Article

An Empirical Study on the Correlation of Undergraduate Teaching Evaluation

Qin Chen[1], Fangxiao Tang[2] and Xiankun Zhang[3,*]

[1] Teaching Quality Monitoring and Evaluation Center, Tianjin University of Science and Technology, Tianjin, China
[2] Academic Affairs Office, Tianjin University of Science and Technology, Tianjin, China
[3] College of Artificial Intelligence, Tianjin University of Science and Technology, Tianjin, China
* Correspondence: Xiankun Zhang, College of Artificial Intelligence, Tianjin University of Science and Technology, Tianjin, China

Abstract: Teaching evaluation serves as a fundamental approach to enhancing pedagogical quality and fostering continuous improvement in higher education. However, the current paradigms of student evaluation of teaching and supervisor evaluation have long been questioned regarding their empirical validity and reliability. Using an extensive dataset comprising five years of teaching records from a university college, this comprehensive study adopts grouped statistics, correlation analysis, and linear regression models to systematically explore the internal relationships among student academic performance, student evaluation results, and supervisor evaluation metrics. The empirical results reveal no significant linear correlation between these three distinct sets of evaluation data. Instead, only minor local differences and complex non-linear characteristics are observed within stratified samples. Additionally, course types and disciplinary attributes exert markedly heterogeneous effects on evaluation outcomes, demonstrating that correlation patterns vary substantially across different academic courses. Fundamentally, these findings suggest that different evaluators hold distinct perspectives and adopt entirely separate observation dimensions when assessing instructional quality. Consequently, the existing evaluation framework fails to align with the core requirements of Outcome-Based Education (OBE), and thus cannot truly reflect the actual effectiveness of classroom teaching. This research critically identifies structural biases within the diversified teaching evaluation system and provides robust empirical evidence for universities to optimize their evaluation mechanisms, ultimately guiding the development of more scientific, equitable, and effective systems for teaching quality assessment.

Keywords: teaching evaluation; academic performance; correlation analysis; higher education; quality assessment; outcome-based education

Received: 22 May 2026
Revised: 10 July 2026
Accepted: 24 July 2026
Published: 31 July 2026

1. Introduction

Classroom teaching is the core link of talent cultivation in universities, and teaching evaluation constitutes a vital component of the teaching quality assurance system. Currently, most universities adopt a multi-stakeholder evaluation model that includes students, supervisors, and peer instructors. Teaching evaluation aims to continuously enhance teaching quality by identifying existing problems and promoting targeted improvements [1]. However, student evaluation and supervisor evaluation remain controversial in practical implementation. For instance, student evaluation scores tend to be inflated and lack distinction; biased and superficial evaluations are prevalent; evaluation results are susceptible to course difficulty and assessment rigor, making it difficult to objectively reflect instructors' actual teaching proficiency. Supervisory evaluation is constrained by outdated evaluation concepts and mismatched professional expertise, while evaluation standards overemphasize formal classroom requirements and neglect the quality of teaching content and the actual educational effectiveness of the course.

As the concept of Outcome-Based Education (OBE) in engineering education accreditation has been increasingly adopted by universities, teaching evaluation has begun to shift from "evaluating teaching" to "evaluating learning" with its indicator system becoming more professional and refined, and evaluation models have also become increasingly diverse. Theoretically, a scientifically designed evaluation system should accurately reflect course teaching effectiveness, and student grades should correlate strongly and positively with evaluation scores. However, existing research shows that the reliability and validity of evaluation results are still contested, and a policy briefing released by the Northern Illinois University Teaching Center in 2025 raised the critical question of whether student evaluations measure teaching effectiveness at all [2].

The bias in student evaluation results mainly arises from various student-level factors. First, evaluation participation affects the credibility of evaluation outcomes. When students who are dissatisfied with instruction, perform poorly academically, or hold negative views of the evaluation system are required to participate in evaluations through administrative measures, they may be more likely to provide superficial responses, thereby undermining the credibility of the results. Students can deliver authentic and valuable feedback only when they believe evaluations help improve teaching quality. Low participation rates can lead to serious selective bias in student evaluation data, which fails to reflect actual teaching conditions objectively. Second, course-related factors also exert a profound influence on student evaluation results. For instance, evaluation scores often rise when student performance exceeds expectations; small-group seminars usually receive higher scores than large-group lectures, and instructors in STEM fields generally get lower overall evaluation scores than those in the humanities and social sciences [2]. Finally, the evaluation system itself exerts a notable influence on evaluation results. For example, revising the question "Is the teacher's organization clear?" to "Are the learning objectives of this course clear?" can redirect the evaluation focus from instructors' personal teaching styles to course design. Moreover, minor differences in rating scales and anchor labels can also alter the distribution of evaluation results.

To reveal the potential problems underlying teaching evaluation results, this paper examines the existence of statistical correlations among student academic performance, student evaluations, and supervisor evaluations from a multi-dimensional evaluation perspective. It investigates whether the three types of evaluation outcomes mutually corroborate each other or reflect diversified dimensions of classroom teaching. Based on the findings, this paper summarizes practical implications for teaching management and teacher development. The research results can provide empirical data support and decision-making basis for universities to optimize their evaluation systems, deliver targeted teaching guidance, and promote high-quality teacher development [3].

2. Research Design and Methods

2.1. Data Sources

This study focuses on undergraduate courses as its research object and categorizes three types of evaluation data. Student performance represents teachers' assessment of students' understanding of course content, while student evaluation and supervisor evaluation assess teaching quality from the perspectives of students and teaching supervisors, respectively. The study aims to analyze the inherent and underlying correlations and patterns among these three types of evaluation results to provide empirical references for optimizing and improving teaching practices [4]. Undergraduate teaching data from a secondary college of a university, spanning the years 2020 to 2025, were selected as the research sample. This dataset encompasses all courses offered to first- to fourth-year undergraduate students in the college. The large sample size enhances the comprehensiveness and reliability of the research findings. To ensure that students' evaluation results are not influenced by emotions related to final examination scores, the university implements a management mechanism where teaching evaluations are conducted prior to final examinations. This approach maximizes the fairness and validity of student evaluation results. The dataset includes 1004 courses, 2321 students, and 46 teaching supervisors, comprising 849 supervision evaluation records, 65,508 student performance records, 150,216 student evaluation records for theoretical courses, and 43,843 student evaluation records for practical courses.

2.2. Data Analysis Methods

This study employs various statistical methods, including grouped statistics, correlation analysis, linear regression, segmented regression, fixed-effects regression, and non-parametric fitting, to investigate the internal relationships among different evaluation indicators. The specific analytical techniques are outlined as follows. Grouped statistics summarize the fundamental distribution characteristics of the sample data. Pearson and Spearman correlation coefficients are applied to preliminarily evaluate the pairwise correlations between variables. Intraclass correlation coefficients are utilized to assess consistency, measuring the evaluation uniformity of the same course across different evaluators. Multiple linear regression and hierarchical regression models are used for regression estimation. Additionally, grouping and heterogeneity analyses are conducted to determine whether variable correlation patterns significantly differ under heterogeneous conditions, such as varying course types and teachers' academic ranks.

3. Research Results and Analysis

3.1. Relationship between Student Evaluation and Student Grade

3.1.1. Group Statistics

Student grades are divided into five intervals, and the corresponding student evaluation scores within each interval are analyzed statistically [4]. The detailed findings are summarized in Table 1.

Table 1. Student Evaluation Statistics by Student Grade Interval

Student Grade Interval	Mean Evaluation Score	Number of Records
0-60	93.457348	1052
60-70	93.397141	2417
70-80	92.955470	6802
80-90	93.790179	12237
90-100	94.401761	7437

The table demonstrates that student grades approximately follow a normal distribution, with the majority of scores concentrated in the range of 80 to 90. Overall, evaluation scores across different grade intervals show no significant variation. Student evaluation scores are consistently high, exhibiting a clear clustering trend toward higher scores [5].

3.1.2. Correlation Analysis

The Pearson and Spearman correlation coefficients between student grades and student evaluation scores were 0.00399 and 0.01519, respectively [6]. Both coefficients are close to zero, indicating no significant linear or monotonic correlation between the two variables. Specifically, students with high academic performance do not necessarily assign high course evaluation scores; conversely, student evaluation scores cannot directly reflect their understanding of course content.

3.1.3. Regression Analysis

Using student grades as the dependent variable and student evaluation scores as the independent variable, an OLS (Ordinary Least Squares) regression model was developed [5]. The detailed regression results are shown in Table 2 and Table 3.

Table 2. Student Evaluation Statistics by Student Grade Interval

Indicators	Values
R-squared	0
Adj. R-squared	0
F-statistic	0.4867
Prob (F-statistic)	0.485

Table 3. Variable Coefficients of the Regression Model

Variables	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	80.6958	0.883	91.437	0.000	78.966	2.426
Student Evaluation Scores	0.0065	0.009	0.698	0.485	-0.012	0.025

Table 2 indicates that the model is statistically insignificant overall (F = 0.4867, p = 0.485), with an adjusted R-squared of -0.000. This suggests that the model has negligible explanatory power, implying that student evaluation scores do not have a significant linear effect on student grades [7].

As presented in Table 3, the intercept term is 80.70 (p < 0.001), representing the baseline level of student performance when the student evaluation score is zero [5]. The coefficient for student evaluation scores is 0.0065, indicating that, holding other factors constant, a one-point increase in evaluation scores corresponds to a minimal 0.0065-point increase in student grades, which is practically insignificant. Consistent with the overall model results, the variable coefficient is not statistically significant (p = 0.485). Therefore, student evaluation scores do not have a meaningful statistical or practical impact on student academic performance.

To further investigate the relationship between student grades and student evaluations of teaching, this study uses student evaluation scores as the dependent variable and student grade intervals as independent variables (with the low-score segment as the baseline group) to construct an OLS regression model [5]. The results are detailed in Table 4 and Table 5.

Table 4. Overall Situation of Regression Model

Indicators	Values
R-squared	0.003
Adj. R-squared	0.002
F-statistic	19.74
Prob (F-statistic)	3.04e-16
Observations	29945

Table 5. Regression Model Variables

Variables	coef	p-value
Intercept	93.4573	<0.001
Grade Interval(ref:.Low)	Baseline group	-
Grade Interval(Low to Medium)	-0.0602	0.870
Grade Interval(Medium)	-0.5019	0.129
Grade Interval(Medium to high)	0.0328	0.918
Grade Interval(High)	0.9444	0.004

Table 4 demonstrates that the model is highly statistically significant overall (F = 19.74, p < 0.001). However, the adjusted R-squared is only 0.002, indicating that student grade segments have very limited explanatory power for student evaluation scores [8].

Table 5 reveals that the average evaluation score for students in the low-scoring segment is 93.4573, confirming that evaluation scores across all segments are consistently high. The high-scoring segment exhibits significantly higher evaluation scores than the low-scoring baseline group (β = 0.9444, p < 0.01). In contrast, the differences between the low-medium, medium, and high-medium segments and the baseline group are not statistically significant. This suggests that only students in the highest grade segment display a systematic difference in evaluation behavior compared to low-performing students. Overall, the influence of academic performance on student evaluation scores remains relatively minimal [5].

3.1.4. Nonlinear Model Analysis

LOWESS (Locally Weighted Scatterplot Smoothing) is a nonparametric regression technique used to smooth data through local weighting without pre-setting a global function, revealing underlying trends. GAM (Generalized Additive Model) is another statistical approach for nonlinear modeling. It captures nonlinear associations between independent and dependent variables by incorporating a smoothing term for each predictor [3]. This study adopts LOWESS smoothing and GAM fitting to examine potential nonlinear relationships between variables. The corresponding results are presented in Figure 1 and Figure 2. As shown in Figure 1, the smoothed curve remains relatively stable across different score intervals, indicating that student grades do not systematically influence student evaluation scores, which is consistent with the OLS regression results. The scatter plot reveals that evaluation scores are mostly clustered at high values. Even among students with low academic performance, high evaluation scores are prevalent. This phenomenon indicates that student ratings tend to be generally lenient. The evaluation results primarily reflect students' subjective perceptions of course experience, classroom atmosphere, teaching approaches, and workload, rather than being determined purely by academic performance. Compared to the LOWESS smoothed curve, the GAM fitted line captures more subtle fluctuations, yet its overall trend remains flat. Slight variations are observed within the 0--60 score range, where the curve rises mildly at first and then declines. However, such fluctuations are minimal and carry no practical or statistical significance. Together, both methods confirm that no statistically significant relationship exists between student grades and student evaluation scores.

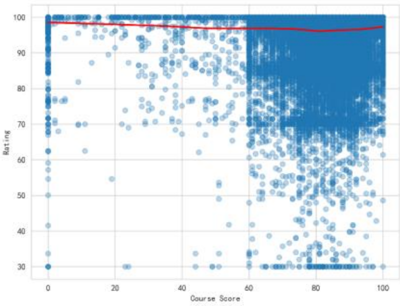

Figure 1. Lowess Fitting Curve

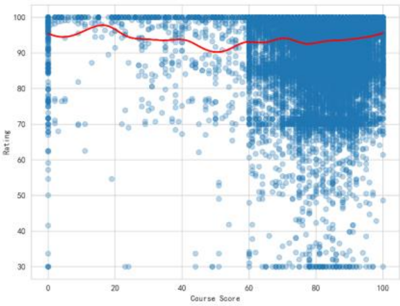

Figure 2. GAM Fitting Curve

3.2. Relationship between Student Evaluation and Supervisor/peer Evaluation

To examine the relationship between student and supervisor evaluation scores, student evaluation data were aggregated by semester and course code, with the mean calculated for each course [9]. These results were subsequently matched with supervisor evaluation data to create a course-level comparison dataset.

3.2.1. Statistical Analysis

This study conducts a statistical analysis of 694 pairs of matched data, as detailed in Table 6.

Table 6. Statistical Analysis of Student and Supervisor Evaluations

Indicators	Values	Indicators	Values
Sample size (count)	694	25%	-2.3670
Mean (mean)	1.3206	50%	0.8273
Standard deviation (std)	4.8739	75%	4.9031
Minimum value (min)	-16.2136	Maximum value (max)	22.9012

The mean score difference is 1.3206, and the median is 0.8273. Both values are positive, indicating that, on average, student evaluation scores are slightly higher than supervisor evaluation scores. The standard deviation is 4.8739, and the range is 39.1148, reflecting substantial variability in score discrepancies across individual courses. The 25th percentile is -2.3670, signifying that 25% of the samples show student evaluation scores more than 2.3670 points lower than supervisor scores. The 75th percentile is 4.9031, indicating that another 25% of the samples have student evaluation scores more than 4.9031 points higher than supervisor scores. The quantile distribution highlights significant variation in both the direction and magnitude of the differences. This suggests that students tend to apply either stricter or more lenient scoring standards compared to supervisors.

3.2.2. Correlation Analysis

The correlation analysis results indicate that the Pearson correlation coefficient and Spearman correlation coefficient between student evaluation scores and supervisor evaluation scores are -0.0123 and -0.0233, respectively, both of which are close to zero [10]. This demonstrates an absence of significant correlation between the two sets of scores. In essence, student evaluation scores do not serve as reliable predictors of supervisor evaluation scores, nor do the latter effectively reflect the former.

3.2.3. Regression Analysis

This study uses supervisor evaluation scores as the dependent variable and the average student evaluation score as the explanatory variable, constructing an OLS regression model to examine how the average student evaluation score influences supervisor evaluation scores [11]. The regression results are detailed in Table 7 and Table 8.

Table 7. Overall Regression Model

Indicators	Values
R-squared	0
Adj. R-squared	-0.001
F-statistic	0.1048
Prob (F-statistic)	0.746

Table 8. Regression Model Variables

Variables	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	95.0807	6.489	14.653	0.000	82.340	107.821
Average Student Evaluation Score	-0.0223	0.069	-0.324	0.746	-0.157	0.113

Table 7 indicates that the overall model is not significant ($F=0.1048$, $p=0.746$), with an adjusted R2 of -0.001. This demonstrates that the model lacks meaningful explanatory power, suggesting that the average student evaluation score does not have a significant linear effect on supervisor evaluation scores [12].

As presented in Table 8, when the average student evaluation score is 0, the predicted supervisor evaluation score is 95.08, representing the baseline level of supervision scores and indicating that overall supervisor evaluation scores are very high. The coefficient of the independent variable "average student evaluation score" is -0.0223, implying that for every 1-point increase in the average student evaluation score, the supervisor evaluation score decreases by 0.0223 points—an effect with negligible practical significance [13]. With a p-value of 0.746 ($p > 0.05$), this effect is not statistically significant. These findings confirm that the average student evaluation score neither has a significant statistical impact on supervisor evaluation scores nor effectively explains variations in these scores.

To investigate the possibility of a nonlinear relationship between student evaluation scores and supervisor evaluation scores, a quadratic regression model was constructed by introducing the squared term of the average student evaluation score into the baseline linear model. The regression results reveal that the model is not statistically significant overall ($F=0.6180$, $p=0.539$), with an adjusted R^2 of -0.001, indicating minimal explanatory power (Table 9 and Table 10).

Table 9. Overall Regression Model

Indicators	Values
Dep. Variable	Supervision Score
R-squared	0.002
Adj. R-squared	-0.001
F-statistic	0.6180
Prob (F-statistic)	0.539
Observations	694

Table 10. Regression Model Variables

Variables	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	-5.7432	95.024	-0.060	0.952	-192.313	180.827
Student Evaluation Scores	2.1728	2.065	1.052	0.293	-1.882	6.228
I (Student average evaluation score** 2)	-0.0119	0.011	-1.064	0.288	-0.034	0.010

The constant term has a p-value of 0.952, indicating that the intercept (i.e., the predicted supervision score when the average student evaluation score is 0) is not statistically significant. The coefficient of the linear term is 2.1728 ($p=0.293$), showing that the linear effect of the average student evaluation score on supervisor evaluation scores is not significant [2]. The coefficient of the quadratic term is -0.0119 ($p=0.288$), which theoretically suggests an inverted U-shaped relationship; however, with $p=0.288 > 0.05$, this nonlinear effect is also not statistically significant. Overall, both the linear and quadratic terms confirm that the average student evaluation score does not have a significant statistical effect on supervisor evaluation scores.

To further analyze the relationship between different segments of student evaluation scores and supervisor evaluation scores, a dummy variable regression model was constructed using the low-score segment as the baseline group [13]. The regression results are summarized in Table 11 and Table 12.

Table 11. Overall Regression Model

Indicators	Values
Dep. Variable	Supervision Score
R-squared	0.005
Adj. R-squared	0.001
F-statistic	1.269
Prob (F-statistic)	0.284
Observations	694

Table 12. Regression Model Variables

Variables	coef	p-value
Intercept (constant term)	91.828	0.00
C (Student evaluation score segment) [T. Low]	Baseline group	-
C (Student Evaluation Range) [T. Low to Medium]	0.0939	0.964
C (Student Evaluation Range) [T.Medium to High]	1.4064	0.462
C (Student Evaluation Range) [T. High]	0.9729	0.611

Table 11 demonstrates that the model is not statistically significant overall (F=1.269, p=0.284), with an adjusted R^2 of only 0.001. This indicates that the model has minimal explanatory power, suggesting no significant systematic association between different student evaluation score segments and supervision scores [13].

Table 12 reveals that the average supervisor evaluation score for the low-score segment is 91.828, indicating that overall supervisor evaluation scores are very high. The differences in supervision scores between the low-to-medium, medium-to-high, and high-score segments and the low-score segment are not statistically significant. This suggests that different segments of student evaluation scores do not have a significant systematic impact on supervisor evaluation scores [10].

3.2.4. Nonlinear Model Analysis

We employed a LOWESS smoothing curve and a generalized additive model (GAM) nonlinear fitting to examine the potential existence of a nonlinear relationship [9]. The findings are illustrated in Figure 3 and Figure 4.

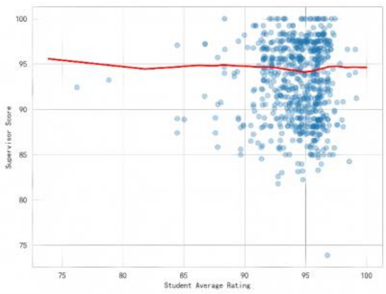

Figure 3. Lowess Fitting Curve

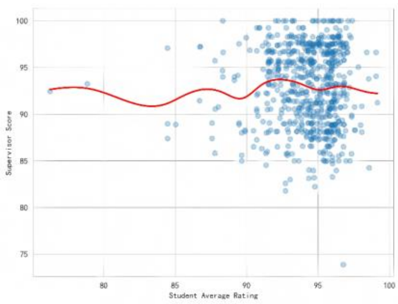

Figure 4. GAM Nonlinear Fitting Curve

The analysis of the LOWESS and GAM curves reveals that, despite both types of scores being generally high, no significant linear or nonlinear statistical correlation exists between them. This suggests that student evaluation scores are not reliable predictors of supervisor evaluation scores, highlighting notable differences in the perspectives and criteria of the two evaluation systems, as well as a lack of alignment.

3.3. Relationship between Supervisor Evaluation and Student Performance

3.3.1. Group Statistics

Student performance is categorized into five statistical intervals, and the supervision scores for each interval are analyzed statistically, as shown in Table 13. The data indicates that, after grouping student performance into these intervals, supervision scores across all groups remain consistently high with only minor fluctuations. Variations in student performance intervals do not result in consistent increases or decreases in supervision scores, further supporting the absence of a clear correlation between the two sets of scores.

Table 13. Statistics of Supervision Scores for Performance Ranges

Student Performance Intervals	Average Supervision Score	Number of Courses
0-60	92.70	52
60-70	93.08	45
70-80	94.08	184
80-90	91.72	248
90-100	93.17	50

3.3.2. Correlation Analysis

The Pearson and Spearman correlation coefficients between average student performance and supervision scores are 0.0634 (p=0.1273) and -0.1158 (p=0.0053), respectively. Both values are close to zero, suggesting no significant linear relationship between the two indicators [9]. While the Spearman coefficient reaches statistical significance (p=0.0053), its absolute value is very small, reflecting only a weak negative monotonic association. That is, courses with higher average student scores do not tend to obtain higher supervision scores, and vice versa.

3.3.3. Regression Analysis

The ordinary least squares (OLS) linear regression model constructed with supervision score as the dependent variable and average student grade as the independent variable is shown in Table 14 and Table 15. Table 14 shows that the overall model fit is very poor, with both R^2 and adjusted R^2 being -0.001, indicating that the average student performance can hardly explain the variation in supervisor scores; the model's F-statistic is only 0.1777, with a corresponding p-value as high as 0.673, which is much greater than the significance level of 0.05, further indicating that the regression equation as a whole is not statistically significant.

Table 14. Overall Regression Model

Indicators	Values
Dep. Variable	Supervisor Rating
R-squared	0.000
Adj. R-squared	-0.001
F-statistic	0.1777
Prob (F-statistic)	0.673
Observations	906

Table 15. Regression Model Variables

Variables	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	93.1776	0.771	120.904	0.000	91.665	94.690
Average Score	-0.0042	0.010	-0.422	0.673	-0.024	0.015

Table 15 shows that when the independent variable "average student performance" is 0, the predicted supervisor score is 93.1776, indicating that overall supervisor evaluation scores are at a very high level [3, 5]. The regression coefficient of the independent variable "average student performance" is -0.0042, with a corresponding p-value of 0.673, indicating no significant linear effect of student average performance on supervisor evaluation scores.

3.3.4. Nonlinear Model Analysis

The LOWESS smoothing curve and GAM nonlinear fitting were employed to examine the nonlinear relationship between course average student grades and supervisor evaluation scores [2]. The findings are illustrated in Figure 5 and Figure 6.

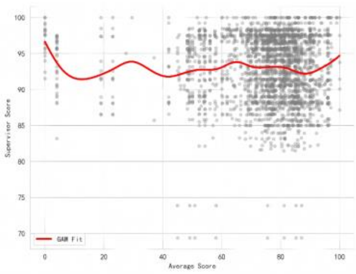

Figure 5. GAM Fitting Curve

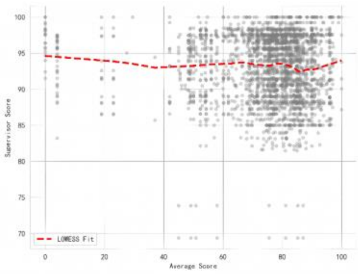

Figure 6. Lowess Fitting Curve

LOWESS and GAM analyses revealed a distinct pattern. The GAM model effectively identified the nonlinear interaction between the two variables: when student average performance falls within the low range of 0–20, supervisor evaluation scores exhibit a significant downward trend as average performance increases; in the 20–40 range, supervisor evaluation scores rebound sharply with rising average scores; and once average scores exceed 40, supervisor evaluation scores stabilize, fluctuating between 90 and 95 without a clear upward or downward trend. This demonstrates that the relationship between student average performance and supervisor evaluation scores is not linear but follows a complex nonlinear pattern characterized by "initial decline, followed by a rebound, and then stabilization." Importantly, the relationship in the low-score range contradicts the linear assumption. While traditional linear models fail to capture this critical feature, the GAM model successfully reveals the nonlinear interaction between the two.

The LOWESS local smoothing fit further corroborated the nonlinear trend identified by the GAM model. Its curve displayed a similar pattern: a gradual decline, followed by a slight rebound, and then stabilization [7]. Within the 0–30 range of student average performance, supervisor evaluation scores exhibit a slow downward trend as scores increase; between 30–40, the curve begins to rise slightly; and once scores exceed 40, supervisor evaluation scores stabilize, fluctuating within a narrow range of 92–94. This pattern aligns closely with the GAM results. Compared to the GAM model, the LOWESS curve is smoother, devoid of abrupt fluctuations, and presents the overall trend more clearly.

These findings confirm that a nonlinear association between student average performance and supervisor evaluation scores is consistently observed across different fitting methods. This eliminates the possibility of random bias from a single model and reinforces the robustness of the nonlinear relationship between the two.

3.4. Multidimensional Regression Analysis

This study examines the relationships among student evaluation scores, supervisor evaluation scores, and student grades, incorporating control variables such as course nature and course category for further analysis [4].

3.4.1. Relationship between Student Evaluations and Grades for Courses of Different Natures

First, in terms of correlation analysis, the Pearson coefficients reveal a very weak linear association between evaluation scores and grades across all three course types. The coefficients are −0.0696 (p=0.1914) for elective courses, −0.0054 (p=0.4071) for required courses, and 0.0224 (p=0.0664) for limited elective courses. None reach statistical significance, demonstrating no stable linear relationship between the two variables. By contrast, the Spearman coefficients indicate weak but statistically significant monotonic correlations. The values are −0.1564 (p=0.003169) for elective courses, −0.0137 (p=0.03554) for required courses, and 0.0442 (p=0.0002923) for limited elective courses. These results point to a mild monotonic trend rather than a strict linear relationship.

Second, OLS regression results show that the R^2 values for the three course types are all near zero: 0.007 for electives, 0.002 for required courses, and 0.001 for limited electives. The models exhibit nearly no explanatory power for variations in grades, and the regression coefficients of evaluation scores are not statistically significant. This suggests that linear models fail to capture the relationship between the two variables.

Finally, LOWESS and GAM fitting results indicate that evaluation scores and grades for all three course types present a mildly fluctuating nonlinear trend. Elective courses show the most volatile curve with no consistent monotonic pattern. While required and limited elective courses display a slight upward trend, the magnitude of change is minimal, resulting in limited practical influence [5]. Relevant results are presented in Figure 7 and Figure 8.

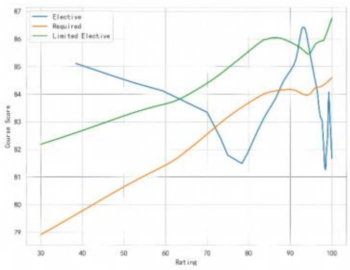

Figure 7. Lowess Fitting Curve

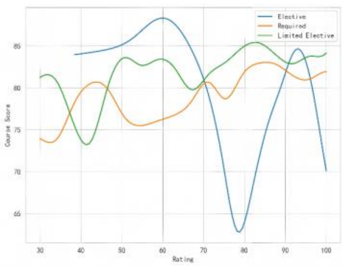

Figure 8. GAM Fitting Curve

In summary, no significant linear or monotonic correlation exists between student evaluation scores and student grades, and supervisor evaluation scores cannot serve as a valid predictor of student grades. Their relationship may be influenced by other unincluded factors, such as course difficulty, students' prior academic proficiency, and assessment methods [12].

3.4.2. Relationship between Student Evaluations and Grades for Courses of Different Types

First, correlation analysis reveals a very weak correlation between student evaluation scores and grades for all course types. The Pearson correlation coefficients are −0.0144 for major courses, 0.0682 for other courses, 0.0193 for practical courses, −0.0311 for extended courses, and 0.0160 for theoretical courses. Despite minor positive and negative deviations across categories, all values are close to zero, suggesting a negligible linear association. The Spearman coefficients likewise indicate weak monotonic correlations. Extended courses (0.1907) and theoretical courses (0.0182) present a slightly stronger positive trend, while major courses (0.0141) and practical courses (0.0159) yield very low coefficients [4]. These results demonstrate mild monotonic trends rather than stable linear relationships.

Second, LOWESS and GAM fitting results show that student evaluation scores and grades for all course types follow a fluctuating nonlinear pattern. The curve for extended courses fluctuates sharply, with a distinct peak and subsequent drop when student evaluation scores reach approximately 90 points. Courses in the "Other" category generally show a gradual upward trend. By contrast, major, practical, and theoretical courses display mild fluctuations with marginal changes and no consistent monotonic pattern. Relevant results are presented in Figure 9 and Figure 10.

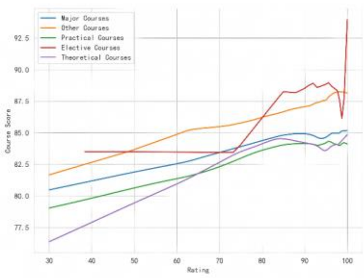

Figure 9. Lowess Fitting Curve

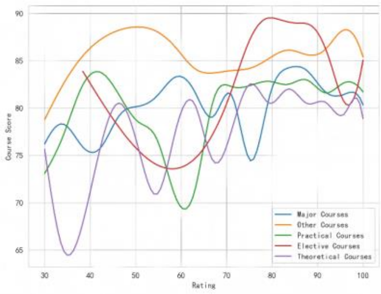

Figure 10. GAM Fitting Curve

In summary, across different course types, there is no significant linear or monotonic correlation between student evaluation scores and student grades [1]. Linear models offer almost no explanatory power for their relationship, and the observed nonlinear trends lack a consistent pattern. These findings indicate that course type exerts no substantial influence on the association between the two variables, and student evaluation scores cannot be used to predict student grades for any course category.

3.4.3. Relationship between Supervisor Evaluation Scores, Course Type, and Student Evaluation Scores

Pearson correlation analysis reveals that only theoretical courses exhibit a statistically significant weak-to-moderate negative correlation with supervision scores ($r=-0.2849$, $p<0.001$). The coefficients for major courses, other courses, and practical courses are all near zero and lack statistical significance [2, 12]. Due to limited sample sizes, the results for extended courses are deemed unreliable. Overall, a meaningful linear correlation is observed solely in theoretical courses.

The OLS regression analysis produces an R^2 value of just 0.001, indicating the model has negligible explanatory power for variations in student evaluation scores. The regression coefficient for supervisor evaluation scores is −0.0133 ($p=0.532$), while the dummy variable coefficient for practical courses is 0.0030 ($p=0.995$). Neither coefficient achieves statistical significance. These findings suggest no significant linear relationship between supervisor evaluation scores, practical course categorization, and student evaluation scores.

LOWESS curves illustrate distinct trends in student evaluation scores across course types as supervisor evaluation scores vary. For theoretical courses, scores initially decline and then rise slightly with increasing supervisor evaluation scores. Major courses remain relatively stable, while practical courses exhibit a gradual downward trend. Extended courses show pronounced fluctuations due to their small sample size. The GAM model further highlights evident nonlinear characteristics [4, 12]. Student evaluation scores across all course types display complex fluctuating patterns relative to supervision scores, without a consistent monotonic trend. Relevant results are depicted in Figure 11 and Figure 12.

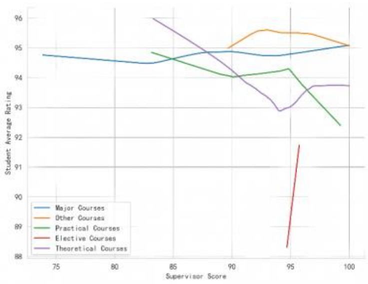

Figure 11. Lowess Fitting Curve

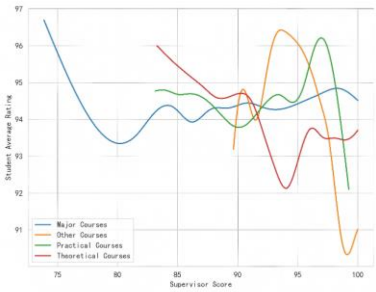

Figure 12. GAM Fitting Curve

In conclusion, a weak but statistically significant negative correlation between the two indicators is observed only for theoretical courses. Overall, no universal linear relationship or consistent nonlinear pattern is identified between supervisor evaluation scores and student evaluation scores. Additionally, course type does not exert a systematic or significant influence on this relationship. These findings indicate that supervisor evaluation scores are not effective predictors of student evaluation levels.

3.4.4. Relationship between Supervisor Evaluation Scores, Course Type and Student Evaluation Score

First, Pearson correlation analysis indicates no significant linear correlation between supervisor evaluation scores and student evaluation scores across elective courses (r=-0.0405, p=0.596), required courses (r=-0.0058, p=0.914), and restricted elective courses (r=-0.0956, p=0.148). This reveals the absence of a stable association between the two indicators for different course types [11].

Second, to further explore their relationships by course type, separate OLS regression models were established for electives, required courses, and limited electives. The results are presented in Table 16 and Table 17.

Table 16. Summary of Regression Models by Course Nature

Course type	R2	Adj. R2	F-statistic	p-value (F-test)	Sample size
Elective courses	0.002	-0.004	0.2823	0.596	174

compulsory courses	0	-0.003	0.01164	0.914	349
Limited electives	0.009	0.005	2.102	0.148	230

Table 17. Regression Coefficients by Course Nature

Course type	Coefficient coef	p-value	Significant
Elective courses	-0.0263	0.596	NS
compulsory courses	-0.0029	0.914	NS
Limited electives	-0.0458	0.148	NS

Table 16 and Table 17 show that the model fit for elective courses is very low, with R2= 0.002, an overall p-value of 0.596, and a supervisory rating coefficient of -0.0263, which is not significant. For compulsory courses, R^2 is close to 0, with an overall p-value of 0.914 and an independent variable coefficient of -0.0029, also lacking statistical significance. For limited elective courses, R^2= 0.009, with an overall p-value of 0.148 and a supervisory rating coefficient of -0.0458, still failing the significance test. This indicates that regardless of the course type, there is no significant linear correlation between supervisor evaluation scores and student evaluation scores; student evaluation scores are generally at a very high level and do not show systematic differences with changes in supervisory ratings.

Finally, LOWESS local weighted regression analysis revealed differences in the nonlinear trends of the three types of courses: the student evaluation scores for compulsory courses showed a clear downward trend as the supervisor evaluation score increased, gradually declining from nearly 96 points; elective courses exhibited a fluctuating pattern of first decreasing, then increasing, and then decreasing again; while limited elective courses showed an inverted U-shaped fluctuation of "increasing-decreasing-increasing again." The GAM model further revealed complex nonlinear characteristics, with the curves of all three types of courses exhibiting multi-peak fluctuations and no uniform monotonic pattern. Relevant results are presented in Figure 13 and Figure 14.

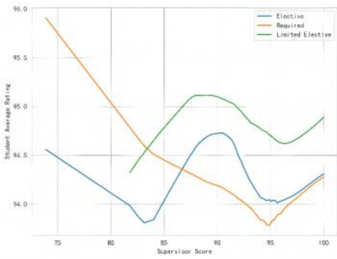

Figure 13. Lowess Fitting Curve

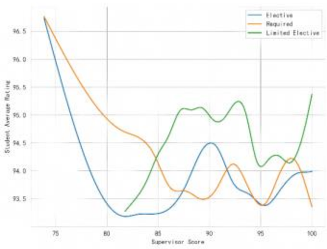

Figure 14. GAM Fitting Curve

In summary, there is no significant linear relationship between supervisor evaluation scores and student evaluation scores, nor is there a general stable nonlinear trend. Course type has a significant impact on student evaluation scores, and the nonlinear relationship between the two shows drastically different forms in compulsory courses, elective courses, and limited elective courses [8]. This indicates that supervisor evaluation scores cannot effectively predict the level of student evaluation scores, but course type is one of the important factors affecting student evaluation scores.

4. Discussion and Conclusions

4.1. Key Findings

This study takes university courses as the research object and collects three types of real-world data: student academic performance, student evaluation scores, and supervisor evaluation scores. A range of empirical methods are employed, including correlation analysis, multiple linear regression, piecewise regression, quadratic nonlinear regression, LOWESS smoothing, and generalized additive models (GAM), to explore the intrinsic correlations and heterogeneous patterns among these three teaching evaluation outcomes. The core conclusions are as follows:

First, there is no statistically significant correlation between student evaluations and student performance. Specifically, student evaluation scores are generally high, exhibiting a strong ceiling effect with scores clustered near the upper end of the scale. Secondly, although no significant linear correlation was found in the full sample, subgroup analysis revealed that high-achieving students gave higher evaluations than low-achieving students. This suggests that grades exert some influence on evaluation results, but the overall effect size is relatively small. Finally, the fitted results show that even among low-performing students, evaluations remain predominantly high. This indicates that student evaluations tend to be lenient and are more reflective of subjective experiences such as course satisfaction, classroom atmosphere, teaching methods, and perceived workload, rather than being primarily based on academic performance.

Second, student evaluation scores and supervisor evaluation scores show poor consistency. Both regression and segmented statistical analyses confirm no correlation between the two, suggesting that raters from the two groups adopt independent judgment criteria. Overall, student evaluation scores are marginally higher than supervision scores, while the score differences between them are widely dispersed. Specifically, 25% of observations show student scores exceeding supervision scores by more than 4.9 points, and another 25% show student scores falling more than 2.4 points below supervision scores. These findings reveal fundamental divergences between students and supervisory experts in evaluation perspectives, assessment dimensions, and value criteria, which means the two sets of results can hardly serve as mutual references or corroborate one another.

Thirdly, the overall correlation between supervisor evaluation scores and student grades is weak, and nonlinear features are observed only within specific score intervals. Although the correlation coefficient reaches statistical significance, the association is limited, meaning student grades cannot effectively explain variations in supervision scores. Multiple fitting analyses further confirm a nonlinear relationship between the two, which follows a distinct pattern of decline, rebound, and stabilization. To be specific, supervisor evaluation scores drop markedly at the lowest grade range, rise rapidly in the low grade range, and level off across medium and high grade ranges. This pattern cannot be captured by conventional linear regression models, which further indicates that students' final examination performance is not adopted as a core indicator in teaching supervision evaluations.

Fourthly, course nature leads to heterogeneous evaluation outcomes, while the overall correlation pattern is not bounded by course categories. No stable linear or nonlinear correlation exists between student evaluation scores and grades across courses of different natures and categories. The linkage trends between student evaluations and supervisor evaluation scores also differ by course category: for theoretical courses, student evaluation scores decline first and then rise slightly as supervisor evaluation scores increase; major courses maintain an overall stable trend; practical courses present a gradual downward trend; and extended courses show dramatic fluctuations due to the small sample size. In terms of course category, the relationship between student evaluations and supervisor evaluation scores also varies substantially: student evaluation scores for required courses decrease markedly with rising supervisor evaluation scores; elective courses follow a fluctuating trajectory of decline, rise, and then decline again; and limited electives display an inverted N-shaped trend of rising, falling, and rising once more. These findings demonstrate that course attributes exert a notable influence on student evaluation scores, and the nonlinear characteristics of the two evaluation systems differ across required courses, elective courses, and limited electives. Supervisor evaluation scores cannot effectively predict student evaluation levels, and course nature serves as a key factor affecting student evaluation results.

4.2. Discussion

The experimental results reveal significant discrepancies between current teaching evaluation outcomes and the principles of the OBE philosophy, which are reflected in the following aspects:

4.2.1. Systemic Bias in the Evaluation System

The analytical results indicate that there is no statistically significant correlation between student evaluation scores, supervisor evaluation scores, and student academic performance. This finding highlights a misalignment between the current classroom evaluation system and the principles of the OBE philosophy. Specifically, the evaluation process does not adequately focus on the actual learning outcomes of each course, resulting in weak correlations among these three sets of evaluation metrics.

4.2.2. Differences in Judgment Stances and Evaluation Demands of Evaluators

Students, as the primary participants in classroom teaching, evaluate based on their subjective learning experiences, emphasizing intuitive aspects such as classroom appeal, teaching style, course workload, and assessment difficulty. However, they often overlook deeper educational effectiveness and the development of comprehensive knowledge systems. Their scoring tends to be lenient, leading to a common upper-limit effect with scores clustering at the higher end. In contrast, teaching supervisors prioritize evaluations of instructional performance and the teaching process, focusing on professional aspects such as procedural standardization, the robustness of instructional design, consistency in classroom delivery, and the achievement of course objectives, applying rigorous and objective criteria. Due to the inherent differences in evaluation perspectives, dimensions, and priorities between these two groups, their final evaluation results naturally diverge.

4.2.3. Misalignment between Evaluation Content and Observation Dimensions

Most existing student evaluation indicators emphasize teachers' in-class performance and are limited to basic assessments, with insufficient focus on students' learning outcomes, such as knowledge acquisition and skill development. While supervision evaluations address the teaching process and curriculum design, they lack a strong connection to students' academic achievements. Additionally, student grades primarily reflect final examination results. The observation dimensions of these three evaluation approaches are fragmented and fail to form a cohesive system aligned with talent cultivation objectives. This fragmentation leads to inconsistent evaluation outcomes and, in some cases, negative correlations in theoretical courses [12].

4.3. Conclusion

This study uses undergraduate course data from a single university for empirical analysis; therefore, the sample has certain contextual limitations. The relationships among the three evaluation indicators across universities with different institutional levels and disciplinary backgrounds still require further examination. In addition, this study controls only for basic course characteristics and does not incorporate other heterogeneous factors, such as teachers’ academic ranks and class sizes, leaving room for further refinement of the variable framework. Nevertheless, based on real and large-scale teaching data, this study clearly identifies the internal relationships and structural inconsistencies among student grades, student evaluations, and supervision evaluations. The findings remain reasonably representative and can provide empirical support and practical insights for universities seeking to improve diversified teaching evaluation mechanisms, strengthen an outcome-oriented approach to teaching, and optimize teaching quality assessment systems under the OBE concept.

Funding: This work is supported by the 2026 Annual Research Projects on Xi Jinping Thought on Culture of Tianjin University of Science and Technology (KDWH202601-20); Tianjin Undergraduate Teaching Quality and Teaching Reform Research Program (A250120001, A251005702).

References

1. Asare, S., and Daniel, B. K., "Factors influencing response rates in online student evaluation systems: A systematic review approach," in *eLearn: World Conference on EdTech*, Association for the Advancement of Computing in Education (AACE), Oct. 2017, pp. 537-541.
2. Kogan, V., Genetin, B., Chen, J., and Kalish, A., "Students' Grade Satisfaction Influences Evaluations of Teaching: Evidence from Individual-Level Data and an Experimental Intervention," EdWorkingPaper No. 22-513, Annenberg Institute for School Reform at Brown University, 2022.
3. Clayson, D. E., "Student evaluations of teaching: Are they related to what students learn? A meta-analysis and review of the literature," *Journal of Marketing Education*, vol. 31, no. 1, pp. 16-30, 2009.
4. Sullivan, D., Lakeman, R., Massey, D., Nasrawi, D., Tower, M., and Lee, M., "Student motivations, perceptions and opinions of participating in student evaluation of teaching surveys: a scoping review," *Assessment & Evaluation in Higher Education*, vol. 49, no. 2, pp. 178-189, 2024.
5. Quansah, F., Cobbinah, A., Asamoah-Gyimah, K., and Hagan Jr, J. E., "Validity of student evaluation of teaching in higher education: a systematic review," in *Frontiers in Education*, vol. 9, p. 1329734, Frontiers Media SA, Feb. 2024.
6. Marzano, R. J., and Toth, M. D., *Teacher evaluation that makes a difference: A new model for teacher growth and student achievement*, ASCD, 2013.
7. Diseth, Å., "Students' evaluation of teaching, approaches to learning, and academic achievement," *Scandinavian Journal of Educational Research*, vol. 51, no. 2, pp. 185-204, 2007.
8. Sánchez, T., Gilar-Corbi, R., Castejón, J. L., Vidal, J., and León, J., "Students' evaluation of teaching and their academic achievement in a higher education institution of Ecuador," *Frontiers in Psychology*, vol. 11, p. 233, 2020.
9. Milanowski, A., "The relationship between teacher performance evaluation scores and student achievement: Evidence from Cincinnati," in *Assessing teacher, classroom, and school effects*, Routledge, 2014, pp. 33-53.
10. Bembenutty, H., "Teaching effectiveness, course evaluation, and academic performance: The role of academic delay of gratification," *Journal of Advanced Academics*, vol. 20, no. 2, pp. 326-355, 2009.
11. Hoge, R. D., and Coladarci, T., "Teacher-based judgments of academic achievement: A review of literature," *Review of Educational Research*, vol. 59, no. 3, pp. 297-313, 1989.
12. Steinberg, M. P., and Sartain, L., "Does teacher evaluation improve school performance? Experimental evidence from Chicago's Excellence in Teaching project," *Education Finance and Policy*, vol. 10, no. 4, pp. 535-572, 2015.
13. Blazar, D., "Teacher and teaching effects on students' academic performance, attitudes, and behaviors," Harvard University, 2016.