. The difference between the two can be expressed as the extreme difference between the sizes of the intersection, in order not to let the size of the intersection of the existence of too large or too small, then here we want to make the value of the extreme difference smaller, so the optimization objective is to minimize the extreme difference.

For the above analysis and the established optimization model, we can first generate a series of solutions $A = \{a_1, a_2, \dots\}$ randomly according to the constraints, and then compute the value of the objective function $minZ$ for each solution a_i if the series of solutions satisfy the constraints, and then select some of the better solutions, and finally choose one of the better solutions randomly as the solution for distributing the works to the judges. A single solution a_i can be viewed as a matrix X_{MN} of M rows and N columns consisting of 0-1 decision variables x_{ij}, and then some of the better solutions are filtered out, and one of them is randomly selected as the solution for distributing the work to the judges.

3. T-score method

3.1. Differences in scoring between experts

In order to standardize the scores of different experts, we first set up a formula for calculating the standard score. Now suppose that an expert has scored m pieces of work, then the sample of scores of this expert is $\{a_1, a_2, \ldots, a_m\}$, and we can calculate the mean of the scores given by him/her $\bar{a} = \frac{\sum_{i=1}^{n} a_i}{n}$ and the standard deviation $s = \sqrt{\frac{\sum_{i=1}^{n}(a_i - \bar{a})^2}{n-1}}$, by using these two statistics, the formula for defining the standard score is.

$$b_i = 50 + 10 \times \frac{a_i - \bar{a}}{n} \tag{7}$$

But there are limitations to calculating standardized scores in this way. First, we plotted the distribution of all raw scores in the first round of judging, as shown in Fig. 1.

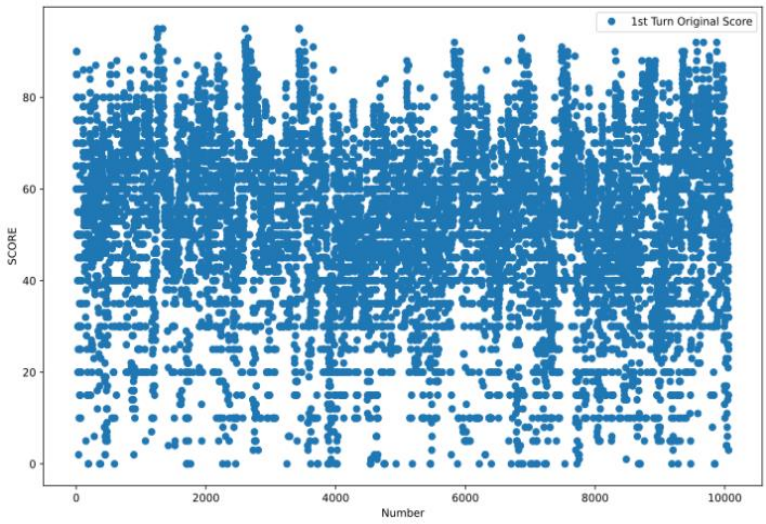


Fig 1. Distribution of raw scores in the first round of evaluation.

In Fig. 1, the x-axis indicates how many times all the experts scored, and it can be observed that the experts scored about 10,000 times, and the y-axis indicates the numerical value of the scores, and the distribution of the numerical value of the scores in the image indicates that the scores of the experts are very discrete, and the number of low scores and high scores is also not small. Then we examine the average of the scores given by each expert in the first round of evaluation, and also plot them, as shown in Fig. 2.

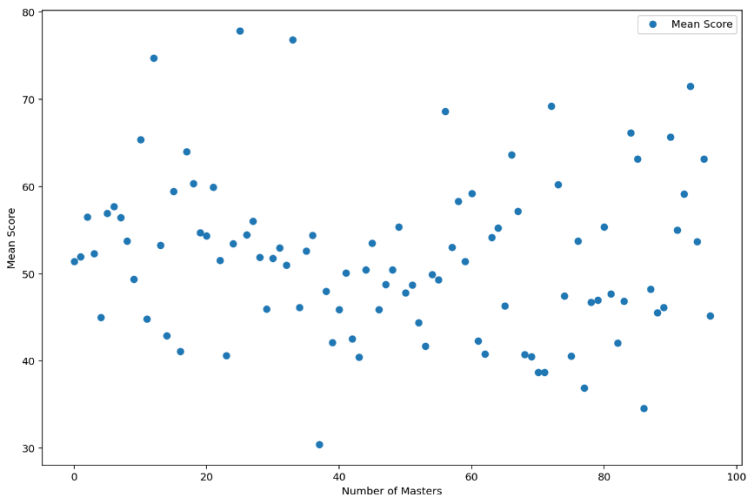


Fig 2. Distribution of mean values of all scores given by each expert in the first round of evaluation.

Fig. 2, the x-axis represents one by one experts, it can be observed that there are about 100 experts participated in the evaluation, y-axis represents the average of all the scores of a certain expert, from the distribution of the average scores in the image, there are great differences in the scores of the experts, that is to say, it means that the different experts on the evaluation of the academic level of the entries is not uniform.

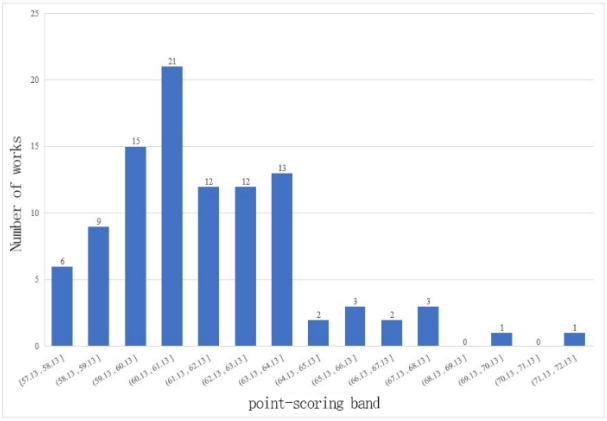
3.2. Differences in different scoring criteria

Now we select a part of the entries, each work is scored by five experts, and all the original scores given by the experts by formula (7) to calculate the standard score, respectively, take the average of the original score and the standard score of the entries ranked, part of the specific scores and rankings as shown in Table 2:

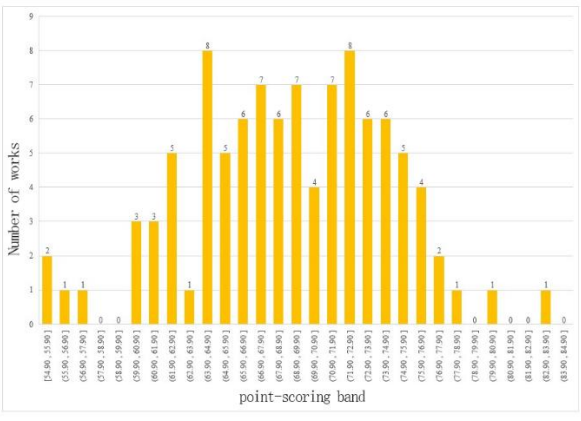
Table 2. Works rankings.

Raw score	Standard Score Ranking	Expert 1		Expert 2		Expert 3		Expert 4		Expert 5		First review	
		raw score	standard scores	raw score	standard scores	raw score	standard scores	raw score	standard scores	raw score	standard scores	Average raw score	Average standardized score
10	1	74.00	62.61	75.00	70.81	85.00	78.12	67.00	66.81	75.00	71.73	75.20	70.02
6	3	85.00	71.14	87.00	64.96	68.00	64.72	70.00	69.20	73.00	70.35	76.60	68.07
46	40	83.00	63.45	70.00	67.95	64.00	57.17	54.00	60.45	77.00	59.88	69.60	61.78
8	25	89.00	64.76	59.00	59.16	60.00	60.51	86.00	64.07	85.00	66.78	75.80	63.06
1	13	80.00	65.78	80.00	63.96	78.00	62.54	92.00	60.15	83.00	67.51	82.60	63.99
41	52	82.00	57.54	80.00	63.56	66.00	63.68	62.00	55.96	62.00	64.13	70.40	60.97
87	23	61.00	62.15	69.00	56.44	54.00	62.71	62.00	70.66	73.00	63.59	63.80	63.11
49	27	70.00	67.11	83.00	69.40	65.00	57.87	43.00	53.72	86.00	64.86	69.40	62.59
19	44	55.00	51.99	76.00	62.93	70.00	62.01	80.00	63.71	82.00	66.82	72.60	61.49
12	11	80.00	63.43	72.00	65.83	65.00	63.91	80.00	63.56	76.00	64.62	74.60	64.27
12	73	80.00	63.43	72.00	47.15	70.00	56.77	70.00	64.82	81.00	67.13	74.60	59.86

From the comparison of the two left columns of Table 2, we can see that the ranking difference of the same work under different scoring criteria may be very big, for example, if the 10th work is ranked according to the average of raw scores and average of standard scores, the former is ranked at 19, while the latter is ranked at 44. The common statistical model only considers the relationship between the data, but the data hides a lot of subjective factors among experts, including the different academic levels among experts, personal preferences, and so on. Common statistical models only consider the relationship between the data, but the data hide many subjective factors among the experts, including the different academic level of the experts and their personal preference [3].



(a) Distribution of score bands with mean raw scores.



(b) Distribution of score bands With standardized score averages.

Fig 3. Distribution of score bands for different scoring criteria.

For different scoring criteria, as shown in Fig. 3, the score bands of two different scoring criteria are demonstrated. As shown in Fig. 3(a), if the average of raw scores is taken as the criterion directly, there will be many scores concentrated in a small range, and the extreme deviation of all the scores is 26.20; as shown in Fig. 3(b), if the average of processed standardized scores is taken as the criterion, the concentration of scores will be relatively smaller than that of the raw scores, and the scores will be distributed in a wider range, and the extreme deviation of all the scores is 12.10, which is obviously better than the average of raw scores here. As shown in Fig. 3(b), if the mean of the processed standardized scores is taken as the standard, the concentration of the scores is relatively smaller than that of the raw scores, and the scores are distributed in a wider range, and the extreme deviation of all the scores is 12.10, which is obviously better than the method of the mean of the raw scores.

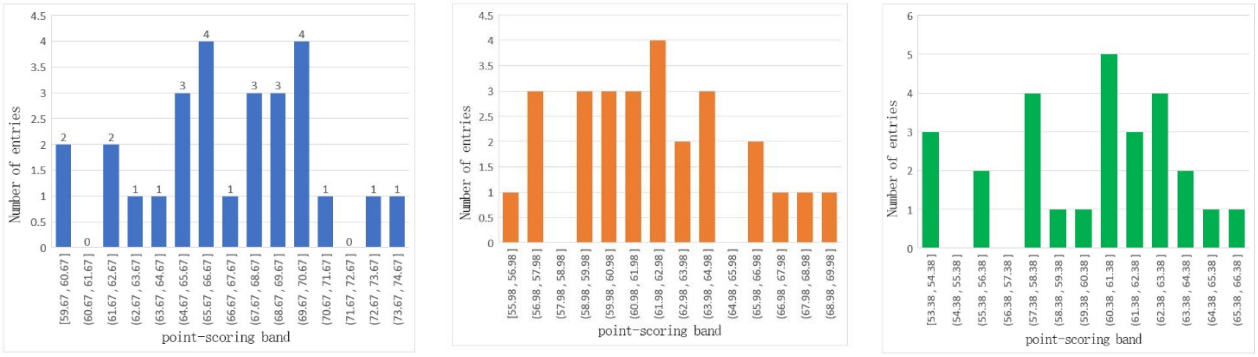
3.3. Modeling solutions

Combined with the above analysis, we can understand that the mean value of different experts' scores will not be at a uniform level, so we consider that the mean score and standard deviation of different experts' scores will be shifted to the same level, so we convert the original scores according to the formula (8).

$$y_{ik} = s \times \frac{a_{ik} - \overline{a_i}}{s_i} + \mu \tag{8}$$

This method is called the T-score method, and the score calculated by this method is called the T-score [4]. In Equation (8), a_{ik} denotes the original score of expert i for work k, $\overline{a_i}$ denotes the mean of all scores scored by expert$i$, and s_i denotes the standard deviation of all scores scored by expert i. The T-score method can be used to flatten the mean and standard deviation of different scores by different reviewers to our specified level. Using this T-score method, the mean and standard deviation of the scores of different reviewers can be leveled up to the level we specify, which is set to be$\mu = 60, s = 10$.

Since the ranking of the first prize entries selected in the second judging stage of a general competition is agreed upon by the experts, the model uses this section as experimental data to test the model's performance. Meanwhile, we compare the score bands resulting from the three scoring methods, as shown in Fig. 4.



(a) Distribution of score bands with mean raw scores

(b) Distribution of score bands with standardized score averages

(c) Distribution of score bands for the mean of T-scores

Fig 4. Histogram of the distribution of score bands for the three different scoring criteria.

From the information in Fig. 4, the following comparative results can be seen in Table 3.

Table 3. Experimental comparison of three different scoring criteria.

	Score Polarity	Evenness of score distribution
raw score averaging	14.00	more diffuse
SSA	13.27	more centralized
T-score averaging	12.59	Lowest extreme values, most concentrated

From the above comparison results, it can be seen that the T method has the lowest extreme value among the three methods, and the distribution of the scores is more uniform, which can further reduce the influence of various subjective factors on the scores among different experts, and effectively control the subjective errors, so that the comparability of the scores is improved, and the accuracy of the rankings is also improved.

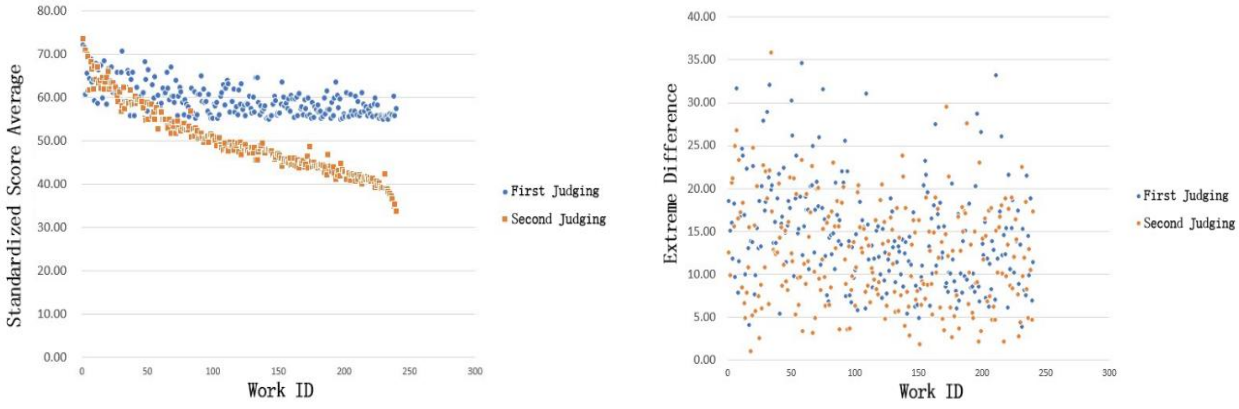
4. "Extreme Difference" Modeling

First of all, we understand that there is no standard answer for the works of innovative competitions, so a judgment on the innovation of the works (innovation can be understood as the main criterion of the results) will be affected by the evaluation of different experts, and the result of this influence is that the same work has a large gap between the ratings of different experts, so the current problem to be solved is the problem of the differences between the experts, which can improve the science and consistency of the judging results. Therefore, the problem of differences between experts should be solved, so as to improve the scientificity and consistency of the evaluation results.

For the works with big difference in scores, we can also find that there are works with big difference in low score and big difference in high score. The big difference in low score will not affect the evaluation of excellent works, so it is not a big impact on the evaluation work, while the evaluation of works with high score directly affects the evaluation of the award level, so the works with high score often need to enter the second stage of the review, which is more authoritative than the first stage of the review, so the scores of works with big difference in high score need to be adjusted in order to ensure the credibility of the evaluation result. Therefore, it is necessary to adjust the scores of the works with large differences in the high score band to ensure the credibility of the judging results.

Considering that there may be a correlation between high score extremes and high innovativeness of a work, it is important to be able to deal with the problem of extremes [5].

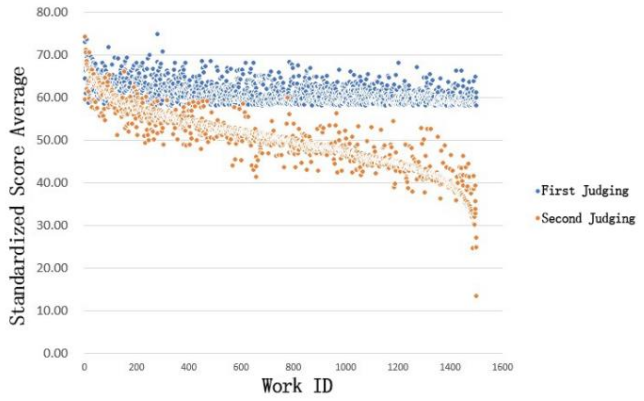
First of all, for the given two groups of innovative competition results of data processing, which the first group of competition data contains 240 works of the results of the sample, the second group of competition data contains 1,500 works of the results of the sample, we are the two stages of the standardized scores for the mean and the extreme deviation, the first group of competition data for the results of the results of the second group of data for the results of the competitions are as shown in Fig. 5 shown in Fig. 6:



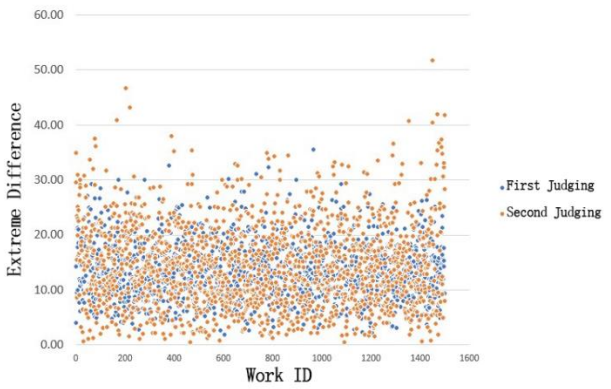
(a) Scatterplot of the distribution of mean standardized scores

(b) Scatterplot of the distribution of extreme variances

Fig 5. Analysis of the scoring results of the first group of works.



(a) Scatterplot of the distribution of mean standardized scores



(b) Scatterplot of the distribution of extreme variances

Fig 6. Comparison of the analysis of the scoring results of the works in the second group.

In terms of the average standardized score, the results of the second evaluation are more widely distributed than the first one, which can better reflect the differences between the works. In terms of the extreme deviation, the results of the second evaluation have more works with lower extreme deviation than the first one, that is, the extreme deviation is generally lower, which shows that the second evaluation is more reasonable. If we only carry out a one-time evaluation scheme, the advantage is that the time and labor costs will be lower, but the differentiation of the high scoring works will be reduced, so that the final results of the comparability of the data; if we carry out a phased evaluation scheme, the evaluation work we may need to spend more time and labor costs, but for the results of the competition, it will be more scientific than a one-time evaluation, so that the results are more differentiated. But for the results of the competition, it will be more scientific than the one-time evaluation program, so that the results of the data are more differentiated to reflect the gap between different works.

For the treatment of "extreme difference", the first step we enter the second evaluation of all the works, after the second evaluation, the extreme difference of its statistics, to find out the extreme difference of the part of the work for reconsideration, here set a parameter η , used to indicate the number of works need to be reconsidered of all the works of the second evaluation of the percentage of η. Assuming that we set $\eta = 0.15$, it means that the scores of the top 15% of the entries with the highest to the lowest extreme deviation need to be reconsidered for all the entries that enter the second evaluation.

In the second step, for the works that need to be reconsidered, we firstly count the scores given by the experts in the second review, find the median among these scores, revise the scores that are larger than the median, lower the scores that are too high, and raise the scores that are too low, in order to lower the extreme deviation value of the scores.

$\mathbf{A_i}$ denotes the i-th entry into the second round of evaluation, M denotes the total number of entries into the second round of evaluation, a_{ij} denotes the score of the i-th entry into the second round of evaluation by expert$j$, and K denotes the number of experts needed to evaluate each entry into the second round of evaluation, based on the above assumptions, we use a set of the way to express the scores of each work in the second round of evaluation, such as Equation (9).

$$\mathbf{A}_i = \{a_{i1}, a_{i2}, \dots, a_{iK}\}, \quad i \in \{1,2, \dots, M\} \tag{9}$$

In summary, we can summarize the "extreme variance" model in the following steps.

(1) Count all the scores of each work$\mathbf{A}_i$, and form a set containing all the scores, as in Equation (9).

(2) The extreme difference r_i of all the scores of each work $\mathbf{A}_i$ is calculated as follows.

$$r_i = \max\{a_{i1}, a_{i2}, \dots, a_{iK}\} - \min\{a_{i1}, a_{i2}, \dots, a_{iK}\}, \quad i \in \{1,2, \dots, M\} \tag{10}$$

(3) The extreme values r_i of all the works form a set of extreme values**R**, $\mathbf{R} = \{r_1, r_2, \ldots, r_M\}$;

(4) Set the parameterη, and calculate the number of works to be reconsidered by$\eta \times M$.

(5) Sort the elements in the set $\mathbf{R} = \{r_1, r_2, \ldots, r_M\}$ in descending order, and find the work that corresponds to the element $\eta \times M$ in the first place.

(6) For a work to be reviewed, find the median a_m from the set of scores, assuming that the original standardized scores were$\{a_1, a_2, \ldots, a_K\}$.

(7) Set the parameterθ, which is used to indicate which scores need to be modified, and judge each element in the set of scores$\{a_1, a_2, \ldots, a_K\}$. The judgment conditions and processing methods are shown in Equation (11).

$$\begin{cases} \text{Raise the score appropriately,} & \text{if} \quad a_i < (1-\theta) \times a_m, \\ \text{No adjustment required,} & \text{if} \quad (1-\theta) \times a_m \le a_i \le (1+\theta) \times a_m, \\ \text{Appropriately lowered scores,} & \text{if} \quad a_i > (1+\theta) \times a_m, \end{cases} \tag{11}$$

(8) For entries that require a review of the scores, the final ranking will be based on the reviewed scores as the standardized scores, and the entries will be ranked and selected to receive the various awards.

4.1. First Review Stage "Large Extreme Difference" Modeling

First of all, we need to set two thresholds T_{1H}, T_{1L}, which are the judgment thresholds for high scoring works and the judgment thresholds for low scoring works, and the specific judgment conditions and processing methods are as shown in Equation (12).

$$\begin{cases} \text{The work is judged to be "low scoring",} & \text{if Highest score} < T_{1L}, \\ \text{The work is judged to be "high grade",} & \text{if Minimum score} > T_{1H} \end{cases} \tag{12}$$

For the works judged by the above two indicators, the following actions are performed, as shown in Fig. 7.

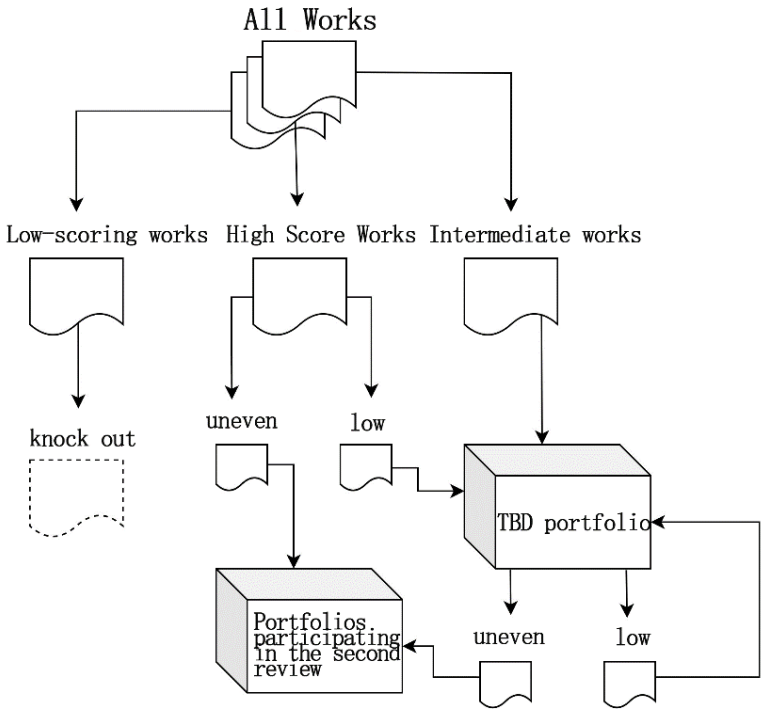


Fig 7. Schematic diagram of rating segment management.

According to the above operation, as the whole process does not take into account the effect of the extreme differences in the results, the first stage of the non-high and non-low scores of the "large extreme" can be better dealt with, and to ensure that the quality of the work into the second stage of the evaluation of the work.

5. Conclusion

Through this study, we propose and validate a set of optimization schemes for several key problems in the evaluation of science and technology innovation competitions. First, the work distribution model designed by using stochastic simulation method successfully realizes the balanced distribution of expert judging workload and effectively reduces the intersection of judged works. Second, the standard score calculation is optimized by introducing the T-score method, which significantly reduces the subjectivity and polarity of scoring. Again, the implemented segmentation management strategy rationally categorizes the works by setting score thresholds, effectively highlighting and screening out high-quality innovative projects. Taken together, the evaluation scheme proposed in this study has improved in terms of science, fairness and efficiency, and has high practical value. Future work can further consider more dimensional evaluation criteria and smarter evaluation tools to further improve and enhance the overall effectiveness of the evaluation system.

References

[1] Christina Li, Wei Jiangyao, Zhao Ningbo, et al. Construction and application of lean procureme nt review system in power grid material enterprises[C]//China's power enterprise management in novation practice (2022). Materials Company of State Grid Shaanxi Electric Power Co Ltd;,202 4:3.DOI:10.26914/c.cnkihy.2024.001212.

[2] Yang Jianqiang, Wang Shi,Lai Yunbing, et al. Competitive procurement of equipment contract supervision innovative method system of (XV) cross-evaluation method[J]. China Military to Civilian, 2023, (17):40-41.

[3] FENG Pengfei, ZHOU Inhibition, LI Zhixuan, et al. A probabilistic statistical model of measured ground shaking based on generalized extreme value distribution [J/OL]. Journal of Shanghai Jiao Tong University,1-22[2024-05-16].https://doi.org/10.16183/j.cnki.jsjtu.2023.558.

[4] GUO Dongwei, DING Genhong, MAO Juncheng, et al. Weighted T-score method for group decision-making essay-based competition rankings [J]. China Science and Technology Paper, 2015, 10(17):2059-2063.

[5] Huang Peiyi, Peng Qiuzhi, Zhu Dan, et al. The effect of removing county population density extremes on the relationship between topographic factors and population density [J]. Journal of Inner Mongolia Normal University (Natural Science in Chinese), 2023, 52(05):489-495.