Development and Validation of a Multimodal Evaluation Indicator System for Higher Educational Instruction

Yuhang Li *, Huajie Qu, Yinghui Ge, Lili Zhang, Tingting Liu

School of Mechanical and Electrical Engineering, Beijing Polytechnic University, Beijing 100176, China

*liyuhang@bpu.edu.cn

Abstract. Traditional higher education evaluation paradigms suffer from metric fragmentation, superficial data usage, and an inability to dynamically adjust to changing pedagogical contexts. This study introduces a comprehensive framework designed to bridge the gap between raw multimodal classroom data and structured instructional insights. Utilizing a rigorous design process involving the Delphi method and Analytic Hierarchy Process (AHP), we developed a 4-dimensional developmental evaluation matrix featuring 16 secondary indicators and 42 fine-grained multimodal behavioral anchors. To validate this metric configuration, a non-intrusive multi-sensor spatial network was deployed to collect a 120-hour synchronized dataset spanning 20 faculty profiles in authentic higher educational settings, followed by strict multi-pass human expert annotation. Exploratory data analysis reveals significant correlations between instructional demographics and specific behavioral dimensions, while content validity is established through an alignment matrix mapping classroom behaviors directly to industry-demanded skill profiles. This foundational methodology provides a scalable blueprint and gold-standard dataset for building closed-loop, context-aware artificial intelligence evaluation architectures in higher education.

Keywords: Multimodal Evaluation, Higher Educational Instruction, Indicator System, Human-Annotated Dataset.

1 Introduction

With the global acceleration of digital transformation in higher education, leveraging artificial intelligence (AI) to optimize instructional quality has transitioned from a theoretical concept to an operational necessity [1, 2]. Traditional evaluation methodologies, relying heavily on static, end-of-term student evaluation of teaching (SET) questionnaires, are constrained by subjective biases and a lack of formative, actionable feedback loops [3]. While recent advances in computer vision, automatic speech recognition (ASR), and natural language processing (NLP) have enabled automated classroom data collection [4], current systems fail to provide scalable, pedagogically sound interventions [5].

Z. Ge et al. (eds.), *Proceedings of the 2026 6th International Conference on Education, Information Management and Service Science (EIMSS 2026)*, Atlantis Highlights in Computer Sciences 25,
https://doi.org/10.2991/978-94-6239-766-8_29

This failure stems from three critical blindspots in existing educational informatics literature: (1) Metric Fragmentation: Most platforms focus on single-scene or isolated modality tracking (e.g., speech fluency or student hand-raises), lacking a unified framework linking instructional design, real-time spatial interaction, emotional resonance, and digital literacy into a cohesive narrative [6]. (2) Pedagogical Decontextualization: Extant models apply rigid, uniform scoring rules across highly diverse environments, failing to adaptively calibrate metric weights between theoretical lectures and practical laboratory sessions. (3) Absence of High-Fidelity Gold Standards: Deep learning models for educational scene understanding are bottlenecked by a lack of high-quality, domain-specific labeled data within complex, occluded classroom environments [7, 8].

To address these systemic challenges, this paper presents a rigorous framework for developing and validating a multimodal instructional evaluation system [9, 10]. We detail the methodological pipeline used to establish a 4-dimensional, 42-anchor metric matrix via expert consensus and hierarchical weighting. Furthermore, we present the structural engineering behind a 120-hour human-annotated dataset captured across 20 distinct higher education instructors to establish a baseline for future automated, context-aware AI evaluation loops.

2 Development of the Multimodal Evaluation Indicator System

To ensure that the proposed indicator system balances theoretical rigor with empirical computability, a sequential mixed-methods design translated principles of Outcome-Based Education (OBE), Flanders Interaction Analysis, and cognitive load theory into trackable multimodal features.

2.1 The Derivation Process: Delphi Method and AHP

The synthesis of the indicator matrix followed a strict three-stage pipeline:

Step 1 (Initial Item Pool Generation): A comprehensive review of regional quality evaluation standards, institutional guidelines, and academic literature compiled a broad initial pool size of 128 granular behavioral observation items.

Step 2 (The Delphi Consensus Phase): A panel of 15 educational experts and computer science researchers executed a multi-round Delphi consensus study. Items were rated on a 5-point Likert scale. To guarantee robust item retention, strict consensus thresholds were enforced: items failing to meet a mean score threshold of ≥ 4.2 or exhibiting a Coefficient of Variation (CV) > 0.2 were systematically eliminated, condensing the pool into 4 primary dimensions and 16 secondary indicators.

Step 3 (AHP Weighting): The expert panel completed pair-wise comparison matrices to calculate baseline priority weights. The Consistency Ratio ($CR = CI/RI$) for all pair-wise comparison matrices was strictly below 0.1 ($CR < 0.1$), confirming the mathematical validity and logical consistency of the weight calculations.

2.2 Indicator Taxonomy and Behavioral Anchors

The finalized indicator system is organized into four symmetrical macro-dimensions, mapping 16 secondary metrics to 42 computable multimodal behavioral anchors (Figure 1):

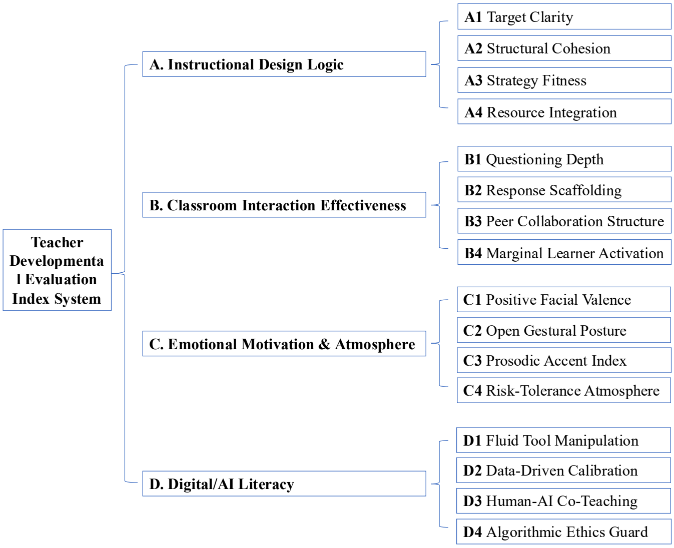

Fig. 1. Comprehensive indicator classification

Dimension A (Instructional Design Logic): Measures structural orchestration, conceptual clarity, and resource alignment. It comprises Target Clarity (semantic matching density of syllabus milestones), Structural Cohesion (topology extraction of slide content via knowledge graphs), Strategy Fitness (instructor-to-student timeline ratios), and Resource Integration (media shift counts).

Dimension B (Classroom Interaction Effectiveness): Captures real-time communication depth and distribution. It includes Questioning Depth (semantic classification of interrogation sentences), Response Scaffolding (acoustic/textual transition mapping), Peer Collaboration Structure (multi-channel audio segmentation), and Marginal Learner Activation (gaze-direction tracking vectors to peripheral zones).

Dimension C (Emotional Motivation & Atmosphere): Quantifies teacher enthusiasm and classroom climate via non-verbal features, including Positive Facial Valence (90-dimensional facial mesh tracking), Open Gestural Posture (kinematic joint velocity), Prosodic Accent Index (F_0 pitch variance), and Risk-Tolerance Atmosphere (temporal tolerance windows during student errors).

Dimension D (Digital/AI Literacy): Tracks technology-enhanced learning orchestration, comprising Fluid Tool Manipulation (system latency tracking), Data-Driven Calibration (dashboard glance frequencies), Human-AI Co-Teaching (tutoring platform telemetry logs), and Algorithmic Ethics Guard (keyword spotting for data privacy boundaries).

3 Dataset Acquisition and Annotation Protocol

To validate the framework, a high-fidelity dataset capturing real-world teaching behaviors was constructed within smart classrooms using a visual tracking array (4K UHD cameras tracking student expressions and instructor kinematics), an acoustic array (ceiling-mounted beamforming microphones), and background telemetry log streamers (Figure 2).

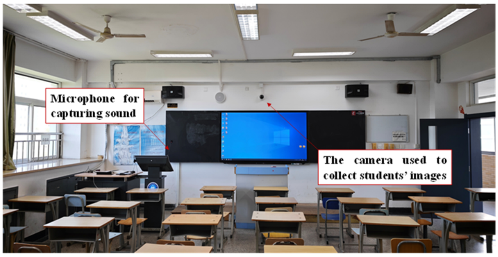

Fig. 2. Classroom collection equipment

The raw streams were pre-processed via automatic speech recognition (ASR) and frame-rate normalization to yield a total corpus of 120 hours. The specified dataset composition spans 20 unique higher education instructors (purposely distributed across full professors, associate professors, and lecturers) across 70 hours of theoretical lectures and 50 hours of practical laboratory sessions, covering basic engineering, general education, and specialized engineering courses.

For the annotation protocol, a team of 6 graduate student coders executed fine-grained sequence tagging (marking timestamps, indicator categories, and intensities) using a custom spatial-temporal labeling tool. To eliminate subjective drift, every hour of classroom data was labeled independently by two blind coders. The human annotation protocol achieved an average Fleiss' Kappa coefficient of $\kappa = 0.84$ across the 6 graduate student coders, indicating excellent statistical agreement and establishing the dataset as a gold standard for algorithmic training.

4 Empirical Verification and Results

To establish the descriptive validity of our metrics, comprehensive statistical analyses were performed directly on the human-annotated gold-standard dataset across three distinct age-based faculty seniority cohorts: Junior (Age ≤ 35, N=7), Mid-Career (Age 36–50, N=8), and Senior (Age > 50, N=5).

4.1 Demographic Factor Analysis: Faculty Seniority vs. Behavioral Focus

As Figure 3 shown, the empirical distribution reveals that Junior teachers exhibit a significantly higher density of anchor B3 (Peer Collaboration Structure) and C1 (Positive

Facial Valence), indicating a preference for high-energy, student-centered group dynamics. Conversely, Senior faculty lean heavily toward anchor A2 (Structural Cohesion) and B1 (Questioning Depth), spending 65% of classroom time in deep, high-order concept-mapping and logical evaluation loops. Under Dimension D (Digital/AI Literacy), Junior and Mid-Career faculty maintained superior technical fluency, demonstrating low operational latencies under anchor D1 (Fluid Tool Manipulation) by averaging less than 3 seconds per system switch, whereas Senior faculty exhibited higher transition latencies but scored higher in D2 (Data-Driven Calibration) by waiting for structural student analytics before adjusting instructional pacing.

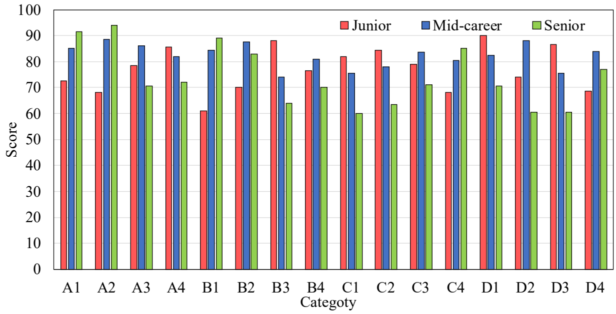

Fig. 3. The behavioral differences of different levels of teaching staff (junior, mid-aged, senior) across various dimensions.

4.2 Content Validity and Industry Alignment

To establish robust content validity for downstream application, a structural alignment matrix correlated the 42 classroom observation anchors against industry-demanded skill profiles gathered from an empirical survey of 2 enterprise technical leads and 25 dual-qualified institutional instructors. The matrix confirmed that behavioral traces under Dimension B map precisely to a "T-shaped" competency demand curve, directly capturing **collaborative task management**, while high-order questioning (B1) and feedback loops (B2) align tightly with **adaptive debugging logic** and **communication clarity under stress**.

5 Conclusion

This paper detailed the development and validation of a multi-modal evaluation framework for higher educational instruction. By combining the Delphi method and AHP (CR < 0.1), we established a balanced, 4-dimensional indicator system validated through a 120-hour gold-standard dataset (κ= 0.84). Exploratory analysis confirmed that the framework successfully captures distinct multi-generational behavioral patterns while maintaining explicit alignment with workplace competency needs, providing a scalable foundation to transition higher education evaluation into a responsive, data-driven paradigm.

Acknowledgments

This project was supported by General Project, Beijing Education Sciences Planning Program during the 14th Five-Year Plan (CDGB25546).

Disclosure of Interests

The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Chi Zhou, Fenghua Shao, Wenqi Cai, and Min Chen. 2025. Classroom Quality Evaluation in Higher Education: A Multimodal Learning Analytics Approach. In Proceedings of the 2025 International Conference on Intelligent Education (ICIE '25). IEEE, 1–10.
2. Man C. E. Chan, Xavier Ochoa and David Clarke. (2020). Multimodal Learning Analytics in a Laboratory Classroom. In: Virvou, M., Alepis, E., Tsihrintzis, G., Jain, L. (eds) Machine Learning Paradigms. Intelligent Systems Reference Library, vol 158. Springer, Cham. https://doi.org/10.1007/978-3-030-13743-4_8
3. Jewoong Moon, Sheunghyun Yeo, Seyyed Kazem Banihashem, and Omid Noroozi. 2024. Using multimodal learning analytics as a formative assessment tool: Exploring collaborative dynamics in mathematics teacher education. Journal of Computer Assisted Learning 40, 6 (Dec. 2024), 2753–2771. https://doi.org/10.1111/jcal.13028
4. Ting Cai, Yu Xiong, Chengyang He, and Lulu Chen. 2026. TBRNet: A Multi-Modal Network for Teacher Behavior Recognition with Cascaded Collaborative Attention and Dynamic Query-Driven. Electronics 15, 2 (Jan. 2026), 460. https://doi.org/10.3390/electronics15020460
5. Michail Giannakos and Mutlu Cukurova. 2023. The role of learning theory in multimodal learning analytics. British Journal of Educational Technology 54, 5 (April. 2023), 1247–1270. https://doi.org/10.1111/bjet.13320
6. Xinrui Ma and Jingxia Chen. 2024. Analysis of Classroom Teaching Behaviors Based on Multimodal Data Model. Journal of Intelligence and Knowledge Engineering 2, 4 (Dec. 2024), 34–42. https://doi.org/10.62517/jike.202404407
7. Alvaro Becerra, Ruth Cobos, and Roberto Daza. 2026. A Multimodal Dataset of Student Oral Presentations with Sensors and Evaluation Data. arXiv preprint arXiv:2601.07576 (Jan. 2026). Science Data Bank DOI: https://doi.org/10.57760/sciencedb.33655
8. Chenglei Huang, Jia Zhu, Yilong Ji, Weijie Shi, Min Yang, Hanghui Guo, Jianxia Ling, Pasquale De Meo, Zilong Li, and Zhangze Chen. 2025. A Multi-Modal Dataset for Teacher Behavior Analysis in Offline Classrooms. Scientific Data 12, 1115 (July 2025). https://doi.org/10.1038/s41597-025-05426-6
9. Lunan Fan, Guiping Cheng and Na Yan. 2025. Assessing local cultural awareness in university EFL learners: A Delphi and AHP-based index framework. PLOS ONE 20(10): e0332233. https://doi.org/10.1371/journal.pone.0332233
10. Tiantian Zhao, Long Yu, Xinyue Jiang, Shaoxiong Yang, Xuehui Zhang, Min Zhang, Yan Zhang, Tong Yan, Huixin Wu and Fang Xu. 2025. Construction of a public health practice teaching quality evaluation system based on the CIPP model using the Analytic Hierarchy Process (AHP) and entropy weight method. BMC Medical Education, 25, 368. https://doi.org/10.1186/s12909-025-06943-z