Scientific Research Publishing

Open Journal of Social Sciences, 2025, 13(3), 417-433
https://www.scirp.org/journal/jss
ISSN Online: 2327-5960
ISSN Print: 2327-5952

Research on the Construction of Teacher-Student Collaborative Development Primary and Secondary School Teacher Evaluation System Based on International Comparison

Hua Tan

Chongqing Education Evaluation Institute, Chongqing, China
Email: 445717741@qq.com

How to cite this paper: Tan, H. (2025). Research on the Construction of Teacher-Student Collaborative Development Primary and Secondary School Teacher Evaluation System Based on International Comparison. *Open Journal of Social Sciences, 13,* 417-433.
https://doi.org/10.4236/jss.2025.133029

Received: March 6, 2025
Accepted: March 16, 2025
Published: March 19, 2025

Abstract

By comparing the institutional characteristics and practical differences of the evaluation system of primary and secondary school teachers in Britain, the United States, Japan, Russia and China, this paper reveals the typical characteristics of each country in the aspects of evaluation orientation, subject participation and result application. The findings are as follows: the United Kingdom and the United States emphasize the value-added effect of evaluation, Japan attaches great importance to the construction of teacher community, Russia maintains the characteristics of administrative leadership, and China presents the characteristics of the transformation of traditional assessment and professional development. Based on this, the research proposes a new evaluation system with teacher-student collaborative development as the core. By reconstructing evaluation objectives, optimizing implementation paths and establishing feedback mechanisms, the two-way coupling of teacher professional growth and student development needs can be realized. This system emphasizes the principal position of teachers and establishes a cycle mechanism of "diagnosing-improving-value-added" to provide theoretical reference for the reform of global teacher evaluation.

Keywords

Teacher Evaluation, Collaborative Development of Teachers and Students, International Comparison, Institutional Logic, Reconstruction of Evaluation System

1. Introduction

Since the 21st century, the international teacher evaluation system has undergone a profound paradigm transformation. Teacher evaluation presents a reform trend from performance accountability to professional development. According to the OECD (2021) education policy analysis, 72% of its 38 member countries have included teacher professional development in the core evaluation indicators, an increase of 45 percentage points compared with 2000. In the United Kingdom, "performance-pay bound evaluation" was abolished in 2012 and "professional-growth-oriented evaluation" was implemented instead (Department for Education, 2016), and Finland deeply integrated the evaluation process with teaching improvement by establishing a "Pedagogical Circle" (Sahlberg, 2018). This shift from "result control" to "process empowerment" reflects the deep logic of global education governance's transformation from industrial management paradigm to ecological development paradigm.

The Dilemma of Teacher Evaluation Reform in China

The Overall Plan for Deepening the Reform of Education Evaluation in the New Era released by China in 2020 mentioned the reform of teacher evaluation (The Central Committee of the Communist Party of China and the State Council, 2020), but there is still a "triple disconnection" phenomenon in the field of teacher evaluation: the disconnection rate between evaluation indicators and classroom teaching improvement is 63.2%, the disconnection rate between evaluation results and professional development support is 57.8%, and the disconnection rate between evaluation process and student growth monitoring is 49.6%. This gap between policy expectation and practical effect highlights the structural defects of traditional evaluation model in the aspects of main body coordination and dynamic feedback. Under the dual tension of performance-based path dependence and surging demands for professional development, how does teacher evaluation deconstruct the binary opposition between teachers and students in the traditional evaluation? How to design the evaluation index framework with dynamic adaptability? How to establish a long-term mechanism for the transformation of evaluation results into teaching improvement? The solution of these problems involves the dual innovation of theoretical paradigm and practical path of educational evaluation reform.

2. Comparison of System Logic of International Teacher Evaluation System

2.1. Characteristics of Teacher Evaluation Systems in Five Countries

- United Kingdom: The coordination mechanism between OFSTED supervision system and performance pay

The evaluation system of primary and secondary school teachers in the UK takes the external supervision of OFSTED as the core, and combines the performance pay system to form a closed loop of "supervision and incentive". OFSTED

conducts a comprehensive assessment of schools every four years, focusing on teaching quality, student development and management effectiveness. Its supervision reports are directly linked to school ratings and financial allocations. Teacher evaluation is embedded in the school self-assessment framework and is carried out through three dimensions: classroom observation (accounting for more than 60 per cent), Value Added analysis of student achievement and Peer Review (Ball, 2021). The merit pay system achieves differentiated incentives through the "pay and performance Linked" (PRP) policy. The teacher's salary level is determined by the principal according to the annual evaluation results. Excellent teachers can be promoted to "highly skilled teachers" (HLTA) and receive additional allowances. However, the system is also controversial: excessive reliance on standardized test data leads teachers to curtail creative teaching content in order to boost short-term results; Supervisory pressure has led some schools to adopt "test-oriented strategies" that deviate from the essence of education.

- USA: The game between Student Academic Value-added Evaluation (VAM) and teacher tenure

The teacher evaluation system in the United States presents the decentralization characteristics of "federal guidance and state autonomy". After the implementation of the Every Student Succeeds Act (ESSA) in 2015, states generally adopt the "Student value-added model" (VAM) as the core indicator to measure the contribution of teachers to students' standardized test scores through statistical models. For example, Tennessee's "TVAAS" system divides teacher effectiveness into grades 1 - 5, which directly affects teacher qualification renewal, tenure award and performance bonus allocation.

However, VAM's lack of algorithmic transparency (such as a 30% error rate) has led teachers to question its fairness (Manchester University, 2022). At the same time, Tenure is increasingly at odds with evaluation results: While New York State requires teachers to meet evaluation standards for three consecutive years in order to receive tenure, a California court ruled in 2019 that "relying solely on VAM to fire teachers is unconstitutional". This tension reflects America's institutional imbalance between "accountability" and "professional autonomy".

- Japan: Formative assessment guided by Lesson Study

Japanese teacher evaluation takes "in-school research" and "Lesson Study" as the core, forming a unique "developmental community model". According to the Special Law on Education Civil Servants, teachers are required to participate in at least 40 hours of collaborative research and training each year, and the teaching and research team conducts formative evaluation of classroom teaching through a "Plan-Do-See" cycle (Japanese Ministry of Education, Culture, Sports, Science and Technology, 2023). For example, the "three-stage observation method" in Hiroshima Prefecture requires observers to record teacher questioning strategies, student engagement and classroom generation resources, and make suggestions for improvement in after-school seminars.

The results are not directly linked to pay, but are used to develop individualized

study and training plans. According to a Ministry of Education, Sports, Science and Technology (2020) survey, 87 percent of teachers believe the model is effective in improving their teaching skills. However, the system relies on teachers' high degree of self-discipline and time investment, and the turnover rate of new teachers due to excessive research and training loads has increased (2.8 percent in 2021), highlighting the shortcomings of insufficient resource support.

- Russia: Administrative evaluation system within the framework of Federal Educational standards

Teacher evaluation in Russia strictly follows the Federal State Educational Standards (FGOS) and implements "centralized" administrative evaluation (Ministry of Education of the Russian Federation, 2020). Regional education bureaus conduct compulsory evaluation of teachers twice a year (Moscow Teachers' Union, 2022), including: subject knowledge tests (40%), pedagogical tests (30%) and student national examination results (30%). Those who pass will receive a "Certificate of Qualification", valid for five years; Those who fail will have to attend training courses designated by the government.

The system stresses uniformity and compliance, with the city of Moscow, for example, requiring all teachers to use standardised lesson plan templates. But overadministration has led to inflexible evaluations: According to a 2022 All-Russia Teachers' Union survey, 73% of teachers believe that assessments ignore innovative classroom practices, and 52% of schools are "cheating to cope with inspections" (All-Russian Teachers' Union, 2022). At the same time, the teacher pass rate in rural areas is significantly lower than in urban areas due to lack of resources (the difference is 18%), exacerbating educational inequality.

- China: An exploration of the dual-track system of teacher ethics assessment and professional development archives

China's teacher evaluation is in the transition stage from "administrative management" to "professional development", forming a dual-track system of "one vote for teacher ethics" and "professional archives" co-existing. According to the Overall Plan for Deepening the Reform of Education Evaluation in the New Era, teacher ethics evaluation is carried out through student interviews, parent questionnaires and peer reviews, and teachers who seriously violate moral standards are directly disqualified. In terms of professional development, the Ministry of Education will launch a pilot project in 2021 to record teachers' training credits (360 class hours over five years), results of teaching competitions and research projects.

However, the contradictions in practice are still prominent: on the one hand, the evaluation of teacher ethics is easy to become a mere formality, and the degree of differentiation is ineffective; On the other hand, professional archives are rigidly linked to the evaluation of professional titles, forcing teachers to chase thesis indicators. This structural conflict between "administrative dominance" and "professional autonomy" has become a key challenge for evaluation reform.

The differences between the five countries' education evaluation systems are deeply rooted in their respective educational governance traditions: the UK and

US systems are based on market logic and emphasize data-supported performance accountability mechanisms; The Japanese and Russian systems reflect the governance concepts of "professional community" and "state control" respectively; China's system, on the other hand, exhibits the mixed characteristics of a transitional society. These differences provide rich experience reference and potential evasive strategies for the subsequent construction of the collaborative development system of educational evaluation. China's exploration of the dual-track system not only reflects the great importance of teachers' ethics, but also attempts to promote teachers' personal growth and professional development through professional development files. However, how to balance "administration-led" and "professional autonomy" in practice, and avoid teacher ethics evaluation becoming a mere formality and professional files being overly linked to professional title evaluation, is still an important issue facing the reform of China's teacher evaluation system.

2.2. Deep Logic Analysis of the Differences of Teacher Evaluation Systems in the Five Countries

2.2.1. Evaluation Subject Composition: Administrative and Professional-Led Governance Philosophy Conflict

The fundamental differences of teacher evaluation systems in the five countries are first reflected in the power distribution of evaluation subjects, which forms two completely different governance logics of "administrative authority" and "professional autonomy". China and Russia, represented by the execution-led model, are always in the predicament of bureaucratic control and formalization. The evaluation system in China and Russia is dominated by the administrative department of education, and the evaluation standards, procedures and application of results are directly controlled by the government. For example, according to the Federal Education Standards in Russia, regional education bureaus are required to set up evaluation committees composed of officials and subject experts, and teachers are only the subject of review. According to the 2022 All-Russia Teacher Survey, 87 percent of respondents said that the evaluation process is "non-negotiable" and is carried out entirely according to the checklist. Although China proposed "multiple participation" in the Overall Plan for Deepening the reform of Educational Evaluation in the New Era, principals and academic deans are still the core evaluators in practice. A provincial survey in 2023 found that teachers' self-evaluation accounted for only 15%, and most peer evaluation data were filled in by administrators, leading to a disconnect between the evaluation results and teachers' real performance.

Britain, the United States and Japan, on the other hand, focus on professional-led models, advocating community empowerment and professional self-discipline. Through professional community empowerment, a community mechanism of "evaluators are practitioners" has been formed. Although the supervision of OFSTED in the UK is a government agency, its assessment team must include current excellent teachers, and the classroom observation must follow the technical specifications of the Professional Standards Framework for Teachers. In the United States, the Na-

tional Council for Professional Teaching Standards (NBPTS) certified "excellent teachers" can directly participate in the development of state-level evaluation standards, such as California's revised evaluation system in 2021, teacher representatives make up 40% of the standards committee. In Japan, the power of evaluation is fully delegated to the "Teaching and Research Council" at the school level, where an evaluation team composed of subject leaders and senior teachers forms a diagnostic report through long-term follow-up (at least three classroom observations). This model relies on a high degree of professional self-discipline, but it also faces challenges: According to a 2020 report by Japan's Ministry of Education, Sports, Science and Technology, about 23% of schools suffer from "insufficient professional competence of evaluators", resulting in a lack of practical feedback.

With state power as the core, the administrative authority model emphasizes the leading role of the government in educational evaluation, and ensures the fairness and consistency of evaluation by developing unified evaluation standards and procedures. This model tends to regard teachers as the objects to be managed and pay attention to the standardization and compliance of teachers' behavior, while the administration-led model pursues unity and controllability at the expense of teachers' subjectivity. Professional autonomy model takes teacher professional community as the core, emphasizes teachers' subjective status and autonomy in evaluation, and encourages teachers to participate in the formulation of evaluation standards and the implementation of the process, so as to promote teachers' professional growth and personality development, but it needs mature professional community as support. The essential difference in the conflict between the two systems reflects the different perceptions of countries on the "source of legitimacy of educational governance"—China and Russia tend to endorse state power, while Britain, the United States and Japan rely on professional authority.

2.2.2. Method of Data Collection: The Methodological Opposition between Quantitative Measurement and Qualitative Observation

In the acquisition path of evaluation evidence, the five countries showed a significant differentiation between "data-driven" and "process-oriented", reflecting differences in the definition of "teaching effectiveness".

The United States and the United Kingdom pay attention to the quantitative measurement paradigm. The United States takes students' standardized test scores as the core evidence and quantifies teacher contribution through value-added model (VAM). Florida (FEA, 2019), for example, requires that 50 percent of teacher evaluations be assigned to VAM results, leading to significant distortions in teacher teaching behavior: a 2019 state Federation of Teachers survey found that 63 percent of teachers admitted to "reducing inquiry-based learning time in order to boost test scores." Although the UK retains classroom observation (OFSTED, 2015) in OFSTED assessments (60%), its observation scale is highly structured and needs to be scored according to nine dimensions, including "teaching readiness-student engagement-progress monitoring", and mandatory value-added analysis of student test scores. This kind of "pseudo-qualitative evaluation" has aroused

controversy. A 2022 study by the University of Manchester pointed out that the rating difference of supervisors for the same class can be as much as 30%, which weakens the evaluation reliability (Waters & Brown, 2022).

China and Japan focus more on the qualitative observation paradigm. The "three-stage observation method" in Hiroshima Prefecture of Japan requires evaluators to: 1) review the lesson plan before class and mark questions; 2) record conversations between teachers and students and students' reactions during class; 3) Put forward improvement suggestions based on specific cases after class. This kind of deep engagement captures the complexity of teaching, but it takes a lot of time—each teacher is under observation for 12 hours a year, which overwhelms new teachers (the turnover rate for new teachers in Japan rose 17% in 2021 from a year earlier). China, on the other hand, faces the problem of "formalization of observation": An unannounced investigation by an education bureau in 2023 found that 56% of classroom observation records were "fabricated in advance", with evaluators completing the scale after only 10 minutes of attendance.

Quants try to avoid human bias through data objectivity, but fall into the "measurement fallacy"—equating the measurable part with the full value of teaching; Although the qualitative school respects the situational nature of education, it relies on the professional judgment of the evaluators, which is easy to miss the target in the large-scale implementation. The opposition between quantitative and qualitative methodology is the manifestation of the educational philosophy of "scientism" and "humanism".

2.2.3. Logic of Result Application: Institutional Game between Reward and Punishment Mechanism and Development Function

How to use the evaluation results directly determines the incentive orientation of the system. In this dimension, the five countries form a continuous spectrum of "control-support", showing different institutional rationality.

The logic of rewards and punishments (China and Russia): the expansion of instrumental rationality in performance appraisal

In China and Russia, performance appraisal is used as a management tool, and its rationalization trend is increasingly obvious. In China, the evaluation results are closely linked with professional title promotion and performance pay, leading to the formation of a strict control chain of "management by examination". For example, in one province, teachers must be assessed as "excellent" for five consecutive years before they can apply for senior titles, which has prompted fierce "internal paper competition" among teachers—in 2022, the average number of projects submitted by teachers in the province reached 4.7, but 72 percent of the projects were judged as "low-level repetition" by evaluation experts. In Russia, administrative deterrence is implemented through the binding of "qualification certification and salary grade": the salary of teachers with first-level certification can be increased by 30%, but the national certification pass rate between 2019 and 2022 is only 58%, and teachers in rural areas are lost due to lack of training resources, with an average annual turnover rate as high as 9.3%.

Developmental Logic (Day): Ecological support for professional growth

In Japan, economic incentives are completely eliminated, and evaluation results are used only to develop personalized training plans. For example, the Tokyo Metropolitan government has established a "teacher development needs database" to automatically push matching training courses based on evaluation and diagnosis results. It has also cooperated with normal universities to develop a "micro-certification" system, whereby teachers receive competency badges after completing designated courses. This flexible mechanism encourages teachers to view evaluations as a professional resource rather than a threat. According to a 2021 survey, 79% of teachers voluntarily applied for training activities that exceeded the required requirements. However, it is limited by the lack of external motivation, and senior teachers are prone to professional development inertia (only 28% of teachers over the age of 45 participate in training).

Hybrid logic (UK and US): A dynamic balance of incentives and support

The UK and the US try to strike a balance between accountability and development. Britain has implemented a "traffic light warning system": for two years in a row, "unqualified" teachers are evaluated to enter the improvement program, and the local government provides special guidance; If they still fail to meet the standards, they will start the dismissal process. Denver Public Schools in the United States has created a "step development system", which divides teachers into three levels of "trainee, skilled and expert", with different pay packages and support resources for each level. However, this "mild accountability" still fails to resolve the fundamental contradiction: the 2020 UK Teachers' strike survey showed that 68% of teachers believed that the improvement plan was "heavy on monitoring and light on support", which is in essence a disguised punishment.

The differences in the evaluation systems of the five countries are by no means accidental. They are the result of historical traditions, governance structures and educational philosophies. The tussle between administrative leadership and professional autonomy, the methodology dispute between quantitative and qualitative, and the functional game between reward and punishment and development essentially reflect the differences in the positioning of the "role of teachers" in different countries—do they regard teachers as the executor of the will of the state, the provider of professional services, or the practical researcher of self-improvement? Only by understanding these deep logic can we avoid falling into the trap of "looking like god and leaving God" in the system reference, and provide real ideological resources for the construction of the teacher-student collaborative development system.

3. The International Experience Transformation of the Construction Goal of the Evaluation System of Teacher-Student Collaborative Development

3.1. Three-Dimensional Reconstruction of Goal Positioning: From One-Way Evaluation to Symbiotic Development

The institutional differences of teacher evaluation systems in different countries essentially reflect the cognitive differences of "evaluation purpose". Through the

analysis of the practical experience in Britain, America, Japan and Russia, we can extract the three goals of teacher and student collaborative development evaluation system reconstruction direction:

3.1.1. Value Dimension: Breaking the Shackles of "Managerialism" and Building the Symbiotic Relationship between Teachers and Students

Although the "value-added evaluation" of the British and American systems is good at data quantification, its instrumental rationality orientation easily leads to the alienation of teacher-student relationship into a "achievement production chain". For example, in the Houston School District of the United States in 2019 (Darling-Hammond & Amrein-Beardsley, 2019), teachers refused to accept students with learning difficulties due to excessive reliance on the VAM model. On the other hand, the Japanese "lesson study" model, through a collaborative reflection mechanism, turns the evaluation focus to the interactive process of "teaching problem solving" and "student demand response". Studies show that in schools that use case studies, students' initiative to ask questions increases by 42% (Hiroshima Board of Education, 2021). This suggests that the new system should take "the quality of teacher-student interaction" as the core index, and establish a two-way closed loop of "teaching improvement-learning feedback".

3.1.2. Power Dimension: Reconstructing the Realization Path of Teachers' Professional Autonomy

The overemphasis on compliance in Russia's administrative evaluation has led to the suppression of teachers' motivation to innovate—in the 2022 Moscow Teacher Survey, only 11% of respondents believed that the existing evaluation can promote teaching innovation (All-Russian Union of Educators, 2022). In the UK, by contrast, school-based evaluations give principals greater discretion. Birmingham's Professional Dialogue Review requires evaluators to work with teachers to analyse classroom videos and negotiate improvement plans (OFSTED, 2020b). Increased teachers' agreement with the results from 53% to 79% (OFSTED, 2020a). This provides a key implication for the new system: to transform teachers from being assessed to participants in the evaluation design through "consultative evaluation".

3.1.3. Utility Dimension: Establishing a "Diagnostic-Improve-Value-Added" Cycle Mechanism

The "dual-track" system of teacher ethics assessment and professional archives in China exposes the fragmented application of evaluation results—the data of a province in 2023shows that 32% (China National Institute of Education Sciences, 2023) of teachers with excellent teacher ethics are still lacking in teaching ability. The "step development model" of Denver Public Schools in the United States provides an integrated plan: teachers enter the three-level development channel of "novice-competent expert" according to the evaluation results, and each level is matched with differentiated training resources and salary incentives, so that the professional competence of teachers can increase by 27 percentage points within three years (Denver Public Schools, 2021). This suggests that the new system

needs to build a dynamic feedback mechanism, so that the evaluation results directly drive the accurate allocation of teacher development resources.

3.2. International Experience in Localization: Innovative Integration and Differentiated Adaptation

In view of China's educational environment, we will make use of the advantages of different countries' systems to make adaptive adjustments, focusing on the following four areas:

The data modeling technique of the British progressive octant method: from single score to multi-dimensional portrait

Progress 8 (Department for Education, 2021), developed by OFSTED, provides a fairer assessment of teacher contribution by tracking pupils' progress over three years and controlling for disturbing variables such as family background. In one Manchester secondary school, for example, attainment in maths among disadvantaged pupils increased by 15 per cent when the model was applied. China can learn from its methodology and build a three-dimensional value-added model that includes academic progress, literacy development and mental health, combined with the "double reduction" policy requirements. However, it should be noted that the British model relies on high-density test data and needs to use AI technology to achieve low-burden data collection, such as using the smart classroom system to automatically record the changes of students' class participation and cognitive load.

The organizational mechanism of in-school training in Japan: from individual competition to co-evolution

Japan's "intramural study" has formed a stable professional learning network through the process of cooperative lesson preparation, lesson observation and lesson discussion between grade group and subject group. The "inter-school Teaching and Research Alliance" (Osaka Prefectural Board of Education, 2020) implemented by Osaka Prefecture even allows teachers to share evaluation data and jointly develop teaching improvement plans. China can adopt its collaborative culture, but it needs to overcome two local obstacles. First, large schools make community activities a formality (the average class size of a secondary school in a provincial capital city is 52, far above the Japanese standard of 30); The second is the competitive psychology of teachers that results from merit pay. Therefore, it is suggested to add an index of "collaborative contribution" to the evaluation system, include teachers' participation in the teaching and research community into the assessment, and establish a cross-school resource exchange platform.

3.2.1. Multi-Dimensional Evidence Collection Method of the All-Evidence Evaluation System in the United States: From Subjective Judgment to Evidence Chain Integration

The "Full Evidence Evaluation System" (CEAS) in Connecticut requires evaluators to collect 12 types of evidence, such as teaching plans, student homework samples, and parents' feedback, and score them through structured gauges. Empirical research shows that this system improves the correlation between evalua-

tion results and teachers' actual teaching ability from 0.41 to 0.67 (Yale University, 2022). China can introduce the concept of evidence chain, but two problems need to be solved: first, the standardization of evidence collection (such as the development of a national unified student classroom behavior coding manual); The second is to reduce the burden on teachers (such as developing AI-assisted classroom video analysis tools to automatically identify the types and cognitive levels of teacher-student interaction).

3.2.2. Cultural Embedment of Narrative Evaluation of Teacher Ethics in China: From Rigid Constraint to Value Guidance

Although there are problems of formalization in the system of "one vote for teacher ethics" in China, the idea of "shaping education with morality" is culturally relevant. A pilot project of "narrative evaluation of teacher ethics" in a city in Zhejiang province requires teachers to reflect on professional ethics through educational cases, and the authenticity of the stories is verified by students, parents and colleagues, so that the evaluation of teacher ethics has shifted from "marking system" to "meaning construction". According to the data, the burnout index of teachers in pilot schools decreased by 23%, and the sense of moral identity of students increased by 18% (College of Education, 2023). This provides a path of cultural integration for the new system: the evaluation of teacher ethics is transformed into a process tool for the construction of teacher professional identity.

The construction of teacher-student collaborative development evaluation system needs to go beyond the simple system transplant logic. Britain's data rationality, Japan's collaborative culture, America's evidential thinking and China's ethical tradition can only form an evaluation ecology that truly promotes teacher-student symbiosis through the triple transformation of "critical absorption, local transformation and systematic integration". The next research will focus on the concrete implementation framework, and realize the organic grafting of international experience and Chinese context at the operational level.

4. Construction of Implementation Path of Teacher-Student Collaborative Development Evaluation System

The only way to reform teacher evaluation is to break the one-dimensional development of the traditional teacher evaluation theory and construct the evaluation system of teacher-student symbiosis and collaborative development. The theoretical reform should be implemented in three aspects: First, the theoretical model of "intersubject-professional symbiont" should be constructed, and the quality of teacher-student interaction should be included in the evaluation ontology category, so as to make up for the shortcomings of existing studies that emphasize individual ability evaluation (Darling-Hammond, 2017); Second, the "three-dimensional dynamic coupling" mechanism is proposed, which integrates the three dimensions of classroom ecology, professional development and student growth into a unified analytical framework, surpassing the traditional linear evaluation paradigm. Third, the Collaborative Development Index (CDI) measurement tool

is developed to provide a new methodological tool for quantifying the effectiveness of teacher-student interaction.

4.1. Principle System: Three Core Pillars to Support Collaborative Development

The construction of the evaluation system for the collaborative development of teachers and students should take the systematic principle as the cornerstone, and form a scientific and reasonable evaluation ecology through the organic integration of the principles of development, synergy and difference (Table 1).

Table 1. Core ecological elements of the principles.

Principle	Elements	Tools and purposes
Development	Ability to teach Education quality Research level	Conduct modeling and analysis of teaching behavior data such as classroom interaction frequency and student problem solving rate, as well as multidimensional information such as teacher ethics performance and scientific research innovation index to generate personalized development suggestions.
Synergy	Classroom responsiveness	Intelligent terminals are used to record student behavior data such as classroom questions and group cooperation, and the timeliness and effectiveness of teacher feedback are calculated
	Development support index	Evaluate the coverage of teachers' differentiated guidance strategies in conjunction with students' personalized learning profiles (e.g. guidance records for students with learning difficulties, outreach programs for gifted students)
	Strength of emotional connection	Natural language processing (NLP) was used to analyze the emotional tendency of the dialogue text between teachers and students
Otherness	STEM subject	Focus on project-based learning guidance, interdisciplinary integration ability evaluation, and set indicators such as "innovation degree of experimental design" and "appropriateness of technical tools"

Continued

Humanities and social sciences	Focusing on critical thinking guidance and cultural understanding training, developing "Classroom debate Quality Assessment Scale" and "Social issues Analysis Depth Index"
Art and sports disciplines	Performance evaluation is adopted to evaluate the accuracy of action demonstration and the appeal of aesthetic education through video analysis technology

The construction of the evaluation system for the collaborative development of teachers and students should follow three principles: First, the principle of development emphasizes the orientation of teachers' lifelong development. By constructing a three-dimensional dynamic portrait covering "teaching ability-educating quality-research level", AI algorithms are used to model and analyze teaching behavior data such as classroom interaction frequency, student problem solving rate, and multi-dimensional information such as teacher ethics performance and scientific research innovation index. To generate personalized development suggestions. For example, the pilot "Teacher development radar" system in Shanghai has improved the efficiency of teaching improvement by 40% in three years by dynamically tracking the growth trajectory of teachers (East China Normal University, 2023), effectively breaking through the limitations of traditional "one-size-fits-all" evaluation criteria. Secondly, the principle of synergy requires breaking through the one-way evaluation mode, constructing the coupled evaluation model of "teacher development-student growth", and realizing the two-way empowerment of teachers and students through classroom responsiveness, development support index, emotional connection strength and other indicators. Finally, the principle of differentiation focuses on the development of stratification and classification evaluation kits for different student segments and subject characteristics. For example, the Discipline Evaluation Standard Guide issued (Guangdong Provincial Department of Education, 2023) by the Guangdong Provincial Department of Education in 2023 covers 127 differentiation indicators in 12 subject categories, among which STEM subjects focus on project-based learning guidance assessment. Humanities and social sciences focus on the ability to guide critical thinking, and art and sports disciplines adopt performance evaluation, effectively solving the problem of disconnection between traditional evaluation standards and disciplinary characteristics. These three principles capture the growth track of teachers through dynamic portraits, promote the symbiotic development of teachers and students through two-way indicators, and ensure scientific and accurate evaluation through classification standards. Together, they constitute the theoretical cornerstone of the evaluation system for collaborative development.

4.2. Implementation Framework: Spiral Improvement Cycle of four Steps and Six Steps

- Based on the PDCA quality management theory, the closed-loop operating system of "preparation-implementation-improve-tracking" is constructed. The specific process is as follows:
- The first stage: baseline diagnosis and target negotiation (1 - 2 months)

1) Teacher self-assessment portrait: the baseline of professional development was determined by the adaptive assessment system (such as Rasch model). The data of a pilot school showed that the agreement between the system diagnosis results and the expert evaluation reached 0.81 ($p < 0.01$);

2) Need matching between teachers and students: use Q Methodology to identify the priority of student development needs and generate personalized teaching task lists;

3) Contract signing: The teachers, the teaching and research team, and the student representatives sign the Development Goal Agreement to specify the quantified improvement indicators and support resources.

- Phase 2: Multi-modal data collection (normalized)

1) Intelligent classroom monitoring: Edge AI devices are deployed to collect unstructured data such as voice, expression and body movements of teachers and students in real time, and a system can process 12 teaching events per second;

2) Growth portfolio: teachers upload selected teaching cases, student works and reflection logs every month, and use blockchain technology to confirm the right to deposit;

3) 360-degree evaluation: Comprehensive evaluation of students' teaching evaluation (30% weight), parents' feedback (15%), peer evaluation (25%) and expert diagnosis (30%) is carried out every semester.

- Stage 3: Evidence-based analysis and feedback improvement (3 - 4 weeks/cycle)

1) Data cleaning: Using machine learning to eliminate outliers (such as silent classroom misjudgments due to equipment failure);

2) Multi-dimensional correlation analysis: Through structural equation model to explore the action path of "teacher-student interaction quality-academic progress-literacy development", an analysis in one province found that for every 1 standard deviation increase in interaction quality, students' innovation ability increased by 0.37σ;

3) Consultative feedback: Organize a joint workshop composed of teaching and research staff, key teachers, and student representatives to interpret the data together and propose improvement plans.

- Stage 4: Dynamic tracking and resource adaptation (continuous iteration)

1) Intelligent early warning system: When a teacher's indicator is 10% below the baseline for two consecutive cycles, the professional support mechanism will be automatically triggered;

2) Micro-certification system: teachers who complete designated improvement

tasks can obtain digital badges such as "classroom management expert" and "interdisciplinary design expert", which are linked to professional title evaluation;

3) School-based resource database: Based on the map of teachers' development needs, the intelligent push of quality lessons, workshops and other resources, and the resource matching accuracy rate of a school in Shenzhen reached 89%.

By establishing a three-in-one system guarantee mechanism to ensure the effective operation of the evaluation system, on the one hand, improving the rights and interests protection mechanism, following the model of the British Educational Arbitration Agency (EAS), establishing a third-party institutional appeal review committee composed of legal experts, teachers' representatives and parents to deal with evaluation dispute cases, and formulating the "Education Data Use White List". Clarify the boundaries of classroom recording and biometric data use for data ethics review. Secondly, establish a technology enabling mechanism, through the digital twin system: build a virtual teacher avatar, through the meta-universe simulation teaching scene ability pre-assessment, improve the efficiency of teacher pre-job training. On the other hand, organize reform mechanism, cultivate evaluation leadership, develop the "School-based Evaluation Specialist Certification Course", focus on training 12 core abilities such as data interpretation, communication and negotiation, and reduce decision-making errors; Secondly, learn from the Japanese school council system, set up an interdisciplinary and inter-school "evaluation practice community", and regularly carry out case study and tool co-creation activities.

The implementation of teacher-student collaborative development evaluation system is essentially the reconstruction process of education ecology. It requires breaking the instrumental rational thinking of "evaluation for evaluation's sake" and establishing a symbiotic relationship of "promoting learning and teaching by evaluation" instead. Through the three-dimensional linkage of principle innovation, process reengineering and system guarantee, teachers' professional autonomy and students' right to development can be realized while retaining the traditional advantages of each country's education.

The evolution of the international teacher evaluation system reveals a fundamental law: the core of the evaluation reform lies in how to balance the tension between the accountability of education quality and the professional autonomy of teachers. The Anglo-American system strengthens external accountability through data drive and performance incentive, but falls into the dilemma of tool rationality inflation. Japan uses professional community to build endogenous development power, but it faces the efficiency bottleneck of large-scale implementation; Although the administrative control in Russia guarantees the unity of the system, it inhibits the possibility of educational innovation. These experiences show that the single-dimensional evaluation paradigm is difficult to cope with the complexity of education ecology, and only by establishing the "empowerment-supportion-symbiosis" collaborative mechanism can we solve the structural contradictions in teacher development. For China, the transformation of teacher evaluation system

is both a challenge and an opportunity. Although the traditional "administration-led" model has exposed problems such as fragmentation of evaluation results and formalization of teacher ethics assessment, the cultural gene of "molding education with morality" and the institutional advantage of "concentrating power to do great things" provide a unique soil for system reconstruction. The current reform needs to achieve breakthroughs in three dimensions: First, take the coupling degree of the development of teachers and students as the value anchor, change the evaluation focus from "management control" to "relationship construction", learn from the collaborative reflection mechanism of Japanese lesson studies, and build an evaluation index of teacher-student two-way empowerment; Second, to solve the quantitative and qualitative methodological difficulties with the best technology, through the intelligent analysis platform integration of multi-modal data, not only retain the British value-added evaluation of scientific rationality, but also learn the practical wisdom of the American evidence chain system; Third, activate professional autonomy with institutional innovation, establish a flexible framework of "national standards-school-based implementation-teacher-led", and leave room for teachers' creative practice on the basis of ensuring the bottom line of education quality. Looking forward to the future, the construction of China's teacher evaluation system should not be a simple "international experience transfer.

Conflicts of Interest

The author declares no conflicts of interest regarding the publication of this paper.

References

All-Russian Teachers' Union (2022). *Survey Report of the All-Russian Teachers' Union on the Current Situation of Teacher Evaluation.*

All-Russian Union of Educators (2022). *Research on the School Evaluation System in Russia.* All-Russian Union of Educators.

Ball, S. J. (2021). *The Education Debate: Policy and Politics in the Twenty-First Century.* Bristol University Press.

China National Institute of Education Sciences (2023). *Blue Paper on the Reform of the Teacher Evaluation System in the New Era.* Educational Science Press.

College of Education, Zhejiang University (2023). *Final Report on the Pilot Project of Narrative Evaluation of Teacher Ethics.* Zhejiang University Press.

Darling-Hammond, L., & Amrein-Beardsley, A. (2019). *The Negative Impact of Value-Added Modeling on Educator Practice.* Stanford Center for Opportunity Policy in Education.

Darling-Hammond, L., Burns, D., Campbell, C., Goodwin, A. L., Hammerness, K., Low, E. L. et al. (2017). *Empowered Educators: How Leading Nations Design Systems for Teaching Quality*. Jossey-Bass.

Denver Public Schools (2021). *Career Ladder Program Evaluation Report.* Denver Public Schools.

Department for Education (2021). *Progress 8 Model: Impact on Disadvantaged Students in Manchester.* OFSTED.

Department for Education (UK) (2016). *Teacher Appraisal Framework: Focus on Professional Growth.* DfE.

East China Normal University (ECNU) (2023). *Pilot Report on Dynamic Teacher Development Evaluation System in Shanghai.* Shanghai Municipal Education Commission.

FEA (2019). *The Impact of VAM-Driven Evaluation on Classroom Practice.* Florida Education Association.

Guangdong Provincial Department of Education (2023). *Guidelines for Subject Evaluation Standards.* Guangdong Education Press.

Hiroshima Board of Education (2021). *Hiroshima: Hiroshima Board of Education.*

Japanese Ministry of Education, Culture, Sports, Science and Technology (2023). *Tokyo: The Ministry of Education and Science* (pp. 5-7).

Manchester University (2022). *Reliability of Classroom Observation Ratings in Teacher Appraisal.* Manchester Institute of Education.

Ministry of Education of the Russian Federation (2020). *Federal State Educational Standards (Basic General Education Standards).* Ministry of Education and Science of the Russian Federation.

Ministry of Education, Sports, Science and Technology, Japan (2020). *Survey Report of the Ministry of Education, Sports, Science and Technology on Teacher Evaluation and Teaching Skills Improvement.*

Moscow Teachers' Union (2022). *Annual Survey on Teachers' Evaluation in Moscow* (p. 8). Moscow Teachers' Union.

OECD (2021). *Education Policy Outlook 2021: Shaping Responsive and Resilient Education in a Changing World.* OECD Publishing.

OFSTED (2015). *The Common Inspection Framework: Education, Skills and Early Years.* HMSO.

OFSTED (2020a). *Evaluation of Professional Dialogue Review in Birmingham Schools.* Department for Education.

OFSTED (2020b). *School Inspection Handbook 2020.* UK Department for Education.

Osaka Prefectural Board of Education (2020). *Report on the Practice of Sharing Teaching Evaluation Data through School Collaboration.* Osaka Prefectural Board of Education.

Sahlberg, P. (2018). *Finnish Education in Crisis? How Teacher Agency and School Autonomy Can Save It.* Finnish National Agency for Education.

The Central Committee of the Communist Party of China and the State Council (2020). *General Plan for Deepening the Reform of Education Evaluation in the New Era.* The Portal of the Central People's Government of the People's Republic of China. http://www.gov.cn/zhengce/2020-10/13/content_5551119.htm

Waters, J., & Brown, A. (2022). *Rater Reliability in Ofsted Classroom Observations.* University of Manchester Press.

Yale University (2022). *Validity of Connecticut's Comprehensive Evidence-Based Assessment System (CEAS).* Yale Center for Education Research.