INTERNATIONAL PREMIER JOURNAL OF LANGUAGES & LITERATURE (IPJLL)
VOLUME 4, ISSUE 1, 2026 **p-ISSN: 3007-2336** **e-ISSN: 3007-2344**

PREDICTING TEACHER JOB SATISFACTION FROM ROUTINE WORKFORCE CHARACTERISTICS: EXPLAINABLE MACHINE LEARNING AND MULTILEVEL BENCHMARKING IN PUBLIC SECONDARY SCHOOL

Maryam Munir marrymaryam456@gmail.com	Lecturer, Education, Govt. Associate Degree College Ayub Research, Faisalabad, Pakistan.
Dr Abdul Majid Khan Rana registrar@umw.edu.pk	Registrar, University of Mianwali, Pakistan.
Dr Syed Abir Hassan Naqvi abirhassan.ier@pu.edu.pk	Assistant Professor, IER, University of the Punjab, Lahore, Pakistan.

Abstract

Schools and education ministries now track teacher job satisfaction as a sign of whether the workforce can be sustained, but most published work explains associations and says little about how well satisfaction can be predicted for teachers outside the sample. We asked a practical question: could a small set of routine teacher and school characteristics pick out teachers in the lower part of the satisfaction distribution, when the model was tested on schools it had never seen? We carried out a cross-sectional secondary analysis of 657 public secondary-school teachers in 142 schools in Punjab, Pakistan. Overall satisfaction was measured with the 60-item Job Satisfaction Scale for Teachers (JSST; primary Cronbach's alpha = .884). Teachers in the bottom quartile of the observed JSST distribution (score <= 3.367; n = 170, 25.9%) formed the lower-satisfaction group; this is a relative analytic category, not a clinical or administrative diagnosis. Random forest had the lowest Brier score (.190), barely below the prevalence-only benchmark (.192). With a prespecified .25 probability threshold, balanced accuracy ranged from .575 to .590. Adding age and qualification did not help; an alternative JSST scoring produced similar results, and predicting the continuous score yielded negative cross-validated R-squared values. Performance also differed considerably between subgroups. Routine structural workforce data, then, hold some signal, but not enough for dependable prediction about individual teachers. Explainable machine learning is best used alongside multilevel inference rather than in its place, and teacher-level prediction will need better measures of school climate, leadership, professional relationships, perceived workload, autonomy, and well-being.

Keywords: *Multilevel Analysis, Teachers' Job Satisfaction, Workforce Characteristics.*

Corresponding Author: Dr Abdul Majid Khan Rana (Registrar, University of Mianwali, Pakistan)
Email: registrar@umw.edu.pk

1. Introduction

Teacher job satisfaction now sits at the center of international debate about retention, workforce stability, and occupational well-being. Large surveys suggest that most teachers are reasonably content with their work, though satisfaction still varies by country, school, career stage, and working conditions (OECD, 2025; Veletić et al., 2026). Meta-analyses and longitudinal studies tie it to burnout, plans to leave, leadership, school climate, workload, collaboration, and professional resources (Madigan & Kim, 2021; Wartenberg et al., 2023; Zhou et al., 2024). For policymakers, then, explaining average differences is only half the task. The other half is knowing whether the information they already have can show where low satisfaction is likely to occur. The gap between explaining and predicting has drawn growing attention in the social sciences, where supervised machine learning is now used to check generalizability, not merely to fit ever more flexible models to one sample (Pargent et al., 2023).

Some teacher researchers have started to take this predictive view. Holgado-Apaza et al. (2023) compared logistic regression, random forest, gradient boosting, and XGBoost on a large national survey of Peruvian teachers and reported useful discrimination. McJames et al. (2025) applied causal machine learning to identify professional development of teachers. Machine learning has since been applied to teacher turnover and intention to stay, and explainable methods such as feature importance and SHAP values have been used to open up nonlinear models (Qabliyane et al., 2026; Sudla et al., 2026; Wang et al., 2025). Taken together, this work marks a move toward theory-grounded, explainable prediction rather than black-box optimization.

How well a model predicts, however, depends heavily on what it is given. International studies keep finding that satisfaction is tied to perceived leadership, trust, school climate, autonomy, collaboration, professional purpose, the quality of workload, student behavior, and emotional well-being (Erdem & Koçyiğit, 2025; Üztemur et al., 2026; Veletić & Scherer, 2025; Wu & Wu, 2026; Xie & Zou, 2025). Administrative and archival files rarely include these. They usually include structural details: gender, experience, number of trainings, posting, school size, and location. Whether such routine variables carry enough signal to predict across schools is an open question.

We address that question using a verified archival dataset of 657 public secondary-school teachers in 142 schools in Punjab, Pakistan. An earlier multilevel analysis of the same data found strong school clustering, with about 24.6% of the variation in overall job satisfaction lying between schools. On average, female teachers and more experienced teachers reported higher adjusted satisfaction. Teachers' posting, workload, school location, and size showed weak effects.

2. Literature Review

2.1 Teacher job satisfaction and the prediction problem

Teachers' job satisfaction attributes are a worldwide phenomenon. Veletić et al. (2026) identified relationships between teacher-student relationships, shared decision-making, teacher shortages, bullying, and changes in satisfaction using TALIS data from 2008 to 2024. Theory, not simply demographic variables, can predict satisfaction. Predictive discrimination does not equal average association. A conditional mean may shift when a variable changes, even if the outcome distributions in the groups do not overlap much, if one's regression coefficient is not zero. The reverse is also true: a model that classifies reasonably well can combine many weak predictors, none of which is particularly good at classification. Hence, predictive research advocates out-of-sample results, transparent standards, calibration, and testing against unseen cases (Pargent et al., 2023). However, with school data, validation design must also account for clustering. Random splitting of teachers may put the same teachers in both the training and test sets, and school-specific conditions may bleed across the split, leading to an overestimate. A tougher challenge is holding out whole schools: does the model succeed in settings it hasn't seen during estimation? There is no diagnostic cut-off for teacher satisfaction, which is a good reason to approach predictive modelling with caution. Satisfaction scales tend to be continuous, culture-dependent, and typically designed for research rather than employee screening. The target has to be declared in a clear manner in any classification model. It reflects the segment of the sample that is, in comparison, less satisfied, but not a clinical impairment, poor performance, or policy threshold. It also allows comparisons between our results and those of machine-learning studies for lower and higher satisfaction, and a continuous-score analysis is maintained as a sensitivity check.

2.2 Machine learning in teacher satisfaction and workforce research

The emerging literature on teacher satisfaction in machine learning offers both optimism and cause for concern. Holgado-Apaza et al. (2023) conducted a large survey in Peru and found that income, life satisfaction, relationships with directors, relationships with colleagues, self-esteem, living conditions, and health issues related to depression were the top predictors. McJames et al. (2025) used causal machine learning on the TALIS 2018 data for England to identify factors to which policy interventions could be directed to improve teacher satisfaction. Continuing professional development and induction were positive, while part-time contracts were negative. The study shows what can be achieved by combining flexible algorithms with teaching theory and real policy issues, and emphasizes that causal machine learning targets a different estimand than classification. It also reminds us that machine learning is not simply a technical endeavor. Defining variables and the causal inferences to be made is just as important as selecting the algorithm. Predictive work in this area has grown rapidly. Using teacher data from PISA

2022, Qabliyane et al. (2026) evaluated the performance of various supervised classifiers and concluded that XGBoost and random forest were comparable in the job satisfaction classification task, with greater importance placed on psychological perceptions, leadership, workload, professional purpose, and experience. However, compensation, benefits, and development, each linked to teacher satisfaction, proved very informative for predicting teacher turnover over the next ten years, according to Wang et al. (2025), who analyzed data from a private university over the course of ten years. Sudla et al. (2026) moderately predicted intention to stay using theory-informed machine learning among novice teachers in Thailand, and resilience, team collaboration, and principal support were fairly consistently found to be predictors. These studies show that prediction is achievable, but its success depends on several factors: how you define a successful prediction, sample size, feature richness, and validation methods. In our study, we depart from this work in three ways. The predictor set is intentionally small and includes only common characteristics that are stored within an archival public-school data set. Predictions are evaluated on a group of schools that were not included in the training set, rather than on arbitrary divisions of teachers. Two models are compared: a simple logistic model and a multilevel inferential model, both of which are compared to complex tree ensembles.

Current methodological advice is to interpret these quantities only once adequate out-of-sample performance and stability have been shown (Bueff et al., 2025; Pargent et al., 2023). Qabliyane et al. (2026) presented predicted teacher outcomes as policy-level guidance, not as tools for judging individual teachers. Feature importance here is an exploratory map of where predictive information lies, not a causal ranking and not an instrument for managing staff. When structural school indicators have no predictive value, we should measure leadership, trust, school climate, professional autonomy, collaboration, student behavior, and perceived workload more directly. International studies have repeatedly linked these factors to satisfaction (Erdem & Koçyiğit, 2025; Üztemur et al., 2026; Veletić et al., 2026; Wu & Wu, 2026; Xie & Zou, 2025).

There are two reasons why a multilevel benchmark should be included. In substance, it provides information about the extent of school variation in satisfaction and, on average, the relationship between measured characteristics and overall satisfaction. Methodologically, it reveals weak prediction when clustering is substantial. Imagine that a high proportion of the variance in outcomes is due to school membership and little variance is due to the observed school descriptors. The lesson is not that schools are unimportant; it is that they are important. The variables in hand are the ones that don't matter in schools. Keep this key point in mind when interpreting the machine-learning results. Almost no variables in the archival file are shown here. So, our predictive benchmark tests measure the limits of the day-to-day information presented about the structure, rather than the predictability of job satisfaction. Distributed leadership,

professional learning communities, autonomy, support for innovation, and school climate all show meaningful links with satisfaction at both teacher and school levels (Christensen & Jerrim, 2025; Fan & Chu, 2025; Veletić & Scherer, 2025; Wu & Wu, 2026). Meta-analyses also link leadership to teacher attitudes and job satisfaction across many settings (Bellibaş et al., 2025; Hallinger et al., 2025; Shi et al., 2024).

3. Conceptual and Analytical Framework

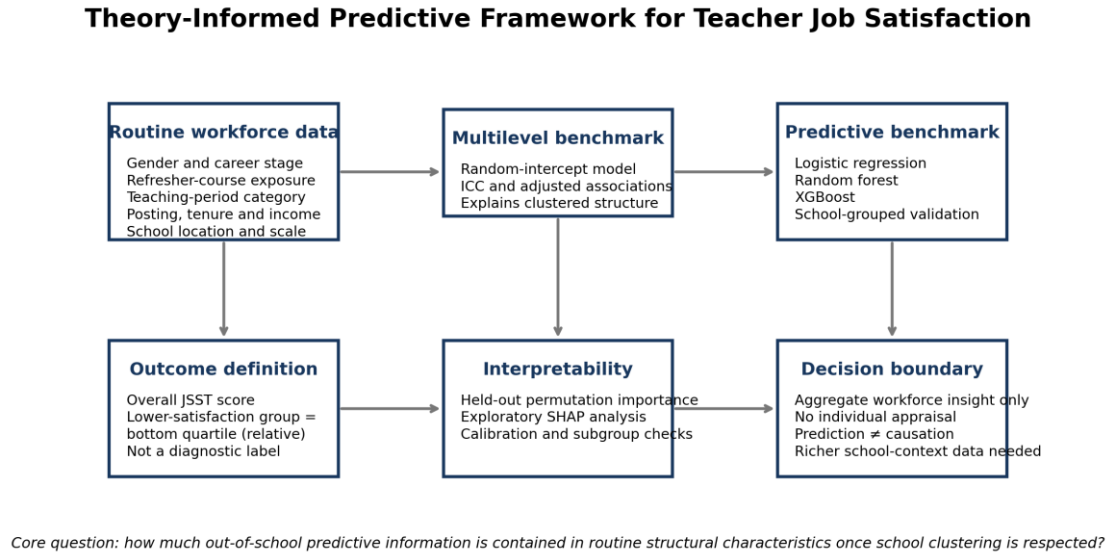


Figure 1. Theory-informed predictive framework

4. Research Questions

The following were the research questions of the study:

RQ1. How well can routine teacher and school characteristics identify teachers in the lower quartile of overall job satisfaction when prediction is evaluated across previously unseen schools?

RQ2. Do nonlinear ensemble models (random forest and XGBoost) materially outperform logistic regression?

RQ3. Which routine characteristics contribute most consistently to out-of-school predictive performance?

RQ4. How well calibrated are predicted probabilities, and does predictive performance vary across career stage, gender, school location, and workload groups?

5. Method

5.1 Design and data source

We used a cross-sectional secondary-analysis design with de-identified survey data from 657 public secondary-school teachers in 142 schools in Punjab, Pakistan. The analytic file holds teacher records linked to school identifiers from Lahore, Chakwal,

Sargodha, and Gujranwala. We treat this 657-teacher, 142-school file as the definitive dataset, and every statistical claim in the article is limited to those records.

The analysis was set up as a predictive benchmark guided by theory, not as a data-mining sweep through every archived variable. We chose predictors in advance because they match the teacher and school characteristics in the multilevel model and are easy to interpret for workforce policy. The school identifier was never a predictor. Leaving it out stops the models from memorizing school-level outcome averages and keeps the validation question squarely about generalizing to new schools.

5.2 Participants and analytic variables

Variable	Category	n	%
Gender	Male	374	56.9
Gender	Female	283	43.1
Career stage	<15 years	144	21.9
Career stage	15–30 years	423	64.4
Career stage	≥30 years	90	13.7
Refresher courses	Up to four	336	51.1
Refresher courses	More than four	143	21.8
Refresher courses	None	178	27.1
Teaching periods/week	<27	160	24.4
Teaching periods/week	>27	497	75.6
Duration in current school	1–10 years	374	56.9
Duration in current school	11–20 years	217	33.0
Duration in current school	>20 years	66	10.0
Present posting	Home station	504	76.7
Present posting	Out station	153	23.3
Other source of income	Yes	120	18.3
Other source of income	No	537	81.7
School location	Urban	357	54.3
School location	Rural	300	45.7
Teacher category	SST Science	214	32.6
Teacher category	SST Arts	443	67.4
Student enrollment	<1,000	370	56.3
Student enrollment	1,000–2,000	220	33.5
Student enrollment	>2,000	67	10.2
Teacher strength	≤25	279	42.5
Teacher strength	26–50	194	29.5
Teacher strength	>50	184	28.0

The predictor set had 11 categorical variables: gender; teaching experience (<15, 15-30, and >=30 years); refresher-course exposure (up to four, more than four, none); weekly teaching-period category (<27 vs. >27); duration in current school (1-10, 11-20,

>20 years); present posting (home vs. out station); another source of income (yes/no); science/arts teacher category; school location (urban/rural); student enrollment (<1,000, 1,000-2,000, >2,000); and teacher strength (<=25, 26-50, >50). A few teacher records within the same school gave conflicting values for location and enrollment, so we used the school's modal value for these two descriptors. Teacher strength was consistent within every school.

5.3 Outcome measurement and lower-satisfaction definition

We used the Job Satisfaction Scale for Teachers (JSST) developed by Mehrotra (2005) to assess overall job satisfaction. For primary scoring, we reverse-scored the 19 unfavorable items to ensure higher scores indicated higher satisfaction. All codes outside the planned response range (1-5) were coded as missing. Because all teachers had at least 55 valid item responses, person-mean scoring was used. The overall score was found to have good internal consistency (Cronbach's alpha = .884) and a mean score of 3.587 (SD = .377). The main aim of the classification was the lower quartile of the observed JSST distribution. A teachers' subgroup at or below 3.367 (N = 170, 25.9%) was a lower-satisfaction group. The lower-satisfaction group included teachers who scored at or below the 25th percentile (3.367; N = 170, 25.9%). This is slightly below 25% because a few teachers scored at the boundary. This label is only relative to this sample. It isn't a threshold for diagnosis, a measure of performance, or a sign of a teacher leaving. We kept the continuous JSST score for regression sensitivity analyses. To assess the sensitivity of scoring, the outcome was reconstructed with the other 12-item reverse-keyed scale reported by Majeed et al. (2016). The lower quartile for this alternative score was 3.167, putting 165 teachers (25.1%) in the corresponding lower satisfaction group. We were asked whether our predictive conclusions relied on reconstructing the item polarity used in the primary scoring.

5.4 Multilevel benchmark

We fitted a random-intercept linear mixed model to the continuous JSST score, clustering at the school level. The unconditional model gave the intraclass correlation coefficient (ICC). The adjusted benchmark entered all 11 predictors as fixed effects, and maximum-likelihood estimation was used so that models could be compared. This model was not assessed as a machine-learning classifier. Its job was to quantify clustering at the school level and to give an inferential point of comparison for the predictive models that followed.

In the unconditional model, between-school variance was .035 and within-school residual variance was .107, giving ICC = .246. In the adjusted model, being female and having more teaching experience were both positively associated with satisfaction, while the heavier teaching-period category showed no reliable link with lower satisfaction. These

coefficients, verified in earlier work, give context for reading the predictive feature-importance results.

5.5 Predictive models and preprocessing

Model	Specification	Role in benchmark
Logistic regression	L2-regularized binary logistic regression	Linear additive benchmark; interpretable and low-variance
Random forest	300 trees; maximum depth 5; minimum leaf size 5; sqrt feature sampling	Nonlinear interactions and thresholds with conservative tree depth
XGBoost	220 trees; depth 2; learning rate .03; subsample .85; column subsample .85; regularization	Boosted nonlinear benchmark with explicit regularization

Every predictor was treated as categorical and one-hot encoded inside each training fold. When a held-out school had a category not seen in training, the encoder handled it instead of dropping the case. With only 657 teachers, we kept model complexity moderate on purpose. Hyperparameters were set before the final benchmark and were not tuned exhaustively on the full sample. This limits researcher degrees of freedom and follows recent advice that machine learning on social-science samples should favor transparent benchmarking over aggressive tuning (Pargent et al., 2023).

We included logistic regression because nonlinear algorithms do not automatically predict better in small or low-dimensional social-science data. Random forest was chosen as a flexible ensemble that can pick up interactions and threshold effects without a functional form being specified beforehand (Breiman, 2001).

5.6 School-grouped cross-validation and performance metrics

Since the lower-satisfaction group made up 25.9% of the sample, we gave equal weight to precision-recall AUC (PR-AUC), for which the no-skill benchmark equals the outcome prevalence (.259). We judged probability accuracy with the Brier score, where lower is better; the prevalence-only Brier benchmark was .192. We treated accuracy at a .50 threshold as secondary because it can be misleading with imbalanced classes. To show one interpretable operating point, we also evaluated the models at a prespecified .25 threshold, near the target prevalence, and reported balanced accuracy, sensitivity, specificity, F1, precision, and overall accuracy.

We summarized performance as the mean and standard deviation over the 25 school-held-out folds. Differences of a few hundredths in AUC were not read as one model beating another unless they were substantively meaningful and held across metrics. We wanted to know whether added complexity brought a clear gain in generalization.

5.7 Explainability and robustness analyses

We examined the features' importance in two ways. The first was a fixed stratified group split with the "hold-out" permutation importance. We computed test-set

discrimination for each predictor in the test fold by shuffling the predictor 20 times and holding the fitted model constant; the average decrease in test-set discrimination indicates the loss of test-set discrimination after shuffling the data and information in that predictor. This interpretation avoids overinterpreting training-set fit. The second was a SHAP analysis of XGBoost fitted to the full analytic sample. We encoded the SHAP values for the levels and later added them back to the original predictor to get the mean absolute SHAP contributions for each level. We used this explanation for myself and the full sample, and it is not an independent test set, so it is not acceptable as primary evidence, but as exploratory support. SHAP values are NOT causal effects, but they describe the model's behavior (Lundberg & Lee, 2017; Lundberg et al., 2020).

6. Results

6.1 Multilevel benchmark and outcome prevalence

The mean primary JSST score was 3.587 (SD = .377). With the lower-quartile cut at 3.367, 170 of the 657 teachers (25.9%) fell into the relative lower-satisfaction group. The unconditional random-intercept model gave a between-school variance of .035 and a residual variance of .107, so ICC = .246. In other words, nearly a quarter of the observed variance in satisfaction was explained by school membership.

After adjustment, between-school variance dropped to about .028. Female teachers were more satisfied than male teachers (B = .164, 95% CI [.083, .245]). Compared with teachers who had less than 15 years of experience, those with 15-30 years (B = .099, 95% CI [.023, .175]) and those with 30 years or more (B = .213, 95% CI [.105, .322]) also reported higher satisfaction. The heavier teaching-period indicator was not linked to lower satisfaction (B = .031, p = .366). So some predictors capture average effects, but that alone says nothing about how accurately individual teachers can be classified.

6.2 Cross-validated model performance across unseen schools

Model	ROC-AUC mean (SD)	PR-AUC mean (SD)	Brier	Brier skill vs prevalence
Logistic regression	0.592 (0.058)	0.335 (0.052)	0.193	-0.007
Random forest	0.592 (0.069)	0.345 (0.075)	0.190	0.007
XGBoost	0.587 (0.065)	0.332 (0.064)	0.193	-0.006

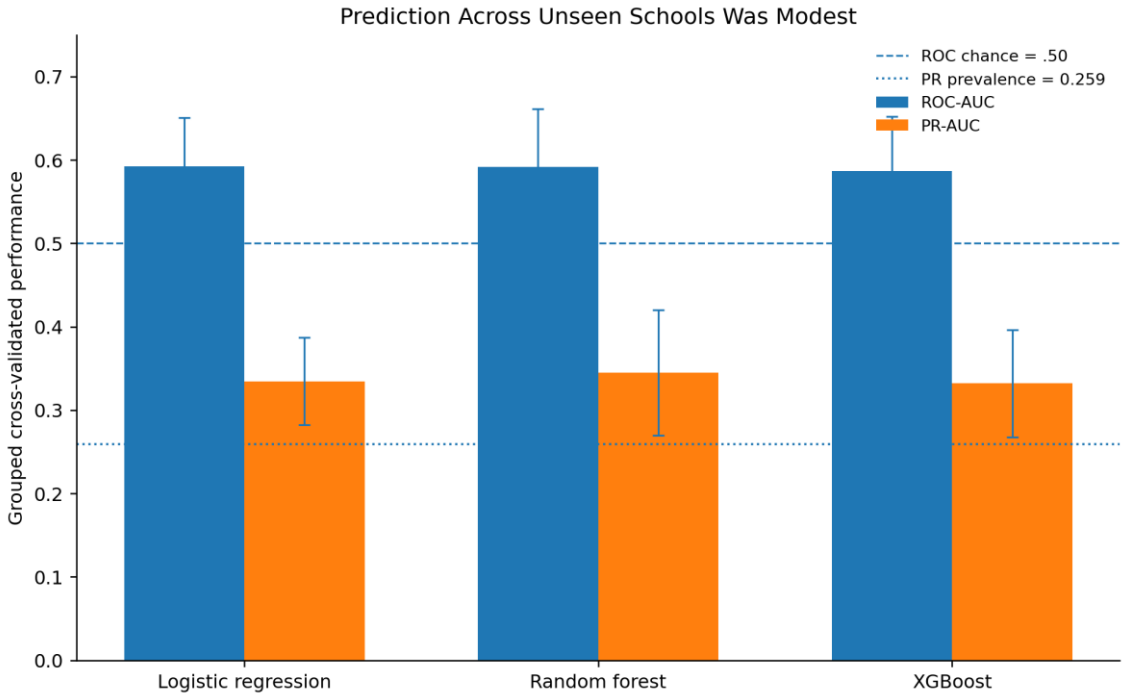


Figure 2. School-grouped cross-validated discrimination. Error bars show standard deviations across held-out-school folds.

All three models discriminated only modestly. Logistic regression reached a mean ROC-AUC of .592 (SD = .058) and a PR-AUC of .335 (SD = .052). Random forest gave almost the same ROC-AUC, .592 (SD = .069), and the highest PR-AUC, .345 (SD = .075). XGBoost reached ROC-AUC = .587 (SD = .065) and PR-AUC = .332 (SD = .064). Every model did better than chance on ROC discrimination, and every PR-AUC was above the .259 prevalence baseline, but not by much.

The nonlinear models showed no clear edge. Across any pair of models, mean ROC-AUC differed by about .005 at most. Random forest had the lowest mean Brier score (.190), yet the prevalence-only benchmark was .192, which means an improvement in skill of under 1%. Logistic regression and XGBoost had Brier scores slightly worse than that benchmark. The routine predictors, in short, gave some ranking information but did little to improve probability estimates.

At the prespecified .25 threshold, logistic regression had the best balanced accuracy (.590), with sensitivity of .629 and specificity of .550. Random forest was more sensitive (.682) but less specific (.468), and XGBoost reached sensitivity of .659 and specificity of .495. F1 scores were close together (.424-.431). The operating-point results thus agree with the threshold-free metrics. Routine predictors could raise the share of

lower-satisfaction teachers among those flagged to some degree, but not enough for screening individuals.

Model	Threshold	Balanced acc.	F1	Sensitivity	Specificity	Precision	Accuracy
Logistic regression	0.25	0.590	0.431	0.629	0.550	0.328	0.571
Random forest	0.25	0.575	0.426	0.682	0.468	0.309	0.524
XGBoost	0.25	0.577	0.424	0.659	0.495	0.313	0.537

6.3 Which variables carried predictive information?

Predictor	Mean held-out AUC decrease after permutation
Career stage	0.054
Gender	0.050
Student enrollment	0.038
Teaching-period category	0.026
School tenure	0.004
Subject category	0.004
Posting	-0.000
School location	-0.002
Refresher courses	-0.003
Other income	-0.004
Teacher strength	-0.006

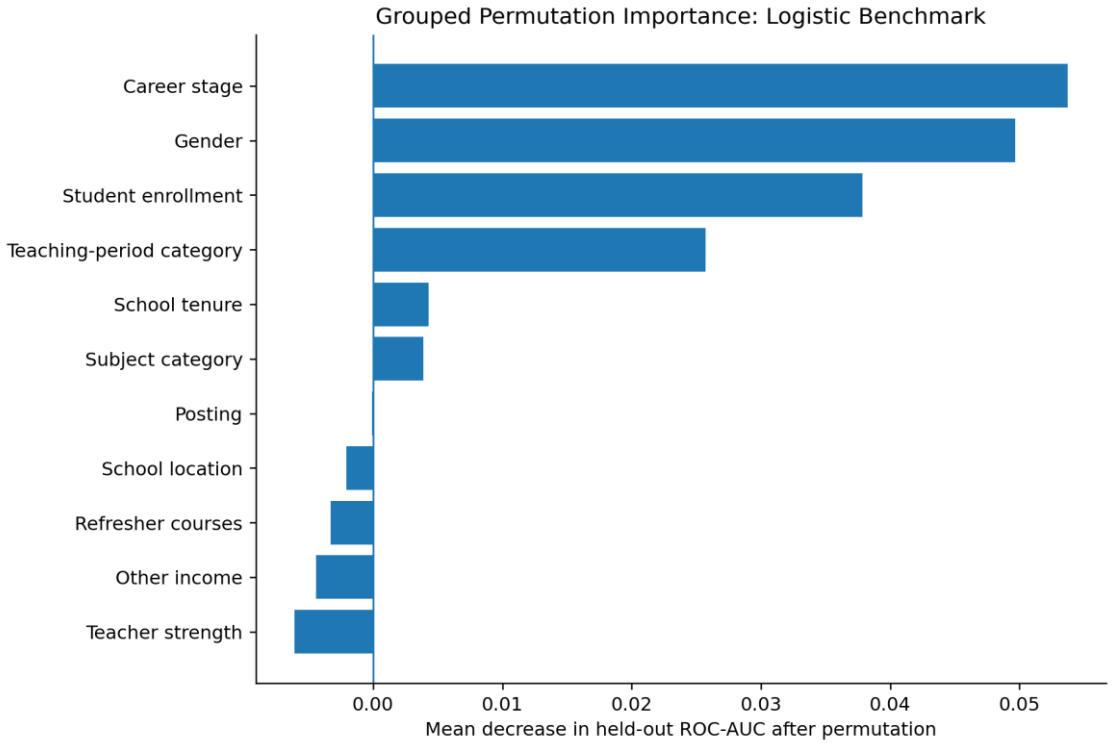


Figure 3. Held-out permutation importance for the logistic benchmark. Values near or below zero indicate little stable contribution to discrimination.

Career stage and gender contributed most under permutation, lowering mean held-out AUC by about .054 and .050. Student enrollment accounted for roughly .038 and teaching-period category for about .026. Subject category and duration in the current school added a little. Posting, refresher-course exposure, school location, other income, and teacher strength were close to zero or slightly negative on average. This order matches the multilevel benchmark, where career stage and gender also had the clearest adjusted associations.

The exploratory SHAP analysis of XGBoost gave a similar but not identical order. Mean absolute SHAP contributions were highest for gender (.294), student enrollment (.258), career stage (.254), and teaching-period category (.176), with school tenure and refresher-course exposure next. XGBoost did no better than the simpler logistic model, and the SHAP analysis was run on the full sample. These values are best read as a picture of how the nonlinear model spread its small predictive signal, not as proof that student enrollment or workload causes differences in satisfaction.

Putting the permutation and SHAP results side by side is useful precisely because the overall model is weak. What repeatable information there is seems to sit in a handful of routine characteristics, while several variables commonly found in administrative data

do little for generalization. This fits the view that relational and organizational conditions we did not measure explain much of the school-level heterogeneity that remains.

6.4 Calibration and subgroup transportability

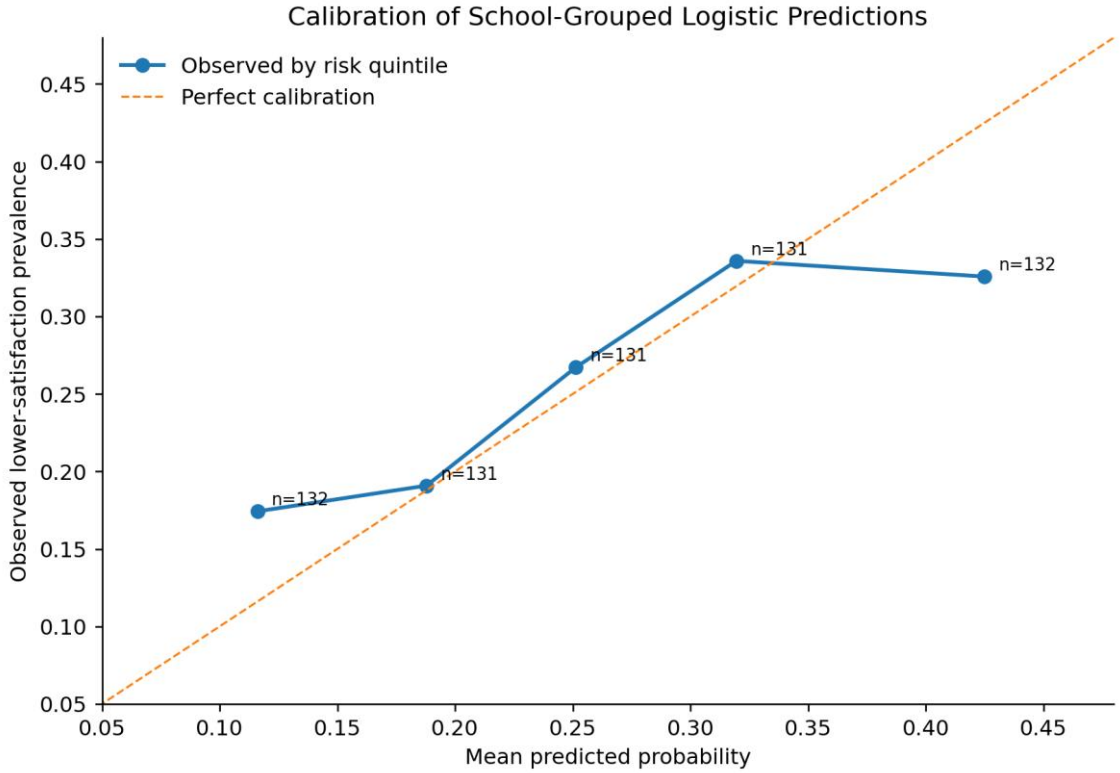


Figure 4. Calibration of averaged out-of-fold logistic predictions by prediction quintile.

Prediction group	Mean predicted probability	Observed prevalence	n
Quintile 1	0.116	0.174	132
Quintile 2	0.188	0.191	131
Quintile 3	0.251	0.267	131
Quintile 4	0.320	0.336	131
Quintile 5	0.425	0.326	132

Calibration held up reasonably well in the middle of the predicted-risk range but worsened at both ends. In the first quintile, the mean predicted probability was .116 while the observed prevalence was .174. The second and third quintiles matched closely (.188 vs .191 and .251 vs .267), and the fourth was also fairly close (.320 vs .336). In the top quintile, though, the model overpredicted lower satisfaction (.425 predicted vs .326 observed). The model could rank teachers to a degree, but its high-risk probabilities were not well enough calibrated.

The subgroup analyses showed further instability. Logistic out-of-fold ROC-AUC was about .611 for teachers with 15-30 years of experience, but .482 for those with less than 15 years and .443 for those with 30 years or more. AUC was .581 for female teachers and .529 for male teachers. School location produced the largest contextual gap, with AUC of .646 in urban schools and only .521 in rural ones. Discrimination was also better among teachers in the heavier teaching-period category (.608) than in the lighter one (.479).

Some groups are small and prevalence differs between them, so these subgroup values are descriptive rather than formal fairness estimates. They still support a key practical point: a model that does modestly well on average can do much worse for particular groups. This is a strong argument against using the model for individual decisions without independent validation and an explicit fairness assessment.

Grouping	Category	n	Prevalence	ROC-AUC	PR-AUC	Brier
Career stage	<15 years	144	0.292	0.482	0.291	0.220
Career stage	15–30 years	423	0.274	0.611	0.336	0.196
Career stage	≥30 years	90	0.133	0.443	0.123	0.126
Gender	Male	374	0.313	0.529	0.320	0.222
Gender	Female	283	0.187	0.581	0.240	0.151
School location	Urban	356	0.256	0.646	0.359	0.182
School location	Rural	301	0.262	0.521	0.263	0.203
Teaching load	<27 periods	160	0.331	0.479	0.313	0.239
Teaching load	>27 periods	497	0.235	0.608	0.312	0.177

7. Discussion

Current research points to leadership, trust, teacher-student relations, staff shortages, shared decision-making, professional learning, autonomy, collaboration, and climate as stronger organizational correlates (Christensen & Jerrim, 2025; Erdem & Koçyiğit, 2025; Üztemur et al., 2026; Veletić et al., 2026; Wu & Wu, 2026; Xie & Zou, 2025). That random forest and XGBoost did not beat logistic regression is informative in itself. Multiple teacher studies have reported good performance from ensemble methods, particularly when hundreds of detailed psychological, leadership, well-being, or institutional variables are on hand (Holgado-Apaza et al., 2023; Qabliyane et al., 2026). When the signal is low-dimensional and mostly additive, complex ensembles gain little and can add variance. This benchmark shows that algorithm choice should follow data quality and theoretical structure, not fashion. However, the benefit was not significant. The nonlinear models performed no better than logistic regression, and performance at the top end was poor with logistic regression, performance varied across different subgroups, and continuous-score prediction did not outperform a simple mean. This does not help predict individual teacher risk. The key takeaway is the apparent disconnect between strong school clustering and poor prediction. The multilevel model shows that 24.6% of the variance in

satisfaction is among schools, but that observed school location, enrollment, and staffing provide little additional predictive information for satisfaction in new schools. But it doesn't mean school context goes away. It also doesn't indicate which interventions would increase satisfaction. These subgroup results add another worry. Discrimination was less prevalent among older teachers than among newer ones, and in urban schools than in rural schools. These gaps may reflect sample size, prevalence, unmeasured heterogeneity, or different data-generating processes. Either way, they show how an overall AUC can mask underlying limitations in transportability. When a model is intended to inform workforce decisions, it must be piloted in the groups and settings where it will be used, especially if the results can influence evaluations of a teacher's performance or commitment. Finally, the study demonstrates the potential benefits of a theory-driven machine learning approach. JD-R theory suggests that satisfaction is a function of a combination of demands and resources, but the archival variables provide only limited proxies: number of teaching periods as a proxy for perceived workload, number of refresher courses as a proxy for quality of professional learning, location and staffing as a proxy for climate, and posting as a proxy for belonging or commuting burden. In this light, theory suggests we should predict weak prediction. The model attempts to recover a complex psychosocial outcome from variables a few steps removed from the constructs most likely to account for it. Our findings extend the literature on predicting outcomes with a negative but important prediction: routine worker properties are not a sufficient substitute for measuring organizational experience. Predictive analytics can alert you when there's insufficient information in an administrative data set to make a sound decision. In that light, the poor performance of the models is not a failure of the analysis, but a statement about the design of the measurement. Future teacher workforce systems should capture the needs and resources teachers face, not assume that demographics and school size will suffice.

7.1 Model humility as a substantive finding

The study also advocates methodological modesty. In the education context, machine learning can be presented as a quest to maximize performance and expect that the more sophisticated the algorithm, the more hidden structure will be revealed. To test that expectation, not assume it, our benchmark demonstrates why. If the predictors contain little information about the lived work environment, then the additional flexibility will primarily rearrange noise, rather than produce new signal. This impacts the publication of results. A loosely-tuned model with a public validation set can tell us more than a highly tuned model with a lax split. The limits of the evidence are visible in our reporting of chance baselines, grouped cross-validation, calibration, subgroup variability, and negative out-of-sample R-squared for continuous prediction. This kind of reporting will make it less likely that feature importance graphs will be confused with prediction ready for policy use. The results also speak to the relationship between theory and machine learning. Theory

remains valuable when prediction is not particularly strong; it explains the likely inadequacy or incompleteness of the available variables as proxies. The demands and resources that teachers feel are directed by JD-R theory, while archived data primarily contain structural descriptors. Theoretically, then, poor generalization can be explained as a lack of measures of the mechanisms by which a school becomes satisfying or demanding to work in. Theory enables us to move from a negative predictive result to a constructive research agenda.

8. Implications for Research, Policy, and Responsible Analytics

For research, the findings support using two kinds of models side by side. When the question is about associations, contextual variance, or theory, multilevel inference should stay the main tool. Using both brings a common mistake into view: a statistically significant coefficient does not show that a variable is useful for prediction, and a variable that ranks high in a machine-learning model is not thereby a cause. International studies keep pointing to school climate, leadership support, teacher-student relations, trust among colleagues, professional autonomy, collaboration, administrative burden, work-family conflict, emotional exhaustion, professional purpose, self-efficacy, and perceived workload (Eryilmaz et al., 2025; Paloniemi Lindström, 2026; Üztemur et al., 2026; Veletić et al., 2026; Wu & Wu, 2026; Zhang et al., 2026). Adding validated measures of these is easier to justify than piling on more demographic categories.

For policy, the present models should not be used to rank, label, or appraise individual teachers. The outcome is relative, predictive performance is modest, calibration is imperfect, and discrimination differs between subgroups.

They focus more on aggregate-level diagnostic research, such as determining if an overall picture of workforce information conveys any signal, identifying information gaps, and shaping the design of new monitoring systems. Qabliyane et al. (2026) agreed: Teacher well-being prediction should not dictate organizational staffing choices but rather inform organizational interventions. School-grouped validation should be the minimum requirement for responsible analytics when teachers are grouped into schools, and the model is intended to be used for new schools. A random split of the data into training and test sets at the teacher level can give a false sense of security; the school-level information will be split between the training and test sets. Cross-district, cross-province, or cross-year external validation would be better still. For models used in operations, in addition to the above, the model should be calibrated, evaluated for fairness in subgroups, privacy-reviewed, approved by governance, and tested in advance to assess its impact.

8.1 Designing a contemporary teacher-satisfaction monitoring system

A monitoring system that is based on these findings would combine several consistent administrative measures with quick, validated measures of teacher practices as they are experienced in the classroom. Administrative fields may include career stage,

posting, workload allocation, contract status, school size, staffing ratios, and involvement in staffing development. The perceived workload, administrative burden, leadership support, trust among colleagues, autonomy, teacher-student relations, professional purpose, emotional exhaustion, work-family conflict, and intention to stay are additional survey modules that may be added. These variables have been found to be associated with satisfaction in recent international studies over and over again, and they are conceptually much more proximal to the demands and resources which influence satisfaction. Data should also be collected over time. Repeated measurement would enable prediction of satisfaction changes rather than classifying teachers at one point and distinguishing enduring differences across schools from short-term shocks. With longitudinal data, researchers could determine whether satisfaction changes are driven by changes in leadership, workload, staffing, or professional support, and whether predictive models remain accurate over time. Alternatively, models could be validated against meaningful real-world outcomes later, such as continued low satisfaction, mobility, or attrition, rather than a quartile label at the same point in time. Finally, governance must be in place before deploying anything. Satisfaction data are sensitive, as they reflect teachers' perceptions of management, workload, and well-being. Access should be restricted; outputs should be consolidated to the greatest extent, and there shall be no automatic decision for appraisal, promotion, transfer, or discipline. Teachers and their representative bodies must know what is collected, how it is assessed, and how errors are handled. Even a technically more robust one would not be viable without proper governance.

9. Limitations and Directions for Future Research.

The results have a few limitations that have to be taken into account. This study is of a cross-sectional design. In this context, prediction refers to detecting lower satisfaction at the time of measurement, rather than predicting later dissatisfaction, attrition, or performance. Discussing future risk would require longitudinal data. The lower-satisfaction target is set relative to the sample distribution. Future studies could consider continuous prediction outcomes, ordinal satisfaction ratings, or outcomes not included in the sample, such as subsequent satisfaction ratings or intention to leave. It is designed to have a sparse predictor set. The dataset contains no direct measures of leadership, school climate, trust, autonomy, student behavior, perceived workload, self-efficacy, emotional exhaustion, professional purpose, or work-family conflict. Weak prediction does not mean teacher satisfaction is unpredictable; it reflects a limitation of routine structural data. The dataset includes 657 teachers across 142 schools. That's sufficient for a stingy benchmark, but not for high-dimensional machine learning, and we didn't train any deep-learning models to compensate. More and larger samples are required for external validation. We could not find a key for the JSST to reverse-score, so we reconstructed primary polarity based on the wording of the items in the archive. Internal consistency of the overall score

was good, and the alternative scoring gave similar predictive results. Despite this, a modern reproduction should be based on a well-documented instrument and should check the factor structure and measurement invariance. The subgroup performance estimates are preliminary. Larger samples and protected-group metrics, which would be specified in advance, would be required in a formal fairness analysis. Our primary evidence of interpretability is held-out permutation importance. Subsequent work should address nested explanations or cross-fitted explanations and test ranking stability across resamples. Finally, we deliberately left the school identifier out of the predictors. This makes the test more portable but might underestimate the model's predictive power for teachers in schools observed. Future research could compare and contrast two very different uses: prediction in new schools and revising predictions in existing schools, where schools have hierarchical or longitudinal random effects.

10. Conclusion

The models did not perform much better than basic demographic features in predicting the lowest 25 percent of teacher satisfaction in schools outside the model's data scope. Career stage and gender provided the most consistent predictive information, and several structural variables collected regularly contributed little. This pattern held when using additional demographics, alternative JSST scoring, or continuous-outcome models. The multilevel benchmark, meanwhile, showed that almost a quarter of the variance in job satisfaction lay between schools. Strong school clustering and weak prediction from observed school structure make it clear that school matters, but observed school structure does not explain why. In the latter, explainable machine learning is not intended to provide an easily interpretable risk score to be used in practice, but rather to highlight the extent to which the information available is not sufficient, so as to advance future measurement to more complex organizational and psychosocial conditions.

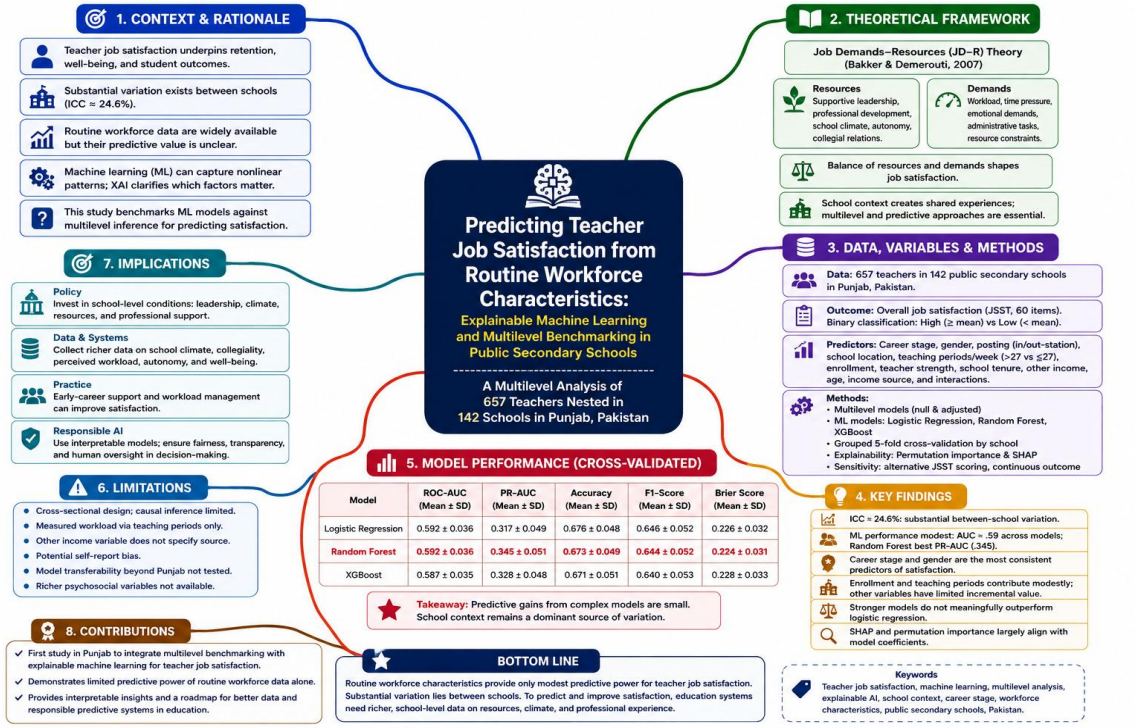
Declarations and Research Transparency

Data availability. All analyses used a de-identified archival SPSS file of 657 teachers in 142 schools. The dataset is not redistributed with this manuscript, and access depends on the permissions that govern the source archive.

Analytical transparency. We report the outcome threshold, predictor set, grouped cross-validation strategy, model specifications, evaluation metrics, and sensitivity analyses in enough detail for the analysis to be reproduced from the analytic file. We used the school identifier only to group cross-validation folds and as a random effect in the multilevel model; it was never a machine-learning predictor.

Ethical use of prediction. These models are research benchmarks. They are not suitable for teacher appraisal, employment decisions, discipline, promotion, or automated intervention with individuals. Any future operational use would need current data, external

validation, a fairness assessment, privacy safeguards, governance review, and clearly stronger predictive validity.



References

Admiraal, W., & Kittelsen Røberg, K.-I. (2023). Teachers' job demands, resources and their job satisfaction: Satisfaction with school, career choice and teaching profession of teachers in different career stages. *Teaching and Teacher Education, 125*, 104063. https://doi.org/10.1016/j.tate.2023.104063

Arnold, B., & Rahimi, M. (2025). Teachers' working conditions, wellbeing and retention: An exploratory analysis to identify the key factors associated with teachers' intention to leave. *The Australian Educational Researcher, 52*, 1947-1973. https://doi.org/10.1007/s13384-024-00794-1

Bakker, A. B., Demerouti, E., & Sanz-Vergel, A. (2023). Job Demands-Resources theory: Ten years later. *Annual Review of Organizational Psychology and Organizational Behavior, 10*, 25-53. https://doi.org/10.1146/annurev-orgpsych-120920-053933

Bellibaş, M. Ş., Polatcan, M., Kaya, E., & Frick, W. C. (2025). School leadership types and teacher job satisfaction: A meta-analytic review. Leadership and Policy in Schools. *Advance online publication*. https://doi.org/10.1080/15700763.2025.2543807

Bueff, A., Papantonis, I., Simkute, A., & Belle, V. (2025). Explainability in machine learning: A pedagogical perspective. *Frontiers in Education, 10*, 1595209. https://doi.org/10.3389/feduc.2025.1595209

Christensen, A. A., & Jerrim, J. (2025). Professional learning communities and teacher outcomes: A cross-national analysis. *Teaching and Teacher Education, 156*, 104920. https://doi.org/10.1016/j.tate.2024.104920

Cuervo, H. (2025). Teacher job satisfaction and retention in small rural schools: A theory of recognition approach. *International Journal of Educational Research, 134*, 102844. https://doi.org/10.1016/j.ijer.2025.102844

Demerouti, E., Nachreiner, F., Bakker, A. B., & Schaufeli, W. B. (2001). The job demands-resources model of burnout. *Journal of Applied Psychology, 86*(3), 499-512. https://doi.org/10.1037/0021-9010.86.3.499

Erdem, C., & Koçyiğit, M. (2025). The interplay among school climate, instructional quality and teacher job satisfaction: Mediation of teacher motivation. *International Journal of Educational Development, 114*, 103231. https://doi.org/10.1016/j.ijedudev.2025.103231

Eryilmaz, N., Bellibaş, M. Ş., & Berkovich, I. (2026). Distributed leadership and teacher job satisfaction over time: A large-scale trend analysis of TALIS 2013 and 2018. *Studies in Educational Evaluation, 88*, 101555. https://doi.org/10.1016/j.stueduc.2025.101555

Eryilmaz, N., Kennedy, A. I., Strietholt, R., & Johansson, S. (2025). Teacher job satisfaction: International evidence on the role of school working conditions and teacher characteristics. *Studies in Educational Evaluation, 86*, 101474. https://doi.org/10.1016/j.stueduc.2025.101474

Fan, X., & Chu, Z. (2025). The influence of distributed leadership on Chinese teachers' job satisfaction: The chain mediation of teacher collaboration and teacher self-efficacy. *International Journal of Environmental Research and Public Health, 22*(4), 507. https://doi.org/10.3390/ijerph22040507

Fütterer, T., van Waveren, L., Hübner, N., Fischer, C., & Sälzer, C. (2023). I can't get no (job) satisfaction? Differences in teachers' job satisfaction from a career pathways perspective. *Teaching and Teacher Education, 121*, 103942. https://doi.org/10.1016/j.tate.2022.103942

Ha, C., Pressley, T., & Marshall, D. T. (2025). Teacher voices matter: The role of teacher autonomy in enhancing job satisfaction and mitigating burnout. *PLOS ONE, 20*(1), e0317471. https://doi.org/10.1371/journal.pone.0317471

Hallinger, P., Liu, S., & Chen, L. (2025). Principal instructional leadership and teacher attitudes: A meta-analytic review, 1989-2024. *European Journal of Education, 60*(4), e70199. https://doi.org/10.1111/ejed.70199

Holgado-Apaza, L. A., Carpio-Vargas, E. E., Calderon-Vilca, H. D., Maquera-Ramirez, J., Ulloa-Gallardo, N. J., Acosta-Navarrete, M. S., Barrón-Adame, J. M., Quispe-Layme, M., Hidalgo-Pozzi, R., & Valles-Coral, M. (2023). Modeling job satisfaction of Peruvian basic education teachers using machine learning techniques. *Applied Sciences, 13*(6), 3945. https://doi.org/10.3390/app13063945

Hussain, S., Saba, N. U., Ali, Z., Hussain, H., Hussain, A., & Khan, A. (2022). Job satisfaction as a predictor of wellbeing among secondary school teachers. *SAGE Open, 12*(4). https://doi.org/10.1177/21582440221138726

Jerrim, J. (2024). Are satisfied teachers better teachers? International evidence from the TALIS video study. *Teaching and Teacher Education, 148*, 104687. https://doi.org/10.1016/j.tate.2024.104687

Jerrim, J., & Sims, S. (2021). When is high workload bad for teacher wellbeing? Accounting for the non-linear contribution of specific teaching tasks. *Teaching and Teacher Education, 105, 103395. https://doi.org/10.1016/j.tate.2021.103395*

Kang, H., & Mavrogordato, M. (2025). Examining the relationship between multiple facets of professional development and teacher job satisfaction in the United States. *Educational Studies, 51*(4), 503-528. https://doi.org/10.1080/03055698.2023.2256921

Li, J., Xue, E., & Liu, Y. (2024). The chain mediating role of teachers' job stress in the influence of distributed school leadership on job satisfaction: Evidence from China, the United States, England, and Australia. *Behavioral Sciences, 14*(4), 279. https://doi.org/10.3390/bs14040279

Liu, S., Keeley, J. W., & Sui, Y. (2023). Multilevel analysis of factors influencing teacher job satisfaction in China: Evidence from the TALIS 2018. *Educational Studies, 49*(2), 239-259. https://doi.org/10.1080/03055698.2020.1837615

Lundberg, S. M., & Lee, S.-I. (2017). A unified approach to interpreting model predictions. *Advances in Neural Information Processing Systems, 30*, 4765-4774.

Lundberg, S. M., Erion, G., Chen, H., DeGrave, A., Prutkin, J. M., Nair, B., Katz, R., Himmelfarb, J., Bansal, N., & Lee, S.-I. (2020). From local explanations to global understanding with explainable AI for trees. *Nature Machine Intelligence, 2*, 56-67. https://doi.org/10.1038/s42256-019-0138-9

Madigan, D. J., & Kim, L. E. (2021). Towards an understanding of teacher attrition: A meta-analysis of burnout, job satisfaction, and teachers' intentions to quit. *Teaching and Teacher Education, 105*, 103425. https://doi.org/10.1016/j.tate.2021.103425

McJames, N., Parnell, A., & O'Shea, A. (2025). Factors affecting teacher job satisfaction: A causal inference machine learning approach using data from TALIS 2018. *Educational Review, 77*(2), 381-405. https://doi.org/10.1080/00131911.2023.2200594

Mehrotra, A. (2005). *Leadership styles of principals: Authoritarian and task-oriented.* Mittal Publications.

OECD. (2025). *Results from TALIS 2024: The state of teaching.* OECD Publishing. https://doi.org/10.1787/90df6235-en

Paloniemi Lindström, M. (2026). Exploring the role of school climate, teacher workload, and job satisfaction in student achievement: Insights from PIRLS and TIMSS. *Educational Assessment, Evaluation and Accountability, 38*, 5-31. https://doi.org/10.1007/s11092-025-09476-7

Pargent, F., Schoedel, R., & Stachl, C. (2023). Best practices in supervised machine learning: A tutorial for psychologists. *Advances in Methods and Practices in Psychological Science, 6*(3). https://doi.org/10.1177/25152459231162559

Poloni, R., Oliveira, C., & Lopes, J. (2025). Individual and organizational variables in job satisfaction of first- to ninth-grade teachers: A PRISMA-guided scoping review. *Education Sciences, 15*(4), 451. https://doi.org/10.3390/educsci15040451

Qabliyane, B., Khoudi, Z., Elouafi, A., Salhi, A., & Baslam, M. (2026). Supervised machine learning-based multiclass classification and interpretable feature importance analysis of teacher job satisfaction. Information, 17(4), 377. https://doi.org/10.3390/info17040377

Shi, X., Fan, Q.-Z., Zheng, X., Qiu, D.-F., Sindakis, S., & Showkat, S. (2024). Optimal leadership styles for teacher satisfaction: A meta-analysis of the correlation between leadership styles and teacher job satisfaction. *Journal of the Knowledge Economy, 15*(4), 20459-20490. https://doi.org/10.1007/s13132-023-01697-9

Sudla, W., Surawatakul, L., Wancham, K., & Boonta, N. (2026). All you had to do was stay? Roles of personal and contextual factors in predicting novice teachers' intention to stay. *Teaching and Teacher Education, 182*, 105805. https://doi.org/10.1016/j.tate.2026.105805

Üztemur, S., Gökalp, A., İlğan, A., & Dinç, E. (2026). Leadership, autonomy, and organizational trust as predictors of teacher wellbeing and job satisfaction: A cross-cultural study based on PISA 2022 data. *Frontiers in Psychology, 16*, 1703458. https://doi.org/10.3389/fpsyg.2025.1703458

Veletić, J., & Scherer, R. (2025). The Nordic perspective on the association between distributed leadership and teacher job satisfaction: A multilevel SEM approach. *International Journal of Educational Research, 130*, 102559. https://doi.org/10.1016/j.ijer.2025.102559

Veletić, J., Steinmann, I., & Nilsen, T. (2026). Teacher job satisfaction: International stability over 16 years and associations with school climate changes. *Teaching and Teacher Education, 181*, 105670. https://doi.org/10.1016/j.tate.2026.105670

Wang, J., Liu, Y., & Yang, X. (2025). Predicting teacher turnover in private universities: A machine learning approach based on 10 years of data and satisfaction factors. *Frontiers in Psychology, 16*, 1670195. https://doi.org/10.3389/fpsyg.2025.1670195

Wartenberg, G., Aldrup, K., Grund, S., & Klusmann, U. (2023). Satisfied and high performing? A meta-analysis and systematic review of the correlates of teachers' job satisfaction. *Educational Psychology Review, 35*, 114. https://doi.org/10.1007/s10648-023-09831-4

Wu, Y., & Wu, H. (2026). Unpacking Chinese teacher job satisfaction: Multilevel mediation and moderation of collaborative professional learning, efficacy, and innovation support. *Large-scale Assessments in Education, 14*, 2. https://doi.org/10.1186/s40536-025-00275-6

Xie, Z., & Zou, X. (2025). How perceived teaching autonomy predicts job satisfaction in China: The mediating role of self-efficacy and the moderating role of school climate. *Studies in Educational Evaluation, 87*, 101498. https://doi.org/10.1016/j.stueduc.2025.101498

Zhang, J., Su, H., Pu, D., & Li, P. (2026). Job satisfaction among Chinese secondary school teachers: A two-wave longitudinal study using the job demands-resources model. *Journal of School Psychology, 116*, 101558. https://doi.org/10.1016/j.jsp.2026.101558

Zhang, Y., Djekourmane, D., & Qin, Y. (2026). Understanding the relationship between teacher collaboration beliefs and job satisfaction via collaborative attitudes: A moderated mediation model of classroom self-efficacy from PISA 2022. *Teaching and Teacher Education, 174*, 105411. https://doi.org/10.1016/j.tate.2026.105411

Zhou, S., Slemp, G. R., & Vella-Brodrick, D. A. (2024). Factors associated with teacher wellbeing: A meta-analysis. *Educational Psychology Review, 36*, 63. https://doi.org/10.1007/s10648-024-09886-x