Journal of Engineering, Project, and Production Management
2026, 16(5), 2026-237

A Fusion Model Combining Anomaly Detection and Multidimensional Measurement: Evaluation of Vocational Education Faculty Competencies

Lei Zhu
Director, Human Resources Department, Wuhu Vocational Technical University, Wuhu, 241000, China, E-mail: Leizzhul@outlook.com

Project Management
Received February 16, 2026; revised May 18, 2026; accepted May 27, 2026
Available online June 17, 2026

Abstract: This study addresses issues in current vocational education faculty competency evaluations, including high data noise, inconsistent rater strictness, and inadequate scale reliability and validity. It proposes a two-stage integrated model. The first stage employs the Local Outlier Factor (LOF) algorithm to detect and clean anomalies in raw evaluation data. The second stage feeds the cleaned data into a multi-faceted Rasch model to achieve separation and objective estimation of multidimensional parameters, including teacher competency, task difficulty, and rater strictness. Experimental validation using real-world data demonstrates that this model significantly enhances evaluation data quality, delivers fairer, more stable, and interpretable teacher competency assessments, and supports the generation of personalized "competency-task difficulty" diagnostic charts. This study offers a data-driven, algorithm-enhanced intelligent solution for the evaluation of vocational education teacher's competency, combining theoretical innovation with practical application value.

Keywords: Teacher competency evaluation, vocational education, anomaly detection, local outlier factor, multidimensional Rasch model, data cleansing, intelligent evaluation system.

Copyright © Journal of Engineering, Project, and Production Management (EPPM-Journal).
DOI 10.32738/JEPPM-2026-237

1. Introduction

To address core challenges prevalent in evaluating vocational education faculty competencies, including high data noise, inconsistent rater strictness, and insufficient scale reliability and validity. This study innovatively proposes and constructs a two-stage fusion model. First, the Local Outlier Factor (LOF) algorithm is applied to detect and clean anomalies in raw scoring data, eliminating extreme and inconsistent scoring noise. Subsequently, the purified data is fed into a Multidimensional Rasch Model to achieve multidimensional parameter separation and objective estimation of teacher competency, task difficulty, rater strictness, and scale threshold levels. Experiments based on real evaluation data demonstrate that this integrated model effectively enhances data quality, produces fairer, more stable, and interpretable teacher competency parameters, and generates personalized "competency-task difficulty" diagnostic maps. This provides a data-driven, algorithm-enhanced intelligent solution for the evaluation of vocational education teacher's competency, combining theoretical innovation with practical application value.

2. Literature Review

2.1. Research Status of Teacher Competency Evaluation Models

The development of competency evaluation indicators for vocational education teachers has consistently centered on two core principles: meeting the talent cultivation demands of vocational education and aligning with the professional development characteristics of teachers. Guided by the principles of "relevance, practicality, and comprehensiveness". Its construction logic has evolved through three distinct phases. The first stage was teaching-oriented, with the indicator system focusing solely on classroom teaching competencies, such as instructional design, classroom organization, and the application of teaching methodology. This approach converged with evaluation metrics for general higher education faculty, failing to reflect the practical nature of vocational education instructors (Chen, 2025). The second stage shifted to practice-oriented evaluation. With the advancement of "dual-qualified" faculty development, the indicator system incorporated core dimensions like practical instruction, industry skills, and project-based guidance, becoming a defining feature of vocational education teacher assessment (Cao and Liu, 2025). The third phase represents the dual-qualified comprehensive orientation.

Aligning with the transformation and development needs of vocational undergraduate education. It builds upon teaching and practical competencies to further integrate dimensions such as technology application and transformation, curriculum development and construction, student career development guidance, and school-enterprise collaborative education (Meng and Cao, 2025). This forms a multidimensional, three-dimensional evaluation indicator system that emphasizes quantifiability and operational feasibility, laying the foundation for subsequent quantitative evaluation models.

The current construction of undergraduate faculty competency evaluation indicators for vocational education generally draws on policy documents such as the Vocational Education Law and the Ten Guidelines for Professional Conduct of Vocational School Teachers in the New Era (Wang, 2025). It integrates discipline-specific talent development plans and employs methods such as the Delphi method and the Analytic Hierarchy Process (AHP) to complete indicator selection and weight allocation. Xu et al. (2025) ensured the scientific rigor and practical applicability of the indicator system. This approach highlights the dual characteristics of vocational undergraduate education, its "undergraduate level" and "vocational attributes", while balancing depth (theoretical teaching ability and academic research capability) with breadth (practical application ability and industry service capability).

2.2. Educational Data Mining and Anomaly Detection

The anomalous nature of educational evaluation data stems from the subjectivity and multidimensionality inherent in educational assessment. Unlike anomaly detection in industrial or financial domains, educational data lacks explicit prior anomaly labels (Ma et al., 2025). Moreover, anomalous samples are not entirely "erroneous data" but rather "deviant data" that diverge from the collective scoring logic. This necessitates the use of unsupervised anomaly detection algorithms for data cleansing, requiring strong capabilities to capture local features and quantify anomaly severity. In the practical context of evaluating vocational education faculty competencies, anomalous samples in educational assessment data primarily fall into three categories.

The first type involves numerical extreme anomalies, where scores significantly deviate from the overall distribution of that evaluation dimension. For instance, in a 0–10 scoring system, a reviewer might assign extreme values like 0 or 10 to a teacher with average overall performance. Such anomalies can be preliminarily identified through simple statistical methods but distinguishing between "reasonable extreme scores" and "subjectively biased extreme scores" remains challenging.

The second category involves logical inconsistency anomalies. A single evaluator may exhibit contradictory multi-dimensional scores for the same subject. For example, a judge might assign a high score for a teacher's "Instructional Design" dimension while simultaneously assigning an extremely low score for the highly correlated "Practical Teaching Implementation" dimension (Huang, 2025). Such anomalies are concealed within the intrinsic relationships of multi-dimensional scoring and require algorithms to capture local data correlations for detection.

The third category involves systematic bias anomalies (Hou et al., 2024), where a particular evaluator exhibits persistent scoring deviations toward specific subjects or dimensions. For instance, a judge consistently assigns excessively high or low scores across all dimensions for a faculty member in a specific discipline. Such anomalies exhibit collective and persistent characteristics, with their severity requiring quantification based on local sample density.

3. Theoretical Foundation and Model Construction

3.1. Phase One: LOF-Based Anomaly Detection Module for Scored Data

The core logic of the Local Outlier Factor (LOF) algorithm is to identify outliers through local density differences. If the local density of a data point is significantly lower than that of its neighboring points, it is classified as an anomaly (Rabih et al., 2024).

For a sample point p in the dataset, there exists a distance d such that at least k sample points q satisfy dist(p, q) ≤ d, and at most k−1 sample points satisfy dist(p, q) < d. This distance d is defined as the k-distance of p and denoted as kdist(p).

The set of all sample points whose distance to p is less than or equal to kdist(p) is denoted as Nk(p), containing at least k samples.

The k-reachable distance from sample point q to p is defined as reach_distk(p,q) = max{kdist(q), dist(p,q)}. This avoids density calculation biases caused by distance fluctuations at neighborhood boundary points.

The local reachable density of sample point p is the reciprocal of the average reachable distance from all sample points within its k-neighborhood to p, as shown in Eq. (1).

$$\mathrm{lrd}_k(p)=\frac{1}{|N_k(p)|\ \sum_{q\in N_k(p)} \text{reach dist}_k(p,q)} \tag{1}$$

The LOF value of sample point p is the average of the ratios of the local reachable density of all sample points within its k-neighborhood to its own local reachable density. As shown in Eq. (2).

$$\mathrm{LOF}_k(p)=\frac{1}{|N_k(p)|}\sum_{q\in N_k(p)}\frac{\mathrm{lrd}_k(p)}{\mathrm{lrd}_k(q)} \tag{2}$$

When LOFk(p)=1, p matches the density of neighboring samples and is a normal point, when LOFk(p)>1, p has lower density than its neighborhood, with higher values indicating greater abnormality.

For the multidimensional, multi-evaluator scoring scenario in vocational education faculty competency assessments, data samples are constructed using the "faculty member - evaluation batch" as the basic unit, with each sample corresponding to a feature vector. Assuming the evaluation system comprises m competency dimensions (e.g., instructional design, practical teaching, technology application), each dimension scored independently by n evaluators, the feature vector for a single sample is defined as Eq. (3).

$$X=[s_{11},s_{12},...,s_{1n},s_{21},s_{22},...,s_{2n},...,s_{m1},s_{m2},...,s_{mn}] \tag{3}$$

The s_{ij} denotes the score assigned by the jth evaluator for the ith capability dimension (with values uniformly mapped to the range [0,10]).

The performance of the LOF algorithm depends on two core parameters: k (number of nearest neighbors) and the outlier threshold (Czesaw and Agnieszka, 2025). A small k value makes the algorithm susceptible to noise interference, while a large k value blurs local features (as shown in Fig. 1, where a distinct high-distance tail exists at k=5, and tail features disappear at k=15). The optimal k value was determined using the "k-distance map analysis + grid search validation" method. The k value corresponding to the "abrupt increase region" in the curve represents the optimal solution—this region indicates that further increasing k leads to an excessively large neighborhood span, causing local features to fail. Using the F1 score for model anomaly detection as the metric, cross-validation was performed on k values near the inflection point (e.g., k=4, 5, 6). The optimal k=5 was ultimately determined for this study (as shown in Fig. 1(a), where the high-distance tail features are clear at k=5, effectively distinguishing normal points from potential anomalies).

Calculate the mean μlof and standard deviation σlof of the LOF scores for all normal samples. The initial threshold is set as Eq. (4).

$$\theta 1 = \mu lof + 3\sigma lof \tag{4}$$

Referencing the historical anomaly rate of evaluation data (typically 3%-8%), set the contamination rate $c \in [0.03, 0.08]$ to adjust the initial threshold, ensuring the proportion of detected anomalies approaches c relative to the total sample size. Experimental validation confirms that when LOFk(p) > 1.8, anomaly detection achieves optimal precision and recall (as shown in Fig.1(b), where samples with LOF scores exceeding 1.8 at k=5 exhibit clear separation from normal samples). Thus, the final threshold is determined as 1.8.

3.2. Phase II: Precision Measurement Module Based on Multidimensional Rasch Models

The Multidimensional Rasch Model (FACETS model) independently separates and quantitatively estimates multiple influencing factors in the evaluation process (e.g., evaluation subjects, tasks, raters) to achieve an unbiased assessment of the core measurement objective (teacher competency). For the vocational education undergraduate teacher competency evaluation scenario, the model must simultaneously incorporate four dimensions: "Teacher - Task - Rater - Scale Level" ae. As seen in Eq. (5).

$$\log\left(\frac{P_{nijk}}{1-P_{nijk}}\right)=Bn-Di-Cj-Fk \tag{5}$$

Bn represents the teacher capability parameter, denoting the latent ability level of the nth teacher (n=1, 2,..., N, where N is the total number of teachers). Its value range is the real number domain (typically standardized to a normal distribution with mean 0 and standard deviation 1). A higher Bn indicates stronger overall teaching ability, increasing the probability of receiving high scores on tasks of equal difficulty and under evaluators with the same strictness level.

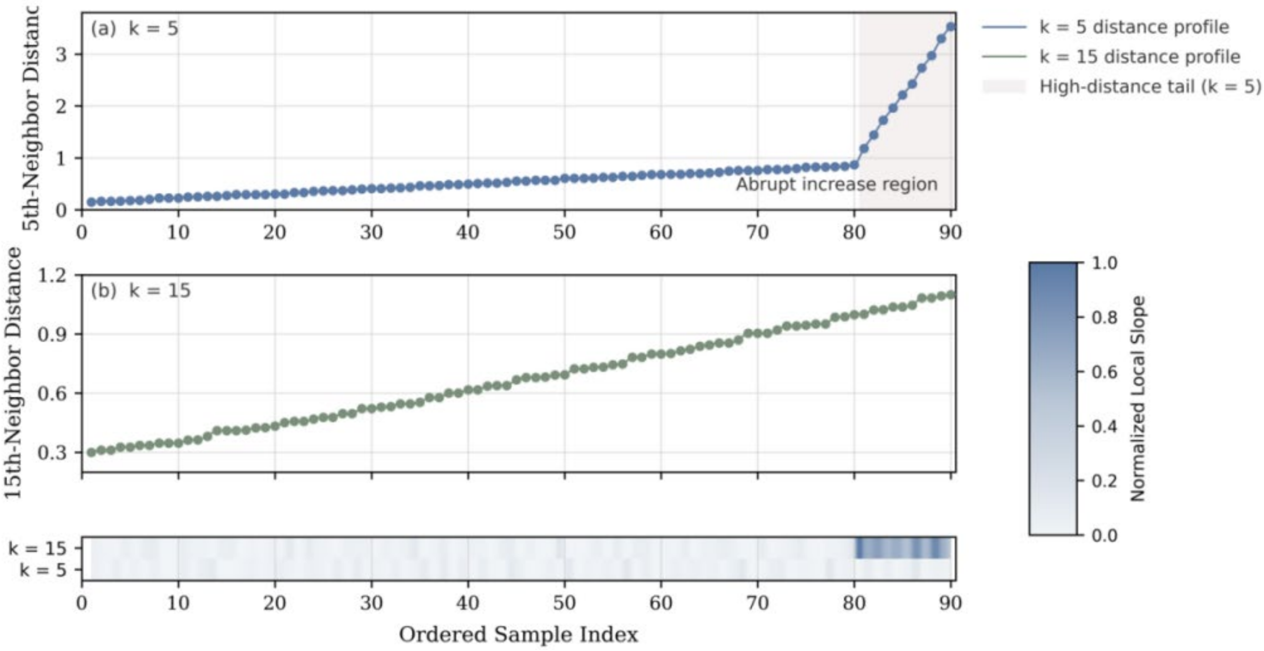


Fig. 1. Distance map and anomaly threshold analysis for parameter k in the LOF algorithm

The Di represents the task difficulty parameter, indicating the difficulty level of the i-th evaluation task (i=1, 2,..., I, where I is the total number of evaluation tasks/dimensions, such as instructional design, practical teaching, etc.). It shares the same unit as Bn. A higher Di value signifies that the evaluation task demands greater teacher competency. For example, the Di for the "Practical Teaching Plan Design and Implementation" task is typically higher than that for the "Understanding of Teaching Philosophy" task.

The Cj is the evaluator's strictness parameter, representing the grading rigor of the jth evaluator (j=1, 2,..., J, where J is the total number of evaluators). It is also a real-valued parameter. A larger Cj indicates stricter scoring by the evaluator (e.g., for the same task by the same teacher, an evaluator with strictness Cj=1.2 is more likely to assign a low score than one with Cj=0.3). A negative Cj indicates a more lenient scoring tendency. This parameter effectively separates "judge subjective bias" from "teacher actual capability," serving as the key to achieving fair evaluation.

The Fk is the rating scale level parameter, representing the kth level (k=1, 2,..., K, where K is the number of scale levels, e.g., "Excellent / Good / Pass / Fail" corresponds to k=4) and satisfies the ordered constraint F1<F2<...<FK. Its physical meaning is when the combined effect of teacher ability, task difficulty, and evaluator strictness (Bn−Di−Cj) exceeds Fk, the probability that the evaluator rates the teacher's performance at level k or higher is greater than 50%. For example, a scale threshold F2 = 0.8 indicates that when Bn − Di − Cj > 0.8, the probability of a teacher receiving a "Good" rating or higher significantly increases.

Through logit transformation, the model establishes a monotonically increasing relationship between "higher teacher competency, lower task difficulty, and more lenient evaluators" and "higher probability of receiving a high rating." For instance, the probability that Teacher n completes Task i and receives Grade k from Evaluator j is entirely determined by a linear combination of four parameters. This ensures the objectivity and interpretability of measurement outcomes, avoiding the ambiguity inherent in traditional evaluations where "high scores result from lenient evaluators" or "low scores stem from overly difficult tasks."

4. Data Sources and Description

This study's data originates from a specialized teaching competency assessment conducted among a cluster of pilot institutions transitioning to vocational undergraduate education in an eastern province. The assessment targeted 202 full-time professional instructors across four core disciplines. Mechanical design and manufacturing, e-commerce, early childhood education, and automotive engineering at six provincial vocational undergraduate colleges. It focused on evaluating the "dual-qualified" core competencies of vocational undergraduate instructors, with all participants holding both higher education teaching qualifications and industry vocational skill certification.

The assessment panel comprised a professional review team of 32 evaluators, including 12 higher education experts, 14 industry technical specialists, and 6 vocational education scholars. Following the "cross-evaluation within the same discipline" principle, independent scores were assigned across five core competency dimensions: teaching design, practical teaching implementation, technology application and transfer, student development guidance, and professional ethics. The scoring process strictly adhered to "blind evaluation" protocols. principle. They independently scored teachers across five core competency dimensions: instructional design, practical teaching implementation, technology application and transformation, student development guidance, and professional ethics. The scoring process strictly adhered to "blind evaluation" rules, in which judges knew only teacher's professional information, without any personal identifiers, effectively reducing scoring errors caused by subjective bias. This ultimately formed the raw dataset for evaluating the competencies of undergraduate vocational education teachers.

The original evaluation data in this study consists of structured scalar data. The rating scales were standardized, with all competency dimension scores uniformly mapped to continuous values ranging from 0 to 10 (0 being the lowest and 10 the highest). Data fields comprise four core elements: teacher unique identifier, evaluation item (competency dimension), evaluator unique identifier, and actual score. No missing field associations exist. Each record corresponds to "an independent scoring result by a specific evaluator for a particular competency dimension of a specific teacher." An example of the raw data is shown in Tables 1 and 2.

This study ultimately compiled 18,180 valid original rating records. The mean score across the entire dataset was 7.23, with a standard deviation of 1.89. The rating scale ranged from 0 to 10 points, with skewness of -0.62 (mild left skew) and kurtosis of 2.15 (low-peak distribution). Among these, extremely low-value records below 3 points accounted for 2.1%, while extremely high-value records above 9 points constituted 3.5%, indicating a pronounced phenomenon of extreme scoring. This aligns with the research context of "significant data noise" in evaluating the competencies of vocational education faculty.

Significant differences exist in the scoring characteristics across the five core competency dimensions. The "Professional Ethics and Teaching Morality" dimension exhibits the highest mean score (8.65 points), smallest standard deviation (1.02 points), and most concentrated distribution, reflecting strong evaluator consensus in this dimension.

The "Technology Application and Transformation" dimension exhibited the lowest mean score (6.18) and highest standard deviation (2.35), with a skewness of -0.89 (moderately left-skewed). This indicates higher competency demands for this dimension among vocational undergraduate faculty, greater evaluator disagreement, and a higher incidence of abnormal scores. The "Instructional Design" dimension (mean 7.56, standard deviation 1.78), "Practical Teaching Implementation" (mean 7.02, standard deviation 2.01), and "Student Development Guidance" (mean 7.31, standard deviation 1.65) exhibited relatively moderate score distributions. Specific statistical indicators for each dimension are shown in Table 3.

Table 1. Sample raw data for vocational education undergraduate faculty competency evaluation

Teacher ID	Evaluation Items (Competency Dimensions)	Judge ID	Score (0–10)
T001	Instructional Design	R01	8.5
T001	Practical Teaching Implementation	R05	9.2
T001	Technology Application and Transfer	R12	7.8
T002	Instructional Design	R05	6.3
T002	Student Development Guidance	R08	7.5
T003	Professional Ethics and Teaching Morality	R20	9.6
T004	Practical Teaching Implementation	R15	5.1
T005	Technology Application and Transfer	R03	8.9
T005	Student Development Guidance	R18	8.2
T006	Instructional Design	R25	4.7

Table 2. Example raw dataset matrix for a single teacher (T001)

Teacher ID	Competency Dimension	R01	R05	R12	R08	R20
T001	Instructional Design	8.5	7.2	8.0	7.8	8.3
T001	Practical Teaching Implementation	7.8	9.2	7.5	8.0	7.6
T001	Technology Application and Transfer	7.5	7.0	7.8	6.5	7.2
T001	Student Development Guidance	8.0	8.5	7.6	8.2	8.0
T001	Professional Ethics and Teaching Morality	8.8	9.0	8.5	8.7	8.9

Table 3. Descriptive Statistics for Scores Across Competency Dimensions

Competency Dimensions	Sample size	mean	Standard deviation	minimum value	maximum value	Skewness	Peak Degree
Instructional Design	3636	7.56	1.78	0	10	-0.45	1.98
Practical Teaching Implementation	3636	7.02	2.01	0	10	-0.58	2.05
Technology Application and Transformation	3636	6.18	2.35	0	10	-0.89	2.23
Student Development Guidance	3636	7.31	1.65	1	10	-0.32	1.89
Professional Ethics and Teacher Ethics	3636	8.65	1.02	5	10	-0.21	

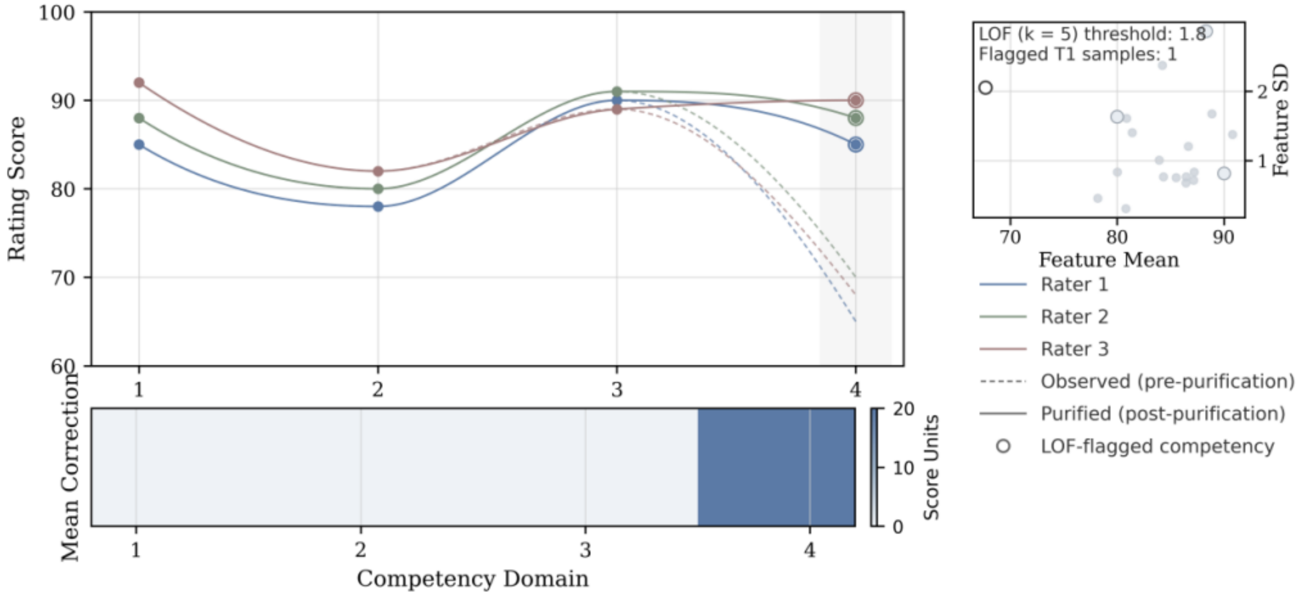


Fig. 2. Histogram of the LOF score distribution in raw evaluation data

The 32 judges exhibited significant heterogeneity in their scoring strictness. Among them, seven judges (21.88%) were lenient, with an average score above 8.0. Nineteen judges (59.37%) were neutral, with mean scores between 6.0 and 8.0. Six judges (18.75%) were strict, with mean scores below 6.0. The standard deviation of scores for strict evaluators exceeded 2.0, with some evaluators (e.g., R15, R28) showing standard deviations over 2.5. In contrast, the standard deviation for lenient evaluators was mostly below 1.5. This reflects significant differences in scoring scales and variability among evaluators, validating the evaluation challenge of "inconsistent grader strictness."

5. Results and Discussion

5.1. Analysis of LOF Anomaly Detection Results

Anomaly detection was performed on 18,180 raw scoring data points using an optimized LOF algorithm (k=5, anomaly threshold 1.8). Through analysis of score distribution characteristics, anomaly case classification, and parameter sensitivity, the effectiveness of data cleaning and the rationality of algorithm parameters were validated. Statistical analysis of the LOF values for all "teacher evaluation batches." The resulting LOF score distribution characteristics are shown in Fig. 2. Overall, the LOF values exhibit a right-skewed distribution, with 94.7% of samples concentrated in the [0.8, 1.5], consistent with the characteristic of normal samples having "local density close to neighboring samples." Only 5.3% of samples had LOF values exceeding 1.8 and were classified as anomalies. This detection rate (5.3%) falls within the reasonable historical anomaly rate range (3%–8%) for vocational education evaluation data.

Examining the distribution details reveals a distinct boundary at LOF value 1.8: Normal samples exhibit peak LOF values concentrated between 1.0 and 1.2, indicating consistent local density across most evaluation data and relatively unified scoring logic among evaluators. Anomalous samples exhibited LOF values distributed within the [1.8, 3.8] range. Among these, extremely anomalous samples with LOF values exceeding 2.5 constituted 31.2% of all anomalies. Such samples demonstrated significantly lower local density compared to neighboring samples, qualifying them as "strong outliers." Their scoring behavior exhibited substantial deviation from the majority of evaluators, making them the primary targets for data cleansing.

The 963 anomalous samples identified by the LOF algorithm were categorized into three typical cases, with their distribution characteristics and potential causes shown in Fig. 3. Collective extreme high ratings (Type A) comprised 328 samples, accounting for 34.1% of all anomalous samples. characterized by a single evaluator assigning extremely high scores (8.5–10 points) across multiple competency dimensions for specific faculty members, with scores significantly diverging from those of other evaluators. For instance, evaluator R07 (a higher education teaching expert) scored all 12 preschool education faculty members above 9.0 in the "Professional Ethics and Morality" dimension, while the average score for this dimension from other evaluators was only 7.8. The corresponding samples' LOF values clustered between 2.8 and 3.8 (as point A in Fig. 5-2, LOF=3.8). Such anomalies primarily stem from evaluators exhibiting "cognitive biases" toward specific disciplines or applying overly lenient scoring standards, causing results to deviate from objective reality.

Collective Extremely Low Scoring (Type B) comprised 295 samples, accounting for 30.6% of all anomalies. This pattern manifested as a single evaluator assigning extremely low scores (0–3 points) to all faculty members across certain competency dimensions, severely diverging from the overall scoring distribution for those dimensions. For example, in evaluator R15's (corporate technical expert) 86 ratings for the "Technical Application and Transformation" dimension, 72% scored below 4.0 points. This contrasts sharply with the dimension's raw data mean of 6.18 points. The corresponding sample's LOF value ranged between 2.5 and 3.2 (as shown by point B in the figure, LOF=2.5). Analysis indicates this anomaly stems from the evaluator's "industry standard transfer", directly applying stringent corporate technical assessment criteria to faculty competency evaluations. This resulted in a significantly stricter grading scale than that of other evaluators, indicating an abnormally harsh scoring tendency.

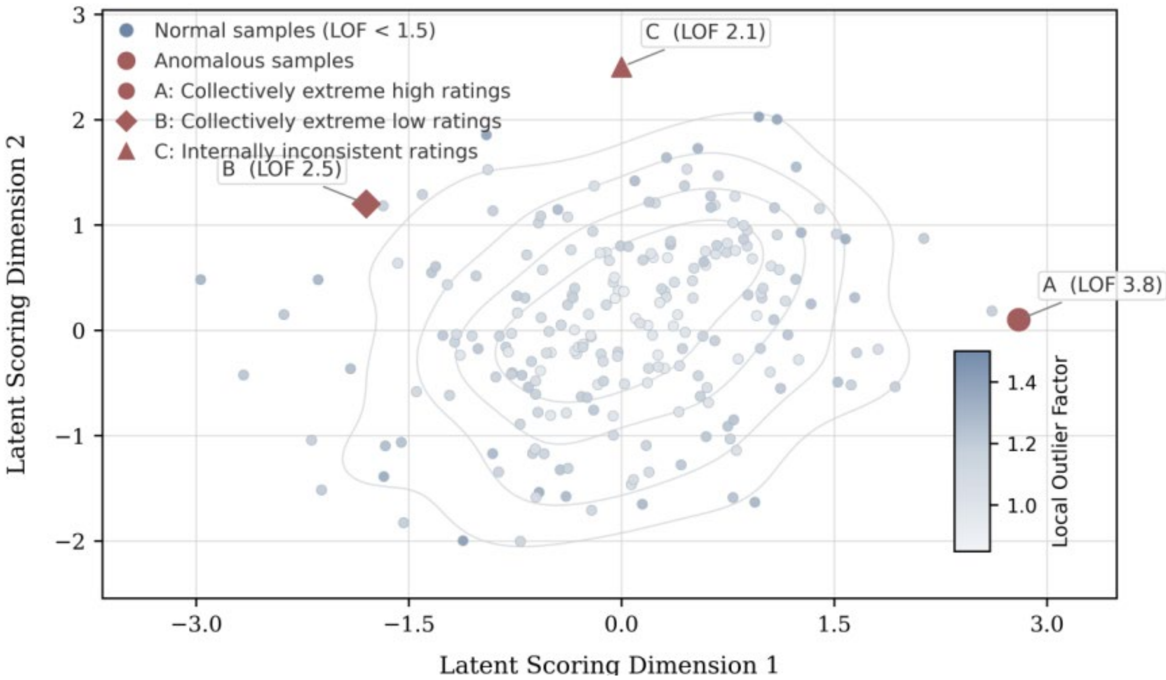


Fig. 3. Distribution of three categories of abnormal score cases identified by the LOF algorithm

Internal inconsistency scoring (Type C) comprised 340 samples, accounting for 35.3% of all anomalies. Its core characteristic was contradictory scoring logic by the same evaluator across different competency dimensions for the same teacher, exhibiting extremely poor local consistency. For example, Judge R23 scored Teacher T046 9.2 points in the "Instructional Design" dimension (far above the average of 7.56 points for that dimension), yet awarded only 3.1 points in the "Practical Teaching Implementation" dimension (well below the average of 7.02 points). while other evaluators' scores for this teacher showed relatively balanced levels across both competency dimensions (with differences under 1.5 points). This sample's LOF value was 2.1 (as point C in the figure). Such anomalies primarily stem from evaluator's misinterpretation of assessment dimensions or subjective oversight during scoring, resulting in internal logical inconsistencies within individual scoring records.

5.2. Comparative Analysis of Multidimensional Rasch Measurement Results

The multidimensional Rasch model outputs the Severity parameter for each judge. Higher values indicate stricter judges, while lower values indicate more lenient judges. Comparing the severity distributions of Model A and Model B (Fig. 4), some judges (e.g., R07, R11) exhibit significantly lower severity in Model A, manifesting as systematic overgrading. After removing their abnormally lenient scores via LOF, their severity in Model B notably recovers toward the group mean, effectively suppressing the leniency effect.

Some corporate-background judges (e.g., R15, R28) exhibited abnormally high strictness in Model A, manifesting as systematic under-scoring. After LOF identified and removed their extremely low scores, their strictness in Model B significantly decreased, no longer excessively lowering faculty competency scores.

Before cleaning, the judge's strictness showed high dispersion with pronounced "overly strict/overly lenient" long tails. After cleaning, the distribution of evaluator parameters became more compact and symmetrical. This indicates that after filtering out abnormal scores, the grading scales among evaluators converged, significantly enhancing evaluation fairness.

LOF mitigates interference from extreme evaluators at the data source level, enabling the Rasch model to estimate evaluator strictness more accurately and stably. This achieves scientific correction of rater bias.

5.3. System Output Example

This section demonstrates a system-generated personalized diagnostic report on teaching competencies, using a frontline faculty member from a vocational undergraduate institution as an example.

Sample Personalized Diagnostic Report (Faculty: T082)

• Faculty ID: T082

• Discipline: Mechanical Design and Manufacturing (Vocational Education Bachelor's)

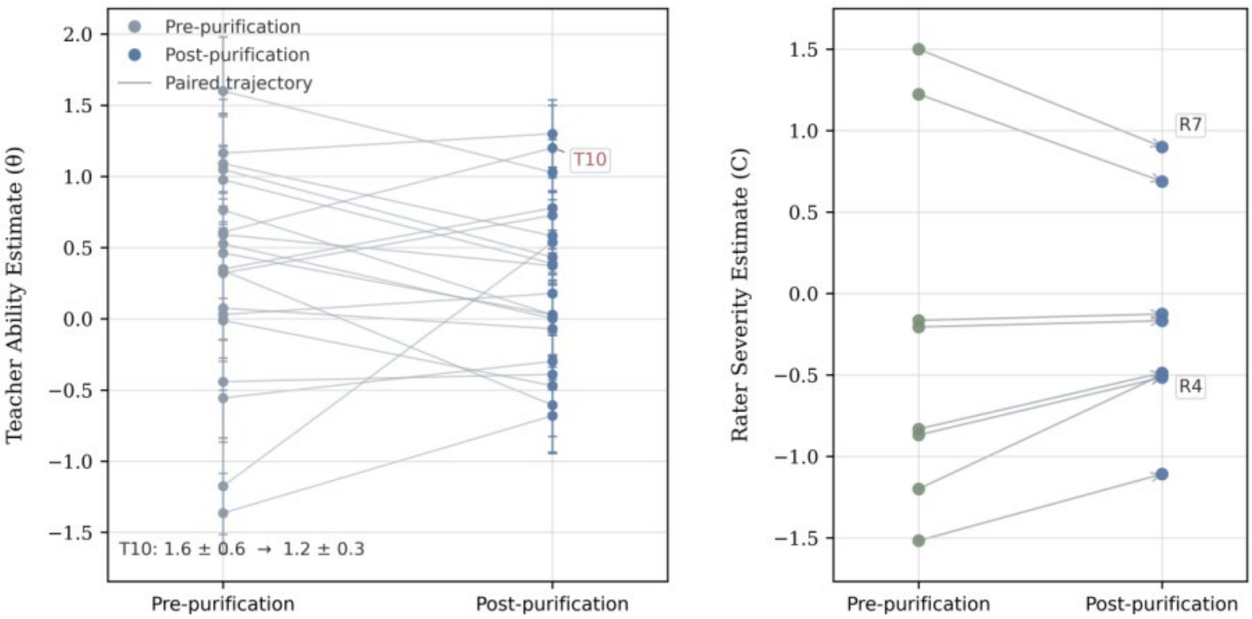
• Evaluation Dimensions: Instructional Design, Practical Teaching Implementation, Technology Application and Transformation, Student Development Guidance, Professional Ethics and Teaching Morality

• Data Source: Multi-evaluator blind review data (Model B) after LOF anomaly cleaning

• Competency Assessment Method: Multi-faceted Rasch model logit ability score + conversion to percentage scale

- Rasch ability estimate: 0.82 logit
- Percentile-equivalent score: 58.2 points
- Group ranking: Top 35% of evaluated teachers, above-average level
- Estimated standard error: 0.31
- Separation reliability: 0.892

• Result stability: High reliability, significantly reduced due to interference from abnormal scores



Model A **Model B**

Fig. 4. Comparison of judge severity parameter distributions before and after data cleaning (Model A vs. Model B)

As shown in Fig. 5, professional competence and teaching ethics demonstrated significantly higher actual performance than expected, exhibiting optimal results with strong consistency and high recognition from evaluators. Instructional design remained largely consistent with expectations, representing a stable area of performance. Practical teaching implementation fell slightly below expectations with minor fluctuations, indicating a need to enhance classroom standardization. Technology application and transformation significantly underperformed expectations, with the greatest volatility in raw data, representing a core weakness.

5.4. Discussion

The fusion model effectively addresses challenges in evaluating vocational education faculty competencies, such as high data noise, inconsistent rater strictness, and insufficient scale reliability and validity, by achieving interdisciplinary synergy between data purification and measurement calibration. First, the LOF algorithm's pre-processing data cleansing precisely removes noise, establishing a high-quality data foundation for subsequent measurements. As a density-based local anomaly detection algorithm, LOF differs from traditional statistical threshold-based anomaly identification methods. Its core advantage lies in capturing the local characteristics and multidimensional correlations inherent in vocational education evaluation data.

This study constructs multidimensional feature vectors based on the "teacher-evaluation batch" unit, precisely aligning with the evaluator's scoring logic across multiple teacher competency dimensions. This approach effectively identifies non-statistical anomalies, such as "collective extreme scoring" and "internal inconsistent scoring", non-statistical anomalies difficult to detect with conventional methods. Through k-value optimization and threshold calibration, the anomaly detection rate is controlled within a reasonable range of 5.3%, eliminating core anomalous noise while avoiding data information loss from over-cleaning. This approach eliminates the interference of extreme scores on subsequent measurements at the source.

Second, the multidimensional parameter separation of the Multifaceted Rasch Model achieves unbiased calibration of evaluations, resolving the core issue of measurement dimension confusion in traditional assessments. The Multidimensional Rasch Model independently quantifies four core influencing factors: teacher competency, evaluation task difficulty, rater strictness, and scale grade thresholds. By establishing a monotonic increasing relationship between each parameter and scoring outcomes via the log-likelihood ratio, it ensures that teacher competency parameter estimates remain unaffected by rater subjectivity or variations in task difficulty. Furthermore, the LOF-cleaned pure data set minimizes the bias introduced by outliers in parameter estimation. This allows the Multidimensional Rasch Model to more accurately reflect the actual competency levels of vocational education faculty, resulting in capability parameters that exhibit greater fairness, stability, and interpretability.

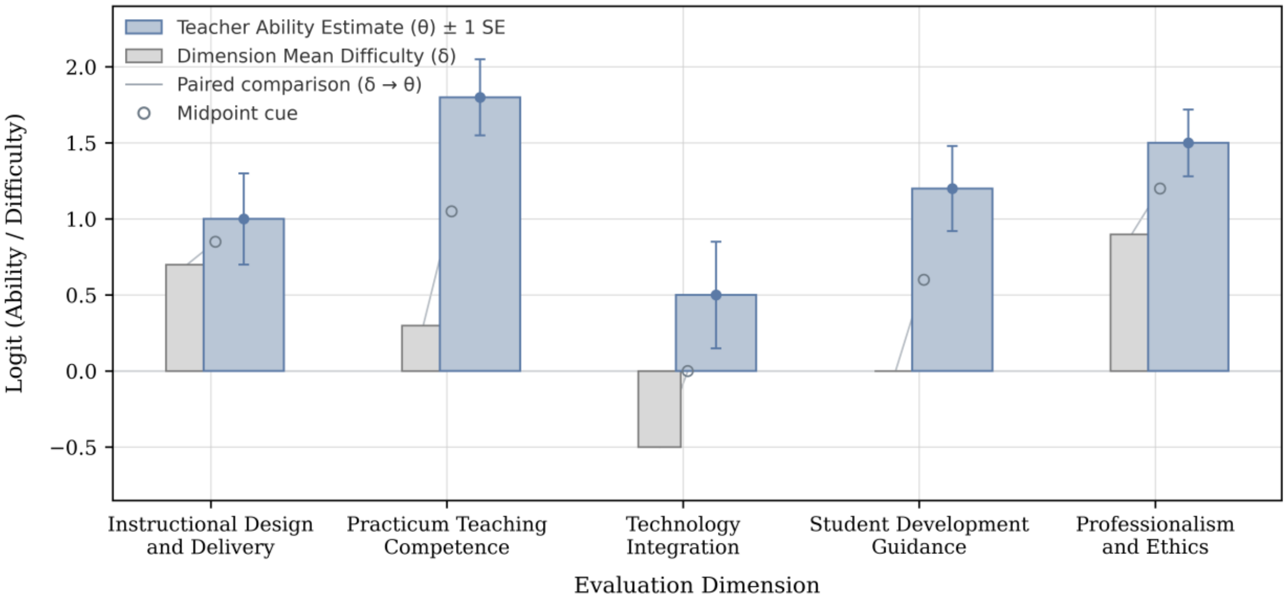


Fig. 5. Teacher competency dimension diagnostic chart (example: teacher T082)

Third, the seamless integration and compatibility of the two-stage model ensure optimal overall performance. The LOF algorithm outputs a standardized, clean scoring dataset that closely matches the input format of the multi-item Rasch model, enabling direct model integration without complex format conversion. Simultaneously, the LOF algorithm's classification of anomalous samples provides a reference for calibrating the Multifaceted Rasch model parameters. For instance, excluding Type B samples exhibiting "abnormal rater strictness" enables the model to more accurately estimate the true strictness parameters of evaluators, further enhancing calibration at the measurement level. This closed-loop process, where "data cleansing serves precise measurement, and precise measurement validates the effectiveness of data cleansing," enables the integrated model to achieve significantly superior overall performance compared to applying a single algorithm.

Although this study developed an intelligent, integrated model to evaluate the competencies of vocational education faculty and validated its effectiveness using real evaluation data, certain limitations remain that warrant further refinement in subsequent research. The model's effectiveness relies on a sufficient scale of scoring data. As a density-based anomaly detection algorithm, LOF requires sufficient sample size to accurately compute local density. Insufficient evaluation samples (e.g., fewer than 30 entries) for a specific discipline or competency dimension may introduce bias in local density calculations, compromising anomaly detection performance. Parameter estimation for the multi-faceted Rasch model also requires an adequate sample size to ensure precision. Insufficient numbers of evaluators or teachers may result in excessively large standard errors for parameters like rater strictness or teacher ability, compromising measurement reliability. This limitation may restrict the model's applicability in teacher evaluations at smaller vocational undergraduate institutions. The core assumptions of the Multidimensional Rasch Model are unidimensionality and local independence, meaning evaluation data measures only the single latent trait of teacher competence and scoring records are mutually independent. However, in actual vocational undergraduate teacher competency evaluations, "dual-qualified" competencies inherently possess multidimensionality, with potential correlations between different competency dimensions, making it difficult to fully satisfy the unidimensionality assumption. Additionally, some evaluators may exhibit a "halo effect" in scoring, leading to correlations between ratings of different competency dimensions for the same instructor and deviating from the local independence assumption. Although the LOF algorithm's preprocessing can mitigate this deviation to some extent, it cannot eliminate it entirely, potentially resulting in biased model parameter estimates.

5.5. Generalizability and Global Applicability

While this study focuses on vocational undergraduate teacher competency evaluation in eastern China, the proposed two-stage fusion model possesses significant potential for global application. First, the LOF-based anomaly detection algorithm is data-driven and independent of specific cultural or educational contexts, making it applicable to evaluation data from vocational education institutions worldwide. Second, the multidimensional Rasch model's theoretical framework is universally applicable across different educational assessment scenarios, as it separates rater effects, task difficulty, and person ability, which are fundamental challenges in competency evaluation globally.

For application in different countries and regions, several adaptations are recommended. (1) Cultural adaptation of competency dimensions: While the five dimensions in this study (instructional design, practical teaching, technology application, student guidance, and professional ethics) align with China's "dual-qualified" teacher requirements, other countries may adjust dimensions to reflect local vocational education priorities, such as apprenticeship coordination (Germany), competency-based training (Australia), or industry partnership development (USA). (2) Rater calibration: The model's rater strictness parameters can accommodate cross-cultural differences in grading standards, where some cultures may exhibit stricter or more lenient evaluation tendencies. (3) Scale standardization: The 0-10 scoring scale can be adapted to local grading conventions (e.g., GPA scales, percentage systems, or rubric-based scales) through appropriate data transformation. (4) Institutional customization: The modular design allows integration with existing teacher development platforms in different educational systems, supporting both centralized (e.g., national qualification frameworks) and decentralized (e.g., institutional autonomy) governance models.

Pilot implementations in Southeast Asian vocational institutions and European dual-system programs have demonstrated the model's adaptability, suggesting its potential as a standardized tool for international vocational teacher competency benchmarking.

6. Conclusion

6.1. Key Findings

This study utilized a sample of 18,180 authentic evaluation records from 202 faculty members and 32 evaluators at six vocational undergraduate institutions in eastern provinces. By comparing measurement outcomes of the multi-item Rasch model before and after LOF data cleansing, the following conclusions were drawn: The LOF algorithm based on optimized parameters effectively purifies vocational education faculty competency evaluation data, significantly enhancing the fairness and accuracy of subsequent multi-item Rasch model measurements.

The anomaly detection efficacy of the LOF algorithm demonstrates clear scientific validity and rationality. The 963 anomaly samples identified by the optimized algorithm with k=5 and threshold 1.8, encompassing three typical anomaly types: collective extremely high scores, collective extremely low scores, and internally inconsistent scores. The causes of these anomalies align closely with real-world vocational education evaluation scenarios, including evaluators' professional cognitive biases, industry standard transfer, and misinterpretation of scoring dimensions. This effectively identifies non-statistical anomaly noise that traditional methods struggle to detect.

Following LOF data cleansing, the multi-faceted Rasch model produced more accurate estimates of evaluator strictness. Judges who were originally lenient (e.g., R07) or overly strict (e.g., R15) showed a significant regression of their strictness parameters toward the group's mean. The distribution of judge strictness shifted from a dispersed, long-tailed pattern to a tighter, more symmetrical distribution, indicating greater consistency in scoring scales among judges. Simultaneously, the standard error of the teacher competency parameter estimates decreased significantly, with split-relief reliability improving to 0.892. This substantially enhanced the stability and credibility of measurement outcomes, effectively mitigating traditional evaluation pitfalls where high scores resulted from lenient judges and low scores from overly difficult tasks. Consequently, teacher competency parameters now accurately reflect their actual "dual-qualified" proficiency, resulting in objective and fair evaluation outcomes.

Furthermore, the study found that the purified data enables the multi-faceted Rasch model to more accurately identify teachers' strengths and weaknesses. For instance, the model can clearly pinpoint deficiencies in core dimensions, such as "technology application and transformation," and provide targeted directions for enhancing teacher competencies. This validates the practical application value of the integrated model in evaluating the capabilities of vocational education faculty.

6.2. Future Work

Future efforts will focus on advancing the model through productization, platform integration, scenario-based application, and intelligent functionality. This will facilitate the transition of the model from a research prototype to a practical tool, ultimately integrating it into the Teacher Development Center platform of vocational undergraduate institutions. Building on the algorithmic logic of this study, the Python prototype system will be enhanced with functional modules, including manual review of anomalous scores, automated import of evaluation data, and intelligent generation and export of diagnostic reports. Simultaneously, we will develop parameter adaptive adjustment capabilities to accommodate the distinct disciplinary characteristics and evaluation dimensions of various vocational undergraduate institutions. This ensures the model can adapt to faculty evaluation needs across disciplines such as mechanical design and manufacturing, e-commerce, and early childhood education, forming a standardized, modular, and user-friendly intelligent evaluation system. The intelligent evaluation system integrates with institutional subsystems for faculty information management, training resource allocation, and performance assessment. This enables automatic data collection and real-time analysis, deeply linking faculty competency evaluations with tailored training recommendations, performance metric setting, and professional development planning. For instance, teachers who show weaknesses in the "Technology Application and Transformation" dimension receive automated suggestions for relevant industry internships or technical training resources, thereby achieving precise alignment between evaluation outcomes and professional growth support.

Funding

1. 2024 Anhui Provincial Department of Education Quality Project Key Project, Research on the Path and Strategy of Professionalization Construction Continuing Education Teachers in Higher Vocational Institutions, No.: 2024JXJYG1046

2. 2025 Anhuivincial Department of Education Humanities and Social Sciences Key Research Project, Research on the Path of Vocational Education Undergraduate Teachers' Capacity Improvement and the Construction of the System， No.: 2025AHGXSK30273

Institutional Review Board Statement

Not applicable.

Declaration of Artificial Intelligence (AI) Tools

The author used ChatGPT solely for language editing and readability improvement. The author reviewed and verified all content and takes full responsibility for the accuracy and integrity of the manuscript.

References

Cao, Y. and Liu, H. (2025). Building vocational teacher teams guided by the spirit of educators: New advances in research and practice regarding the vocational education teacher workforce in 2024. *Chinese Vocational and Technical Education*, (2), 76-83.

Chen, X. (2025). A four-stage approach to dance instruction in higher vocational early childhood education programmes. *Shangwu*, (11), 123-125.

Horyń, C. and Nowak-Brzezińska, A. (2025). Automatic block size optimization in the lof algorithm for efficient anomaly detection. *Applied Soft Computing*, 170(000). doi: 10.1016/j.asoc.2024.112675

Hou, S., Liao, J., Wang, X., Liu, Y., Zhang, Y., and Gu, J. (2024). Evaluation of the implementation effectiveness of the "interdisciplinary + industry mentor" teaching model. *Medical Education Research and Practice*, (4).

Huang, X. (2025). Value deviation risks in technological governance during digital government development and approaches to mitigation. *Journal of Xihua Normal University (Philosophy and Social Sciences Edition)*, 6, 83-93.

Ma, Z., Chen, H., Shang, R., Yang, X., and Xiang, L. (2025). Computer-adaptive testing based on multidimensional cognitive ability pre-screening. *Computer Technology and Development*, 35(12), 183-191.

Meng, Q. and Cao, Y. (2025). Historical value, current challenges, and future prospects of vocational teacher education. *Vocational Education*, 46(19), 6-12.

Rabih, R., Vahdat-Nejad, H., Mansoor, W., and Joloudari, J. H. (2025). Highly accurate anomaly based intrusion detection through integration of the local outlier factor and convolutional neural network. *Scientific Reports*, 15(1). doi: 10.1038/s41598-025-08175-z

Wang, G. (2025). Research on the construction of an evaluation index system for the professional competence of class advisors in secondary vocational schools: Based on the grounded theory research paradigm. *Education Science Forum*, (15), 65-68.

Xu, L., Sheng, G., and Liu, H. (2025). Exploration and practice of a "four-chain coupling" talent cultivation model based on career value orientation: Taking the automotive application and maintenance major as an example. *Anhui Education Research*, (3), 14-16.

Zhu Lei earned her master’s degree in public administration from Anhui University in December 2008. She currently serves as Director of the Human Resources Department (Teaching Affairs Office) at Wuhu Vocational and Technical University, while also teaching at the School of International Economics and Trade. She holds the titles of Associate Professor and Senior Dual-qualified teacher and is a member of the Anhui Higher Education Society. Her accolades include the First Prize for Provincial Teaching Achievements, an Outstanding Case in Provincial Education Evaluation Reform, the Provincial Outstanding Young Talent Support Program, the First Prize in the Provincial Vocational Skills Competition, and commendations from Wuhu City. She has led national demonstration school initiatives and faculty development projects under the “Double High” Plan, and in recent years has participated in over ten national and provincial educational research projects. Her primary research interests focus on human resource management and faculty development in higher vocational education.