Scientific Research Publishing

Open Journal of Social Sciences, 2023, 11, 47-66
https://www.scirp.org/journal/jss
ISSN Online: 2327-5960
ISSN Print: 2327-5952

Research on the Causes and Correction Strategies of Group Peer Assessment Performance Bias in Online Collaboration

Xiulin Ma, Shumin Tian, Shijing Luo, Xue Jiang

Faculty of Education, Beijing Normal University, Beijing, China
Email: maxl@bnu.edu.cn, t18139869165@163.com, 202111010027@mail.bnu.edu.cn, jx19981115@163.com

How to cite this paper: Ma, X. L., Tian, S. M., Luo, S. J., & Jiang, X. (2023). Research on the Causes and Correction Strategies of Group Peer Assessment Performance Bias in Online Collaboration. *Open Journal of Social Sciences, 11,* 47-66.
https://doi.org/10.4236/jss.2023.117005

Received: June 10, 2023
Accepted: July 9, 2023
Published: July 12, 2023

Abstract

Group peer assessment is essential to online collaborative learning. Quality group peer assessment can effectively promote collaborative knowledge construction and maximize the quality of collaborative group learning. However, in concrete teaching practice, group peer assessment can suddenly reveal serious biases that affect learners' motivation. In organizing online learning in a collaborative group, the author's sudden mutual evaluation performance bias directly affected educational equity and seriously demotivated most students. In this paper, we analyzed and corrected the assessment data based on the quality assessment strategy and the Z-score standardization technique and explored the factors that led to the narrow assessment results to establish an effective organization strategy for online collaborative learning assessment activities, a quality assessment system for the assessment data and a correction strategy to avoid the deviation of mutual assessment results in online collaboration and to solve the motivation decay caused by the deviation of mutual evaluation.

Keywords

Collaborative Learning, Group Peer Assessment, Performance Deviation, Correction of Deviation

1. Introduction

1.1. Online Collaborative Learning Is an Important Guarantee of the Quality of Online Education

With the continuous development and popularization of internet technology, online education has advantages such as resource sharing, personalized teaching,

improving learning efficiency, and reducing educational costs. Among them, online collaborative learning is a process of knowledge construction through negotiation and interaction under the guidance of teachers, with the Internet as the medium and the technology platform as the primary tool for communication and resource sharing, which can promote students' collaborative cognition, develop collaborative skills and mutual emotions, and promote optimal learning performance (Peng, 2010). It has become a meaningful way to cultivate collaboration literacy in the "5C model of 21st-century core literacy" (Xu et al., 2020). As a pedagogy centered on group discussion, online collaborative learning can remedy the problems of insufficient sense of belonging and low persistence of learning expected in asynchronous online learning environments and guarantee high-quality development of online education.

At present, the focus of online collaborative learning research has shifted from the effectiveness of online collaborative learning, the analysis and evaluation of the interaction process in online collaborative learning to the focus on group meaning construction and the evaluation of collaborative learning effects (Chai, & Li, 2010). Exploring the effectiveness of collaborative learning is the basis for further designing, organizing, and carrying out the process of group collaboration, further feeding back the quality of collaborative group interactions and improving the quality of online collaboration.

Traditional online evaluation methods are usually conducted by teachers, while students lack opportunities for participation and feedback. Group mutual evaluation, as a way to evaluate the effectiveness of online collaborative learning, has the function of measuring the level, value or quality of other groups' work and students' contribution to that work (Zhao & Li, 2000), which can solve the problem of single evaluation subject and evaluation form in online learning. In the online collaborative learning environment, more pedagogical exploration and empirical support are needed to further exploit the advantages of group peer assessment, cultivate learners' critical thinking, enhance learners' sense of social presence, and improve the effectiveness of online collaborative learning.

1.2. The Emergence of Deviation in the Performance of Group Peer Assessment

In September 2022, in a hands-on online teaching activity for an information technology class organized by the author's team, we organized a collaborative learning activity for students with the learning strategy of group collaboration. The activity divided all students into eight groups, with 6 - 8 students in each group. It required students to collaborate in groups to complete open-ended tasks with specific requirements for different themes (image processing, video production, animation, web development). Once groups were divided, the instructor provided each student with the collaborative tasks, requirements, and other activity support services through the CEN platform. After receiving the tasks, students discussed online, shared resources, and assigned tasks with group

members regarding the difficulty and completion of the tasks. Since group members can see each other's interactions and participation on the platform, it will invariably urge each member and, to some extent, reduce group members' feelings of isolation and negativity. To give full play to group members' agency and avoid undesirable problems such as hitchhiking, the instructor developed strict management and monitoring measures for the group collaboration process to encourage active intra-group knowledge sharing and healthy inter-group competition. To ensure the diversity and fairness of evaluation, the instructor also used a combination of evaluation methods such as student self-assessment, group peer assessment, teacher evaluation, intra-group evaluation, group leader evaluation, process assessment, and summative evaluation.

To promote group collaboration and progress, the instructor required an online debate and critique once a week for each learning theme, with each group's debriefer presenting the group's work. After the debrief, the group needed to accept inquiries from other groups. In addition to the debriefer, the questioner can choose any one of the debriefing groups to answer the questions. If the debriefing group cannot answer the questions comprehensively, points will be deducted from the other groups, which increases the student's sense of social and interactive presence. At the end of the review activity, each group of students was required to fill out a group peer assessment form. Based on the data recording function in the platform, the instructor summarized each group's evaluation form and calculated the arithmetic mean of each group's score as the basis for ranking the group's mutual evaluation score.

In terms of teaching practice, the intergroup discussion in this session promoted diversified communication and sharing among students, solved the shortage of simple group discussion, and gave full play to the learning initiative of each group member through the comprehensive evaluation of excellent works. It also enabled students to avoid their weaknesses in subsequent activities. However, during the teaching practice, the author found a very peculiar phenomenon: among the eight collaborative groups, group 3's work did not stand out, and the group members were often stuck when answering the queries of other groups.

Therefore, both the author and the teaching assistant thought that the performance of group 3 could have been worse. However, the three rounds of collaborative learning revealed that the arithmetic mean of group 3 was consistently higher, and its ranking was consistently at the top of the class. However, the data in the table found that the group received terrible grades in each evaluation. What could have caused this phenomenon?

1.3. Deviation in Group Peer Assessment Scores Affects Educational Equity and the Motivation of Other Groups to Learn

The deviation of the group peer assessment was outstanding and not only alerted the author and the teaching assistant but also attracted the attention of the stu-

dents: After the two rounds of online collaborative learning and group peer assessment results were announced, some students privately communicated with the author that they questioned the ability of group 3 to obtain such an excellent comprehensive assessment and hoped that the instructor could find out the reason why group 3 could obtain a high rating despite its poor work and poor reporting.

After comparing the data, the author found that: during the group peer assessment process, Group 3 gave the other groups generally harsher scores, and the scores given by Group 3 to the other groups differed from the other groups by about 8 points in mean value compared to the scores given by the other groups. For this reason, the other groups' final mean scores fell, which led to Group 3 moving up in the rankings.

The students felt that this deviation seriously compromised educational equity, could harm hardworking students, and was detrimental to the sustainability of collaborative learning.

1.4. Research Problem

Based on the phenomenon of deviations in group peer assessment in online learning, the author believes that we should start from three aspects: teaching process management, evaluation of the quality of mutual evaluation data, and correction of deviation to solve the problem of performance deviation in group peer assessment.

1) How to evaluate the quality of group peer assessment data to promptly identify problems in group peer assessment?

2) What strategies should be adopted to correct the bias in the already problematic group peer assessment data to avoid the frustration of learning motivation due to the evaluation bias?

3) What effective strategies should instructors adopt to avoid the deviation of teaching evaluation from the source for online collaborative learning?

2. Literature Review

2.1. Evaluation of Collaborative Learning

Collaborative learning assessment is an essential means of measuring individual learning outcomes and the learning performance of collaborative groups. The effectiveness of collaborative learning needs to be judged by collaborative learning assessment, and the assessment feedback of collaborative learning can effectively motivate students to participate, so collaborative learning assessment is critical to promoting further development of online collaborative learning. The assessment orientation of collaborative learning can be divided into formative and summative assessments. Formative assessment focuses on various elements of the collaborative process, including learners' knowledge, skills, and emotions. The summative assessment focuses on students' cognitive outcomes, and the formative elements of the collaborative process are not included in the assessment.

The assessment methods for collaborative learning vary at different learning times. Before collaborative learning, Questionnaires are usually used; during collaboration, Textual Analysis, Group Collaborative Learning Profiles, Social Network Analysis, and Content Analysis are generally used; after collaboration is completed, Outcome Assessment, Analytic Hierarchy Process, joint group testing, Weighted Sum Method, and Weighted Average Method are usually used (Yu & Zheng, 2015).

Currently, collaborative learning assessment focuses on the deep-level interactions of students in the collaborative process, such as knowledge construction, level of interaction, and metacognition. However, research on the quality assessment of collaborative learning has yet to be in-depth. There needs to be more research on the relationship between group peer assessment and learning feedback and how to improve the quality of learners' feedback.

2.2. A Study on the Effectiveness of Peer Assessment

Group peer assessment, generally referred to as peer evaluation, also known as peer assessment and peer feedback (Kate, 1992), is the process by which learners evaluate the learning work and outcomes of other peers with the same learning background (Topping, 1998), grading and commenting are the main ways in which peer assessment is achieved. The primary purpose of peer assessment is to provide feedback to learners, which is more direct, effective, and personalized than teacher feedback (Nicol et al., 2014). Peer assessment also promotes collaborative interaction and deep knowledge construction among students through the dialogue process of assessment and feedback (Xu & Zhu, 2022). Interactive behaviors such as arguing and questioning during mutual assessment can also promote the development of students' reflection and critical thinking (Zhang et al., 2022). When students conduct multiple evaluations, the gap in assessment ability between students and teachers gradually narrows. Students gain experience viewing other works from the evaluator's perspective and distinguishing between high and low-quality works (Seifert & Feliks, 2019). Moreover, collaborative peer assessment is more accurate than self-assessment, increasing accuracy with the number of assessments (Rico-Juan et al., 2022). In addition, peer assessment based on assessment scaffolding can also improve assessment consistency and rubric quality and increase learning effectiveness and learner recognition (Ma et al., 2022). However, some studies have shown that in peer assessment, students are often not accustomed to switching their roles from that of the evaluated to that of the evaluator, often unsure of their own or others' ability to assess, as well as the accuracy of feedback they provide or receive from others (Vu & Dall'Alba, 2007).

In summary, when learners evaluate and reflect on peer works, they can comprehensively play the role of "guiding, identifying, diagnosing, regulating, and improving", achieving the learning function of assessments (Cai et al., 2021). However, how to organically combine group peer assessment with online colla-

borative learning and improve the reliability and validity of group peer assessment still need to be explored in depth.

2.3. The Impact of Collaborative Learning Evaluation on the Quality of Collaboration

2.3.1. Online Collaborative Learning Theory

The online collaborative learning theory focuses on social learning, communication, and collaboration and argues that the process of collaborative learning includes: 1) generation of ideas, where students express different views through brainstorming; 2) organization of ideas, where cognitive conflicts, opposing views, and interpretations among students trigger deeper thinking about the problem; and 3) intermingling of minds, where students co-construct knowledge by understanding the ideas generated and transforming them into explicit conclusions (Harasim & Xiao, 2015). The three processes are iterative, with discussion and student collaboration facilitating learning. The design of evaluation in online collaborative learning can lead to high-quality collaborative learning, and practical collaborative learning evaluation can stimulate learning behavior with potential motivational value, thus promoting the efficient advancement of collaborative quality.

In other words, learning motivation is one of the critical factors that affect the quality of collaborative learning and has a decisive impact on motivating, maintaining, and strengthening students' learning behavior (Ma et al., 2019). Unjust and unreasonable collaborative evaluation can negatively affect students' participation enthusiasm, emotional state, and innovation ability and make it challenging to maintain students' motivation, which can lead to "social loafing" and other undesirable problems in the collaborative learning process, resulting in the reduction of group performance and knowledge acquisition.

Therefore, comprehensive and scientific evaluation is an intrinsic motivation for collaborative learning activities and a key measure to improve the quality of education and teaching.

2.3.2. Group Dynamics Theory

Group dynamics theory places great emphasis on the primacy of democratic leadership, emphasizes the criticality of the participation of members within the team in decision-making and collaborative atmosphere, and states that individual behavior is influenced by both internal needs and external environmental forces, while interactions between individuals have a duality (Xie et al., 2009). Many factors, such as team members' values, norms, group cohesion, and role task, affect group activities' effectiveness and individual and group development (Liao & Zhuang, 2005).

Group interaction in collaborative learning occurs through sharing, communication, discussion, and interaction between teachers and students and students and students. The interaction of various factors inside and outside classroom teaching will be transformed into individual and group dynamics, thus promot-

ing the development of collaborative knowledge construction and learning socialization. Assessing the performance of collaborative learning can effectively eliminate group-slacking behaviors, help students develop a sense of identification with the group, and form interdependent relationships within and outside the group. In turn, it stimulates collaborative group dynamics, motivates team members to create more team cooperation behaviors under the incentive of the collaborative group, creates a democratic and harmonious evaluation atmosphere, and maximizes the quality of collaborative group learning.

3. Exploration and Correction Strategies for the Deviation of Group Peer Assessment Results

To circumvent the influence of missing values, the author normalized the collected raw data. On this basis, the validity of the current round of group peer assessment was discerned, and how the evaluation quality of the mutual evaluation data and the reasons for the problematic data quality were analyzed. Then the standardized processing technique was used to correct the deviated scores to reflect the group peer assessment results objectively.

3.1. The Raw Data and Standardized Processing of Group Peer Assessment Results

3.1.1. Raw Results of Group Peer Assessment

After the third round of collaborative activities and group peer assessment, the instructor organized and collected the evaluation forms of each evaluation group and compiled the scores given and scored by each group, as shown in Table 1.

In Table 1, the initial scores given, scores received, and means for each group were shown, where each row refers to the group's score received, and each column is the score given by the group to the other groups. For example, the "Group 1" row (i.e., the second row) is the score that Group 1 received from the other evaluation groups, while the "First group" column (i.e., the second column) is the score given by the first group to other reporting groups.

Table 1. Raw scores of each group's mutual evaluation.

Group Number	First group	Second group	Third group	Fourth group	Fifth group	Sixth group	Seventh group	Eighth group	Mean
Group 1		83.5	63.5	80.5	84.2	84.8	71.5	79.5	78.2
Group 2	81.5		75.5	80.5	82.8	85	63.9	76.5	78
Group 3	81.8	82.5		81.5	81.5	80.5	72.5	82.5	80.4
Group 4	80.5	79.5	65.5		77.5	84	62	74.5	74.8
Group 5	81.5	89	62.5	79.8		89.5	69	82.5	79.1
Group 6	90.5	91.5	68.5	85.5	88.5		77.5	88.7	84.4
Group 7	83.5	90.8	70.5	83.2	82.9	89.8		80.3	83
Group 8	81.5	89	64.3	79.8	79.7	87.3	66.5		78.3

Since there was a separate group self-assessment session, each group did not participate in the evaluation of their group during the mutual evaluation session, Hence, a blank data band from the top left to the bottom right corner, i.e., a series of missing values, appears in Table 1.

3.1.2. Standardized Handling of Group Mutual Assessment

As can be seen from Table 1, missing values are present in every row and every column. If the case with the missing value is directly prohibited from participating in the subsequent data analysis process, then there will be no valid data to analyze. To solve this problem, the author decided to fill the unique missing value in each column with the mean value of the data in that column so that the missing value is no longer "missing". In other words, this paper will use the mean value of all scores given by a specific group as the self-assessment value of the group. For example, in Table 1, row 1 and column 1's positions have a null value, which will be replaced by the mean value of the other 7 data in column 1. The normalized data is shown in Table 2.

For the standardized data, the author did a single sample K-S test on all eight columns of data. The results confirmed that: the Sig values of all eight columns of data were more significant than 0.05, indicating that the data being tested obeyed a normal distribution. The distribution of these data was still excellent.

3.2. Quality Analysis of Group Peer Assessment Results

The quality analysis of the group peer assessment results can be carried out from two perspectives: firstly, the differentiation of group scores, which is the core indicator to determine the validity of mutual evaluation results. If the evaluation cannot clearly distinguish the ranking of the groups, the evaluation is invalid. The second is to check whether there is a strong consistency in the scores given by the evaluators. If the scores given by a specific evaluator are significantly inconsistent with other evaluators, there may be a significant deviation in the score given by this evaluator. Therefore the score given by this evaluator should be excluded or weighted.

Table 2. Normalized raw data.

Group Number	First group	Second group	Third group	Fourth group	Fifth group	Sixth group	Seventh group	Eighth group
Group 1	83.0	83.5	63.5	80.5	84.2	84.8	71.5	79.5
Group 2	81.5	86.5	75.5	80.5	82.8	85	63.9	76.5
Group 3	81.8	82.5	67.2	81.5	81.5	80.5	72.5	82.5
Group 4	80.5	79.5	65.5	81.5	77.5	84	62	74.5
Group 5	81.5	89	62.5	79.8	82.4	89.5	69	82.5
Group 6	90.5	91.5	68.5	85.5	88.5	85.8	77.5	88.7
Group 7	83.5	90.8	70.5	83.2	82.9	89.8	69.0	80.3
Group 8	81.5	89	64.3	79.8	79.7	87.3	66.5	80.6

3.2.1. Differentiation Analysis of Group Scores

This paper aims to explore the differentiation level of each group's scores, testing whether there is a significant difference in the overall distribution of the rank scores of the groups' scores with the help of the Friedman and Kendall W coefficient algorithm, a nonparametric test for multiple correlated samples. Based on the Friedman test, the author found that: its chi-square distance was 24.669 and the significance probability of Sig = 0.001, which was much less than 0.05. Therefore, there was a significant difference in the overall distribution of the rank mean of the eight debriefing groups. In addition, the Kendall W value of 0.441 for each debriefing group also indicated that the rank means of the groups' scores were significantly different, i.e., the scores of different groups were distinguishable.

In addition, the rank mean scores of each group after the Friedman test are shown in Table 3.

Observing the rank means of each group's scores reveals that the rank means were evenly distributed in the range of 1 to 7, indicating that the groups' scores were well differentiated. Therefore, the scores of this round of group peer assessment should be valid.

3.2.2. Analysis of Evaluator-Oriented Evaluation Quality

This paper aims to test whether the given scores of the groups have strong consistency in the overall distribution, conducted a non-parametric test based on Kendall W using the given scores of the eight groups as the test variable. The test results revealed that the test probability, Sig = 0.000, was much less than 0.05, indicating a significant difference between the given scores. Thus, the eight groups were not only partially consistent in the overall distribution of given scores. At least one group was given a significantly different score from the other evaluation groups.

The rank means of the scores given by the eight groups were observed, and their data are shown in Table 4.

As can be seen from the rank means presented in Table 4, among the eight

Table 3. Calculation results of Friedman's non-parametric test for scores.

	Group 1	Group 2	Group 3	Group 4	Group 5	Group 6	Group 7	Group 8
Rank Mean	3.94	3.81	5.06	1.75	4.06	6.88	6.75	3.75

Table 4. Distribution of rank mean after non-parametric test with Kendall W algorithm.

	Group 1	Group 2	Group 3	Group 4	Group 5	Group 6	Group 7	Group 8
Rank Mean	5.63	6.89	1.38	4.31	4.81	6.50	1.63	4.56

evaluation groups, the rank mean values given by the third and seventh groups are significantly lower than those of the other six evaluation groups, indicating that the third and seventh groups may have given scores that differed significantly from the other evaluation groups.

After excluding the data from the third and seventh groups, the test of significance of differences based on the "Kendall W" algorithm was performed again. The results showed that the scores given by the other six evaluation groups had Sig = 0.087 > 0.05, indicating that the rank of the scores given by these six evaluation groups was consistent in the overall distribution, and there was no significant difference. Moreover, their Kendall W coefficients were also relatively small, indicating that the rank means of the scores given by these six evaluation groups have a good consistency.

In conclusion, the author believes that: during the mutual evaluation stage of this round of collaborative learning, there was a specific deviation in the scoring of the third and seventh groups, which was the main reason why the quality of group evaluation was questioned.

3.3. Analysis of the Reasons for the Deviation of Group Peer Assessment Results

As can be seen from Table 4, the direct scores of the debriefing group cannot truly reflect the actual level of each group because there are significant differences between the scores given by some groups and the scores given by other groups in the overall distribution, which have already led to the deviation of the scores of the group peer assessment. In other words, it would be unreasonable for each group and the mutual evaluation activity if the scores given by the evaluation group were directly used as the scores of the group peer assessment for ranking.

Further exploring the data in Table 4, the author found that: since the rank mean of the scores given by the third and seventh groups was much lower than the other groups, this indicates that the third and seventh groups gave the other groups a low score in the mutual evaluation stage, which directly led to a lower rank mean of the scores of the other groups. Since the groups did not do self-assessment, then under the same conditions, it would result in the third and seventh groups having a lower score than the other groups, which would make the mean values of these two groups increase and eventually lead to a severe inconsistency in the scores given by the evaluation groups. It is the fundamental reason that leads to a lower reliability of group peer assessment scores and cannot objectively and accurately reflect the actual level of the group.

In summary, if a group generally gives lower scores to other groups in the group peer assessment, it will cause the mean score of other groups to decrease and its mean score to increase. Conversely, if a group generally gives high scores to other groups, it will decrease the mean value of its score and a general increase in the mean value of other groups' scores. Therefore, to ensure the reliability and fairness of the evaluation results, instructors must adopt specific

strategies to correct the already problematic group peer assessment data to avoid frustrating the motivation of some excellent students due to evaluation deviation.

3.4. Correction of Group Peer Assessment Results

3.4.1. Standardized Processing of Group Peer Assessment Scores

In this paper, we use the "score normalization" algorithm to deform the data to solve the deviation of mutual evaluation caused by the inconsistent scoring criteria. Therefore, this paper introduces the standardization method of Z-Score to correct the raw group peer assessment scores that have problems. The equation is:

$$z = \frac{x - \mu}{\delta} \tag{1}$$

Equation (1) measures the standard deviation of the initial value from the mean in units of standard deviation, where z is the standard score, x is a specific value in the data, μ represents the mean of the overall data, and δ represents the standard deviation of the overall data.

In this paper, x is the raw score of the group peer assessment, μ is the mean of each group's score, δ is the standard deviation of each group's score, and z is the standard score. The raw scores, mean value, and standard deviation were substituted into the Equation (1) to obtain the standardized scores for each of the following groups (see Table 5).

The standardized scores were again tested for the difference significance by the "Kendall W" algorithm. It found that after standardization, the Sig values of the differences between the eight evaluation groups were significantly greater than 0.05, indicating no significant difference in the overall distribution of the rank of the scores given by the eight evaluation groups. Further discovery based on rank mean values that the scores given to each evaluation group were relatively consistent. In addition, its Kendall W coefficient is also minimal, indicating that the evaluation groups have high consistency in the overall distribution

Table 5. Standard scores of evaluation data for each group.

Group Number	First group	Second group	Third group	Fourth group	Fifth group	Sixth group	Seventh group	Eighth group	Mean
Group 1		−0.65	−0.80	−0.49	0.50	−0.32	0.47	−0.25	−0.22
Group 2	−0.43		1.80	−0.49	0.10	−0.26	−0.95	−0.90	−0.16
Group 3	−0.34	−0.87		−0.02	−0.27	−1.62	0.66	0.40	−0.29
Group 4	−0.72	−1.51	−0.36		−1.42	−0.56	−1.30	−1.33	−1.03
Group 5	−0.43	0.53	−1.01	−0.83		1.11	0.00	0.40	−0.03
Group 6	2.19	1.07	0.28	1.88	1.74		1.59	1.74	1.50
Group 7	0.15	0.91	0.72	0.79	0.13	1.20		−0.07	0.55
Group 8	−0.43	0.53	−0.62	−0.83	−0.79	0.44	−0.46		−0.31

of the rank given to the scores of different debriefing groups. That is, the standardized processed scores can accurately reflect the actual scores of the groups' mutual assessment to correct the deviation in the scores given by each group as a basis for further ranking.

3.4.2. Final Score and Ranking of Group Peer Assessment

Based on the standardized and processed mutual evaluation data of each evaluation group, the mean score of each debriefing group was obtained and ranked accordingly, as shown in Table 6.

The corrected means were significantly different from the rank means of their raw scores, and the change in ranking occurred in the second and third groups, with the second group giving too high a rating to the other groups and the third group giving too low a rating to the other groups. Therefore, after the correction, the second group's ranking increased while the third group's ranking decreased.

In addition, the data quality analysis from the perspective of the "evaluator" revealed that the "Kendall W" harmony coefficient for the seventh group was very low, and its rank mean value differed significantly from the other groups. However, after correction, it was found that the ranking did not change. After tracking the data, it was found that the evaluation scores obtained were generally high because of the high quality of the works of the seventh group.

Although the ranking of other groups may decrease due to the low score given by this group to other groups, the actual scores of other groups will not threaten the ranking of the seventh group. Therefore, after correction, although the mean value of the other groups has increased, the ranking of the seventh group was still higher than that of the other groups.

3.4.3. Summary

The problem of score deviation in group evaluation was effectively solved by introducing the Z-Score standardization method. After Z-Score normalization, about half of the scores will be less than 0 and the other half more than 0, with a mean score of 0 and a standard deviation of 1, removing the differences in characteristic attributes between different scores. Therefore, in statistics, people often transform the scores given by several evaluators into relative values with the same scale, known as standard scores, and then use the standard scores as the final benchmark for comparison and weighting, achieving effective corrective

Table 6. Final mean and ranking of each group score.

Group Number	Group 1	Group 2	Group 3	Group 4	Group 5	Group 6	Group 7	Group 8
Score rank mean	4.19	3.94	4.31	2.13	4.19	7.38	6.25	3.63
Standardized mean	−0.22	−0.16	−0.29	−1.03	−0.03	1.5	0.55	−0.31
ranking	5	4	6	8	3	1	2	7

measures for performance deviation problems.

4. Discussion and Conclusion

To ensure the objectivity of the group evaluation and the enthusiasm of the cooperation of the group, this paper discusses the factors with unreasonable group peer assessment results and builds an organizational strategy for collaborative learning and mutual evaluation activities, the quality evaluation system of mutual evaluation data, and correction strategies to avoid the impact of undesirable factors in mutual evaluation activities and solve the problem of deviation in mutual evaluation results.

4.1. Key Factors Leading to Unreasonable Group Peer Assessment Results

4.1.1. The Effect of Reciprocity

The reciprocity effect is a common phenomenon in online collaborative mutual evaluation. Group peer assessment is a social, collaborative learning activity, and learning groups are prone to give each other group unequal scores for their contributions due to interpersonal relationships (Zhang et al., 2019). The reciprocity effect may lead to a convergence of scores among groups. If each group negotiates and gives the other a consistent score, there will be no significant difference between the average score of each group and the average score given, resulting in no differentiation in the ranking order of each group. In addition, the reciprocity effect may lead to a specific group giving too high a score, which may negatively affect the fairness of the group peer assessment and discourage inter-group competition.

4.1.2. Influence of Individual Difference Factors of Evaluators

The evaluators' cognitive styles, aesthetic attitudes, and other factors may also give unreasonable scores to the work of the debriefing group. When evaluating each other's works, due to the differences in the cognitive styles of different evaluation groups, the dimensions they focus on may also differ, with some evaluation groups tending to focus on the technical details and complexity of the works. In contrast, some evaluation groups focus on the overall evaluation of works. In addition, the preference of the evaluation group's aesthetic attitude also affects the evaluation results; in general, the evaluations tend to give higher scores on their aesthetic or similar design styles (Yin et al., 2012).

4.1.3. Evaluators Fail to Master the Evaluation Scale Well

At different times of evaluation, different evaluators may have their understanding of the evaluation criteria, resulting in an overly lenient or strict evaluation scale. In addition, at the beginning of the evaluation, the evaluation group as a whole was not yet familiar with the evaluation activities and had a relatively superficial understanding of the works, so it was also possible that a unified evaluation scale could not be formed, resulting in the evaluation scores of the first few debriefing groups fluctuating wildly. However, as the number of subsequent evaluations

increased, the evaluation experience and evaluation ability of the evaluation group stabilized, more objective evaluation guidelines were formed within the evaluation group, and the scores given were more objective and reasonable.

4.1.4. Impact of Implementation Cost Factors

Implementation costs are reflected in the evaluator's scoring time and emotional factors. Usually, rating a large number of works in a short time will often cause cognitive loads to the evaluator, affecting the effective development of internal discussions, evaluation, and feedback of the evaluation group. In addition, the evaluator's emotional state also affects the grading's reliability. When conducting group peer assessments, instructors should try to create a democratic and harmonious atmosphere for grading to avoid the influence of the evaluator's bad mood on the evaluation of the work.

4.2. Establishing a Rigorous Organizational Strategy for Online Mutual Evaluation to Avoid Unreasonable Results in Mutual Evaluation

Based on the problems in the collaborative assessment activities, instructors should explore and build the control rules and organization system for group peer assessment activities and control the interfering factors affecting mutual assessment activities so that the group peer assessment activities of the following themes can be as fair and reasonable as possible.

4.2.1. Developing Rigorous and Standardized Group Peer Assessment Rules

Rigorous and standardized rules and standardized and detailed evaluation scales are the quality assurance of group peer assessment. Accordingly, instructors should formulate clear and precise scoring criteria for participating groups, specifying the score range of each sub-item in detail and what standards to achieve what scores can be obtained. Make each student's score "evidence-based and evidence-supported".

In addition, teachers should develop a professional group peer assessment form that includes scoring information and open-ended questions to check the gains and growth of the group members during the debriefing phase. The open-ended questions aim to record the evaluation and expectations of the evaluation group towards the debriefing group, enhance the student's sense of gain in filling out the form, and discipline their evaluation behavior. With the help of open-ended questions, other students are helped to gain insight into their problems and improve their work in a targeted manner to achieve group knowledge construction and individual knowledge innovation.

All in all, inter-group evaluation should not be a formality; students should think critically about their work and point out the successes and improvements of their work. When students explore the advantages and causes of improvement suggestions, they can gradually form a scientific thinking process, laying the foundation for avoiding or achieving this effect next time.

Based on the problems in the group peer assessment, the author has improved

the group peer assessment form in the teaching of public computer courses at the Beijing Normal University, forming a group peer assessment form as shown in Table 7, which has effectively promoted the quality of knowledge sharing and individual growth in the group peer assessment stage.

4.2.2. Establishing a Rigorous Organizational System for Group Peer Assessment Activities

1) Adopting Second-Order Grouping Strategy

In group peer assessment, good grouping can make students reach the zone of proximal development faster, so this paper adopts the second-order grouping method of the jigsaw model (Huang & Fu, 2010). The first order is intra-group heterogeneous and inter-group homogeneous; the second order is intra-group homogeneous and inter-group heterogeneous. Such a grouping method can make lower-level students get help from high-level peers in the first-order group. The second-order homogeneous grouping mowing can prevent high-achieving students from getting nothing or low-achieving students from having no voice in the heterogeneous groups. Accordingly, in theory, all students can find their direction, gain motivation in different groups, and expand the area and channels of information exchange to achieve information sharing.

2) Designing Evaluation Scaffolds

Evaluation scaffolding is a tool to support students in clarifying and evaluating the evaluation contents (Fen et al., 2008), which can reflect the teaching team's requirements on the quality of the work and is also beneficial for students to use what they have learned to launch a structured evaluation of others' work. At the same time, learners' precise understanding of evaluation scaffolding is also an essential prerequisite for achieving high-quality evaluation feedback. Therefore, in the subsequent group peer assessment session, teachers need to involve students in the process of designing evaluation scaffolds based on the characteristics and requirements of the curriculum theme, which not only helps students

Table 7. Group peer assessment form.

Debriefing Group Title:		
Information of the filling group:	Group leader name: Member name: Scores given:	group number:
What new knowledge, methods, or techniques did you learn from the debriefing group's debriefing?		
What do you think was the greatest success of the debriefing group's work?		
What do you think was unsuccessful or needs improvement in the debriefing group's work?		
What questions would you like to ask?		

clarify their evaluation ideas and make their comments more focused but also enables them to understand the specific requirements and evaluation dimensions of the group tasks and clarify the evaluation levels corresponding to their participation and engagement in the collaborative activities. After designing the evaluation scaffold, instructors need to use the scaffold to guide students to study previous cases of group peer assessment and strengthen the exercise of students' evaluation skills to improve the validity of group peer assessment.

3) Using One-way Anonymous Evaluation

The one-way anonymous evaluation strategy is used in group peer assessment, where the introduction of the debriefing group's work is public. In contrast, the evaluation results of the evaluation group are anonymous. The debriefing group does not know the identity of the evaluation group, which can balance the reciprocal effect of the interpersonal relationship of the evaluation group in the mutual evaluation, make the evaluation group have a positive and pleasant peer-reinforced educational experience and help the debriefing group provide more direct and authentic feedback. Of course, the anonymity of evaluation results is only for the debriefing group, and instructors can obtain the real identity of the evaluation group and judge the credibility of the evaluation results, avoiding the problems that may arise from anonymous evaluation, such as the evaluation process not being rigorous and the evaluation results being unreliable.

4) Standardizing the Quality of Mutual Evaluation Comments

The quality of the comments can be defined as the degree of agreement between the comments and the evaluation criteria (Wang et al., 2019). Quality comments not only improve the debriefing group's understanding of the evaluation results and improve the quality of the learning work; they also regulate and monitor the evaluation group and reduce the arbitrariness of the evaluation. Therefore, teachers can instruct the evaluation group to write targeted comments based on the evaluation criteria, which reflects the evaluation group's thinking and facilitates the debriefing group's understanding and improvement. Suppose the comments the evaluation group gave are not linked to the evaluation criteria. In that case, it means that the evaluation group may have yet to attach importance to the evaluation criteria, and the quality of the comments is not high, affecting the debriefing group's adoption of the comments.

5) Designing a "Reflection-Improvement" Session

After a round of group peer assessment, instructors can set aside time for the evaluation group to review the scores of all debriefing groups to avoid the influence of interfering factors at different evaluation times and to correct inappropriate results in time.

4.3. Establishing Effective Quality Evaluation and Correction Strategies for Mutual Evaluation

For the existing group peer assessment data, the author believes that the quality evaluation system of the mutual evaluation data should be constructed from two

dimensions: analysis and calibration, and corresponding correction strategies should be formulated.

Establishing a Quality Evaluation System for Mutual Evaluation Data

The quality evaluation system of mutual evaluation data is mainly to complete the standardized processing of the raw data. Based on this, evaluate the group's peer assessment data quality.

1) Scientific Evaluation of the Differentiation of Mutual Evaluation Scores

Since each evaluation group will rate the work of other debriefing groups, the rank scores can be obtained based on the rating sequence given to the debriefing groups by different evaluation groups. The average rank of each debriefing group can be obtained accordingly. If there is a significant difference in the overall distribution of the rank scores of each debriefing group, the rank scores given by the evaluation group to each debriefing group are relatively consistent, i.e., there is a high degree of differentiation among the rank scores of each debriefing group, thus proving that the evaluation group's scores are objective and valid. Otherwise, if there is no significant difference in the overall distribution of the debriefing groups' rank scores, it is impossible to distinguish the strengths and weaknesses of the debriefing groups' work. This round of group peer assessment is invalid. The differentiation level of the rank scores of each debriefing group can also be further explored based on the distribution of the rank mean and the magnitude of the Kendall W coefficient.

2) Conduct a Meta-Evaluation of Each Evaluator's Evaluation Quality

The core idea of analyzing the quality of evaluators' evaluations is to verify whether there is a strong consistency in the overall distribution of the scores given by the evaluators and then analyze whether the scores given by specific evaluators are biased. Suppose there is no significant difference in the overall distribution of the rank of the scores given by each evaluation group, which means that the scores given by the evaluation group to the reporting group have consistency in the overall distribution. Then the mutual evaluation data of this round of groups is scientifically valid, and the mutual evaluation scores can accurately reflect the actual scores of the group's peer assessment. Suppose there is a significant difference in the overall rank of the scores given by each evaluation group, which indicates that at least 1 group of evaluation groups has given scores significantly different from other evaluation groups, i.e., in that case, there is some problem with the scores given by the evaluators. Then it is necessary to identify the groups that give deviated scores based on the distribution of the rank mean. Furthermore, by eliminating the scores given by the deviating evaluators, the evaluation scores can reach a more objective and consistent level.

3) Establish a Correction System Based on Standardized Scores

In the raw scoring, there may be deviations in the evaluation scales of different evaluation groups, resulting in different score values for each debriefing group. Therefore, for groups that have yet to undergo quality evaluation, the actual performance of each group should not be directly based on the sum of the

evaluation scores of other groups. In other words, it may not be reasonable to use the cumulative raw scores of different evaluation groups to calculate their scores in the group peer assessment. Because of this, for the group peer assessment in collaborative learning, in addition to establishing the necessary mutual evaluation quality evaluation system, we should also establish the standard of "mutual evaluation performance correction" based on standardized scores. Because the standard scores of different evaluation groups have the same mean and range, these data are additive. When the raw score is converted into a standard score, its distribution pattern does not change. However, the raw score is placed in a relative position among all scores, providing a unified reference value and the same unit, shielding the differences caused by different evaluation scales. Therefore, the cumulative sum of standardized scores can be used to represent the final scores of each group.

In conclusion, the normalization of data based on standardized scores is still effective in concealing the deviation of group peer assessment due to evaluation scales and cognitive biases. Therefore, using standardized scores overcomes the shortcomings of the traditional group peer assessment in which evaluators have different evaluation scales and is an effective method of correcting biased scores, which has a specific promotion value.

4.4. Research Limitations

In this paper, the problem of performance deviation in group peer assessment is circumvented and solved from the top down by starting from three aspects: teaching process management, mutual evaluation data quality evaluation, and deviation correction, and group peer assessment can be conducted more objectively and reasonably. However, in collaborative learning, the participation and completion of group members' tasks are very likely to be unbalanced, and the group peer assessment only gives the group an overall consistent score without providing differentiated scores according to members' individual contribution levels, which is still likely to cause the hitchhiking effect.

Group peer assessment is not the entirety of online collaborative learning evaluation. Teaching is a dynamic and complex system; even a sound evaluation system is only a partial reflection of the learning process, which cannot comprehensively and profoundly describe the learning process. Therefore, the instructor, as the guide of online collaborative learning, instructors are no longer traditional and single knowledge imparters and managers, needs to purposefully select some common or complicated problems in group comments according to the results of group peer assessment, extract the excellent works and high-quality comments shown in group comments, and launch centralized question answering or collective discussion to improve the effectiveness of teaching. At the same time, instructor is also necessary to give full play to students' learning initiative and creativity, provide students with the conditions and environment to evaluate their self-learning effectiveness and the learning outcomes of others, cultivate

and improve their evaluation ability to reduce the error rate of student evaluation, and make them the subject for evaluating learning effectiveness.

Conflicts of Interest

The authors declare no conflicts of interest regarding the publication of this paper.

References

Cai, M. J., Wang, X. Y., Guo, W. R., Li, Y., & Li, M. (2021). A Theoretical and Empirical Research on Online Learners Participation Assessment. *China Educational Technology, No. 3,* 15-23.

Chai, S. M., & Li, K. D. (2010). Research on Collaborative Meaning Making in CSCL from a Dialogue Perspective. *Journal of Distance Education, No. 4,* 19-26.

Fen, X. Y., Zhang, W. Y., & Chen, L. (2008). Research of Application of Scaffolding Education Strategy to Distance Inter-University Collaborative Learning. *Beijing Radio and TV University Journal, No. 1,* 26-30.

Harasim, L., & Xiao, H. J. (2015). Collaborative Learning Theory and Practice—The Fundamental Guarantee of Online Education Quality. *Distance Education in China, No. 8,* 5-16+79.

Huang, J., & Fu, L. (2010). Jigsaw: An Effective Way of Collaborative Learning. *E-Education Research, No. 5,* 98-102.

Kate, M. (1992). Peer Reviews in the ESL Composition Classroom: What Do the Students Think. *ELT Journal, 46,* 274-284. https://doi.org/10.1093/elt/46.3.274

Liao, H. J., & Zhuang, Q. (2005). Application of Group Dynamics in Online Collaborative Learning. *Modern Distance Education, No. 4,* 30-32.

Ma, N., Lu, Y., Guo, J. H., & Liu, C. P. (2022). Research on Effects of Evaluation Scaffolds on Quality of Teachers' Online Peer Assessment. *E-Education Research, No. 2,* 34-41.

Ma, X. L., Liang, j., Li, X. W., & Su, Y. Y. (2019). An Empirical Study on the Effect of Group Perception on Online Collaborative Learning. *E-Education Research, No. 5,* 81-89.

Nicol, D., Thomson, A., & Breslin, C. (2014). Rethinking Feedback Practices in Higher Education: A Peer Review Perspective. *Assessment & Evaluation in Higher Education, 39,* 102-122. https://doi.org/10.1080/02602938.2013.795518

Peng, S. D. (2010). From Face-to-Face Collaborative Learning, and Computer-Supported Collaborative Learning to Blended Collaborative Learning. *E-Education Research, No. 8,* 42-50.

Rico-Juan, J. R., Cachero, C., & Macià, H. (2022). Influence of Individual versus Collaborative Peer Assessment on Score Accuracy and Learning Outcomes in Higher Education: An Empirical Study. *Assessment & Evaluation in Higher Education, 47,* 570-587. https://doi.org/10.1080/02602938.2021.1955090

Seifert, T., & Feliks, O. (2019). Online Self-Assessment and Peer-Assessment as a Tool to Enhance Student-Teachers' Assessment Skills. *Assessment & Evaluation in Higher Education, 44,* 169-185. https://doi.org/10.1080/02602938.2018.1487023

Topping, K. (1998). Peer Assessment between Students in Colleges and Universities. *Review of Educational Research, 68,* 249-276. https://doi.org/10.3102/00346543068003249

Vu, T. T., & Dall'Alba, G. (2007). Students' Experience of Peer Assessment in a Professional Course. *Assessment & Evaluation in Higher Education, 32,* 541-556.

https://doi.org/10.1080/02602930601116896

Wang, Q., Ouyang, J. Y., & Fan, Y. Z. (2019). Study on the Relationship between Reflective Awareness and Learning Outcomes in Peer Assessment of MOOCs. *E-Education Research, 40,* 58-67.

Xie, Y. R., Song, N. Q., & Liu, M. (2009). Exploring the Group Dynamics of Collaborative Knowledge Construction in Online Classrooms. *E-Education Research, No. 2,* 55-58.

Xu, G. X., Wei, R., Liu, J., Li, J. Y., Kang, C. P., Ma, L. H., Gan, Q. L., & Liu, Y. (2020). Collaboration Competence: Part V of the 5Cs Framework for Twenty-First Century Key Competences. *Journal of East China Normal University (Educational Sciences), No. 2,* 83-96.

Xu, W., & Zhu, S. X. (2022). An Empirical Study of Peer Mutual Review on Learners' Knowledge Construction Process—Based on the Epistemic Network Analysis of Time Series. *Modern Education Technology, No. 1,* 44-53.

Yin, B. Y., Liu, J. Q., & Yu, J. M. (2012). Analysis of Factors Influencing Peer Assessment of Electronic Works. *E-Education Research, No. 12,* 58-62.

Yu, J. H., & Zheng, L. Q. (2015). Empirical on the Assessment of Group Performance in Online Collaborative Learning—Based on the Information Flow Approach. *Modern Education Technology, No. 12,* 90-95.

Zhang, H. Y., Chen, M. X., Ma, Z. Q., & Yan, X. J. (2019). Evaluation of Web-Based Collaborative Learning Contribution Based on Self- and Peer-Assessment. *Modern Distance Education Research, No. 2,* 95-102.

Zhang, T., Zhang, S., Gao, Q. Q., & Wang, J. H. (2022). A Study on Promoting Development of Learner's Critical Thinking in Online Peer Assessment. *E-Education Research, No. 6,* 53-60.

Zhao, J. H., & Li, K. D. (2000). Collaborative Learning and Its Collaborative Learning Model. *China Educational Technology, No. 10,* 5-6.