Retraction

Retracted: Abnormal-Aware Multiperson Evaluation System with Improved Fuzzy Weighting

Journal of Function Spaces

This article has been retracted by Hindawi, as publisher, following an investigation undertaken by the publisher [1]. This investigation has uncovered evidence of systematic manipulation of the publication and peer-review process. We cannot, therefore, vouch for the reliability or integrity of this article.

Please note that this notice is intended solely to alert readers that the peer-review process of this article has been compromised.

Wiley and Hindawi regret that the usual quality checks did not identify these issues before publication and have since put additional measures in place to safeguard research integrity.

We wish to credit our Research Integrity and Research Publishing teams and anonymous and named external researchers and research integrity experts for contributing to this investigation.

The corresponding author, as the representative of all authors, has been given the opportunity to register their agreement or disagreement to this retraction. We have kept a record of any response received.

References

[1] S. Ni, "Abnormal-Aware Multiperson Evaluation System with Improved Fuzzy Weighting," *Journal of Function Spaces*, vol. 2022, Article ID 4899831, 8 pages, 2022.

Research Article

Abnormal-Aware Multiperson Evaluation System with Improved Fuzzy Weighting

Shutong Ni [ID]

School of Statistics and Applied Mathematics, Anhui University of Finance and Economics, Bengbu, Anhui 233030, China

Correspondence should be addressed to Shutong Ni; 20193088@aufe.edu.cn

There exists a phenomenon that subjectivity highly lies in the daily evaluation process. Our research primarily concentrates on a multiperson evaluation system with anomaly detection to minimize the possible inaccuracy that subjective assessment brings. We choose the two-stage screening method, which consists of rough screening and score-weighted Kendall-τ distance to winnow out abnormal data, coupled with hypothesis testing to narrow global discrepancy. Then we use fuzzy synthetic evaluation method (FSE) to determine the significance of scores given by reviewers as well as their reliability, culminating in a more impartial weight for each reviewer in the final conclusion. The results demonstrate a clear and comprehensive ranking instead of unilateral scores, and we get to have an efficiency in filtering out abnormal data as well as a reasonably objective weight determination mechanism. We can sense that through our study, people will have a chance of modifying a multiperson evaluation system to attain both equity and a relatively superior competitive atmosphere. A preprint has previously been published (Ni, 2022).

1. Introduction

The evaluation system has long been an indispensable part of measuring the performance of particular behaviour. For years, subjective evaluation and objective assessment have been rather separated in their respective fields. However, with the booming improvement in science and technology, these two indicators are somehow gradually intertwined and have the objective one taken the lead. Even so, subjective evaluation cannot be erased for good, on the account that it has its unique characteristics indeed, which can be generalized as minute scope, fair adaptability, low cost, and high randomness. When it comes to examinations or appraisals, expertise revision of contributions, personnel recruitment, project bidding, judgments on equipment's function or merchandise's quality, and even government's policymaking, subjective assessment operates in every tiny aspect of the society, paving the way for its unremitting upswing.

Practically, the two evaluation strategies have their leanings. To minimize the repercussion of their defects, under a particular circumstance, there are scholars wedded to incorporating the two assessment methods together, in the hope

that the results can be much fairer as well as more reliable [1–3]. Nevertheless, in many scenarios, objective evaluation data is hard to obtain, and many a strength of subjective evaluation make it a more practical means. A simple way to increase the credibility of the evaluation is to summarize the information of multiple reviewers, which will lead to the inconsistency of results so that a significant number of researchers address themselves into how to integrate various information to make the ultimate review authentic as much as possible. To avert a mixture of standards, experts should conduct an evaluation with respect to various indices in different situations. Xing [4] proposes correlation analysis [5] to measure the reliability of reviewers, screening out the discrepant values and, at the same time, streamlining the assessment indices that possess a strong correlation with each other. Regarding that different reviewers have different standards over objects, Xing [4] chooses to apply fuzzy analysis hierarchy process (FAHP) [6] to the problem, in the hope of a comparable and definite weight factor for each evaluation index. That is to say, by forming a fuzzy judgment matrix that consists of experts' appraisals, we can determine the weight factors in the assessment system.

Nevertheless, most works fail to consider the confidence of different reviewers, and they just take average to obtain the score of a specific assessment index for each individual. Moreover, those proposed assessment indices will also be affected both subjectively and professionally, and on a certain condition, we cannot even put forth a scientific index, for instance, when teachers in schools rate their students, the only factor worth referring to is the final score. Thus, our research will prioritize the information-limited problems and attempt to refine the disadvantages of subjective evaluation as below:

To start with, the evaluation standard differs from person to person, so there is no absolute right or wrong, and every authority has his/her own precept and preference. In other words, for a certain wide range of scoring standards, diverse experts have enormous differences in the understanding of the evaluation standards and the grasp of the assessment scales, which simultaneously give rise to a conspicuous contrast. The second lies in the psychological impact of each expert during the evaluation process. When scoring, the experts will inevitably be influenced more or less by the scores of other students he has given. That is to say, subsequent scores will be subject to all of the previous scoring results, which conduce to the essential variation.

According to the imperfections that exist in the subjective evaluation, our study intends to improve the multiperson subjective evaluation method through the lens of mathematical modeling, which is widely applied in various kinds of engineering and simulation problems [7–10]. We endeavour to provide resolutions in the design of assessment procedure as well as assessment approaches, making evaluation results and objective facts coordinate as much as possible. In this case, we may help stamp out the bias and constraints of individual evaluation and demonstrate impartiality as well as authority, so as to shape a superior competitive atmosphere.

In our research process, we use two-step screening as a quick start to examine the anomalous data. Given that the standard Q-test method [11] and 3-σ principle [12] fail to work well when the samples are inadequate, we can utilize these methods for a rough selection and then apply Kendall-τ distance to examine the data winnowed out for the sake of advancing the secondary screening procedure. In light of the drawbacks that original Kendall-τ distance can hardly fully contemplate the differences among values of scores, we rework the idea as score-weighted Kendall-τ distance and regard it as an objective function, screening out the abnormal data which will result in a minor decrease in the objective. Moreover, still in the data preprocessing stage, we propose to mitigate the discrepancy of scores given by the same reviewer between two classes through hypothesis testing. Later, considering the fuzzy relationships among reviewers' judging criteria, we interpret the outcomes from different experts as different judging indices of objects evaluated, so as to use the fuzzy synthetic evaluation model [13–16]. When weighing those experts' reviews, we not only take the weight stemmed from the fuzzy synthetic evaluation into consideration but also calculate another type of weight derived from the scale of anomalous data excluded, which

can tell the reliability of the evaluation. Coupling the two weights with each other, we take the average as the final weight for the index, thereby measuring the accuracy of evaluation more efficiently [17].

In the following sections, we first present the formulation of the entire problem. And then, we narrate the formation process of our evaluation method, including the two-stage screening method, hypothesis test, and fuzzy synthetic evaluation method. Subsequently, we conduct experiments and show the effectiveness of our methods. Furthermore, we ultimately draw a conclusion of the questions and elucidate the notion of our research and modification for a more equitable evaluation system.

2. Problem Formulation

In this section, we introduce standard problem formulation that we aim to tackle via our proposed model. Moreover, we also specify certain assumptions which contribute to the completeness of our definition.

Let g_{ij}^k denote the grade given by reviewer i to student j in class k, where $i = 1, 2, \cdots, m$, $j = 1, 2, \cdots, n_k$, and $k = 1, 2, \cdots, C$. Let π_i denote the ranking of all students, whose papers are reviewed by reviewer i, in descending order of their grades.

We mainly focus on two representative assessment situations. The first is that we are given one class with n students whose papers need to be graded by m reviewers. The second situation is that we are given k classes, where $k \geq 2$, and each of the m reviewers must grade all the classes. In both cases, the task is to estimate the actual grades of all the students and determine their final rankings as fair and objective as possible, without knowing about the evaluation principles or preferences of each reviewer. It is worth mentioning the following assumptions that help avoid certain intricate controversies:

(1) Each reviewer grades papers under the same external conditions

(2) All papers are kept secret before reviewing

(3) Reviewers are not allowed to discuss with each other

(4) Students' rankings are only determined by their final synthesized scores

3. The Proposed Evaluation System

3.1. Anomaly Detection. Since the only difference between the two situations is the number of classes, we decide to tackle this at the end of the analysis. For the screening of the abnormal data, the most common method is indubitably the trimmed mean, which strikes out a certain proportion of the highest and lowest scores and averages the remaining ones. However, in our research, we actually have only a small amount of reviewers for the evaluation task, where the trimmed method cannot operate to its full potential and has low robustness [18]. On the one hand, under the premise that students have only a few scores, removing high

scores or low scores can reduce extreme data to a certain extent, but it also causes a significant loss of data. In this case, for some superficially extreme scores, we decide to winnow them out instead of directly discarding them. This effectively considers that despite the extremity of the lateral comparison among scores given by different reviewers, this kind of extremity can be rather valuable in the whole rankings given by a specific reviewer. Taking this factor into consideration and then designing a rational objective function, we can reach out to a relatively optimal two-step screening method. On the other hand, each reviewer may generally rate students from different classes high or low due to various evaluation standards, and there remains a certain contrast among scoring intervals. Therefore, the scores given by the reviewers do not comply with the normal distribution of their true average score, so the horizontal comparison can be meaningless. As a consequence of that, if we substitute students' scores with their rankings, this new indicator can also make the two-step screening method well-performed.

3.1.1. Rough Screening.
There are plenty of typical means to filter outliers, such as 3-σ principle, quantile method, and Q-test. However, in terms of our research, we notice that there are not enough samples, because each student have gained scores only from few reviewers, and at the same time, the degree of anomaly fails to reach the standards of the methods mentioned above. For the second situation, after we transform their original scores into rankings as discusses above, it will face the same trouble.

Aware of this problem, we alter our perspective from designing an efficient one-round screening method to a two-stage method. In the first step, we roughly calculate the average and variance of students' scores and then subsume data into a set $\mathscr{G}_1$ if its deviation from the average is greater than α times of its variance, that is,

$$\mathscr{G}_1 = \left\{ \left(j, i, g_{ij} \right) : \left| g_{ij} - \mathrm{avg}_j \right| \geq \alpha \bullet \mathrm{std}_j \right\}, \qquad (1)$$

where avg_j and std_j are the average and standard deviation of student j's scores, respectively. We call $\mathscr{G}_1$ the anomaly set after proceeding step i, and α is a hyperparameter required to be fine-tuned. The second step is much more pivotal and will be specified in the following section.

3.1.2. Second Screening via Score-Weighted Kendall-τ Distance.
To conduct the second step of winnowing out abnormal data, we would like to introduce the definition of Kendall-τ distance [19] to you for a quick start.

Definition 1. We define the Kendall-τ distance between τ_1 and τ_2 as below:

$$\delta_K(\tau_1, \tau_2) = \sum_{u \succ_{\tau_1} v} I\left[\left\{ v \succ_{\tau_2} u \right\} \right], \qquad (2)$$

where $\{\cdot\}$ denotes certain events, $I[\cdot]$ denotes the indicative function, and $x \succ_\tau y$ denotes that student x is ranked higher than student y in rank τ.

The ranking of students can be directly obtained through their scores. According to Definition 1, we can sense that if two reviewers' reviews on a particular student differ a great deal, then the Kendall-τ distance will be relevantly larger.

Having a deeper insight into the problem, the given definition of Kendall-τ distance practically remains some drawbacks because it simply contains the order of two students but overlooks the specific difference in scores. However, in our research, the problem encountered has siccar grades statistics, so that we can make a modification and obtain a score-weighted Kendall-τ distance as below:

$$\delta_{SK}(\pi_k, \pi_l) = \sum_{u \succ_{\pi_k} v} \frac{\left[(g_{ku} - g_{kv}) + (g_{lv} - g_{lu}) \right]}{2} \bullet I\left[\left\{ v \succ_{\pi_l} u \right\} \right]$$
$$\bullet I\left[\left\{ u, v \in \pi_k \right\} \cap \left\{ u, v \in \pi_l \right\} \right], \qquad (3)$$

where π_k and π_l are rankings given by reviewers k and l, respectively, and $u \in \pi_k$ denotes that the score of student u from reviewer k has not been filtered out. We hope that the following objective function will become as small as possible after we winnow out the target data.

$$L(\pi_1, \cdots, \pi_m) = \sum_{1 \leq k < l \leq m} \delta_{SK}(\pi_k, \pi_l). \qquad (4)$$

If we, respectively, consider how much the objective function L will decline after removing a subset of the abnormal set $\mathscr{G}1$, it will become an exponential time complexity algorithm, which is pretty impractical. Hence, we apply a greedy method that only needs to figure out how much the objective function will decline when one abnormal score $g_{ij} \in \mathscr{G}_1$ is deleted. Those who cause significant decline are more likely to be eliminated. Here, we do not necessarily use formula (4) for calculation every single time but merely compute the decline value for each g_{ij} as below:

$$\Delta_{ij} = \sum_{k \neq i} \left(\delta_{SK}(\pi_k, \pi_i) . - \delta_{SK}\left(\pi_k \setminus \left\{ g_{kj} \right\}, \pi_i \setminus \left\{ g_{ij} \right\} \right) \right)$$
$$= \sum_{k \neq i} \left(\sum_{j \succ_{\pi_k} v} \frac{\left(g_{kj} - g_{kv} \right) + \left(g_{iv} - g_{ij} \right)}{2} \bullet I\left[\left\{ v \succ_{\pi_i} j \right\} \right] \right.$$
$$\left. + \sum_{j \succ_{\pi_k} v} \frac{\left(g_{kj} - g_{kv} \right) + \left(g_{ij} - g_{iv} \right)}{2} \bullet I\left[\left\{ j \succ_{\pi_i} v \right\} \right] \right) \qquad (5)$$
$$= \sum_{k \neq i} \sum_v \frac{\left| g_{kj} - g_{kv} \right| + \left| g_{iv} - g_{ij} \right|}{2}$$
$$\bullet I\left[\left\{ j \succ_{\pi_i} v, v \succ_{\pi_k} j \right\} \cup \left\{ j \prec_{\pi_i} v, v \prec_{\pi_k} j \right\} \right] ..$$

After sorting $\Delta = \{ \Delta_{ij} : g_{ij} \in \mathscr{G}_1 \}$ in descending order, we select the largest $d\Delta_{ij}$ and find their corresponding $g_{ij} \in \mathscr{G}_1$ to form another abnormal set $\mathscr{G}_2$, where the size of d can be heuristically tuned. Note that if an anomaly g_{ij} has already been blanked out, theoretically, the corresponding
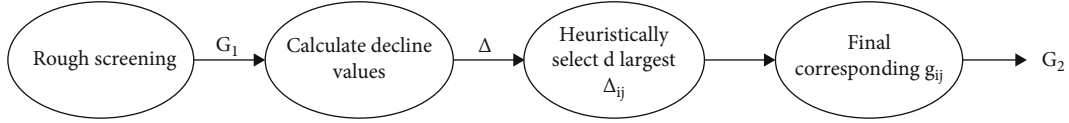
Figure 1: Architecture of two-stage screening method.

decline values of the other anomalies should be recalculated with respect to a smaller scores set. To put it more clearly, if a $g_{i_0 j_0}$ is moved away, terms related to reviewer i_0, or say π_{i_0}, in equation (3) will become zero due to the existence of the second indicative function $I[\{u, v \in \pi_k\} \cap \{u, v \in \pi_l\}]$. However, if we denote the set recalculated decline values as $\Delta' = \{\Delta'_{ij} : g_{ij} \in \mathcal{G}_1 \setminus \{g_{i_0 j_0}\}\}$, we can easily verify that if $\Delta_{i_1 j_1} \leq \Delta_{i_2 j_2}$, then $\Delta'_{i_1 j_1} \leq \Delta'_{i_2 j_2}$. Therefore, getting rid of the anomalies greedily merely with respect to Δ is reasonable. We present a concise architecture in Figure 1 to summarize the two-stage screening.

3.2. Fuzzy Synthetic Evaluation. Based on the problem analysis in Part 2, after the screening of abnormal data, we need to synthesize all the information to determine the final score of each student. Inspired by the methods in [20, 21], we cleverly adapt the fuzzy synthetic evaluation method to our less informatics problem setting, treating each reviewer as an evaluation index. We have the observation matrix $G = (g_{ij})_{1 \leq i \leq m, 1 \leq j \leq n}$, where m is the number of reviewers and n is the number of students. Since the scores given by each reviewer can be understood as a benefit indicator, we can establish a fuzzy benefit matrix $B = (b_{ij})_{1 \leq i \leq m, 1 \leq j \leq n}$, where

$$b_{ij} = \frac{g_{ij} - \min_j g_{ij}}{\max_j g_{ij} - \min_j g_{ij}}. \tag{6}$$

It should be noted that, in accordance with Section 4.1, some g_{ij} has been screened out. We do not fill in the blanks for those missing values but choose to ignore them, which means that if g_{ij} is missing, b_{ij} is also missing in matrix B.

Coming up then, we need to establish the weight w_i, $1 \leq i \leq m$ of each reviewer. Firstly, the coefficient of variation is adopted to fully consider the influence of the size of evaluation intervals on the degree of differentiation of students. The coefficient of variation corresponding to each reviewer is calculated according to formula (7), aiming at fully considering the unknown influence of the size of evaluation intervals on ranking:

$$v_i = \frac{\sigma_i}{\bar{g}_i}, i = 1, \cdots, m, \tag{7}$$

where

$$\bar{g}_i = \frac{1}{n} \sum_{j=1}^{n} g_{ij}, \tag{8}$$

$$\sigma_i = \sqrt{\frac{1}{n-1} \sum_{j=1}^{n} \left(g_{ij} - \bar{g}_i\right)^2}, i = 1, 2, \cdots, n. \tag{9}$$

Note that when calculating σ_i and $\bar{g}_i$, the sample size is taken as the number of scores given by the evaluation reviewer i, excluding the screening ones. Then, v_i is normalized to obtain the first component of the weight:

$$w_i^{(1)} = \frac{v_i}{\sum_{i=1}^{m} v_i}, i = 1, \cdots, m. \tag{10}$$

Secondly, we consider the credibility of each reviewer and use n_i, the number of reviewer i's screened scores, to obtain another part of the weight:

$$w_i^{(2)} = \frac{n - n_i}{\sum_{i=1}^{m} (n - n_i)}, i = 1, \cdots, m. \tag{11}$$

Finally, the weight of each reviewer is determined by the following formula:

$$w_i = \frac{w_i^{(1)} + w_i^{(2)}}{2}, i = 1, \cdots, m. \tag{12}$$

Then, $F_j = \sum_{i=1}^{m} w_i \cdot b_{ij}$ is calculated for student j, and their ranking can be obtained after sorting $\{F_j\}$.

Actually, the ultimate goal of our calculations is to determine the ranking of students. Now that the ranking has been obtained, if we want to output the final score, we only need to select a reference value, for example, $\min_j g_{1j}$, and further plus $F_j (\max_j g_{1j} - \min_j g_{1j})$. Then, the final score f_j of student j is

$$f_j = F_j \left(\max_j g_{1j} - \min_j g_{1j}\right) + \min_j g_{1j}. \tag{13}$$

We also provide a clear flow chart of how to utilize FSE in our problem as shown in Figure 2.

3.3. Normal Hypothesis Tests. For the second situation, we propose to perform the normal hypothesis test before screening abnormal data. For simplicity, we only consider the case with two classes. Since we can give feasible solution, it can be easily generalized to multiclass scenarios by fixing one target class and conducting hypothesis test between the target and other classes one by one. Specifically, we test the mean and variance of the grades given by one reviewer to two classes. We assume that the grades given by reviewer i to class k follow the normal distribution $N(\mu_i^k, \sigma_i^k), i = 1, \cdots, m, k = 1, 2$; we need to test the following two questions:

(1) $H_0 : \mu_i^2 - \mu_i^1 = 0 \leftrightarrow H_1 : \mu_i^2 - \mu_i^1 \neq 0$

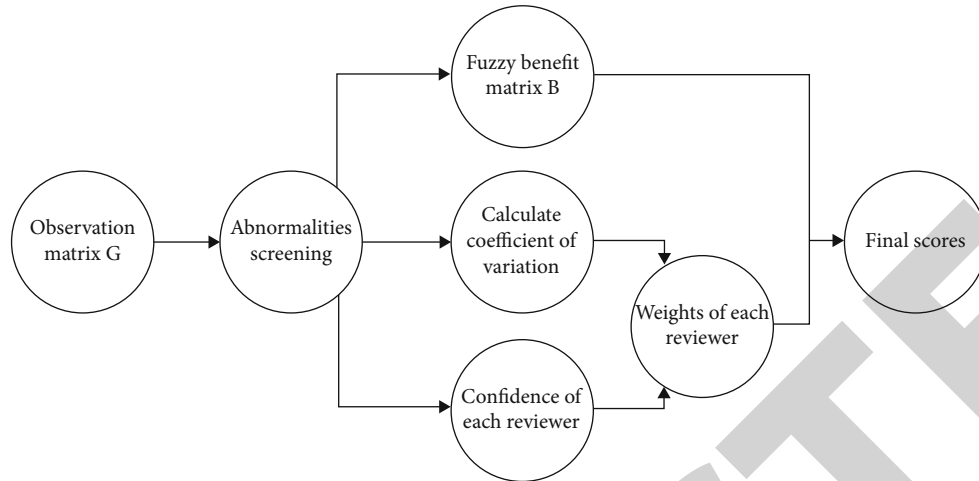(2) $H_0 : \sigma_i^2 / \sigma_i^1 = 1 \leftrightarrow H_1 : \sigma_i^2 / \sigma_i^1 \neq 1$

FIGURE 2: Flow chart of calculating students' final scores via fuzzy synthetic evaluation.

Set the confidence level to α. Note that μ_i^k and σ_i^k are unknown, and here, we have large samples. Therefore, we apply t-test to the first question and F-test to the second, and the rejection region of the test for question (a) and question (b) are, respectively, as below:

$$D_1 = \left\{ \left(g_{i1}^1, \cdots, g_{in_1}^1, g_{i1}^2, \cdots, g_{in_1}^2 \right) : \left| \frac{G_i^2 - G_i^1}{S_i^w} \sqrt{\frac{n_1 n_2}{n_1 + n_2}} \right| \right.$$
$$\left. > t_{n_1+n_2-2}\left(\frac{\alpha}{2}\right) \right\},$$

$$D_2 = \left\{ \left(g_{i1}^1, \cdots, g_{in_1}^1, g_{i1}^2, \cdots, g_{in_1}^2 \right) : \frac{\left(S_i^2\right)^2}{\left(S_i^1\right)^2} \right.$$
$$\left. > F_{n_2-1,n_1-1}\left(1 - \frac{\alpha}{2}\right) \text{or} \frac{\left(S_i^2\right)^2}{\left(S_i^1\right)^2} > F_{n_2-1,n_1-1}\left(\frac{\alpha}{2}\right) \right\},$$

$$\tag{14}$$

where

$$G_i^k = \frac{1}{n_k} \sum_{j=1}^{n_k} g_{ij}^k,$$

$$\left(S_i^w\right)^2 = \frac{1}{n_1 + n_2 - 2} \left[(n_1 - 1)(S_1)^2 + (n_2 - 1)(S_2)^2 \right],$$

$$\left(S_i^k\right)^2 = \frac{1}{n_k - 1} \sum_{j=1}^{n_k} \left(g_{ij}^k - G_i^k \right)^2, \quad i = 1, \cdots, 5, k = 1, 2.$$

$$\tag{15}$$

All the statistics mentioned above can be obtained from the existing data. Based on the results of the two tests as well as the practical meaning of confidence rate, we scale the scores differently:

(1) If both of the test results are rejection, we apply the following transformation to the scores of class 2 given by the reviewer i:

$$g_{ij}^2 = (1 - \alpha)\left[\frac{S_i^1}{S_i^2} \left(g_{ij}^2 - G_i^2 \right) + G_i^1 \right] + \alpha g_{ij}^2 \tag{16}$$

(2) If the result of question (a) is rejection and that of question (b) is acceptance, we apply the following transformation to the scores of class 2 given by the reviewer i:

$$g_{ij}^2 = (1 - \alpha)\left(g_{ij}^2 - G_i^2 + G_i^1 \right) + \alpha g_{ij}^2 \tag{17}$$

(3) If the result of question (b) is rejection and that of question (a) is acceptance, we apply the following transformation to the scores of class 2 given by the reviewer i:

$$g_{ij}^2 = (1 - \alpha)\left[\frac{S_i^1}{S_i^2} \bullet \left(g_{ij}^2 - G_i^2 \right) + G_i^1 \right] + \alpha g_{ij}^2 \tag{18}$$

(4) If both of the test results are acceptance, we keep them unchanged

After rescaling, we can merge the two classes' grades into one table and then conduct abnormal data screening and fuzzy synthetic evaluation.

4. Experiments

In this section, we conduct two experiments that testify to the effectiveness of our methods. We will specifically describe the dataset we use and present the results in detail.

4.1. Settings. For each aforementioned situation in Section 2, we collect two corresponding datasets for evaluating the

TABLE 1: Statistics for the two datasets.

Statistics	Dataset I	Dataset II
Reviewer number (R)	3	5
Class number (C)	1	2
Student number (S)	107	61,50

effect of our proposed method. Table 1 summarizes some critical statistics of the datasets.

Notably, all the scores are rated following the percentage system, ranging from 0 to 100, preventing the uncertainties introduced from other level ranking systems. Apart from that, we do not have any knowledge of the reviewers' criteria for judging.

4.2. Results

4.2.1. Experiment I. In the first scenario, the intervals of the grades given by the three reviewers are 65-95, 62-99, and 60-98, respectively, which are almost identical. Therefore, there is no need to use the student's ranking instead of grades to filter out abnormal data.

We first apply the method proposed in Section 3.1.1 to screen out 33 anomalous scores. After that, we calculate the quantity for each anomalous score based on equation (5) and sort the results in decreasing order as below: $\Delta = \{(91, 3, 2667.0), (57, 3, 1811.5), (6, 3, 1503.5), (53, 3, 1447.5), (7, 3, 1444.5), (62, 3, 1429.5), (60, 3, 1384.5), (94, 1, 1361.0), (8, 2, 1121.5), (68, 3, 1120.0), (74, 3, 1104.0), (82, 2, 1017.0), (90, 2, 952.0), (102, 1, 871.0), (71, 3, 831.0), (48, 3, 816.0), (69, 3, 764.5), (27, 2, 757.5), (20, 3, 731.0), (19, 1, 726.0), (25, 3, 715.0), (47, 3, 696.0), (86, 1, 658.5), (33, 3, 648.0), (73, 3, 645.5), (89, 2, 642.5), (58, 3, 625.0), (51, 3, 624.0), (84, 3, 571.5), (17, 2, 548.5), (2, 1, 452.0), (78, 1, 334.0), (97, 2, 128.0)\}$. "$(j, i, x)$" denotes $\Delta_{ij} = x$.

We set the confidence to 0.80; that is, select the first $33 \times 0.80\% \approx 26$ elements in collection $\{\Delta ij\}$ and take their corresponding g_{ij} as the final anomalous scores: $G_2 = \{(91, 3, 98), (57, 3, 65), (6, 3, 60), (53, 3, 68), (7, 3, 67), (62, 3, 67), (60, 3, 63), (94, 1, 95), (8, 2, 85), (68, 3, 70), (74, 3, 63), (82, 2, 75), (90, 2, 85), (102, 1, 85), (71, 3, 74), (48, 3, 70), (69, 3, 71), (27, 2, 69), (20, 3, 72), (19, 1, 88), (25, 3, 62), (47, 3, 74), (86, 1, 90), (33, 3, 61), (73, 3, 73), (89, 2, 95)\}$. "$(j, i, x)$" denotes $g_{ij} = x$.

The next step is to proceed with the fuzzy synthetic evaluation. The observation matrix can be directly obtained from the dataset, and simple calculation leads to the fuzzy benefit matrix. When these are all done, we can calculate the two types of weights:

$$w^{(1)} = (0.325675, 0.377674, 0.296651),$$
$$w^{(2)} = (0.349153, 0.345763, 0.305085), \quad (19)$$
$$w = (0.337414, 0.361718, 0.300868).$$

Finally, we calculate the final score using formula (13), and the quicksort algorithm [22] can be harnessed to accelerate ranking. Because the weights are all decimals, we assume the final score should be kept to two decimal places. Students with the same score are regarded as the same ranking (for the specific final score table, please refer to Supplementary 1).

4.2.2. Experiment II. There are apparent differences between the evaluation intervals of the five teachers in the second dataset. As mentioned above, we propose to use the students' initial ranking as the score to screen out outliers and then conduct fuzzy synthetic evaluation and other operations. However, the greatest problem here is to mitigate the deviation of evaluation intervals. Thus, we are supposed to first implement the method in Section 3.3. After rescaling, we can fairly compare the grades of different classes; thus, fuzzy synthetic evaluation can be conducted to get final scores and rankings. For the sake of emphasis, we do not present the results of the screening operation. The final scores are presented in Supplementary 2.

5. Discussions

In this section, we reflect on the strengths and weaknesses of our methods in detail.

(1) Based on the experiment results in Section 4.2.1, we can see that the two-stage method shows great performance in screening out the abnormal scores. The scores we selected are abnormal at first glance, matching the criterion of human's intuitive judgement. Moreover, their corresponding decline values are rigorously verified to be large, which goes beyond human intuition and is more trustworthy

(2) Based on the experiment results in Section 4.2.2, we can see that the hypothesis test is suitable for multiclass situations. The distribution of rating is similar to the normal distribution in the real world. The renormalization maintains the scores of different classes in similar level, which is revealed through the rational distributed top students in two classes

(3) We can design concise and efficient algorithms according to the two flow charts we presented in Sections 3 and 4, which demonstrate the feasibility of applying our methods on larger datasets

(4) There are some hyperparameters existing in our methods. We mainly tuned them via heuristic methods, but it contains certain biases. It is better to design a mechanism that can automatically finetune the parameters

(5) When performing the screening of abnormal data, slightly different approaches are used in the two scenarios, respectively, but these methods cannot wholly override all possible abnormalities. Supposing that multiple criterion are used to eliminate the outliers in data for a specific problem and some data appear to be an exception in all methods, we can

more confidently determine the data as abnormal ones

(6) The fuzzy synthetic evaluation (FSE) method is a comprehensive evaluation method based on the problems of fuzziness and uncertainty in the evaluation criteria, evaluation factors, and the problem of difficulty in quantifying qualitative indicators. It can express a fuzzy object with a precise number so that the evaluation of fuzzy events is scientific and reasonable. Nevertheless, the application of the FSE method usually brings about relatively high subjectivity in the determination of indicators, fuzzy relation matrix, weight, etc. Besides, there is no clear or systematic method for determining the membership function, which conspires to a specific difference in results

6. Conclusions

The evaluation problem we concentrate on in our study is based on the fact that for schooling and appraisals, there exist a certain range of situations where no uniform standard is contained. In those cases, assessment results appear to be entirely subjective and divergent. It turns out that the evaluation problem has an inextricable connection with people's daily lives, and this is why our research is intended to center on this question and make an expansion. Our research proposes a modified score-weighted Kendall-τ distance as the judging criterion, adopts FSE and normal hypothesis test to be the principle investigating methods, and uses Python and MATLAB as auxiliary tools for implementation and testing. Under the auspices of fundamental scientific materials, we ultimately get to winnow out anomalies and then synthesize different factors for a comprehensive evaluation, culminating in a relatively equitable judging system. We believe that our work can give inspiration to improve the evaluation system under certain situations lacking in objective criteria. Besides, the two-stage screening method can be further extended to multistage version by properly examining other important characteristics of the reviews, which is left as a future work.

Data Availability

All the data used in this article are available at https://github.com/Ciao-Yvette/Multi-person-Evaluation-System.

Conflicts of Interest

The author declares that there are no conflicts of interest.

Acknowledgments

Supplementary Materials

Table 1: the final scores and rankings of all the students in both classes in the first situation. Table 2: the final scores and rankings of the students in class 1. Rankings are calculated among all the students in both classes. Table 3: the final scores and rankings of the students in class 2. Rankings are calculated among all the students in both classes. (Supplementary Materials)

References

[1] J. L. Hossain, L. W. Reinish, R. J. Heslegrave et al., "Subjective and objective evaluation of sleep and performance in daytime versus nighttime sleep in extended-hours shift-workers at an underground mine," *Journal of Occupational and Environmental Medicine*, vol. 46, no. 3, pp. 212–226, 2004.

[2] C. A. Gumussoy, A. Pekpazar, M. Esengun, A. E. Bayraktaroglu, and G. Ince, "Usability eval- uation of TV interfaces: subjective evaluation vs. objective evaluation," *International Journal of Human–Computer Interaction*, vol. 38, no. 7, pp. 661–679, 2021.

[3] D. A. Crolla, D. C. Chen, J. P. Whitehead, and C. J. Alstead, "Vehicle handling assessment using a combined subjective-objective approach," *SAE Transactions*, vol. 107, pp. 386–395, 1998, http://www.jstor.org/stable/44740965.

[4] R. Xing, *Research on the Methods of Subjective Evaluation for Handling and Stability of a Car[PhD thesis]*, Jilin University, 2010.

[5] T. K. Koo and M. Y. Li, "A guideline of selecting and reporting intraclass correlation coefficients for reliability research," *Journal of Chiropractic Medicine*, vol. 15, no. 2, pp. 155–163, 2016.

[6] E. E. Haji, A. Azmani, and M. E. Harzli, "Using the FAHP method in the educational and vocational guidance," *International Journal of Modern Education & Computer Science*, vol. 10, no. 12, 2018.

[7] J. B. Liu and X. F. Pan, "Minimizing Kirchhoff index among graphs with a given vertex bipartiteness," *Applied Mathematics and Computation*, vol. 291, pp. 84–88, 2016.

[8] J. B. Liu, J. Zhao, J. Min, and J. D. Cao, "The Hosoya index of graphs formed by a fractal graph," *Fractals*, vol. 27, no. 8, p. 1950135, 2019.

[9] J. B. Liu, J. Zhao, H. He, and Z. Shao, "Valency-based topological descriptors and structural property of the generalized Sierpiński networks," *Journal of Statistical Physics*, vol. 177, no. 6, pp. 1131–1147, 2019.

[10] J. B. Liu, Y. Bao, W. T. Zheng, and S. Hayat, "Network coherence analysis on a family of nested weighted n-polygon networks," *Fractals*, vol. 29, no. 8, p. 276, 2021.

[11] R. B. Dean and W. J. Dixon, "Simplified statistics for small numbers of observations," *Analytical Chemistry*, vol. 23, no. 4, pp. 636–638, 1951.

[12] F. Pukelsheim, "The three sigma rule," *The American Statistician*, vol. 48, no. 2, pp. 88–91, 1994.

[13] Z. Zou, Y. Yi, and J. Sun, "Entropy method for determination of weight of evaluating indicators in fuzzy synthetic evaluation for water quality assessment," *Journal of Environmental Sciences*, vol. 18, no. 5, pp. 1020–1023, 2006.

[14] N. Chang, H. Chen, and S. Ning, "Identification of river water quality using the fuzzy synthetic evaluation approach," *Journal*

of Environmental Management, vol. 63, no. 3, pp. 293–305, 2001.

[15] Y. Xu, J. F. Yeung, A. Chan, D. W. M. Chan, S. Q. Wang, and Y. Ke, "Developing a risk assessment model for PPP projects in China – a fuzzy synthetic evaluation approach," *Automation in Construction*, vol. 19, no. 7, pp. 929–943, 2010.

[16] Y. Wu, W. Jia, L. Li, Z. Song, C. Xu, and F. Liu, "Risk assessment of electric vehicle supply chain based on fuzzy synthetic evaluation," *Energy*, vol. 182, pp. 397–411, 2019.

[17] S. Ni, "Abnormal-aware multi-person evaluation system with improved fuzzy weighting," 2022, http://arxiv.org/abs/2205.00388.

[18] S. M. Stigler, "The asymptotic distribution of the trimmed mean," *The Annals of Statistics*, vol. 1, no. 3, pp. 472–477, 1973.

[19] C. F. Kossack and M. G. Kendall, "Rank correlation methods," *American Mathematical Monthly*, vol. 57, p. 425, 1950.

[20] G. O. Engin, I. Demir, and H. Hiz, "Assessment of urban air quality in Istanbul using fuzzy synthetic evaluation," *Atmospheric Environment*, vol. 38, no. 23, pp. 3809–3815, 2004.

[21] R. S. Lu, S. L. Lo, and J. Y. Hu, "Analysis of reservoir water quality using fuzzy synthetic evaluation," *Stochastic Environmental Research and Risk Assessment*, vol. 13, no. 5, pp. 327–336, 1999.

[22] T. H. Cormen, C. E. Leiserson, R. L. Rivest, and C. Stein, *Introduction to Algorithms*, MIT Press, 3rd edition, 2009.