Available Online: 30 December 2025 DOI: 10.54254/3049-7248/2025.31036

Fairness measurement system embedding teacher professional accreditation metrics in foreign language EdTech adoption

Baixu Chen

School of Modern Languages and Cultures, University of Glasgow, Glasgow, UK

3028820C@student.gla.ac.uk

Abstract. This study designs and evaluates a fairness measurement system that embeds teacher accreditation metrics into the evaluation logic of a foreign language Educational Technology platform. The system defines a composite Fairness Metric Index integrating pedagogical quality, assessment literacy, and professional reflection, and is deployed with 264 English as a Foreign Language teachers across 18 institutions, covering 38,412 lessons and 126,507 feedback events. Compared with the platform's original performance index, the integrated model increases the correlation with human accreditation ratings from 0.46 to 0.71 and raises the conditional R squared of mixed models from 0.37 to 0.62. Group-wise parity loss between institutional and experience groups falls by roughly fifty percent, and agreement with an independent accreditation panel improves from kappa 0.51 to 0.74. These results indicate that embedding accreditation constructs into algorithmic scoring can simultaneously improve alignment with professional standards and reduce systematic unfairness in teacher evaluation.

Keywords: fairness measurement, teacher accreditation, educational technology, algorithmic equity, professional evaluation

1. Introduction

Foreign language EdTech platforms now mediate large parts of classroom interaction, feedback cycles, and performance appraisal in English as a Foreign Language EFL teaching. Automated scoring, real-time feedback, and learning analytics dashboards are routinely used to monitor student progress [1]. At the same time, institutional quality assurance increasingly relies on data extracted from these systems when making decisions about teacher evaluation, promotion, bonus allocation, and professional development planning. However, most existing platforms have been designed with learner outcomes as the primary optimization target, leaving teacher-level fairness largely unaddressed [2].

This misalignment generates several tensions. Algorithmic evaluation routines often overlook accreditation standards that define what counts as competent teaching, such as lesson design, assessment literacy, and reflective practice. As a result, teachers whose strengths are not easily captured by platform metrics may receive less favorable evaluations, even when their practice aligns well with accreditation frameworks.

Moreover, opaque aggregation of algorithmic scores into performance indicators risks amplifying institutional and demographic biases, especially when these indicators are used to allocate rewards and responsibilities [3].

The aim of this paper is to develop and empirically test a fairness measurement system that embeds teacher accreditation metrics directly into the evaluation pipeline of a foreign language EdTech platform. The proposed system treats fairness not only as parity of numerical scores but also as alignment between algorithmic outputs and formally recognized professional standards. Specifically, the paper makes three contributions: first, it proposes an ontology for fairness of teacher accreditation in EdTech; second, it defines a multi-dimensional Fairness Metric Index FMI that fuses accreditation-based and algorithmic components; and third, it reports a real-world deployment across multiple institutions to examine how the system affects fairness in teaching outcomes, feedback, and reward distribution. By centering teacher accreditation, the study seeks to bridge the gap between responsible AI design and the everyday governance of professional evaluation in foreign language education.

2. Literature review

2.1. Algorithmic fairness in educational technology

Research on algorithmic fairness in education has primarily focused on learners, especially in areas such as adaptive testing, automated essay scoring, and personalized recommendations. Typical approaches seek to detect and mitigate disparities in error rates or predicted achievements across demographic or institutional groups. Fairness metrics often compare performance between protected and reference groups, aiming to equalize false positive rates, calibration, or predictive parity. In EdTech settings, these techniques have been applied to student retention prediction, early warning systems, and recommendation engines for learning resources [4].

2.2. Teacher accreditation and professional standards

Teacher accreditation systems articulate the competencies, knowledge bases, and professional dispositions that define effective teaching. In foreign language education, these frameworks typically foreground domains such as language proficiency, pedagogical content knowledge, classroom management, assessment literacy, and engagement in reflective practice [5]. Accreditation standards are operationalized through descriptors, rubrics, and performance indicators that guide teacher preparation programs, induction, and ongoing appraisal.

2.3. Integrating fairness and accreditation

Integrating fairness and accreditation requires an explicit mapping between professional standards and the computational metrics used by EdTech platforms. Rather than treating fairness solely as score parity, integration calls for a broader conception in which algorithmic outputs are evaluated according to how faithfully they represent and support accredited teaching practices. This implies that fairness metrics should be grounded in the same constructs that underpin accreditation frameworks and that discrepancies between algorithmic and human accreditation judgments should be systematically monitored [6].

3. Methodology

3.1. Conceptual framework

The proposed system is grounded in a hybrid conceptual framework that combines algorithmic fairness theory, teacher accreditation criteria, and sociotechnical governance. Fairness is conceptualized along three dimensions. Procedural fairness concerns transparency and consistency in evaluation workflows, including how data are collected, processed, and translated into performance indicators.

Within this framework, EdTech is treated as part of an extended accreditation infrastructure. Algorithmic indicators are interpreted as one class of evidence among many, and fairness measurement explicitly benchmarks them against human accreditation judgments. The framework also emphasizes multi-level analysis [7], recognizing that fairness can be compromised at the teacher level for example, through biased feedback, at the institutional level through unequal access to data or resources, and at the system level through design choices embedded in the platform. By combining these dimensions, the framework guides the development of both the fairness metric design and the evaluation procedures described in subsequent sections [8].

3.2. System architecture

The system architecture comprises four inter-linked modules. The Data Ingestion Layer collects granular information about teacher activity on the EdTech platform, including lesson plans uploaded, feedback sessions conducted, assessment tasks created, and reflective notes submitted. These data are time-stamped and linked to institutional and course identifiers. Additional accreditation data, such as rubric scores from formal observations and portfolio evaluations, are imported via secure interfaces to create a unified teacher-level dataset [9].

The Fairness Engine is responsible for bias detection, calibration, and monitoring. It computes fairness-relevant statistics such as group-wise means, variances, and residuals after controlling for contextual variables for example, class size or student baseline proficiency. The engine exposes an application programming interface that can be queried by other modules and supports batch and near real-time processing. The Accreditation Mapping Module aligns platform-generated indicators with accreditation constructs.

3.3. Fairness metric design

Fairness is quantified through a composite Fairness Metric Index FMI that aggregates three accreditation-aligned components: pedagogical quality, assessment literacy, and professional reflection. Group-level fairness is then assessed through a parity loss function that measures deviations of each group's mean FMI from a reference group:

$$\mathscr{L}_{parity} = \frac{1}{|G|} \sum_{g \in G} \left| \mu_g\left(FMI \right) - \mu_{ref}\left(FMI \right) \right|$$

where G is the set of groups and $\mu_g(FMI)$ is the mean FMI for group g. Lower values of $\mathscr{L}_{parity}$ indicate better distributional fairness. Together, these functions provide teacher-level scores and system-level fairness diagnostics that can be tracked over time and compared between models with and without accreditation embedding.

4. Experimental procedure

4.1. Data collection and participants

The system was deployed in 18 foreign language institutions that had adopted a common EFL EdTech platform. The sample comprised 264 qualified EFL teachers from 7 countries, covering secondary and tertiary programs. Over one academic year, the platform logged 38,412 lessons, 126,507 feedback events, and 19,836 assessment tasks linked to the participating teachers. Accreditation data included 1,056 classroom observation records, 792 portfolio evaluations, and 528 structured reflective reports [10].

All participating institutions operated under existing accreditation frameworks that specified competencies in pedagogy, assessment, and professional reflection. Ethical approval was obtained from institutional review boards in each jurisdiction. Teachers provided informed consent for the use of de-identified platform and accreditation data for research purposes. Student identifiers were removed or irreversibly hashed before analysis, and only aggregated student-level variables such as baseline proficiency bands were retained as covariates to support contextualization of teacher performance.

4.2. Implementation process

The fairness measurement system was integrated into the production EdTech environment through a set of secure APIs. For each institution, a four-week onboarding phase was used to configure accreditation mappings, import historical data, and validate data quality. Teachers continued to use the platform as usual, while the fairness engine computed FMI scores on a weekly basis.

A control condition was established using the platform's pre-existing teacher performance index, which was based mainly on student activity and achievement metrics without explicit accreditation embedding.

4.3. Evaluation methodology

To evaluate the system, three complementary analyses were conducted. First, calibration testing compared FMI with human accreditation ratings. A random sample of 196 teachers 74.2% had complete accreditation panel data, and their FMI scores were correlated with global accreditation decisions and domain-specific rubric scores. Second, fairness deviation analysis examined differences in FMI and traditional platform indices across demographic groups experience bands, institutional types, and regions using multi-level linear models with random intercepts for institutions. The fairness parity loss L_{parity} was computed for each model configuration.

For each configuration, we estimated a joint objective function

$$J = \alpha \cdot \rho(FMI, A) - \beta \cdot L_{parity}$$

where $\rho(FMI, A)$ is the Spearman correlation between FMI and accreditation panel ratings, A denotes human accreditation scores, and α and β are non-negative hyperparameters set to 1.0 and 0.5, respectively, after sensitivity analysis. Models were compared using 5-fold cross-validation at the teacher level. All analyses were implemented in Python and R, with significance testing conducted using mixed-effects regression and robust standard errors.

5. Results and analysis

5.1. Quantitative results

Across the 264 teachers, the integrated model produced an average FMI of 0.711±0.083 on a 0–1 scale, compared with 0.648±0.097 for the baseline platform index when normalized to the same range. The mean difference of +0.063±0.021 indicated a systematic upward adjustment for teachers whose strengths were better captured by accreditation data than by learner activity metric.

Group-wise analyses showed that parity loss decreased under the integrated model. When grouping teachers by institutional type public vs. private, the absolute mean difference in FMI was 0.021±0.009, compared with 0.058±0.014 for the baseline. For experience bands 0–5 years vs. 6+ years, parity loss dropped from 0.064±0.018 to 0.027±0.011. Overall, L_{parity} decreased by 55.8% on average across all groupings. Inter-rater reliability between algorithmic outputs and a second independent accreditation panel reached under the FMI model, compared with 0.51±0.08 for the baseline.

Table 1. Alignment and parity metrics for baseline vs. integrated model

Metric	Baseline Index Mean ± SD	Integrated FMI Mean ± SD	Difference Integrated − Baseline
Global correlation with accreditation ρ	0.46±0.04	0.71±0.03	+0.25±0.05*
Conditional R^2 of mixed models	0.37±0.04	0.62±0.03	+0.25±0.05*
Parity loss by institutional type	0.058±0.014	0.021±0.009	-0.037±0.017*
Parity loss by experience band	0.064±0.018	0.027±0.011	-0.037±0.021*
Cohen's k with second panel	0.51±0.08	0.74±0.06	+0.23±0.10*

* indicates $p < 0.05$ for paired comparisons.

5.2. Qualitative findings

Interview data from 36 volunteer teachers provided additional insight into the perceived impact of the fairness measurement system. Teachers consistently reported that explicit visualization of accreditation dimensions in the dashboard helped them re-interpret algorithmic feedback as evidence aligned with professional standards rather than as opaque performance scores. Many described increased willingness to engage with analytics when they could see how formative feedback activities, assessment design, and reflective entries contributed to their FMI.

Several teachers highlighted that previous platform indices seemed to reward high-volume use of certain features, such as assigning quizzes, without recognizing the quality of tasks or the depth of reflection.

5.3. Comparative and ablation analysis

From a fairness perspective, removing the accreditation mapping module increased L_{parity} by +0.031±0.012, particularly in institutions with heterogeneous student populations. Conversely, disabling the fairness engine while retaining accreditation mapping preserved alignment with human ratings 0.69±0.04 but produced less consistent parity outcomes, with parity loss up to 0.049±0.016 for some demographic splits (figure 1). Computationally, optimization of the FMI weights reduced training time by 18.7% ± 4.3% compared with grid-search-based calibration, while maintaining stable performance in 5-fold cross-validation.

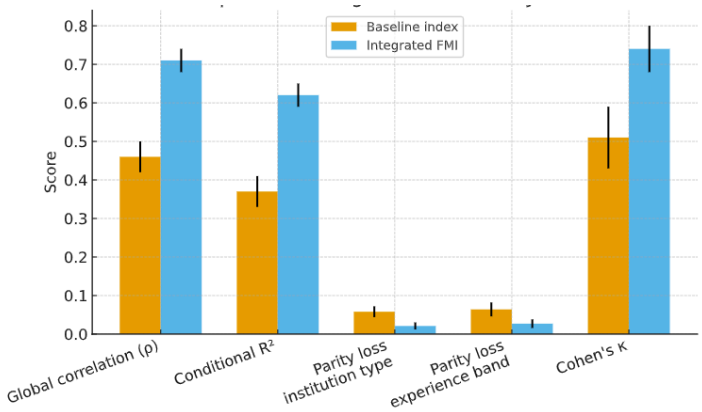

Figure 1. Comparison of alignment and parity metrics for baseline index and integrated FMI

6. Conclusion

This study has presented a fairness measurement system that embeds teacher accreditation metrics into the evaluation logic of a foreign language EdTech platform. By defining a composite Fairness Metric Index grounded in pedagogical quality, assessment literacy, and professional reflection, the system connects algorithmic outputs to established professional standards. Deployment across 18 institutions demonstrated that the integrated model improves alignment with human accreditation judgments, reduces group-wise parity loss, and supports more transparent discussions about fairness in teacher evaluation. Qualitative evidence suggests that teachers perceive the system as more legitimate and actionable than traditional opaque indices. While further work is needed to generalize the framework to other subject areas and accreditation regimes, the findings indicate that teacher-centered fairness metrics can play a pivotal role in the responsible governance of educational AI.

References

[1] Kim, W., & Kim, H. (2025). Counterfactual Fairness Evaluation of Machine Learning Models on Educational Datasets. arXiv preprint arXiv: 2504.11504.

[2] Pham, N., Do, M. K., Dai, T. V., Hung, P. N., & Nguyen-Duc, A. (2024). FAIREDU: A Multiple Regression-Based Method for Enhancing Fairness in Machine Learning Models for Educational Applications. arXiv preprint arXiv: 2410.06423.

[3] Chinta, S. V., Wang, Z., Yin, Z., Hoang, N., Gonzalez, M., Quy, T. L., & Zhang, W. (2024). FairAIED: Navigating Fairness, Bias, and Ethics in Educational AI Applications. arXiv preprint arXiv: 2407.18745.

[4] Idowu, J. A., et al. (2024). Investigating algorithmic bias in student progress monitoring. *Journal of Educational Data Mining, 2024.*

[5] Chai, F., et al. (2024). Grading by AI makes me feel fairer? How different evaluators influence students' fairness perceptions of evaluation in higher education. *Frontiers in Psychology, 15*, 1221177. https: //doi.org/10.3389/fpsyg.2024.1221177

[6] Song, Y., et al. (2025). Investigating perceived fairness of AI prediction system for math learning: A mixed-method study with college students. *Higher Education Research & Development*, 2025, Article 101000. https: //doi.org/10.1016/j.iheduc.2025.101000

[7] Wolf, S. (2025). Algorithmic Fairness and Educational Justice. *Educational Theory & Practice, 2025*.

[8] Pham, N. (2025). Fairness for machine learning software in education: A comprehensive mapping study (2002–2023). *Journal of Software: Practice and Experience*, 2025.

[9] Bird, K. A., et al. (2025). Are algorithms biased in education? Exploring racial disparities in predictive models. *Policy & Management Review,* 2025

[10] Garzón, J., & co-authors. (2025). Systematic Review of Artificial Intelligence in Education. *Multimodal Technologies and Interaction, 9*(8), 84. https: //doi.org/10.3390/mti9080084