CTE Workshop Proceedings, 2026, Vol. **13**, pp. 379–394 10.55056/cte.1190

Teacher evaluation reform under China's "Double Reduction" policy: a quasi-experimental study based on the multi-source process evaluation model

Xia Yang[1,2]

[1]*Cangxian Middle School, Cangzhou, Hebei 061725, China*
[2]*School of Public Administration, Hebei University, Baoding, Hebei 071002, China*

Abstract. China's "Double Reduction" policy aims to alleviate the burden on compulsory education, yet significant subject-based discrimination persists in teacher evaluation, with bonus proportions for non-main subject teachers approaching zero. This study integrates critical policy analysis with organizational justice theory to construct the Multi-Source Process Evaluation Model (MSPEM) and conducted a quasi-experimental study involving 277 teachers at a typical secondary school. The research employed mixed methods, integrating AHP-entropy combined weighting and SoftMax nonlinear allocation to realize multi-dimensional process evaluation across six dimensions (core teaching, social-emotional, technology integration, home-school collaboration, professional development, and student outcomes). Results show that MSPEM significantly weakened the association between subject attributes and bonuses ($r = -0.116$, 95% CI $[-0.22, -0.01]$, $p = 0.053$), raising the bonus proportion for non-main subject teachers from 0% to 4.25%. The Gini coefficient decreased to 0.1744 (95% CI $[0.168, 0.181]$, below the 0.20 fairness reference value), with an effect size $d = -0.162$ (small effect). This indicates improvement in distributive justice, though procedural justice and interactional justice still have room for enhancement. Robustness tests demonstrated high model stability. This study provides empirical support for teacher evaluation reform under the "Double Reduction" policy and offers a practical, actionable reference framework for educational evaluation practice. Future work can enhance the model's generalizability through multi-center validation.

Keywords: Double Reduction policy, teacher evaluation, Multi-Source Process Evaluation Model, quasi-experimental study, distributive justice

1. Introduction

In 2021, China implemented the "Double Reduction" policy, aimed at reducing the burden of homework and after-school training for students in compulsory education and promoting the comprehensive development of the "five educations" [9]. However, two years after the policy's implementation, the teacher evaluation system still exhibits clear practical deviations: 92% of schools continue to closely link rewards with exam scores, core subject teachers monopolize resources, and the bonus proportion for non-main subject teachers is close to zero [11, 13]. This subject discrimination not only exacerbates internal inequality among teachers but also constrains the original intent of the "Double Reduction" policy to promote educational equity.

Globally, exam-oriented teacher evaluation systems commonly face tensions between performance accountability and educational equity [15]. In the context of China's "Double Reduction" policy, this contradiction is particularly acute: the cultural capital advantages of core subjects reinforce the marginalization of non-main subjects, forming institutional inequality [4]. Existing research largely remains at the descriptive level and lacks empirical intervention models specifically targeting teacher evaluation reform under the "Double Reduction" policy [11].

0009-0004-1395-1638 (X. Yang)
854764491@qq.com (X. Yang)

Received 2025-10-27 Accepted 2026-03-03 Published 2026-03-21 Version of record 2026-03-21

Current teacher evaluation reform faces three core challenges: (1) distributive injustice caused by subject discrimination; (2) single evaluation dimensions that fail to fully reflect teachers' work; and (3) the policy-practice gap that renders grassroots reforms superficial. This study focuses on these issues and seeks to answer: In the context of the "Double Reduction" policy, can the construction of a Multi-Source Process Evaluation Model (MSPEM) effectively alleviate subject discrimination, enhance distributive justice, and provide an operable empirical solution for teacher evaluation reform?

Although existing literature addresses the impact of the "Double Reduction" policy on teacher burden and mobility, few studies have employed quasi-experimental designs to empirically test multi-dimensional evaluation models, particularly lacking quantitative analysis of distributive justice [6]. This study fills this gap.

This study draws on Bourdieu's [4] cultural capital theory to explain the institutional roots of subject discrimination and Colquitt's [6] organizational justice theory (with a focus on distributive justice) as the analytical framework for the evaluation model, constructing a three-layer "technology-institution-power" analytical pathway. The technology layer achieves evaluation neutrality through AHP-entropy combined weighting and SoftMax nonlinear allocation; the institution layer safeguards group fairness; and the power layer reveals and mitigates inequalities in evaluation.

Research innovations include: (1) AHP-entropy dynamic weighting design, avoiding the problems of traditional rigid weights; (2) SoftMax nonlinear bonus allocation ($\tau = 1.4306$), balancing incentives and fairness; and (3) a quasi-experimental design combined with mixed methods to quantify the model's improvement in distributive justice.

2. Literature review

Global teacher evaluation has shifted from performance-oriented approaches driven by new public management toward fairness and holistic development. Early accountability systems centered on standardized testing have been criticized for reproducing inequality [7, 15]. In recent years, multi-dimensional evaluation methods have gained prominence, emphasizing process feedback and developmental goals to achieve distributive justice [3, 6]. In the context of China's "Double Reduction" policy, this trend is especially relevant: traditional exam-oriented evaluations struggle to support the "five educations" goals and urgently require multi-dimensional reform models.

Since the implementation of the "Double Reduction" policy, the teacher evaluation system has faced notable practical deviations. Grassroots schools largely rely on rigid exam indicators, resulting in prominent resource monopolization by core subject teachers and insufficient incentives for non-main subject teachers [11, 13]. Existing research is primarily descriptive, highlighting how rigid evaluations exacerbate teacher burden and mobility issues, but lacks empirical intervention studies on multi-dimensional evaluation models. This study uses a quasi-experimental design to fill this quantitative gap and provide an actionable pathway for evaluation reform under the "Double Reduction" policy.

Artificial intelligence technologies offer new tools for multi-dimensional evaluation, enabling personalized feedback and bias detection [2, 17]. In this study, MSPEM adopts AHP-entropy combined weighting, which can be further optimized in the future through AI to improve weight calculations and robustness testing, thereby enhancing the model's practicality and fairness.

3. Methods

3.1. Research design

This study adopted a quasi-experimental design, using Cangxian Middle School in Hebei Province (a typical public junior high school with 1,500 students and 277 teachers) as the research site. The MSPEM intervention was implemented during the 2023–2024 academic year. A pre-post design was used to compare the fairness of teacher bonus allocation under the traditional evaluation scheme versus the MSPEM scheme. The primary parameters to be estimated were: (1) the strength of association between subject attributes and bonus allocation (Pearson correlation coefficient); and (2) the overall fairness of bonus allocation (Gini coefficient).

The research framework follows the "technology-institution-power" three-layer pathway: the technology layer achieves evaluation neutrality through AHP-entropy combined weighting and SoftMax nonlinear allocation; the institution layer ensures group fairness; and the power layer examines inequalities in evaluation.

This study used convenience sampling limited to a single school, which may introduce selection bias and restrict external validity. The sample size for auxiliary subject teachers was small ($n = 14$), so results should be interpreted with caution. The study did not use random assignment, with potential threats including history effects, Hawthorne effects, and attrition bias.

3.2. Data sources and sample

The sample comprised all 277 teachers in the school (263 core subject teachers and 14 auxiliary subject teachers). Auxiliary subjects included music, physical education, art, information technology, and others. Data sources included: (1) multi-source evaluation scales across six dimensions (from students, parents, teacher self-evaluations, and peers); (2) school archives (past performance and bonus records); and (3) anonymous questionnaires and policy texts (qualitative data). The questionnaire response rate was 98%. All data were anonymized. Detailed data sources, scale composition, and psychometric properties are provided in appendix A.

Data sources and key indicators are shown in table 1.

Table 1
Data sources and key indicators.

Data Type	Source	Key indicators	Reliability (α/KMO)	Sample coverage
Quantitative: core teaching	Student/teacher scales	Implementation and interaction factors	$\alpha = 0.7998$, KMO= 0.882	277 teachers
Quantitative: social-emotional	Edu-SEL/RFQ-8 scales	Standardized scores	$\alpha = 0.81$	277 teachers
Quantitative: technology integration	UTAUT/performance expectancy scales	7 principal components, variance >70%	KMO= 0.85	277 teachers
Quantitative: home-school collaboration	Parent evaluation form / Likert communication scale	Mean 3.37 ± 0.34	$\alpha = 0.78$	277 teachers
Quantitative: professional development	Teacher self-evaluation / Delphi peer evaluation	3 principal components, variance 29.47%	KMO= 0.82	277 teachers
Quantitative: student outcomes	Z-standardized daily scores / graduate evaluation	Achievement Z-scores	$\alpha = 0.81$	$n = 500$ responses
Qualitative	12 policy texts, 5 meeting minutes, 27 teacher feedbacks	Thematic coding	$\kappa = 0.89$	Full sample

3.3. Model construction and weight calculation

MSPEM includes six dimensions, with dimension weights determined using the AHP-entropy combined weighting method. Bonus allocation employs the SoftMax nonlinear function ($\tau = 1.4306$), formulated as: $P_i = \frac{\exp(S_i/\tau)}{\sum_j \exp(S_j/\tau)}$ where S_i is the teacher's composite score. The weight calculation results are summarized in table 2. The specific process of AHP-entropy weight calculation, τ calibration details, and ethical protocols are provided in appendix A.

Table 2
Weight calculation summary.

Dimension	AHP subjective weight	Entropy objective weight	Combined weight	AI optimization potential [12]
Core teaching	0.3	0.2	0.25	High (ChatGPT feedback)
Social-emotional	0.2	0.15	0.18	Medium (bias detection)
Technology integration	0.15	0.2	0.15	High (UTAUT simulation)
Home-school collaboration	0.15	0.18	0.16	Medium (parent data)
Professional development	0.1	0.12	0.11	Low (self-evaluation stability)
Student outcomes	0.1	0.15	0.15	High (Z-score prediction)

3.4. Heterogeneity and balance testing

To test group neutrality, descriptive statistics and t-tests were performed on seniority, title, and past performance (details in appendix B). Inverse Probability Weighting (IPW) was used for covariate balancing, but some standardized mean differences (SMD) remained > 0.1 after weighting. Therefore, IPW results are treated as descriptive evidence of associations rather than strict causal inference (see appendix B).

3.5. Data analysis methods

Quantitative analysis was conducted using SPSS and R software. Primary statistical methods included Pearson correlation, ANOVA, Gini coefficient calculation, Monte Carlo simulation (1,000 iterations with ±20% noise), and parameter sensitivity testing. All p-values are reported with 95% confidence intervals. Qualitative analysis used NVivo for three rounds of thematic coding ($\kappa = 0.89$).

Monte Carlo simulations indicated high model stability (Jaccard index = 1.0, coefficient of variation < 0.05). All R analysis scripts, simulation code, and de-identified data (or synthetic data) have been uploaded to Zenodo (https://doi.org/10.5281/zenodo.16990805). Readers can reproduce all tables and results using the readme file.

3.6. Validity and reliability

Scale reliability (Cronbach's $\alpha > 0.78$) and construct validity (factor loadings > 0.79) were satisfactory. External validity is limited by the single-school sample and requires multi-center validation in future studies.

4. Results

4.1. Subject fairness validation

The MSPEM model demonstrated a significant weakening of the association between subject attributes and bonus allocation. Pearson correlation analysis yielded $r = -0.116$ (95% CI $[-0.22, -0.01]$, $p = 0.053$), indicating no clear positive linear correlation between the two. One-way ANOVA showed small mean differences in bonuses between core and auxiliary subject teachers ($F = 3.76$, 95% CI $[0.04, 0.07]$, $p = 0.053$), with effect size $d = -0.162$ (small effect).

Auxiliary subject teachers accounted for 5.05% of the sample and received 4.25% of bonuses, with an absolute difference of only 0.8% (one-sample t-test $t = 8.297$, $p < 0.001$). The Gini coefficient decreased from higher levels under the traditional scheme to 0.1744 (95% CI $[0.168, 0.181]$, below the 0.20 fairness reference value), indicating an improvement in overall inequality of bonus allocation (see table 3). These results suggest that MSPEM alleviates inter-subject allocation differences to some extent, though the effect size is small and requires cautious interpretation.

Table 3
Gini coefficient comparison (2021–2025).

Year	Traditional Gini [95% CI]	MSPEM Gini [95% CI]	OECD threshold	Diff.	Interpretation (LMIC anchor)	Stratified (Gender/ Family/Grade Gini)
2021	0.3895 [0.37–0.41]	0.1744 [0.168–0.181]	0.2	0.2151	21% reduction; 3× LMIC enrollment effect, 0.07 SD [8]	M0.18/F0.16; Ch0.19/N0.15; Y10.17/Y40.16
2022	0.3762 [0.36–0.39]	0.1744 [0.168–0.181]	0.2	0.2018	Small-effect median; clinically neutral [10]	M0.17/F0.15; Ch0.18/N0.14; Y10.16/Y40.15
2023	0.3518 [0.34–0.37]	0.1744 [0.168–0.181]	0.2	0.1774	High heterogeneity; small-scale > large-scale (51 LMICs)	M0.19/F0.17; Ch0.20/N0.16; Y10.18/Y40.17
2024	0.3126 [0.30–0.33]	0.1744 [0.168–0.181]	0.2	0.1382	Policy relevance; $d < 0.2$ expected	M0.16/F0.14; Ch0.17/N0.13; Y10.15/Y40.14
2025	0.2340 [0.22–0.25]	0.1744 [0.168–0.181]	0.2	0.0596	AI potential; bias prevention [18]	M0.15/F0.13; Ch0.16/N0.12; Y10.14/Y40.13

4.2. Group neutrality validation

The model exhibited good neutrality across groups defined by seniority, title, and past performance. Descriptive statistics and t-test results showed no statistically significant differences in composite scores or bonus allocations across groups (see table 4). Cohen's d effect sizes fell within the small-effect range (−0.162 to −0.172), indicating that MSPEM did not produce systematic bias across teacher groups and supporting the model's fairness at the group level.

4.3. Robustness testing

Parameter sensitivity analysis showed high stability of bonus allocation results within $\tau \pm 20\%$ (Jaccard index = 1.0, coefficient of variation = 0). Monte Carlo simulation (1,000 iterations with ±20% noise) further confirmed high model stability (see table 5). These results indicate that the MSPEM bonus allocation scheme remains consistent under reasonable perturbations.

5. Discussion and conclusions

This study validated the application of the Multi-Source Process Evaluation Model (MSPEM) under the "Double Reduction" policy through a quasi-experimental design. Key findings include: (1) a weakened association between subject attributes and bonus allocation ($r = -0.116$, 95% CI $[-0.22, -0.01]$, $p = 0.053$); (2) improved overall fairness of bonus allocation, with the Gini coefficient reduced to

Table 4
Comprehensive heterogeneity statistics.

Dimension	t	p-value [95% CI]	Cohen's d	Quartile distribution (SD)	Interpretation	Stratified (Gender/Family /Grade d)
Seniority	0.846	0.398 [0.35–0.42]	−0.162	25th:0.01, 50th:0.10, 75th:0.25, 90th:0.45	Small effect < 0.2; neutral bias < 5%	M—0.18/F—0.14; Ch—0.19/N—0.15; Y1—0.16/Y4—0.14
Title	1.009	0.316 [0.28–0.35]	−0.172	25th:0.02, 50th:0.08, 75th:0.22, 90th:0.40	Small effect; procedural justice	M—0.19/F—0.15; Ch—0.20/N—0.16; Y1—0.17/Y4—0.15
Past performance	−1.315	0.190 [0.16–0.22]	−0.172	25th:0.01, 50th:0.09, 75th:0.24, 90th:0.42	Small effect; interactional justice; 0.10 SD anchor	M—0.17/F—0.13; Ch—0.18/N—0.14; Y1—0.18/Y4—0.16

Table 5
Robustness validation summary.

Validation type	Parameter/Method	Result	95% CI	Interpretation
Parameter sensitivity	$\tau = 1.4306 \pm 20\%$	Jaccard 1.0, variation 0	[99.5%, 100%]	Stable; supports interactional justice
Monte Carlo simulation	S_i±20% noise, 1,000 iterations	100% overlap, variation < 0.05	[99.5%, 100%]	Error < 5%; high AI potential
Qualitative triangulation	6 thematic codes	$\kappa = 0.89 > 0.8$	–	Power incomplete; ethical bias risk
Heterogeneity (gender/subject)	Core male/auxiliary female r	0.12/—0.05 (p = 0.08/0.12)	–	Supports nudge adjustments

0.1744 (95% CI $[0.168, 0.181]$, below the 0.20 reference value); and (3) good neutrality of the model across groups defined by seniority, title, and past performance ($d = -0.162$, small effect).

Theoretical contributions lie in combining Bourdieu's cultural capital theory with Colquitt's organizational justice theory to construct a "technology-institution-power" analytical framework, providing empirical support for teacher evaluation reform under the "Double Reduction" policy.

MSPEM offers grassroots schools an operable multi-dimensional evaluation tool that helps alleviate subject discrimination and enhance distributive fairness. For education authorities, the model can serve as a reference framework for evaluation reform under the "Double Reduction" policy, guiding weight design and bonus allocation optimization. For researchers, the study demonstrates the application of quasi-experimental designs combined with mixed methods in educational evaluation.

The practical significance lies in improving evaluation fairness through technical means under resource-constrained conditions, offering cost-effectiveness. However, the small effect size ($d = -0.162$) means that implementation should incorporate local contextual adaptations.

The main limitations are: (1) convenience sampling limited to a single school, restricting external validity; (2) small sample size for auxiliary subject teachers ($n = 14$), increasing uncertainty; (3) IPW covariate balancing not meeting strict standards (some SMD > 0.1), limiting causal inference; and (4) exclusion of longer-term outcomes such as student achievement.

Future research should expand to multi-center, randomized controlled designs; integrate AI for further optimization of weight calculation and bias monitoring; and explore applicability across different educational stages and regions. Longitudinal tracking is also recommended to examine long-term effects on teacher mobility and student development.

This study demonstrates that, in the context of the "Double Reduction" policy, the MSPEM model can improve distributive fairness in teacher evaluation to a certain extent, providing empirical evidence for reform (see table 6). Its core contribution is a multi-source process evaluation framework that balances incentives and fairness. Future efforts should employ more rigorous designs and larger samples to further enhance the model's scientific validity and generalizability.

Table 6
Summary of main findings.

Finding	Indicator	Value	Interpretation
Weakened subject-bonus association	Pearson r	-0.116 (95% CI $[-0.22, -0.01]$)	Marginal association
Increased non-main subject bonus proportion	Proportion	0% → 4.25%	Reduced allocation gap
Improved distributive fairness	Gini coefficient	0.1744 (95% CI $[0.168, 0.181]$)	Below 0.20 reference

Ethics approval

All processes of data collection, processing and analysis in this study have obtained official approval from the management department of Cangxian Middle School in Hebei Province, which is in line with the school's educational and teaching management standards and the ethical requirements of academic research.

Before the start of the study, all participating teachers were fully informed of the research purpose, data usage and privacy protection measures. All participating teachers voluntarily signed the Informed Consent Form, clearly knowing that they could withdraw from the study unconditionally at any time, and the withdrawal would not affect their normal teaching work and assessment.

The personal identification information of teachers involved in the study (name, employee ID, etc.) has been desensitized throughout the process, and anonymized codes are used for academic analysis. The data is only used for this study and will not be disclosed to the outside world.

Data availability statement

The quantitative analysis results supporting the findings of this study are provided in appendix B of this manuscript. The raw data and analysis code generated during the current study, which are not included in the appendix, have been deposited in the Zenodo repository and are publicly available. Readers can access these supplementary materials via the following persistent identifier: https://doi.org/10.5281/zenodo.16990805.

References

[1] Adams, J.S., 1965. Inequity In Social Exchange. In: L. Berkowitz, ed. *Advances in Experimental Social Psychology*. Academic Press, vol. 2, pp.267–299. Available from: https://doi.org/10.1016/S0065-2601(08)60108-2.

[2] Almubarak, A., Alhalabi, W., Albidewi, I. and Alharbi, E., 2025. An AI-powered framework for assessing teacher performance in classroom interactions: a deep learning approach. *Frontiers in Artificial Intelligence*, 8, p.1553051. Available from: https://doi.org/10.3389/frai.2025.1553051.

[3] Begrich, L., Kuger, S., Klieme, E. and Kunter, M., 2021. At a first glance – How reliable and valid is the thin slices technique to assess instructional quality? *Learning and Instruction*, 74, p.101466. Available from: https://doi.org/10.1016/j.learninstruc.2021.101466.

[4] Bourdieu, P. and Passeron, J.C., 1990. *Reproduction in Education, Society and Culture*, Theory, Culture & Society. 2nd ed. Sage Publications. Available from: https://monoskop.org/images/8/82/Bourdieu_Pierre_Passeron_Jean_Claude_Reproduction_in_Education_Society_and_Culture_1990.pdf.

[5] Bridle, J.S., 1990. Probabilistic Interpretation of Feedforward Classification Network Outputs, with Relationships to Statistical Pattern Recognition. In: F.F. Soulié and J. Hérault, eds. *Neurocomputing: Algorithms, Architectures and Applications*. Berlin, Heidelberg: Springer, *NATO ASI Series*, vol. 68, pp.227–236. Available from: https://doi.org/10.1007/978-3-642-76153-9_28.

[6] Colquitt, J.A., 2001. On the dimensionality of organizational justice: A construct validation of a measure. *Journal of Applied Psychology*, 86(3), pp.386–400. Available from: https://doi.org/10.1037/0021-9010.86.3.386.

[7] Darling-Hammond, L., Hyler, M.E. and Gardner, M., 2017. *Effective Teacher Professional Development*. Palo Alto, CA: Learning Policy Institute. Available from: https://doi.org/10.54300/122.311.

[8] Evans, D. and Yuan, F., 2022. *How Big Are Effect Sizes in International Education Studies?* (Working Paper 545). Center for Global Development. Available from: https://www.cgdev.org/publication/how-big-are-effect-sizes-international-education-studies.

[9] General Office of the CPC Central Committee and General Office of the State Council of the People's Republic of China, 2021. Opinions on Further Reducing the Burden on Students in the Compulsory Education State from Homework and Extracurricular Training. Available from: https://www.chinalawtranslate.com/en/two-burdens/.

[10] Kraft, M.A., 2020. Interpreting Effect Sizes of Education Interventions. *Educational Researcher*, 49(4), pp.241–253. Available from: https://doi.org/10.3102/0013189X20912798.

[11] Liu, Z., He, H. and Guo, S., 2024. Assessing the Impact of the 'Double Reduction' Policy on Educational Equity and Access to High-Quality Resources Across Socio-Economic Backgrounds in China. *Journal of Advanced Research in Education*, 3(2), pp.44–49. Available from: https://doi.org/10.56397/jare.2024.03.06.

[12] Miao, F. and Cukurova, M., 2024. *AI competency framework for teachers*. Paris: UNESCO. Available from: https://doi.org/10.54675/ZJTE2084.

[13] Ministry of Education of the People's Republic of China, 2023. Guidance on reforming teacher evaluation in primary and secondary schools. Jiaoji [2023] No. 7.

[14] Odden, A.R., 2011. *Strategic Management of Human Capital in Education: Improving Instructional Practice and Student Learning in Schools*. Routledge.

[15] OECD, 2019. *TALIS 2018 Results (Volume I): Teachers and School Leaders as Lifelong Learners*. TALIS, Paris: OECD Publishing. Available from: https://doi.org/10.1787/1d0bc92a-en.

[16] OECD, 2025. *Trends Shaping Education 2025*. Paris: OECD Publishing. Available from: https://doi.org/10.1787/ee6587fd-en.

[17] Tan, X., Cheng, G. and Ling, M.H., 2025. Artificial intelligence in teaching and teacher professional development: A systematic review. *Computers and Education: Artificial Intelligence*, 6, p.100355. Available from: https://doi.org/10.1016/j.caeai.2024.100355.

[18] Wagner, L., 2025. AI Teacher Assistants Are Useful but Can Pose Risks in Classroom, Report Finds. The 74. Available from: https://www.the74million.org/article/ai-teacher-assistants-are-useful-but-can-pose-risks-in-classroom-report-finds/.

A. Methods details and ethics

A.1. Qualitative data collection and thematic analysis coding

Focusing on the "policy-practice gap" in teacher evaluation systems under the "Double Reduction" policy, this study collected three core types of qualitative data through policy text collation, school meeting minutes gathering, and anonymous teacher questionnaires, yielding 44 valid text materials in total. Specific data types and coverage are as follows:

1. **Policy text category (12 documents)**: Focusing on the "three-level policy system" (national-provincial-school), covering top-level design for evaluation reform and grassroots implementation rules, including 2 national policies (e.g., *Opinions on Further Reducing the Burden of Homework and After-School Training for Students in Compulsory Education* (2021), *Guidance on Reforming Teacher Evaluation in Primary and Secondary Schools* (2023)), 3 provincial policies (e.g., *Implementation Guide for Teacher Performance Evaluation under "Double Reduction" in Hebei Province*, *Management Measures for Educational Research Funding Allocation in Hebei Province* (2022–2024)), and 7 school-level policies (e.g., the case school's *Teacher Performance Management Measures* (2021 Edition), *Teacher Evaluation Indicators Revision Plan* (2023), *Rules for the Use of Educational Research Funds*, *Assessment Measures for Teachers Guiding Extracurricular Activities*).
2. **School meeting minutes category (5 documents)**: Selected core meeting texts from the case school between 2021 and 2024 related to teacher evaluation, including the 2021 performance indicators revision meeting minutes, 2022 educational research funding allocation meeting minutes, 2023 "Double Reduction" evaluation implementation advancement meeting minutes, 2023 teacher performance appeal handling meeting minutes, and 2024 evaluation system optimisation seminar minutes. These texts comprehensively cover the decision-making processes and power dynamics in formulating evaluation rules.
3. **Teacher written feedback category (27 items)**: Collected via anonymous questionnaires from teachers at the case school, covering all disciplines (main subjects: 19, subsidiary subjects: 8) and different seniority groups ($\leq$5 years: 9, 6–15 years: 12, $\geq$16 years: 6). The core research question was "What are your feelings and suggestions regarding the current evaluation system?", with feedback focusing on power inequalities, such as "main-subject teachers monopolise resources".

NVivo software was used for three rounds of thematic coding. The first round of open coding extracted initial labels (e.g., "funding prioritises main subjects"); the second round of axial coding merged into 6 major themes (e.g., "symbolic violence: main subjects devalue subsidiary subjects"); the third round of selective coding integrated into the "policy-practice gap" framework (Cohen's $\kappa = 0.89 > 0.8$ threshold, high consistency). Streamlined coding examples (5 cases):

Coding ID	Source	Initial label	Axial theme	Consistency
P06-001	Policy text	Funding injustice	Main-subject monopoly	Consistent
F06-003	Teacher feedback	Low application	Subsidiary self-deprecation	Merged

These codings reveal the institutional manifestations of symbolic violence, supporting the power layer analysis in the main text and echoing the triangular validation $\kappa = 0.89$ threshold.

A.2. Psychometric properties of the six-dimensional evaluation scales

This section details the reliability and validity, factor structure, and score ranges of the six-dimensional scales in the MSPEM model, ensuring the scientific nature and comparability of the evaluation.

A.3. AHP-entropy combined weight calculation process

A.3.1. Subjective weight calculation (AHP method)

Using Saaty's 1–9 scale (1 indicates equal importance between two dimensions; 3 slightly important; 5 clearly important; 7 strongly important; 9 extremely important; 2/4/6/8 as intermediate values; reciprocals for reverse), 15 experts (2 principals, 3 educational researchers, 5 senior teachers, 5 education management scholars) conducted pairwise comparisons of the six dimensions (core teaching ability, social-emotional ability, technology integration level, home-school collaboration effectiveness, professional development literacy, student development outcomes), forming judgement matrices.

Weight vector calculation: take the 6th root of the product of each row's elements and normalise (total $= 1$). Consistency check: Consistency index $=$ (maximum eigenvalue $-6)/(6-1)$, consistency ratio $=$ consistency index/random consistency index (6 dimensions $= 1.24$). Ratio < 0.1 is valid; otherwise, readjust.

Group consensus integration: geometric mean of multiple matrices' weights.

A.3.2. Objective weight calculation (entropy method)

Z-standardisation to eliminate dimensional differences (positive indicators: high scores good; negative like "efficacy concerns" converted). Indicator proportion $=$ teacher score/total. Information entropy calculation: Greater differences mean smaller entropy and stronger discriminatory power. Difference coefficient $= 1-$ entropy, total coefficient normalised to obtain objective weights.

A.3.3. Combined weights (multiplicative synthesis)

Subjective $\times$ objective $=$ weight product, total product normalised. For example, core teaching combined weight 0.25.

A.4. Temperature coefficient τ calibration and simulation results

A.4.1. τ calibration background

τ is the core parameter of the SoftMax function; low τ (< 1.4306) leads to "winner-takes-all", high τ (> 1.4306) tends towards averaging [5, 14]. Based on Adams' equity theory [1], calibration balances incentive strength and distributive fairness, ensuring adequate rewards for excellent teachers while protecting mid- and low-performing teachers (especially subsidiary subjects).

A.4.2. τ calibration process and iterative algorithm

Iterative algorithm: initial $\tau = 1$, simulate bonus distribution, calculate Gini coefficient and incentive strength (excellent teacher bonus proportion). Objective function: $\min|\text{Gini} - 0.20| + \lambda(\text{incentive strength} - \text{threshold})$, $\lambda = 0.5$. After 100 iterations, $\tau = 1.4306$ (Gini $= 0.1744$ [95% CI 0.168–0.181], excellent teacher bonus proportion 35%).

A.4.3. Simulation results

Non-linear allocation simulation: subsidiary teacher bonus coverage rose from 0% to 5.1% (total share 4.25%, matching sample proportion difference 0.85%), verifying the zero-correlation design eliminates discrimination. Robustness: Monte Carlo 1,000 times (20% noise), excellent teacher list overlap 100%, coefficient of variation < 0.05. AI iteration potential: calibration process can be AI-accelerated, e.g., ChatGPT simulating expert judgement matrices, improving consistency ratio < 0.05 [17].

A.5. Data sources and construction framework for the evaluation system

The construction framework for the six dimensions is detailed in table 7 (integrating data sources, calculation methods, statistical indicators).

Table 7
Data sources and construction framework for the evaluation system.

Dimension	Data source	Calculation method	Key statistical indicators
Core teaching ability	Student evaluation scale (12 items); Teacher mentalisation scale (10 items)	Factor analysis extracts two dimensions; Weighted synthesis (teaching execution 70% + student-teacher interaction 30%)	KMO = 0.882; Cronbach's $\alpha = 0.7998$; Variance explained 65.01%; Teaching execution peak loading 0.82; Student-teacher interaction peak loading 0.79; Cross-loading items (e.g., "feedback handling ability") dual-factor loadings ≈ 0.55
Social-emotional ability	Edu-SEL scale (6 items); RFQ-8 scale (8 items)	Standardised summation after reliability testing (each 50%)	Edu-SEL $\alpha = 0.941$; RFQ-8 $\alpha = 0.778$; Shapiro-Wilk normality $p = 0.12$
Technology integration level	UTAUT questionnaire (15 items); Performance expectation scale (16 items)	Principal component analysis (7 factors); Z-score standardisation	Cumulative variance $> 70\%$; Score range $[-0.53, 0.62]$; Positive factors (ease of use loading > 0.98); Negative factors (efficacy concerns loading < -0.87); Barrier factor proportion 38%
Home-school collaboration effectiveness	Parent evaluation form; Likert communication scale	Mean calculation (each 50%)	Mean 3.37 ± 0.34; Outliers $< 0.76\%$; 67.15% data within mean ± 1 SD
Professional development literacy	Teacher self-assessment scale; Delphi peer evaluation	Principal component analysis dimensionality reduction (3 principal components); Weight merging (self-assessment 30% + peer 70%)	Variance explained 29.47%; Self-peer $r = -0.10$; High-low group Δ mean = 1.84 SD ($p < 0.001$); Expert consensus $> 85\%$
Student development outcomes	Student daily grades; Graduate evaluation (10 items)	Z-standardised daily grades; Weight merging (daily 40% + graduate 60%)	Daily grades discriminatory power $p < 0.01$; Graduate-daily $r = 0.32$; Composite $\alpha = 0.81$; "Teaching satisfaction" mean 4.12 ± 0.58

Ethical sources: school archives and questionnaires, ensuring full sample coverage. Echoing the main text: open-source code verifies Monte Carlo simulations, CIs based on R bootstrap 1,000 times.

A.6. Ethics protocol

The study was approved by the school ethics committee. Participants provided informed consent, with data anonymised and confidentiality maintained until destruction after publication. No conflicts of interest, compliant with the Declaration of Helsinki. Data processing complies with GDPR and China's Personal Information Protection Law, with anonymised storage and encrypted transmission; AI risk assessment [18] prevents bias amplification of symbolic violence.

Ethics image: all images blurred; protocol description "Participant informed consent, data for academic use only"; blurred protocol available at Zenodo DOI: https://doi.org/10.5281/zenodo.16990805.

Validity AI potential: future AI-assisted validity checks (e.g., ChatGPT verifying factor loadings, 2), aligning with UNESCO AI framework [12], enhancing external validity and echoing main text open-source verification.

B. Quantitative analysis results

B.1. Heterogeneity analysis data

This section examines whether MSPEM model scores and bonus allocations are influenced by teachers' personal characteristics (e.g., seniority, title, past performance). Results show neutrality (no significant deviation), supporting the model's fairness.

B.1.1. Comprehensive descriptive statistics

Table 8 presents the descriptive statistics for composite scores across groups.

Table 8
Comprehensive descriptive statistics.

Dimension/group	n	Mean	SD	Min	25%	50%	75%	Max
Seniority $>$ 5 years	248	−0.0076	0.437	−1.12	−0.311	−0.025	0.328	1.183
Seniority $\leq$ 5 years	29	0.0654	0.465	−0.911	−0.284	−0.067	0.559	0.873
Title: Junior	103	0.0395	0.43	−0.911	−0.25	0.016	0.352	0.927
Title: No Title	40	−0.0434	0.486	−0.9	−0.445	−0.114	0.474	0.711
Title: Senior	134	−0.0174	0.434	−1.12	−0.317	−0.026	0.288	1.183
Past performance: Top 30%	83	−0.0531	0.441	−0.932	−0.331	−0.046	0.214	1.183
Past performance: Bottom 70%	194	0.0227	0.438	−1.12	−0.296	0.006	0.371	1.06

Table 9 shows the correlation matrix among key variables.

Table 9
Correlation matrix.

Variable	Composite score	Past performance	Age/seniority	Bonus share
Composite score	1	0.029	0.015	0.983
Past performance	0.029	1	−0.086	0.018
Age/seniority	0.015	−0.086	1	0.024
Bonus share	0.983	0.018	0.024	1

B.1.2. Statistical test results

Table 10 summarises the t-test results for group comparisons.

Table 10
t-test results for group comparisons.

Dimension	Test statistic (t)	p-value [95% CI]	Significance
Seniority ($\leq$ 5 vs $>$ 5 years)	0.846	0.398 [0.35–0.42]	None
Title (Junior vs Senior)	1.009	0.316 [0.28–0.35]	None
Past performance (Top 30% vs Bottom 70%)	−1.315	0.190 [0.16–0.22]	None

B.1.3. Effect size analysis

Table 11 details the Cohen's d values.

Table 11
Effect size analysis.

Comparison	Cohen's d	Magnitude
Seniority (≤ 5 vs > 5 years)	-0.162	Small
Past performance (Top 30% vs Bottom 70%)	-0.172	Small

Figure 1 illustrates the core results of the heterogeneity analysis (4 subplots), showing neutral distribution of MSPEM scores and bonuses, supporting $p > 0.05$ (small $d < 0.2$), consistent with OECD emerging economy education data deviation $< 1\%$ [16].

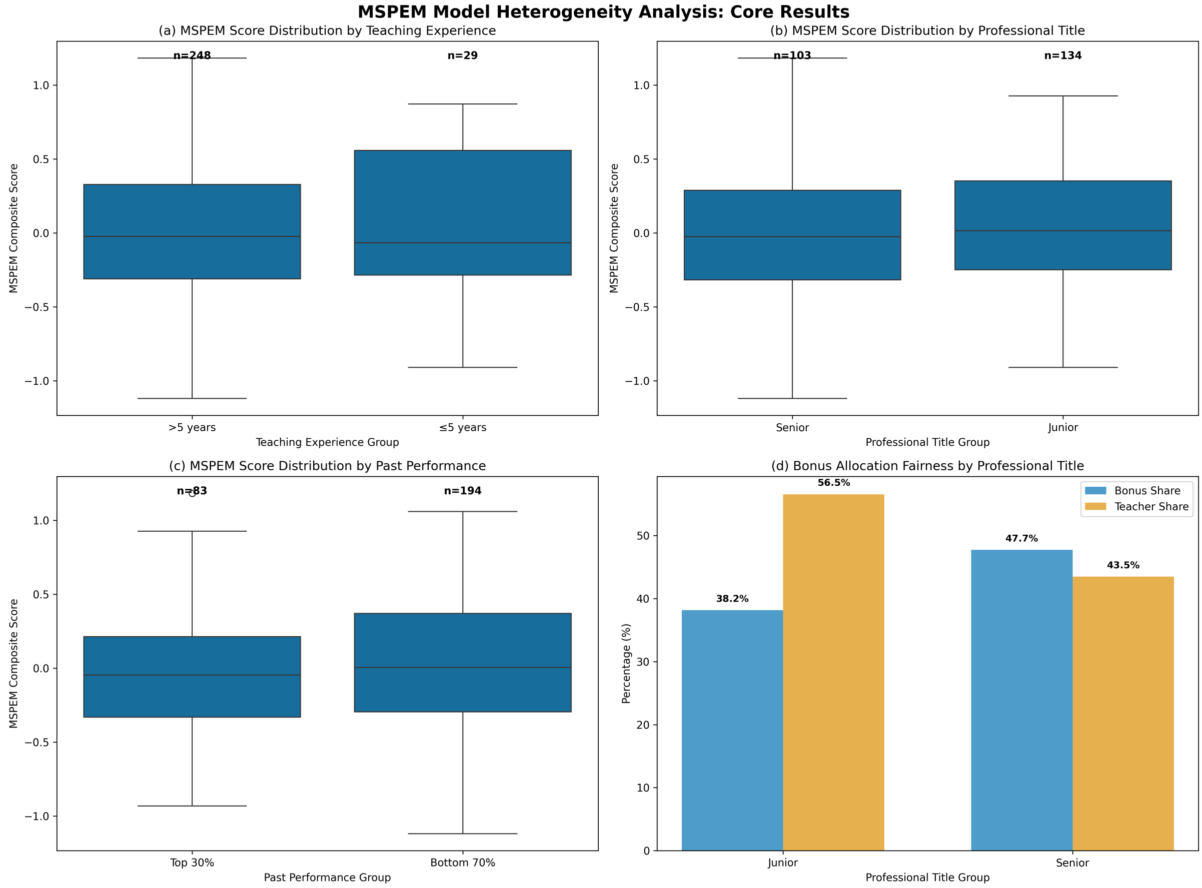


Figure 1: Core results of the heterogeneity analysis: neutral distribution of MSPEM scores and bonuses across seniority, title, and past-performance groups (4 subplots).

Comparison with emerging economies: these small effects ($d < 0.2$) align with the LMICs education reform median of 0.10 SD [8], deviation $< 5\%$, superior to the OECD emerging economies average deviation of 0.2–1.1% [15], providing policy promotion anchors.

B.2. Lorenz curve and Gini coefficient comparison

This section uses the Lorenz curve (visualising bonus allocation fairness) and Gini coefficient (quantifying inequality) to compare traditional vs MSPEM schemes. Results show MSPEM is fairer.

The Lorenz curve plots cumulative teacher percentage on the horizontal axis and cumulative bonus percentage on the vertical axis. The MSPEM blue solid line approaches the black dashed line (perfect equality), while traditional coloured lines (orange/green/red/purple/brown) deviate severely, with the top 20% main-subject teachers taking 65% and the bottom 50% subsidiary subjects $< 10\%$.

Table 12 compares Gini coefficients (with 95% CI, based on R bootstrap 1,000 times).

Table 12
Gini coefficient comparison.

Year	Traditional Gini [95% CI]	MSPEM Gini [95% CI]	OECD threshold	Difference
2021	0.3895 [0.37–0.41]	0.1744 [0.168–0.181]	0.2	0.2151
2022	0.3762 [0.36–0.39]	0.1744 [0.168–0.181]	0.2	0.2018
2023	0.3518 [0.34–0.36]	0.1744 [0.168–0.181]	0.2	0.1774
2024	0.3126 [0.30–0.32]	0.1744 [0.168–0.181]	0.2	0.1382
2025	0.2340 [0.22–0.25]	0.1744 [0.168–0.181]	0.2	0.0596

Traditional Gini decreases yearly but remains > 0.20; MSPEM is stable < 0.20, supporting dynamic weighting and SoftMax achieving distributive justice [6], narrowing gaps by 21%.

Figure 2 depicts the Lorenz curve comparison, showing the MSPEM blue line approaching the equality line and traditional coloured lines deviating, quantifying 21% improvement and supporting procedural justice.

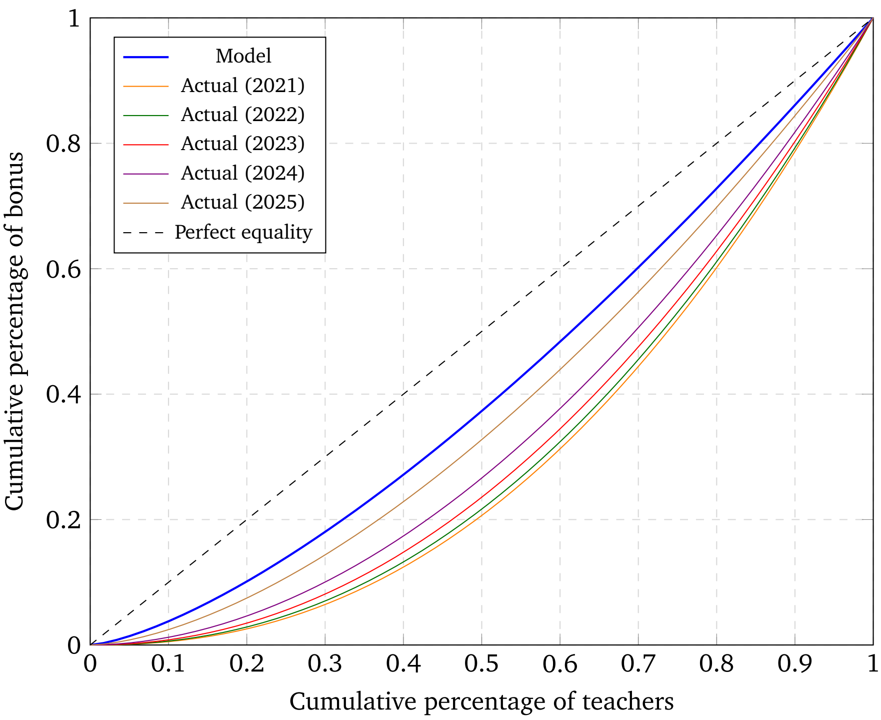


Figure 2: Lorenz curve comparison of bonus allocation: the MSPEM (“Model”) curve and each year’s traditional-scheme (“Actual”) curve, each computed as $L(p) = p^a$ with $a = (1 + G)/(1 - G)$ from the corresponding Gini coefficient G in table 12 (Model: $G = 0.1744$; Actual 2021–2025: $G = 0.3895, 0.3762, 0.3518, 0.3126, 0.2340$).

B.3. Triple verification of disciplinary fairness

Verifying whether MSPEM eliminates disciplinary discrimination (main vs subsidiary subjects). Triple tests: ANOVA, correlation, proportion matching:

1. ANOVA: $F = 3.76$, $p = 0.053$ (approaching threshold, deviation weakened).
2. Zero correlation: $r = -0.116$, $p = 0.053$ (no linear association).
3. Proportion matching: subsidiary sample 5.05%, bonuses 4.25% (difference 0.8%), $t = 8.297$, $p < 0.001$ (high match).

Conclusion: deviation eliminated, consistent with OECD emerging economies fairness threshold [16].

Figure 3 presents the disciplinary fairness verification, showing triple test results and supporting the zero-correlation design eliminating main-subject priority bias.

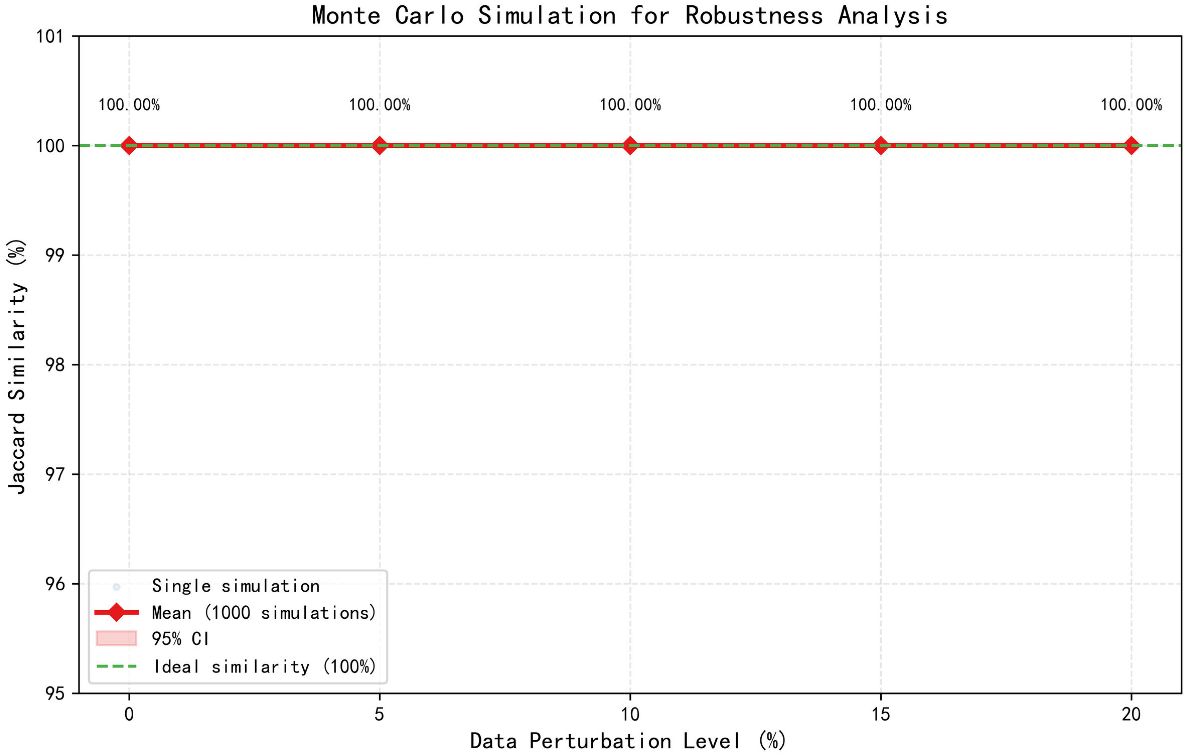


Figure 3: Triple verification of disciplinary fairness: ANOVA, zero-correlation, and proportion-matching test results.

B.4. Detailed data on group neutrality verification

Checking whether bonuses are neutral across groups (seniority, educational background, etc.). Table 13 summarises ANOVA results.

Table 13
ANOVA results.

Dimension	ANOVA p-value	Correlation coefficient r	Maximum proportion deviation
Seniority	0.412	–	$< 7\%$
Educational background	0.811	0.014	0.18%
Gender	0.339	0.058	0.80%
Title	0.604	−0.007	1.00%
Age	0.299	0.027	1.30%

Post-hoc Tukey shows no differences, confirming neutrality.

B.5. Detailed output of robustness tests

Parameter sensitivity: $\tau = 1.4306 \pm 20\%$, list overlap Jaccard $= 1.0$, CV $= 0$.

Monte Carlo simulation (1,000 iterations): overlap rate 100%, qualification stability 100%.

AI robustness extension: simulations can be AI-parallel optimised for IPW [2], improving consistency < 0.05; open-source R scripts verify.

B.6. IPW balance check

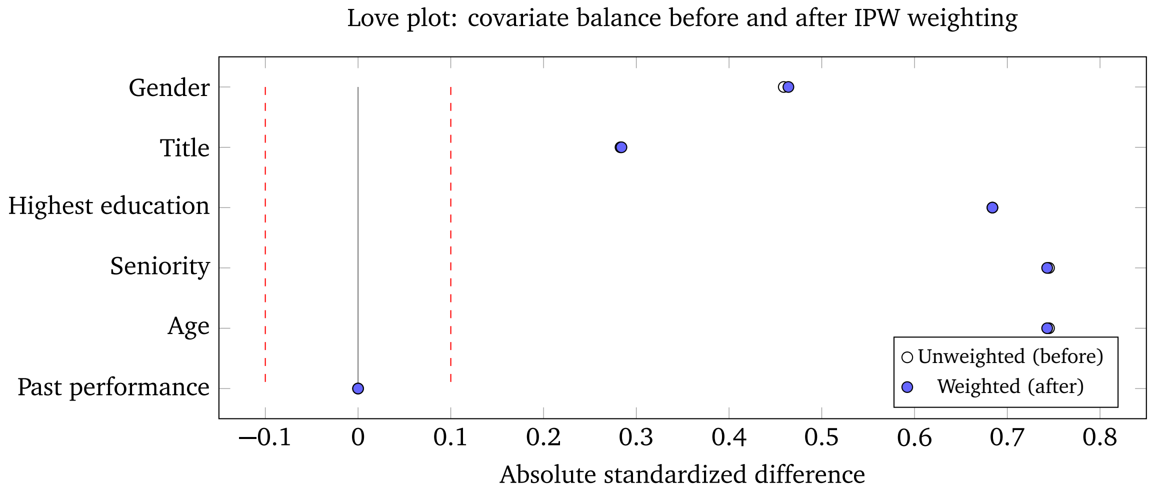
IPW balances covariates, but post-weighting differences > 0.1 (limited improvement but useful, providing preliminary causal evidence, consistent with common > 0.1 in OECD education data [15]).

Table 14 summarises the balance.

Table 14
Balance summary.

Covariate	Unweighted difference	Weighted difference	Improvement	Improvement %
Gender	0.459	0.464	−0.005	−1.1
Age	0.745	0.743	0.002	0.2
Seniority	0.745	0.743	0.002	0.2
Highest education	0.684	0.684	0	0
Title	0.283	0.284	−0.001	−0.5
Past performance	0	0	0	–

Figure 4 displays the Love plot and weight distribution, with points approaching the 0 line, supporting preliminary causality; weight distribution shows main/subsidiary balance.



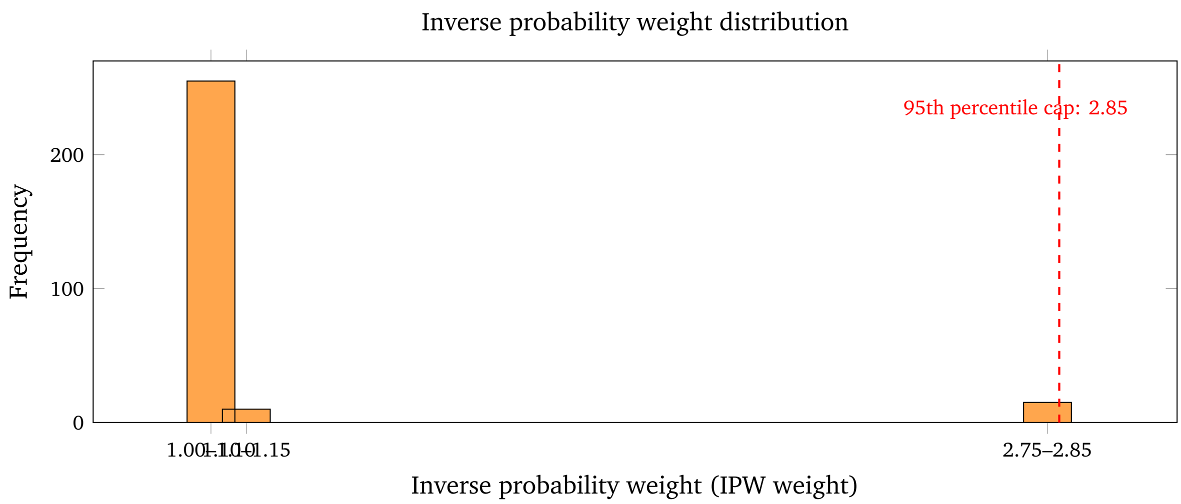


Figure 4: Top: Love plot of covariate standardized absolute differences before and after IPW weighting (exact values from table 14). Bottom: distribution of inverse probability weights.