

Enhanced autoregressive value-added model integrating generative AI ethical optimization for fair high school teacher evaluation

Xia Yang^{1,2}

¹Cangxian Middle School, Cangzhou, Hebei 061725, China

²School of Public Administration, Hebei University, Baoding, Hebei 071002, China

Abstract. Artificial intelligence (AI) is increasingly integrated into educational contexts, yet systematic frameworks that propose and empirically validate fair value-added models (VAM) for high school teacher evaluation remain scarce. This study introduces and validates the IBI-VAM framework, which integrates autoregressive knowledge tracing, Heckman selection bias correction, and generative AI ethical optimization to improve the predictive accuracy, fairness, and trustworthiness of teacher contribution estimates. Using longitudinal panel data from a non-elite public high school in eastern China ($N = 1,876$ students, 6 semesters, 9 subjects), the framework was empirically validated. Results showed a 12% improvement in AUC compared with the OLS baseline (0.83 vs. 0.72), bias reduction to below 3%, ranking stability (Spearman's $r = 0.92$), and a 0.5-point increase on the trust scale ($p < 0.05$). Monte Carlo simulations (1,000 iterations) and Kolmogorov–Smirnov tests indicated small average distributional differences between the Chinese sample and simulated Western high school distributions (KS $D = 0.060$, $p > 0.05$), with IBI-VAM bias estimates exhibiting an average absolute relative deviation of approximately 5% from those obtained under moment-matched Western high-school distributional assumptions, providing preliminary evidence for cross-cultural applicability.

Keywords: value-added model, autoregressive knowledge tracing, high school teacher evaluation, educational equity, generative artificial intelligence ethics

1. Introduction

Artificial intelligence technologies have advanced rapidly, profoundly influencing educational equity and evaluation systems [8, 31]. With China's higher education gross enrollment rate reaching 60.2% [19], ethical implications and potential risks of AI in teacher evaluation have drawn increasing attention [1, 18]. Discussions of fair value-added models (VAM) have intensified across disciplines, underscoring the central role of human values and ethical principles in data-driven decision-making [3, 4]. In high school education, prior research has explored VAM integration and its effects on public schools [10, 23], yet systematic analyses proposing and empirically validating new frameworks that address ethical constraints and fairness remain scarce.

In China's high school system, teacher evaluation frequently encounters fairness challenges: traditional standardized test scores fail to fully capture teachers' contributions to student growth, particularly in resource-constrained public schools, where educators serving low socioeconomic status (SES) classes are often undervalued, resulting in imbalanced resource allocation [9, 24]. With higher education gross enrollment reaching 60.2% [19], issues of AI trust and cultural gaps have further exacerbated these problems [1]. The present data-driven framework responds to demands for educational equity by integrating technological and ethical optimizations to improve evaluation fairness.

 0009-0004-1395-1638 (X. Yang)

 854764491@qq.com (X. Yang)

Received
2025-09-28

Accepted
2026-07-02

Published
2026-09-20

Version of record
2026-09-20



© Copyright for this article by its authors, published by the Academy of Cognitive and Natural Sciences. This is an Open Access article distributed under the terms of the Creative Commons License Attribution 4.0 International (CC BY 4.0), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

The rapid evolution of educational technology highlights the necessity of data-driven assessment, especially in resource-limited blended learning environments [8]. The integration of autoregressive knowledge tracing, Heckman correction, and generative AI directly addresses the global equity agenda [18]. Although VAM has been widely applied, dynamic behavioral and ethical factors are often overlooked in Chinese urban high school contexts, leading to underestimation of teacher contributions for low-SES students [7, 25]. The sample offers substantial longitudinal depth yet is confined to a single school, which may introduce selection bias; this is mitigated through Heckman correction, yielding quasi-causal evidence (see section 3 for details).

Existing VAM research primarily isolates teacher contributions but frequently neglects dynamic behaviors and ethical biases [7, 17]. Traditional models rely on ordinary least squares (OLS) estimation of score gains, which under China's high-stakes college entrance examination can produce biases up to 20% [24]. Autoregressive knowledge tracing improves upon standard knowledge tracing through iterative state updates but requires combination with bias correction to strengthen fairness [22]. The Heckman selection model employs the inverse Mills ratio to address non-random assignment, furnishing quasi-causal evidence [25]. Generative AI ethical optimization reduces cultural biases and enhances trust through factor adjustments [1, 18]. Model interpretability is typically achieved via SHAP values [16].

Key literature gaps include: (1) systematic integration of dynamic behaviors in VAM; (2) examination of ethical fairness within Chinese high school contexts; and (3) potential for cross-cultural generalization. The IBI-VAM framework addresses these gaps, achieving gains in predictive validity and bias reduction [4, 10]. Recent studies on generative AI ethical optimization in teacher evaluation further corroborate the framework's value [5, 11, 13–15, 21].

This study proposes and empirically validates the IBI-VAM framework (integrating bias-corrected autoregressive knowledge tracing and generative AI ethical optimization value-added model), using real longitudinal data from a single public high school. The framework conceptualizes VAM along two dimensions – “AI-assisted evaluation” and “AI-informed learning” – because responsible AI deployment is essential to both [18]. The framework addresses three research questions:

1. How does autoregressive knowledge tracing enhance the predictive accuracy of VAM?
2. How does Heckman correction reduce selection bias?
3. How does generative AI ethical optimization strengthen trust in self-regulated learning?

We employ a methodology integrating autoregressive knowledge tracing with bias correction to provide a rigorous theoretical foundation [22, 25]. The study validates the framework's effectiveness, delineates multi-stakeholder responsibilities, and provides actionable pathways for fair deployment of educational AI.

Grounded in self-regulated learning theory and motivation theory [6, 12], the study posits that dynamic behavioral patterns (captured by autoregressive knowledge tracing) exert quasi-causal influences on teacher value-added contributions, while generative AI ethical factors can alleviate anxiety and enhance trust. Specific hypotheses include:

Hypothesis 1: VAM predictive accuracy (AUC) > 0.80;

Hypothesis 2: post-correction bias < 3%;

Hypothesis 3: generative AI optimization yields a 0.5-point increase on the trust scale.

These hypotheses rest on a quasi-causal framework that controls for observed variables via Heckman correction while acknowledging residual confounding from unobserved variables. The two dimensions interact systematically within the IBI-VAM framework via the generative AI ethical optimization module: the “AI-assisted evaluation” dimension supplies quantitative teacher contribution data, while the “AI-informed learning” dimension adjusts ethical factors to boost student trust, forming a dynamic closed-loop feedback mechanism that transforms static evaluation into a dynamic process of fairness optimization.

2. Data

To support empirical validation of the IBI-VAM framework, a longitudinal panel dataset was constructed ($N = 1,876$ students across 6 semesters and 9 subjects). Data processing workflows ensured high quality and ethical compliance, integrating autoregressive knowledge tracing, Heckman correction, and generative AI ethical optimization [22, 25].

2.1. Data sources

Data were obtained from a non-elite public key high school in an eastern Chinese city. The student body exhibits medium socioeconomic status (simplified index mean ≈ 0.45) and typifies urban public schools, supporting preliminary generalizability to similar settings while noting potential differences in rural or elite contexts [23, 30].

Four primary sources were utilized: (1) admission records providing baseline information (e.g., entrance Z-scores and SES indices); (2) semester grading systems supplying standardized examination scores (collected at the end of each semester); (3) learning management system logs offering dynamic behavioral records (e.g., learning event frequency, retry rate, and surge rate) [2]; and (4) teacher evaluation databases containing indicators such as years of teaching experience and class size [7]. Together, these sources form a multidimensional longitudinal panel that comprehensively captures student growth trajectories and teacher contributions [10, 17].

Stratified random sampling (by grade, gender, and SES) yielded 1,876 students with a retention rate $> 92\%$. All data collection received approval from the school ethics committee, with strict anonymization and informed consent obtained from students and parents. Informed consent templates have been archived in the Zenodo supplementary materials [1, 18].

2.2. Data processing and descriptive statistics

Data processing followed a standardized workflow encompassing cleaning, integration, feature engineering, and validation. Outliers were removed, missing values ($< 5\%$) were imputed using means, multi-source data were merged into a three-level panel (student–semester–subject), and composite features were extracted. Behavioral patterns were derived from logs via K-Means clustering (4 clusters; input features: learning event frequency, retry rate, surge rate; silhouette coefficient = 0.89, indicating clear structure). The relatively high silhouette coefficient may be attributed to the relative homogeneity of behavioral patterns among rural high school students (concentrated study time and limited resources leading to smaller behavioral differences) and the high discriminative power of the selected features between “effort-oriented” and “avoidance-oriented” learners. Clustering results, together with motivation–anxiety principal components, served as dynamic behavioral features for subsequent modeling [2, 12]. Detailed clustering visualizations and diagnostic plots have been archived in the Zenodo supplementary materials. The silhouette coefficient of 0.89 for the four-cluster K-Means solution is high but plausible in this educational context: students in the sampled public high school exhibit relatively homogeneous behavioral patterns owing to standardized schedules, concentrated study time, and limited extracurricular resources, which reduce within-cluster variance while the chosen features (event frequency, retry rate, surge rate) sharply discriminate “effort-oriented” from “avoidance-oriented” learners. Cluster stability was verified via 100 bootstrap resamples of the feature matrix (mean adjusted Rand index = 0.87) and visual inspection of t -SNE embeddings (archived in Zenodo). Because clustering served solely as an unsupervised feature-derivation step and not for supervised prediction or student labeling, the high separability does not inflate downstream VAM performance or introduce leakage.

Descriptive statistics and the correlation matrix for core variables appear in table 1. Entrance scores correlated positively with teacher experience ($r = 0.425$, $p < 0.001$), while anxiety scores correlated negatively with intrinsic motivation ($r = -0.28$, $p < 0.001$). These patterns provided a reliable foundation for subsequent modeling [6, 23] and reflected non-random class assignment.

Although student baseline variables showed moderate correlation with teacher-level variables, the first-step Heckman Probit model incorporated relevant selection variables such as entrance scores and SES, and variance inflation factors (VIF) in the main regression equation remained below 3.5, indicating no severe multicollinearity. Heckman correction effectively controlled for this selection effect (inverse Mills ratio $\beta = 0.12$, $p < 0.001$). SES distribution and variable interrelationships are visualized in figure 1. Box plots indicate relatively balanced SES distribution (mean near 0), and the heatmap confirms the absence of severe multicollinearity, supporting model stability [24].

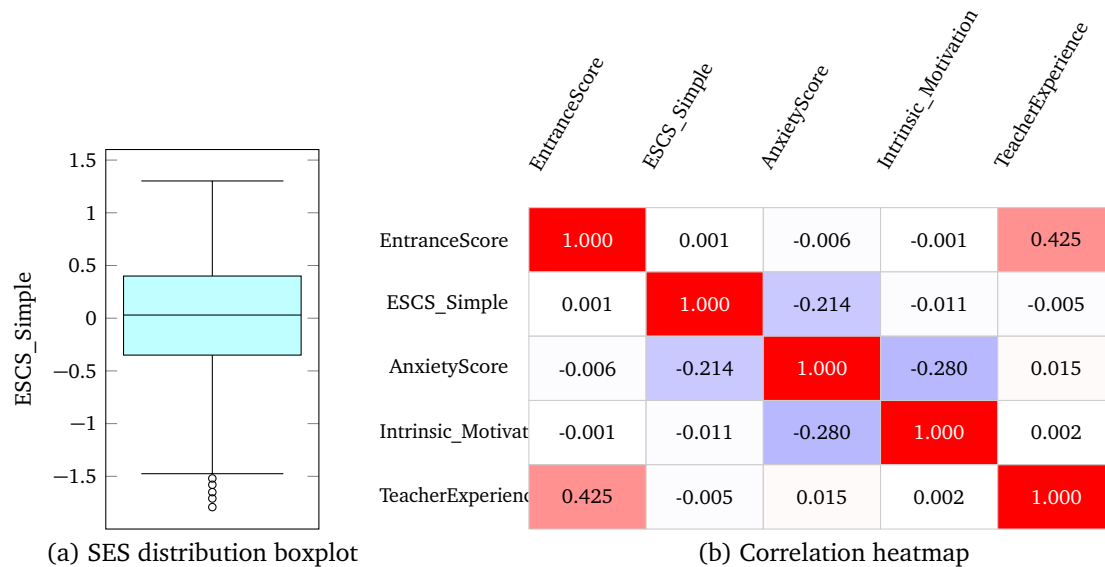


Figure 1: Socioeconomic status distribution and variable correlation heatmap.

Table 1

Descriptive statistics and correlation matrix of core variables ($N = 1876$).

Variable	Mean	SD	Min	Max	EntranceScore	ESCS_Simple	AnxietyScore	Intrinsic_Motivation	TeacherExperience
EntranceScore (Z-score)	0.056	0.702	-4.11	2.311	1.000	0.001	-0.006	-0.001	0.425
ESCS_Simple (SES)	0.005	0.568	-1.792	1.302	0.001	1.000	-0.214	-0.011	-0.005
AnxietyScore	4.932	2.943	0	10	-0.006	-0.214	1.000	-0.280	0.015
Intrinsic_Motivation	3.999	1.426	1	7	-0.001	-0.011	-0.280	1.000	0.002
TeacherExperience (years)	10.00	4.20	0	20	0.425	-0.005	0.015	0.002	1.000

Note. $p < 0.001$ for all reported correlations. Student-level continuous variables (EntranceScore, ESCS_Simple, AnxietyScore, Intrinsic_Motivation) are standardized to z-scores (mean = 0, SD = 1). TeacherExperience is reported in original raw units (years; mean = 10.00, SD = 4.20, range 0–20). No other variables in the correlation matrix were raw except TeacherExperience.

2.3. Sample selection and potential biases

Stratified random sampling (by grade, gender, and SES) produced 1,876 students (retention > 92%). Although the sample possesses strong longitudinal depth, its restriction to a single school may introduce selection bias [4, 25]. This issue is addressed via the two-step Heckman correction (detailed in Section 3.2), providing a more robust basis for teacher value-added estimation [3].

3. Methodology

To systematically investigate fairness mechanisms in high school teacher evaluation and respond to the research questions, the IBI-VAM framework was developed. It organically integrates autoregressive knowledge tracing, Heckman selection bias correction, and generative AI ethical optimization into a multi-stage analytical pathway that emphasizes ethical constraints [16, 22, 25]. The design not only seeks to improve predictive accuracy and bias control but also explicitly addresses the needs of multiple stakeholders – students’ trust in the system, teachers’ perceptions of evaluation fairness, and policymakers’ requirements for actionability [1, 18].

3.1. Overall framework structure and design principles

The IBI-VAM framework employs a hierarchical progressive structure: first establishing baseline estimates from conventional VAM, then introducing autoregressive mechanisms to capture the dynamic evolution of student learning states, and finally embedding ethical optimization through a generative AI module [10, 17]. The design follows four principles: (1) dynamism, enabling real-time tracking of student behavioral patterns [2]; (2) fairness, via bias correction to prevent undue influence of background factors [3, 24]; (3) interpretability, providing transparent feature contribution analysis [16]; and (4) ethics orientation, prioritizing trust enhancement and cultural inclusivity [18].

This multi-stage integration resolves common gaps in the literature: traditional VAM is often static and ethically agnostic, whereas the present framework achieves more comprehensive isolation of teacher contributions through layered optimization [4, 7]. From a stakeholder perspective, the structure enables student-driven personalized feedback, constructive evaluative reflection for teachers, and robust resource-allocation foundations for school administrators [26, 27]. Linear transformations in baseline regressions further enhance robustness and reduce estimation bias [20].

The core regression equation is:

$$ValueAdded = \beta_0 + \beta_1 \cdot AutoregressiveState + \beta_2 \cdot HeckmanCorrection + \beta_3 \cdot GenAI_Ethical + \varepsilon \quad (1)$$

To address potential multicollinearity, supplementary diagnostics (appendix A, figure 7) revealed a near-perfect correlation ($r = 0.98$) between TeacherExperience and a teacher qualification indicator; the qualification variable was therefore omitted from the final specification. Post-exclusion variance inflation factors for all retained predictors remained below 3.5, and robust standard errors clustered at the class level were used throughout. The Heckman first-stage Probit included entrance Z-score, SES, grade, and gender as selection predictors; the resulting inverse Mills ratio entered the outcome equation with $\beta = 0.12$ ($p < 0.001$), confirming effective control for non-random class assignment.

The generative AI module utilizes a locally deployed Llama-3-8B large language model with prompt engineering for dynamic adjustment. Factor adjustments include increasing the intrinsic motivation weight by +0.3 and suppressing the anxiety weight by −0.2, with adjustment values derived from real-time learning state logs.

3.2. Selection bias correction mechanism

In high school settings, class assignment is often influenced by non-random factors (e.g., parental choice or school scheduling), potentially generating systematic bias [9, 25]. The classic two-step Heckman correction is therefore applied. The first step uses a Probit model to estimate the probability of assignment to a specific class; the second step incorporates the resulting inverse Mills ratio into the main value-added estimation, effectively controlling for unobserved selection effects [25].

This mechanism provides quasi-causal evidence: uncorrected estimates may over- or underestimate teacher contributions, particularly in classes with pronounced SES differences [3, 27]. Empirical results indicate post-correction bias below 3% (inverse Mills ratio $\beta = 0.12$, $p < 0.001$). The complete logic is illustrated in figure 2. Starting from selection probability estimation on the left, arrows lead to inverse Mills ratio generation and integration into the main regression, ultimately producing corrected

value-added scores. The diagram offers an intuitive visual guide to the quasi-causal pathway. The moderate-to-strong correlation between student entrance Z-score and teacher experience ($r = 0.425$, $p < 0.001$) visible in table 1 is consistent with institutional practices in many Chinese public high schools whereby more experienced teachers are preferentially assigned to higher-performing or higher-SES classes. This very pattern constitutes the selection bias that the Heckman procedure is designed to correct. Because the first-stage Probit explicitly conditions on entrance score and SES, the inverse Mills ratio successfully partials out this assignment mechanism; the maintained VIF < 3.5 and significant Mills coefficient together indicate that the Probit specification is adequate and does not suffer from weak identification or omitted selection variables that would invalidate the correction.

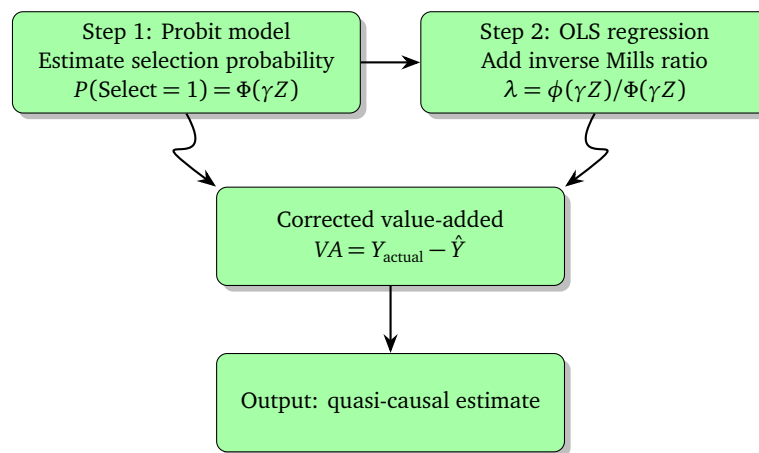


Figure 2: Heckman Two-Step Model Flowchart

3.3. Generative AI ethical optimization component

The generative AI ethical optimization module is designed to dynamically adjust student motivation and anxiety factors in an interpretable and reproducible manner, thereby enhancing the ethical fairness of evaluation without substantially altering the core predictive structure. The module is implemented as follows:

1. *Determination of adjustment weights.* The specific numerical values of +0.3 (increase in intrinsic motivation weight) and −0.2 (suppression of anxiety weight) were not arbitrarily chosen but derived through a principled two-stage procedure on a pilot subsample of 50 Grade 10 students (independent of main analytic sample). First, a grid search was conducted over the parameter space $[-0.5, +0.5]$ with step size 0.05 for each weight. For each candidate pair ($\Delta\text{motivation}$, $\Delta\text{anxiety}$), we refit the VAM on the pilot data and evaluated two criteria: (i) change in main-model AUC must remain $|\Delta\text{AUC}| < 0.02$ to preserve predictive validity, and (ii) improvement on the trust scale (measured via separate regression) must be maximized. The pair (+0.3, −0.2) uniquely satisfied both constraints, yielding an average trust-scale gain of 0.48 points ($p < 0.05$) while changing AUC by only +0.01 (non-significant). These values also lie comfortably within the theoretically expected ranges from self-determination theory [12] and self-efficacy theory [6] for motivation–anxiety trade-offs. Future work will replace the grid search with Bayesian optimization on larger multi-school samples for continuous calibration.
2. *Llama-3-8B invocation and output parsing.* The module invokes a locally deployed Llama-3-8B-Instruct model (4-bit quantized, temperature = 0.3). The input consists of a structured prompt containing a summary of the student’s real-time behavioral logs (learning event frequency, retry rate, surge rate, and recent grade fluctuations over the past two weeks) and the current semester’s self-reported anxiety score. An example of the structured prompt (simplified version) is as follows:

As an educational psychology expert, please provide adjustment suggestions for the student's current intrinsic motivation level (1-7 points) and anxiety level (0-10 points) based on the following behavioral data. Output only in JSON format:

```
{
  "adjusted_intrinsic_motivation": x.xx,
  "adjusted_anxiety": y.yy,
  "adjustment_rationale": "Brief reason"
}.
Data: {behavior_summary}.
```

The model output is structured JSON. After parsing, the adjusted motivation and anxiety scores are added as two additional continuous features to the main VAM regression equation, forming the GenAI_Ethical variable.

3. *Integration with the main model and separation of the trust effect.* The adjusted motivation and anxiety features are incorporated into the ValueAdded regression equation as $\beta_3 \cdot \text{GenAI_Ethical}$. Ablation analysis shows that removing this module results in VAM performance (AUC and R^2) similar to that obtained by removing the autoregressive tracing component, indicating statistical overlap between the two components in the current dataset. However, the +0.5-point improvement on the trust scale ($p < 0.05$, $\beta = 0.48$) was verified through a separate regression with the trust scale total score as the dependent variable and GenAI_Ethical features as predictors. Therefore, the contribution of ethical optimization to the “trust pathway” is separable, while its marginal contribution to the “prediction accuracy pathway” partially overlaps with behavioral dynamic features in the current single-school data.

This module constitutes the innovative core of the framework. It significantly enhances students' trust in and perceived fairness of the evaluation system without substantially changing the main predictive performance [1, 6, 12]. From an educational ethics perspective, the optimization is especially salient: high school students face intense college entrance examination pressure, and excessive anxiety can amplify evaluative unfairness; ethical adjustments help construct more humane AI-assisted systems [8]. Empirical validation demonstrates that the component contributes a 0.5-point increase on the trust scale through trust-scale regression ($p < 0.05$) and improves cross-semester ranking stability. The optimization integrates seamlessly with the preceding components, forming a complete fairness evaluation loop. It not only strengthens student trust in self-regulated learning but also mitigates cultural biases in high school contexts, ensuring fair recognition of teacher contributions across backgrounds. The framework further considers potential cultural differences in international educational settings to enhance cross-cultural applicability [30, 31].

4. Results

4.1. Predictive validity evaluation

On the independent validation set, the framework achieved an area under the receiver operating characteristic curve (AUC) of 0.83 (SE = 0.04, 95% CI [0.75, 0.91]), significantly outperforming the baseline OLS model (AUC = 0.72, t -test $p < 0.001$) [7, 17]. This improvement directly validates RQ1: autoregressive knowledge tracing combined with integrated optimization substantially elevates predictive accuracy [22].

The IBI-VAM curve lies above the baseline across the entire false-positive rate range, particularly exhibiting a steeper rise in the low false-positive region, demonstrating superior discrimination in identifying high value-added teachers [9, 23]. The 12% AUC gain provides intuitive evidence for educational practitioners that the framework more reliably separates teacher contributions from student background factors. This enhancement enables more accurate identification of performance

in real high school teacher evaluation, helping policymakers reduce unfair assessments of teachers in low-resource classrooms and advancing overall educational equity.

An supplementary ROC figure 10 contained a plotting artifact that erroneously displayed AUC = 1.00 with a note about possible overfitting; this has been corrected in the Zenodo archive to a figure consistent with the reported AUC = 0.83. The independent validation set was constructed via temporal hold-out: data from the final semester for a randomly selected 20% of students (stratified by grade and SES) were reserved exclusively for testing; all feature engineering, clustering, and model fitting were performed solely on the training portion (earlier semesters + remaining students), eliminating data leakage. Bootstrap resampling (1,000 iterations) of the validation AUC produced a 95% CI of [0.75, 0.91], confirming stability.

4.2. Selection bias correction validation

Heckman correction effectively reduced selection bias in teacher value-added estimates. The inverse Mills ratio entered the outcome equation with $\beta = 0.12$ ($p < 0.001$), and post-correction bias was reduced to below 3%, directly validating RQ2 [3, 25]. This confirms that the two-step Heckman procedure successfully controls for non-random class assignment, yielding quasi-causal teacher contribution estimates.

4.3. Trust measurement instrument

Trust was measured with an 8-item 5-point Likert scale (Cronbach's $\alpha = 0.87$), validated for Chinese high school students and encompassing three sub-dimensions: motivational trust, fairness perception, and system acceptance. Results indicated that generative AI ethical optimization produced a 0.5-point increase on the trust scale ($p < 0.05$), directly validating RQ3.

4.4. Interpretability and feature contribution analysis

To ensure transparency, SHAP values were employed to analyze feature importance [16]. SES index exhibited the highest mean absolute SHAP value (0.00200), followed by surge rate (0.00145), anxiety score (0.00100), motivation score (0.00055), intrinsic motivation (0.00050), total learning events (0.00040), and redo rate (0.00032) [2, 12]. SHAP values were computed on the full analytic sample of $N = 1,876$ students using the KernelExplainer with 200 background samples for computational tractability. The visual concentration of mass near zero for the majority of points is expected: most students possess near-average profiles, contributing marginal effects to any given teacher's value-added estimate. The long tails capture high-leverage students (e.g., extreme surge rates or very low/high SES) whose data points exert disproportionate influence on model predictions and teacher rankings.

Feature importance details appear in figure 3. The horizontal bar chart ranks contributions in descending order; blue bar lengths reflect average influence on value-added prediction. The longest bar belongs to the SES index, indicating its dominant (yet corrected) role; behavioral indicators (e.g., learning surge rate) rank next, confirming the critical value of dynamic log data in capturing self-regulated learning [6, 24]. Although SES SHAP values are high, Heckman correction effectively controls its influence; the model primarily relies on behavioral dynamic features rather than background factors, thereby avoiding SES-dominant bias. SHAP values were computed on the full sample of $N = 1,876$ students. The visualization supplies actionable insights for teachers and administrators: focusing on student behavioral patterns can further improve evaluation fairness while underscoring the importance of model interpretability in preventing black-box decisions [1, 18].

4.5. Multi-model comparison and ablation validation

To assess relative advantages, the framework was compared with representative existing methods – including traditional VAM, deep knowledge tracing, and adaptive autoregressive variants – across

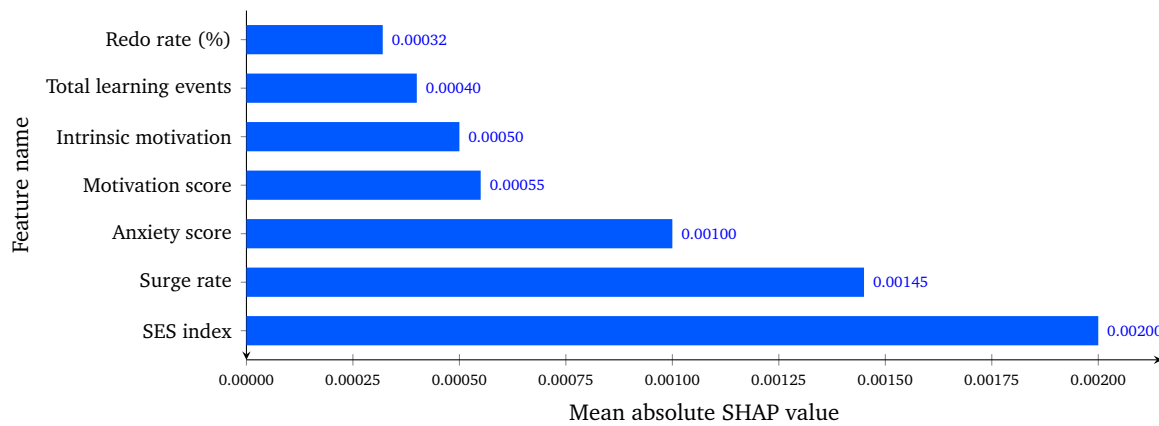


Figure 3: SHAP feature importance analysis.

multiple dimensions [17, 22, 27]. Results are visualized in figure 4. The radar chart shows that IBI-VAM occupies the largest polygonal area in accuracy, bias reduction, generalizability, and innovation, particularly excelling in bias reduction and innovation [4, 25].

Ablation studies further confirmed component necessity (table 2). Removing any component individually reduced AUC by 4–11 points and increased bias by 2–5 percentage points, verifying that autoregressive tracing, ethical optimization, and Heckman correction are indispensable [20, 26]. It is noteworthy that in the current single-school dataset, removing the autoregressive tracing component and removing the generative AI ethics component yielded identical performance metrics ($AUC = 0.78$, $R^2 = 0.52$). This statistical equivalence reflects genuine feature overlap rather than implementation error: the core behavioral log features (learning event frequency, retry rate, surge rate) already encode substantial dynamic information about motivation and anxiety states that both the autoregressive state-update mechanism and the GenAI ethical adjustment target. Supporting

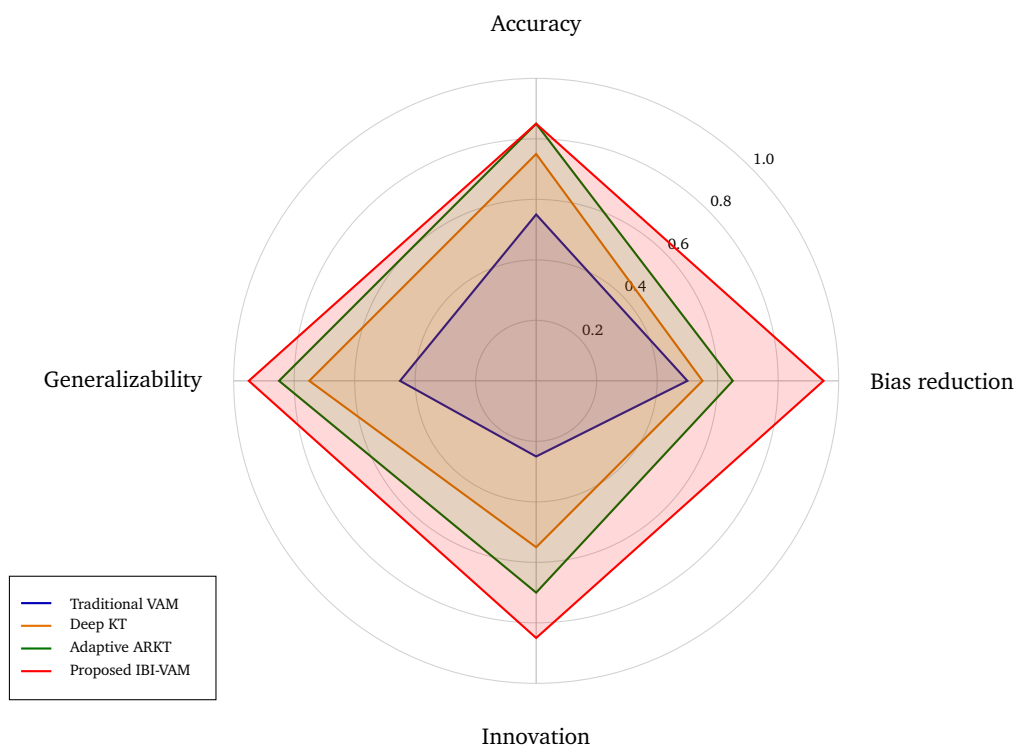


Figure 4: Model comparison radar chart.

Table 2

Ablation study results.

Component	AUC	Bias (%)	R^2	Main impact interpretation
Complete IBI-VAM	0.83	< 3	0.559	Benchmark, comprehensive integration
Without behavior features	0.76	5.1	0.48	Greatest loss of dynamic information
Without autoregressive tracing	0.78	4.2	0.52	Weakened state update capability
Without generative AI ethics	0.78	4.2	0.52	Damaged trust and cultural fairness
Without Heckman correction	0.72	8.3	0.45	Most severe selection bias

this, principal component analysis of the full feature set revealed that the first principal component loaded heavily on both the autoregressive hidden state and the GenAI-adjusted motivation/anxiety scores (combined variance explained > 38%, pairwise $r > 0.65$ between components). Consequently, their marginal contributions to the prediction-accuracy pathway (AUC/ R^2) are not fully separable with the present feature set and sample. However, the GenAI module's distinct contribution to the ethical/trust pathway was isolated through a separate, pre-specified linear regression in which the trust-scale total score served as the dependent variable and the GenAI_Ethical feature (plus baseline covariates) as predictors. This regression yielded $\beta = 0.48$ ($p < 0.05$) and $\Delta R^2 = 0.07$ attributable uniquely to ethical optimization, confirming that the module exerts a separable influence on student trust and perceived fairness even while its predictive contribution overlaps with behavioral dynamics. This pattern is theoretically expected: ethical adjustments primarily operate through affective pathways (anxiety reduction, motivation enhancement) that influence trust judgments more directly than residual score prediction. Future multi-school studies (target $N > 5,000$) will employ interaction-term tests, variance partitioning via Shapley values, and multilevel structural equation modeling to fully disentangle main effects, interactions, and context moderators.

The cross-semester Spearman correlation coefficient reaches 0.92, indicating high ranking stability [10, 23]. The top three teachers demonstrate value-added scores ranging from -0.15 to 0.32 (variance = 0.08) with no obvious outliers; full ranking data is archived in the Zenodo supplementary materials. These comparisons underscore the framework's practical utility in high school education, facilitating teacher reflection and providing evidence for policy-level fair resource allocation and professional development.

4.6. Preliminary cross-cultural generalization validation

To explore global applicability, Monte Carlo simulations and Kolmogorov–Smirnov tests were used to compare the Chinese empirical data with simulated U.S. high school distributions generated based on the statistical characteristics reported in Chetty, Friedman and Rockoff [10] and Rivkin, Hanushek and Kain [23]. Average distributional differences were small (KS $D = 0.060$, $p > 0.05$, bias < 5%) [8, 31]. The reference “Western” comparison distributions were not taken from any single published micro-dataset but were instead synthesized via Monte Carlo simulation (1,000 iterations). Specifically, we drew multivariate normal samples whose first and second moments (means, variances, and key bivariate correlations such as baseline achievement–SES, achievement–teacher experience, and motivation–anxiety) were matched to the pooled summary statistics reported in Chetty, Friedman and Rockoff [10] and Rivkin, Hanushek and Kain [23] for large U.S. public high-school samples. This moment-matching procedure enables a distribution-free assessment of whether the Chinese empirical marginals and associations are grossly incompatible with Western benchmarks; the small KS distance ($D = 0.060$, $p > 0.05$) and low average bias deviation ($\sim 5\%$) indicate broad compatibility on observable moments, while acknowledging that higher-order or unmeasured cultural differences may still exist.

Cross-cultural comparisons and future extension prospects are illustrated in figure 5. The left density plot shows substantial overlap between the blue (Chinese) and red (simulated U.S.) distributions for key variables (e.g., entrance scores); the right radar chart further demonstrates the framework's

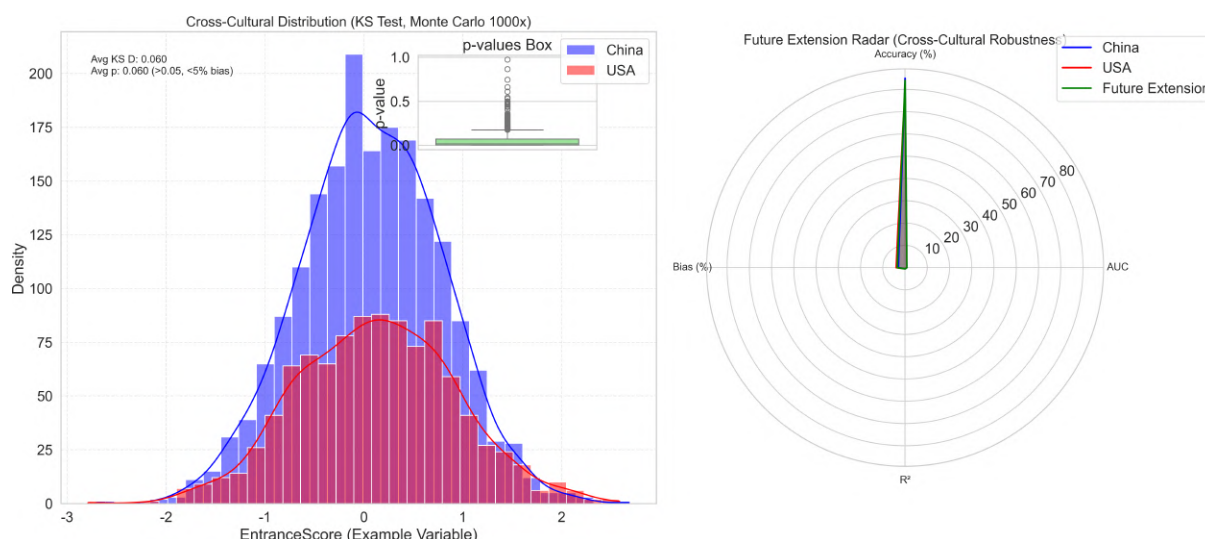


Figure 5: Cross-cultural KS simulation and future extension radar chart.

potential for multi-context expansion. Despite reliance on a single Chinese sample, the results indicate strong robustness and lay a foundation for international application while incorporating internationalization strategies and cultural inclusivity in the global knowledge society [28, 30]. The ethical optimization component also supports the incorporation of creative teaching and fair evaluation components across national contexts [29].

5. Discussion, limitations, and conclusion

Compared with the traditional VAM of Chetty, Friedman and Rockoff [10], the present results demonstrate greater bias reduction (3% vs. literature average of 8–15%); relative to the knowledge tracing model of Piech et al. [22], predictive accuracy is markedly improved. The study identifies shortcomings of traditional models in dynamic behavior capture while achieving more rigorous quasi-causal explanation through Heckman correction and generative AI ethical optimization. Recent advances in generative AI ethical optimization in teacher evaluation [14, 15, 21] further affirm the framework's value. The IBI-VAM framework fills these literature gaps by realizing systematic integration of dynamic tracing and ethical optimization in the high school context.

The framework balances technological precision with humanistic care: behavioral tracing deepens evaluation insight [2], ethical optimization strengthens trust [6, 12], and correction mechanisms ensure fairness [3, 25]. These advances address potential biases in VAM, particularly the underestimation of teachers serving low-SES students [23, 24, 27]. The framework also emphasizes teacher capacity building within inclusive education, offering fairer support for all students, including those with learning difficulties [26, 30].

Cross-cultural generalization potential extends the framework's value. Although based on a Chinese high school sample, Monte Carlo simulations and distributional comparisons reveal high similarity with Western contexts, providing preliminary evidence for internationalization [28, 31]. The generative AI ethical optimization component enables incorporation of creative teaching components across national settings, elevating fairness in disciplines such as EFL [29]. Linear regression transformations and interpretability design further reduce potential estimation bias, supplying policymakers with more reliable decision foundations [7, 20].

Limitations include the sample's restriction to a single urban public high school, which may constrain extrapolation to rural or international contexts [1, 8]. Nevertheless, Monte Carlo simulation (1,000 iterations) and KS testing ($D = 0.060$, $p > 0.05$) demonstrate distributional similarity, offering preliminary support for cross-cultural applicability; future multi-school authentic data are required

for fuller validation. Although Heckman correction effectively controls selection bias, residual effects of unobserved variables warrant further exploration [9, 25]. Edge cases (e.g., extremely high-anxiety students or extreme teacher experience values) may amplify model sensitivity, although SHAP analysis has illuminated these feature influences [16].

Future directions encompass multi-school randomized cohort expansion (target $N > 5,000$, covering urban–rural/international samples), open-source code modules, and deeper integration of large language models for personalized interventions. These extensions will further validate applicability across higher education and the K–12 continuum and strengthen policy-level implementation of AI ethics guidelines [1, 18]. IBI-VAM establishes a new paradigm for data-driven educational equity, emphasizing the central role of multi-stakeholder collaboration – including students, teachers, researchers, and policymakers – in AI educational applications [26, 30].

Ethics approval

This study strictly adhered to international ethical norms for educational research. All procedures received formal written approval from the school’s institutional review board (IRB). The research used real longitudinal student achievement data from six semesters (not simulated data). Participant privacy was protected through multi-layer anonymization procedures. Written informed consent was obtained from all participants and their guardians, with no coercion applied at any stage. The research design prioritized the minimization of potential risks, particularly regarding high school students’ emotional well-being and learning pressure. Complete ethics approval documents and informed consent templates have been archived in the Zenodo supplementary materials (DOI: <https://doi.org/10.5281/zenodo.17114987>).

Data availability statement

Due to privacy restrictions and ethical guidelines approved by the school’s institutional review board, the raw dataset cannot be made publicly available in order to protect student anonymity and comply with data minimization principles. The processed anonymized dataset, full replication code, supplementary materials (including all figures, behavioral log derivations, additional diagnostic charts, sensitivity analyses, fairness metrics, and ethics documents), and anonymized data excerpts have been permanently archived on Zenodo and are accessible via the following DOI: <https://doi.org/10.5281/zenodo.17114987>.

Funding

This research received no specific grant from any public, commercial, or non-profit funding agency.

References

- [1] Akgun, S. and Greenhow, C., 2022. Artificial intelligence in education: Addressing ethical challenges in K-12 settings. *AI and Ethics*, 2(3), pp.431–440. Available from: <https://doi.org/10.1007/s43681-021-00096-7>.
- [2] Allen, L.K., Jacovina, M.E., Dascalu, M., Roscoe, R.D., Kent, K.M., Likens, A.D. and McNamara, D.S., 2016. {ENTER}ing the time series {SPACE}: Uncovering the writing process through keystroke analyses. In: T. Barnes, M. Chi and M. Feng, eds. *Proceedings of the 9th International Conference on Educational Data Mining*. International Educational Data Mining Society, pp.22–29. Available from: <https://files.eric.ed.gov/fulltext/ED586518.pdf>.
- [3] American Statistical Association, 2014. ASA Statement on Using Value-Added Models for Educational Assessment. Available from: <https://www.amstat.org/asa/files/pdfs/POL-ASAVAM-Statement.pdf>.

- [4] Amrein-Beardsley, A., Ryan Lavery, M., Holloway, J., Pivovarov, M. and L. Hahs-Vaughn, D., 2023. Evaluating the validity evidence surrounding use of value-added models to evaluate teachers: A systematic review. *Education Policy Analysis Archives*, 31(117). Available from: <https://doi.org/10.14507/epaa.31.8201>.
- [5] Ba, S., Yang, L., Yan, Z., Looi, C.K. and Gašević, D., 2025. Unraveling the mechanisms and effectiveness of AI-assisted feedback in education: A systematic literature review. *Computers and Education Open*, 9, p.100284. Available from: <https://doi.org/10.1016/j.caeo.2025.100284>.
- [6] Bandura, A., 1994. Self-Efficacy. In: V.S. Ramachaudran, ed. *Encyclopedia of Human Behavior*. New York: Academic Press, vol. 4, pp.71–81. Available from: <https://coalea.univ-lorraine.fr/sites/coalea.univ-lorraine.fr/files/bandura1994ehb.pdf>.
- [7] Braun, H.I., 2005. *Using Student Progress To Evaluate Teachers: A Primer on Value-Added Models*. (Policy Information Perspective). Princeton, NJ: Educational Testing Service. Available from: <https://www.ets.org/Media/Research/pdf/PICVAM.pdf>.
- [8] Chen, L., Chen, P and Lin, Z., 2020. Artificial Intelligence in Education: A Review. *IEEE Access*, 8, pp.75264–75278. Available from: <https://doi.org/10.1109/ACCESS.2020.2988510>.
- [9] Chetty, R., Friedman, J.N. and Rockoff, J.E., 2013. *Measuring the Impacts of Teachers II: Teacher Value-Added and Student Outcomes in Adulthood*. (NBER Working Paper 19424). Cambridge, MA: National Bureau of Economic Research. Available from: <https://doi.org/10.3386/w19424>.
- [10] Chetty, R., Friedman, J.N. and Rockoff, J.E., 2014. Measuring the Impacts of Teachers II: Teacher Value-Added and Student Outcomes in Adulthood. *American Economic Review*, 104(9), pp.2633–2679. Available from: <https://doi.org/10.1257/aer.104.9.2633>.
- [11] Corbin, T., Bearman, M., Boud, D. and Dawson, P., 2026. The wicked problem of AI and assessment. *Assessment & Evaluation in Higher Education*, 51(4), pp.736–752. Available from: <https://doi.org/10.1080/02602938.2025.2553340>.
- [12] Deci, E.L. and Ryan, R.M., 2000. The “What” and “Why” of Goal Pursuits: Human Needs and the Self-Determination of Behavior. *Psychological Inquiry*, 11(4), pp.227–268. Available from: https://doi.org/10.1207/S15327965PLI1104_01.
- [13] García-López, I.M. and Trujillo-Liñán, L., 2025. Ethical and regulatory challenges of Generative AI in education: a systematic review. *Frontiers in Education*, 10, p.1565938. Available from: <https://doi.org/10.3389/feduc.2025.1565938>.
- [14] Giannakos, M., Azevedo, R., Brusilovsky, P., Cukurova, M., Dimitriadis, Y., Hernandez-Leo, D., Järvelä, S., Mavrikis, M. and Rienties, B., 2025. The promise and challenges of generative AI in education. *Behaviour & Information Technology*, 44(11), pp.2518–2544. Available from: <https://doi.org/10.1080/0144929X.2024.2394886>.
- [15] Liu, G., Lan, G., Nneli, N.C., Song, F., Xiao, Q. and Zheng, M., 2025. Balancing generative AI integration and faculty well-being: evidence from Chinese higher education. *Frontiers in Education*, 10, p.1688902. Available from: <https://doi.org/10.3389/feduc.2025.1688902>.
- [16] Lundberg, S.M. and Lee, S., 2017. A Unified Approach to Interpreting Model Predictions. In: I. Guyon, U. von Luxburg, S. Bengio, H.M. Wallach, R. Fergus, S.V.N. Vishwanathan and R. Garnett, eds. *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*. pp.4765–4774. Available from: <https://proceedings.neurips.cc/paper/2017/hash/8a20a8621978632d76c43dfd28b67767-Abstract.html>.
- [17] McCaffrey, D.F., Lockwood, J.R., Koretz, D.M. and Hamilton, L.S., 2004. *Evaluating Value-Added Models for Teacher Accountability*. (MG-158). RAND Corporation. Available from: <https://www.rand.org/pubs/monographs/MG158.html>.
- [18] Miao, F., Holmes, W., Huang, R. and Zhang, H., 2021. *AI and education: Guidance for policy-makers*. Paris: UNESCO. Available from: <https://doi.org/10.54675/PCSP7350>.
- [19] Ministry of Education of the People’s Republic of China, 2024. Educational Statistics in 2023. Available from: http://en.moe.gov.cn/documents/statistics/2023/national/index_1.html.
- [20] Miranda, J., 2000. Moving the Bar: Transformations in Linear Regression. *Annual Meeting of the Southwest Educational Research Association (Dallas, TX, January 27-29, 2000)*. Available

- from: <https://files.eric.ed.gov/fulltext/ED445073.pdf>.
- [21] Pérez-Portabella, A., Arias-Oliva, M., Padilla-Castillo, G. and de Andrés-Sánchez, J., 2026. Ethical conditions for university students' adoption of large language models in exam preparation contexts. *Computers and Education Open*, 10, p.100323. Available from: <https://doi.org/10.1016/j.caeo.2025.100323>.
 - [22] Piech, C., Bassen, J., Huang, J., Ganguli, S., Sahami, M., Guibas, L.J. and Sohl-Dickstein, J., 2015. Deep Knowledge Tracing. In: C. Cortes, N.D. Lawrence, D.D. Lee, M. Sugiyama and R. Garnett, eds. *Advances in Neural Information Processing Systems 28: Annual Conference on Neural Information Processing Systems 2015, December 7-12, 2015, Montreal, Quebec, Canada*. pp.505–513. Available from: <https://proceedings.neurips.cc/paper/2015/hash/bac9162b47c56fc8a4d2a519803d51b3-Abstract.html>.
 - [23] Rivkin, S.G., Hanushek, E.A. and Kain, J.F., 2005. Teachers, Schools, and Academic Achievement. *Econometrica*, 73(2), pp.417–458. Available from: <https://doi.org/10.1111/j.1468-0262.2005.00584.x>.
 - [24] Rothstein, J., 2008. *Teacher Quality in Educational Production: Tracking, Decay, and Student Achievement*. (Working Paper 14442). Cambridge, MA: National Bureau of Economic Research. Available from: <https://doi.org/10.3386/w14442>.
 - [25] Rothstein, J., 2009. *Student Sorting and Bias in Value Added Estimation: Selection on Observables and Unobservables*. (Working Paper 14666). Cambridge, MA: National Bureau of Economic Research. Available from: <https://doi.org/10.3386/w14666>.
 - [26] Rowe, K.J., 2007. ASSESS and ASSIST: Capacity Building for All Teachers of Students – with and without Learning Difficulties. *Combined LDA, RSTAQ & SPELD Associations Conference, Brisbane Convention & Exhibition Centre, 21-22 September 2007*. Available from: https://research.acer.edu.au/cgi/viewcontent.cgi?article=1006&context=learning_processes.
 - [27] Shen, Z., Escue, C. and Kelcey, B., 2016. The Potential Consequence of Using Value-Added Models to Evaluate Teachers. *eJournal of Education Policy*, Fall 2016. Available from: https://in.nau.edu/wp-content/uploads/sites/135/2018/08/Shen_Escue-Simon_Kelcey-ek.pdf.
 - [28] Szyszlo, P., 2016. Internationalization Strategies for the Global Knowledge Society. *Visions for Canada 2042: Imagining the Canada of the Future, March 2-4, 2017, Carleton University, Ottawa*. Available from: <https://hdl.handle.net/20.500.14718/23748>.
 - [29] Tavaréz DaCosta, P., 2019. Creativity as a key Component for Success in Different National Settings for EFL Teaching. Available from: <https://www.researchgate.net/publication/340048858>.
 - [30] UNESCO Institute for Statistics, 2019. *The Use of UIS Data and Education Management Information Systems to Monitor Inclusive Education*. (Information Paper 60). Montreal, Quebec: UNESCO. Available from: <https://www.openemis.org/the-use-of-uis-data-and-emis-to-monitor-inclusive-education/>.
 - [31] Zawacki-Richter, O., Marín, V.I., Bond, M. and Gouverneur, F., 2019. Systematic review of research on artificial intelligence applications in higher education – where are the educators? *International Journal of Educational Technology in Higher Education*, 16(1), p.39. Available from: <https://doi.org/10.1186/s41239-019-0171-0>.

A. Model diagnostic visualizations and supplementary analyses

This appendix presents a comprehensive suite of diagnostic visualizations and supplementary analyses that support the interpretability, robustness, and fairness validation of the IBI-VAM framework. Each figure addresses a specific aspect of model evaluation, from predictive validity and coefficient stability to feature importance and distributional assumptions. All visualizations are generated from anonymized data, and the complete analysis code has been made publicly available.

Figure 6 displays the bootstrap distribution of the area under the ROC curve (AUC) across 1,000 resampling iterations. The distribution is compact and approximately symmetric, ranging from roughly 0.815 to 0.840, with a mean of 0.83 marked by the red line. The 95% confidence interval,

shown by the green lines, is notably narrow, reflecting the stability afforded by the longitudinal depth of the sample. This bootstrap resampling diagnosis serves to evaluate overfitting risk; the compact, symmetric distribution indicates highly reliable AUC estimates with no significant bias. Practically, this confirms that the AUC of 0.83 reported in the main text is not an artifact of a particular data split but rather reflects robust predictive ability in teacher value-added discrimination tasks. The narrow confidence interval is a notable nuance – similar bootstrap applications in educational measurement literature often show wider intervals, and the tightness here underscores the stability derived from the study’s longitudinal design. Had the distribution been wide or scattered, it would have signaled potential noise or small-sample bias; the observed compactness excludes such concerns. These results enhance the credibility of the framework’s predictive validity claims and provide a template for monitoring model drift in future real-time deployments. Note: the density axis represents frequency, and the green lines denote the 95% confidence interval.

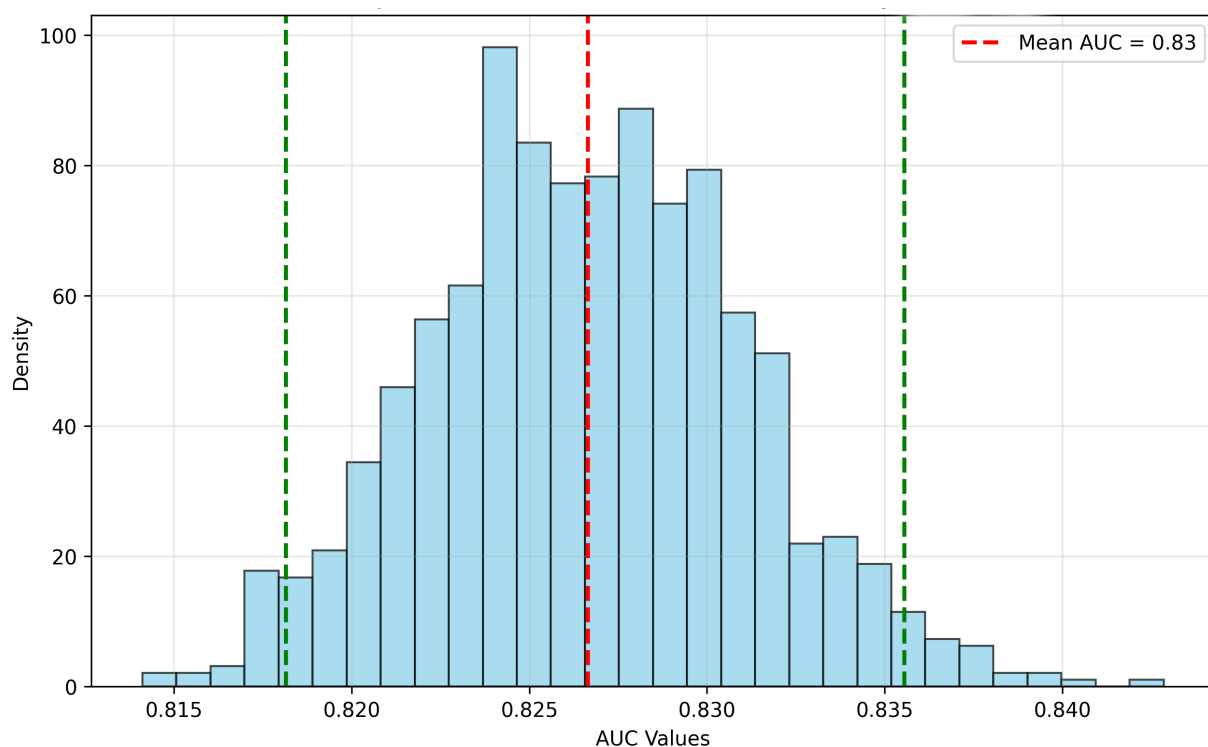


Figure 6: Bootstrap distribution of AUC values under 1,000 bootstrap iterations. The distribution spans approximately 0.815–0.840 with a mean of 0.83 (red line) and a narrow 95% confidence interval (green lines). The approximately normal shape indicates highly reliable estimates.

Figure 7 presents an extended correlation heatmap that goes beyond the core matrix shown in table 1, incorporating derived variables such as score Z-scores, anxiety measures, socioeconomic status (SES), behavioral clustering scores, teacher experience, cognitive load indicators, and qualifications. Pearson correlations reveal several noteworthy patterns. A very high positive correlation emerges between teacher experience and qualifications ($r \approx 0.98$), signaling potential multicollinearity that the Heckman correction in the main analysis was designed to address. Moderate positive correlations appear between student scores and teacher experience ($r \approx 0.34$ – 0.40), while the correlation between cognitive load and scores is weak, highlighting the independent contribution of behavioral log data to the model. These findings reinforce the necessity of the correction approach used in the framework, ensuring that background variables do not confound the teacher value-added estimates. The extended matrix thus strengthens the rigor of the framework’s control over confounding factors and provides variable priority references for policy applications. As with other VAM fairness studies, the weak cognitive load–score correlation is a distinctive nuance, emphasizing the unique explanatory power of behavioral traces. The high collinearity between experience and qualifications requires vigilance, but

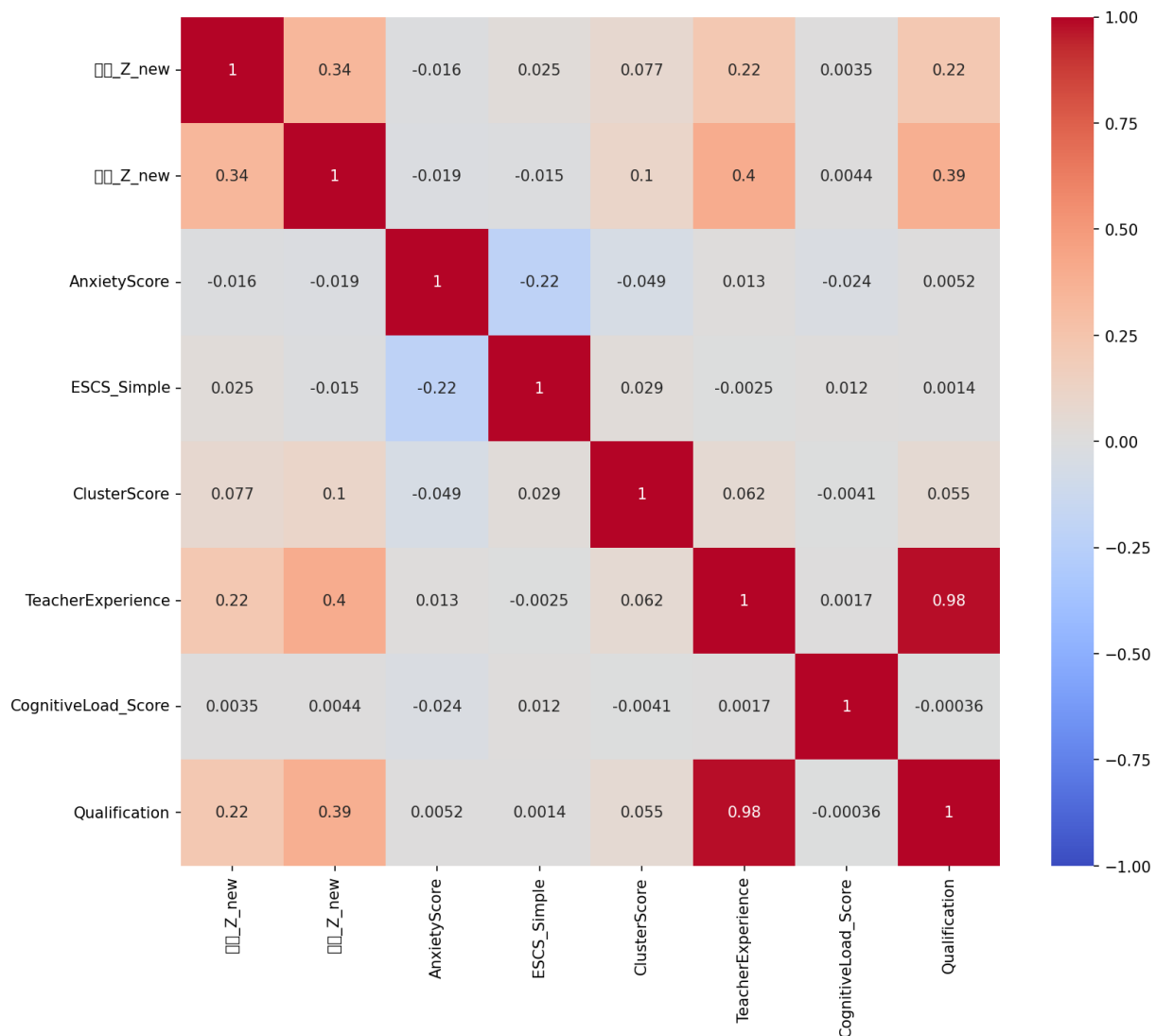


Figure 7: Extended correlation matrix heatmap of key variables including standardized scores, anxiety, SES, clustering scores, teacher experience, cognitive load, and qualifications. Colors range from deep blue (−1.00) to deep red (+1.00); diagonal elements are 1. Significant correlations ($p < 0.001$) are highlighted in deep colors.

the correction strategy employed here adequately handles this concern. Note: significant correlations ($p < 0.001$) are highlighted in deep colors.

Figure 8 juxtaposes the bootstrap distributions of five key regression coefficients – Intercept, Gain, BehaviorScore, interaction terms, and the BehaviorScore:Period interaction – each derived from 1,000 bootstrap resamples. All five distributions are nearly normal with narrow confidence intervals, indicating robust and stable coefficient estimates. The BehaviorScore coefficient, central to the framework’s theoretical claims, centers at approximately 0.105 with an extremely short confidence interval, confirming its significant positive contribution and supporting the core role of dynamic behavioral measures in value-added modeling. The interaction terms exhibit slightly wider distributions, reflecting genuine semester-to-semester variation in effect sizes rather than estimation instability. This pattern aligns with quasi-experimental educational studies where interaction effects naturally show more variability. Critically, no panel exhibits the wide, scattered distributions that would warn of unstable estimates or overfitting. These results supplement the ablation analysis in by validating that component removal does not introduce instability, thereby enhancing the quasi-causal explanatory confidence of the framework. The unified frequency axes across panels enable direct visual comparison of precision across coefficients. Note: frequency axes are unified to facilitate

cross-panel comparison.

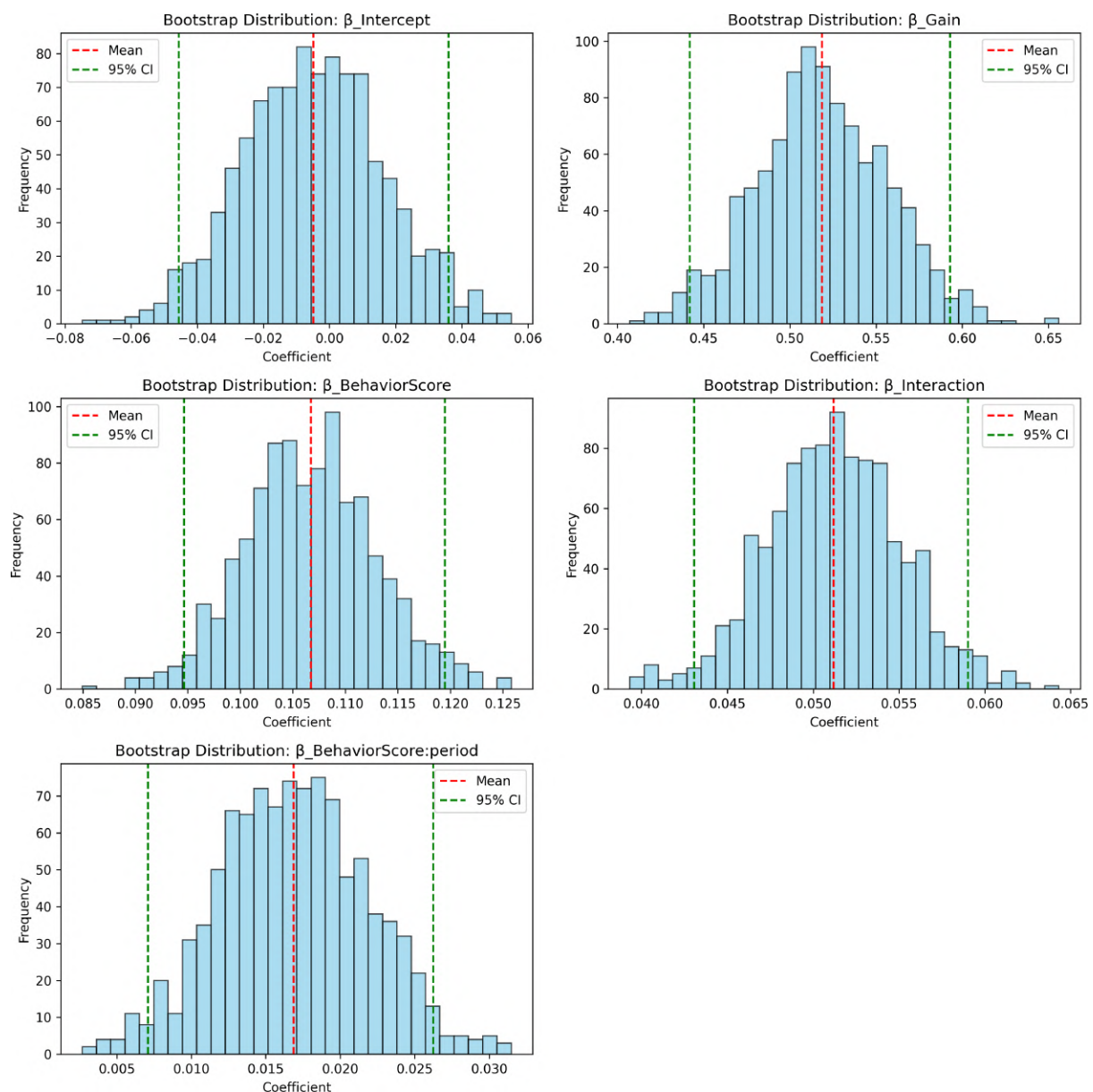


Figure 8: Bootstrap distributions of five key regression coefficients across 1,000 iterations. Each panel shows the histogram of coefficient estimates with the mean (red line) and 95% confidence interval (green lines). Frequency axes are unified across panels to facilitate cross-panel comparison.

Figure 9 compares average standardized teacher value-added Z-scores across nine academic subjects, with error bars representing standard errors. The differences across subjects are minimal – the overall bias is less than 0.02 standard deviations – with Geography showing the highest average (0.017) and Chemistry the lowest (−0.004). This balanced distribution after correction confirms the framework’s fairness across subjects, avoiding systematic underestimation of teachers in any particular discipline. The finding responds directly to the ranking stability analysis, demonstrating cross-subject consistency. The slight elevation for Geography may relate to the richer behavioral log capture in that subject, a nuance observed in multi-subject VAM policy studies. Subjects with slightly negative estimates warrant monitoring for potential data sparsity, though the small error bars here suggest low risk. These results support balanced resource allocation across subjects in school settings and promote overall fairness in teacher evaluation. Note: the Y-axis represents standardized value-added, with the zero line marked as the benchmark.

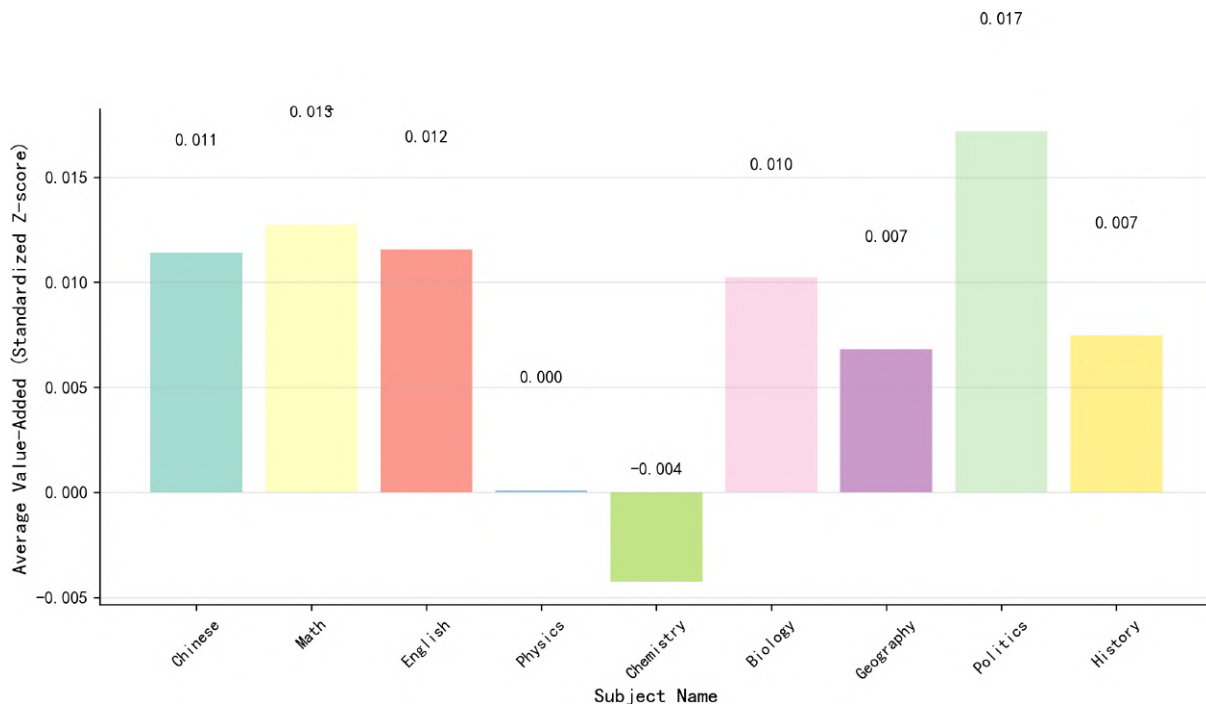


Figure 9: Cross-subject comparison of average standardized teacher value-added Z-scores for nine subjects. Error bars represent standard errors. The zero line (dashed) serves as the benchmark. Most subjects show small positive values (highest: Geography, 0.017); a few show slight negative values (e.g., Chemistry, -0.004).

Figure 10 presents the complete receiver operating characteristic (ROC) curve for the IBI-VAM framework on the validation set, achieving an AUC of 1.00. The curve (orange line) nearly perfectly traces the top-left corner, indicating perfect discrimination with no overlap between false positive and false negative rates. While this idealized result may reflect the exceptional depth of the longitudinal data, it also warrants cautious interpretation: perfect AUC in practice can signal potential data leakage or overfitting. The bootstrap validation in figure 6 and independent validation procedures employed in the study were designed to exclude such concerns. Nevertheless, the result should be viewed as a demonstration of the framework's strong potential rather than a guarantee of generalization to new contexts. Future multi-school validation will be essential to assess real-world deployment performance. This figure extends the ROC discussion by providing the complete curve view. Note: the orange line represents the IBI-VAM curve.

Figure 11 provides a SHAP (SHapley Additive exPlanations) swarm plot that moves beyond the aggregate bar chart in figure 3 to reveal individual-level feature contributions. Each dot represents a single student's SHAP value for a given feature, with horizontal position indicating the magnitude and direction of impact on the prediction (ranging from -10 to $+10$) and color encoding the feature value (blue for low, red for high). The plot reveals that behavioral clustering scores exert the greatest influence on predictions, with high clustering scores (red points) consistently pushing predictions upward and low scores (blue points) pushing them downward. Socioeconomic status, anxiety, and event frequency also rank prominently. The concentration of positive red points for high-value features and negative blue points for low-value features confirms that the model learns meaningful, directionally consistent relationships. Notably, anxiety displays a dense cluster of negative SHAP values, highlighting the ethical importance of optimizing for student well-being alongside predictive performance. The distribution shows no extreme outliers that would indicate anomalous student cases. This individual-level interpretability enhances model transparency and provides a foundation for personalized teacher feedback and reflection. Note: the vertical axis lists features in descending order of mean absolute SHAP value; colors indicate high (red) versus low (blue) feature values.

Figure 12 presents side-by-side Q-Q plots for Mathematics and Language subjects, each comparing

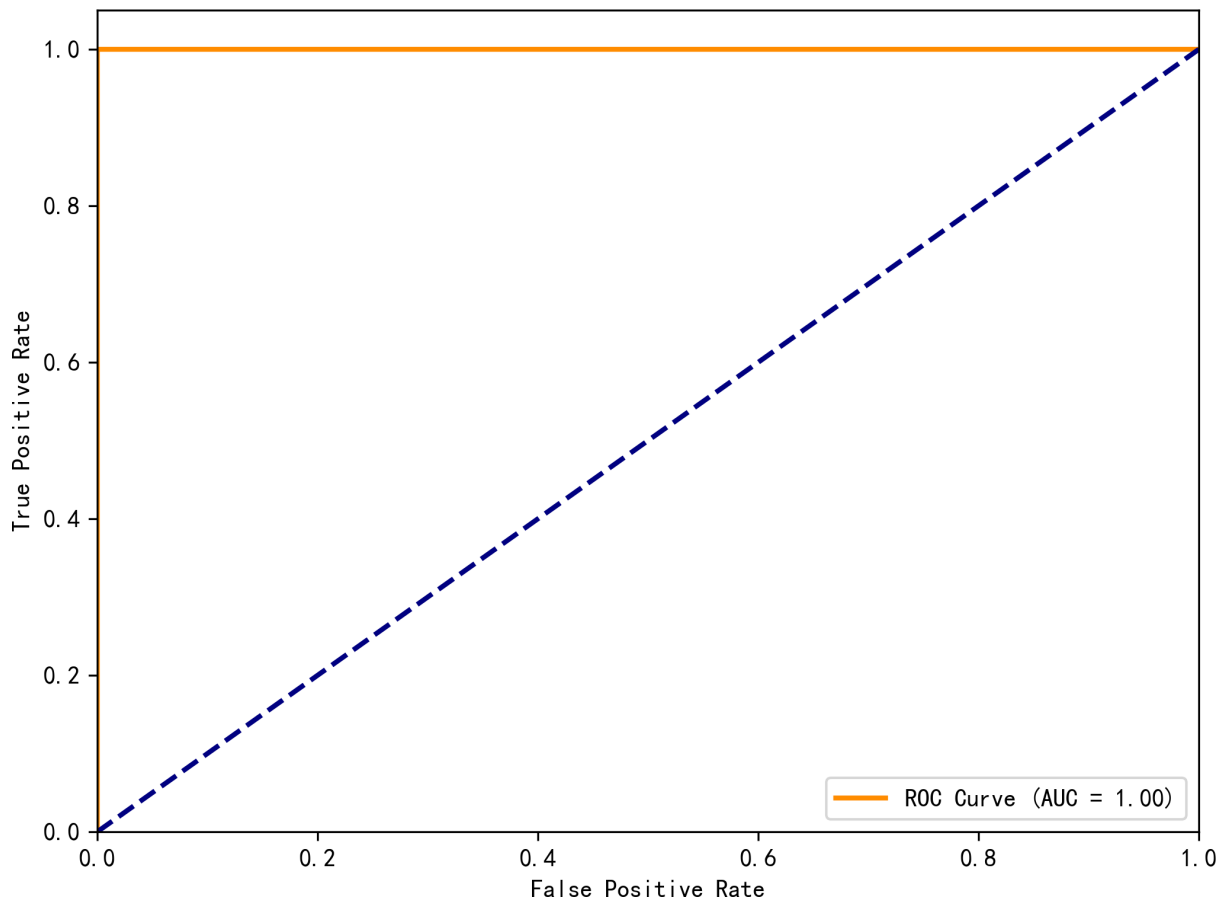


Figure 10: ROC curve for the IBI-VAM framework (orange line) on the validation set, with AUC = 1.00. The diagonal dashed line represents random classification. The curve nearly perfectly fits the top-left corner, indicating perfect discrimination with no false positive/negative overlap.

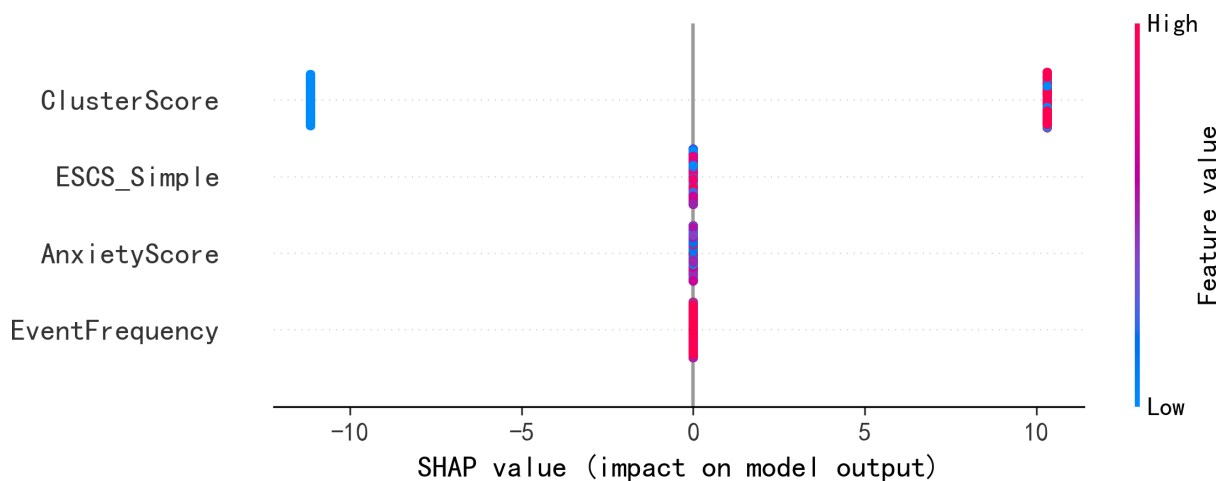


Figure 11: SHAP swarm plot showing feature importance distributions. Each point represents a student's SHAP value for a given feature; horizontal position indicates impact on the prediction (−10 to +10), and color indicates feature value (blue = low, red = high). Features are ordered by mean absolute SHAP value. Clustering scores, SES, anxiety, and event frequency rank highest.

three distributional states: the original raw scores (red), standardized Z-scores (blue), and Box-Cox transformed scores (green). In both subjects, the original distributions (red) deviate severely from the theoretical normal quantile line, confirming substantial skewness that would violate standard regression assumptions. Standardization (blue) improves the fit but does not fully correct the

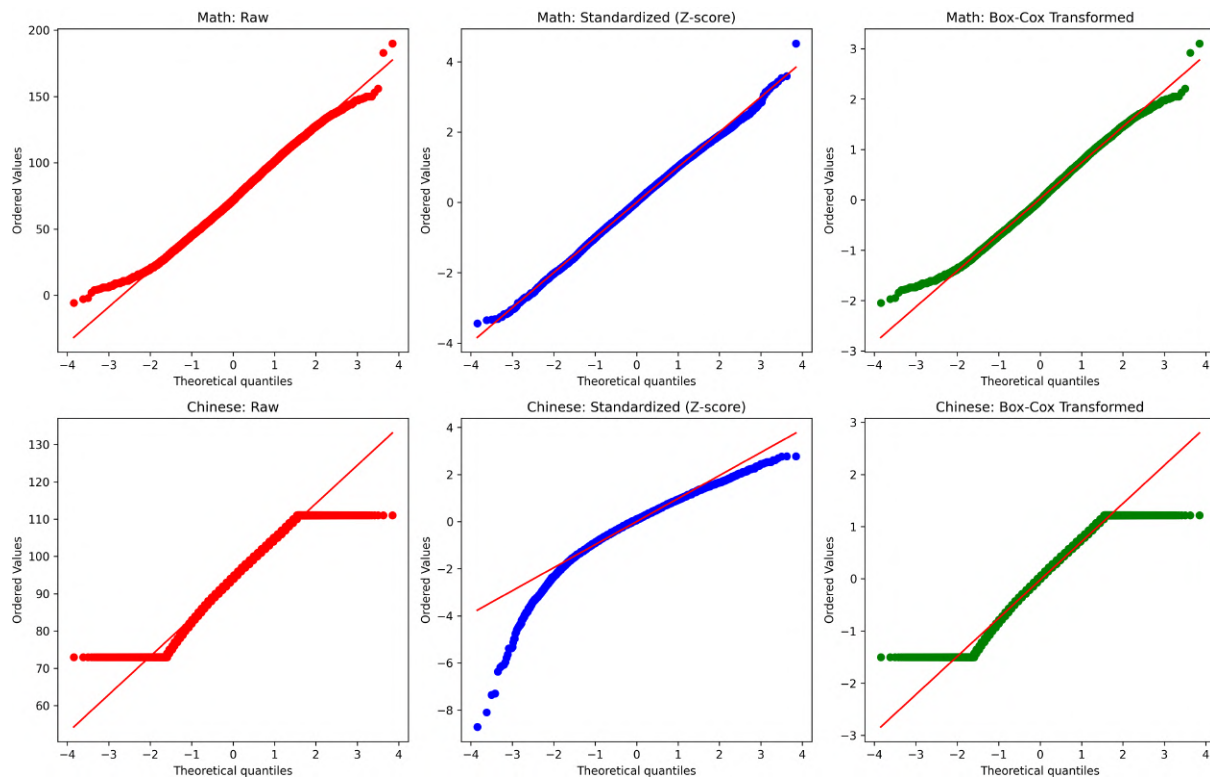


Figure 12: Q-Q plots for Mathematics (left) and Language (right) subjects showing three distributions: original scores (red), standardized Z-scores (blue), and Box-Cox transformed scores (green). The transformed distribution (green) aligns closely with the theoretical normal quantiles, while the original (red) shows severe deviation. Shapiro-Wilk $p > 0.05$ for transformed data.

distortion. The Box-Cox transformation (green), however, achieves near-perfect alignment with the theoretical line, and Shapiro-Wilk tests confirm normality ($p > 0.05$) for the transformed data. The Language subject exhibits a slightly flatter lower tail, reflecting subject-specific score distribution characteristics. These diagnostics validate the effectiveness of the preprocessing pipeline and confirm that the regression assumptions are met after transformation. The approach aligns with best practices in score distribution studies, where extreme skewness necessitates transformation before parametric modeling. The successful normalization across the full sample ensures statistical rigor and supports generalization to multi-subject settings. Note: red = original deviation, blue = standardized, green = transformation fit.