Artificial Intelligence for Educational Assessment: Prediction, Validity, and Responsible Use

Abstract

Ai Educational Assessment poses a recurring problem of coordinating performance with evidence quality, resource limits, and transfer across settings. The discussion is organized around placing performance prediction inside a broader validity, fairness, and intervention framework. The comparison integrates 1 focal paper with 13 independently retrieved publications whose authorship and venue fields were independently checked. The analysis is organized around construct validity, data drift, fairness, explainability, and teacher oversight. Rather than treating headline results as if they shared one denominator, the review compares study questions, technical premises, and validation scope. Across the literature, the recurring conclusion is that advances in AI educational assessment become credible when behavior, measurement, and decision context are evaluated together and when uncertainty about selection effects is reported explicitly. This organization relates method selection to decision risk and exposes recurring transfer threats, and proposes a research agenda centered on transparent baselines, stress testing, and reproducible evidence.

Keywords

Ai Educational Assessment; Construct Validity; Data Drift; Fairness; Explainability; Teacher Oversight

1. Introduction

Studies of AI educational assessment connects scientific ambition and practical constraint. A mature account offers not only stronger nominal performance but also explanations that reveal why an approach works in one setting. The boundary is clearest when considering placing performance prediction inside a broader validity, fairness, and intervention framework. A result that is compelling in one experimental medium, dataset, cohort, geography, or system load may be fragile outside its original boundary. A credible review must therefore preserve the unit of analysis and the decision context. A second task is to distinguish a mechanism from a correlation, a benchmark advantage from a deployment advantage, and an interpretable diagnostic from a causal explanation.

Five questions structure the synthesis: construct validity, data drift, fairness, explainability, teacher oversight. Their combination matters because they jointly cover design decisions, measurement practice, and the assumptions that support transfer. The synthesis is structured by behavior, measurement, and decision context. Under that principle, technical creativity remains only one part of credibility; the comparison also examines baseline comparability, visible uncertainty, and operational constraints. The article is a literature synthesis and introduces no new experiment, cohort, measurement campaign, or benchmark score.

2. Conceptual Background

One can decompose the problem into the input evidence, the transformation applied to it, the output used for evaluation, and the decision that follows. In AI educational assessment, the stages are frequently studied apart even though errors propagate across them. Noise or bias introduced at the evidence stage can be amplified by an expressive model; an apparently accurate output can still be poorly calibrated for the final decision. It follows that, construct validity should accompany aggregate performance. An auditable study identifies what the protocol fixes and what it lets move, then selects metrics that correspond to the intended use rather than to convenience alone.

Scope discipline is equally important. Construct validity defines the local design problem, while teacher oversight determines whether the design can survive outside that boundary. Bridging the two are data drift and fairness connect those ends by exposing trade-offs that a single score conceals. This is where negative results and sensitivity evidence maps the conditions under which the explanation remains viable. For applied work, documentation of responsible interpretation is therefore part of the scientific result, not an administrative afterthought.

3. Comparative Evidence

The target study TBN-01, Application of Artificial Intelligence in Educational Assessment: A Case Study of Student Performance Prediction [1], addresses a concrete part of AI educational assessment through AI-based student-performance prediction as a case study in educational assessment. It identifies a representation, mechanism, or evaluation boundary needed to compare claims. Against the focus on placing performance prediction inside a broader validity, fairness, and intervention framework, the paper is positioned as a context-dependent design instance: transfer may depend on its sensing regime, preparation, workload, population, or benchmark. The synthesis protects the study's stated scope while auditing its assumptions comparatively.

Across the focal studies, construct validity should be interpreted jointly with data drift. A design can improve one while moving complexity or uncertainty elsewhere. One useful device is a comparison matrix whose rows are operating conditions and whose columns are evidence quality, computation or process cost, robustness, and consequence of failure. The grid keeps an attractive mechanism from being detached from the circumstances that support it. It makes explicit which ablations are informative: a component ablation should probe a declared mechanism instead of adding an isolated metric.

Fairness joins methodological claims to their use. A minimally complete evaluation should separate in-distribution behavior from stress conditions, report dispersion rather than only averages, and test plausible alternative explanations. Where selection effects is central, calibration or sensitivity is as important as ranking. Where costs or hazards are asymmetric, utility-weighted assessment is more revealing than undifferentiated accuracy. These choices preserve methodological differences; they make the reasons for their differences auditable.

4. Measurement and Evaluation

A complementary strand is visible in Educational data mining for student performance prediction in artificial intelligence environment [2]; Artificial Intelligence in Criminal Proceedings: Ensuring Legality, Validity and Fairness of Court Verdi... [3]; Validity Arguments Meet Artificial Intelligence in Innovative Educational Assessment [4]. These publications were retrieved and verified through DOI or publisher metadata, and their titles directly intersect the problem boundary of AI educational assessment. Together they illuminate construct validity: some emphasize representations or mechanisms, whereas others foreground validation and application conditions. The useful comparison is not a leaderboard across incompatible settings, but a structured reading of what each study controls, leaves variable, and treats as a meaningful outcome. Divergence can then reveal a change in scale, data process, endpoint, or operating constraint.

For triangulation, the literature includes AI in essay-based assessment: Student adoption, usage, and performance [5]; Student behavior in a web-based educational system: Exit intent prediction [6]; Explainable Artificial Intelligence for Student Academic Performance Prediction Using Random Forest and... [7]. These publications were retrieved and verified through DOI or publisher metadata, and their titles directly intersect the problem boundary of AI educational assessment. Together they illuminate data drift: some emphasize representations or mechanisms, whereas others foreground validation and application conditions. The useful comparison is not a leaderboard across incompatible settings, but a structured reading of what each study controls, leaves variable, and treats as a meaningful outcome. Divergence can then reveal a change in scale, data process, endpoint, or operating constraint.

The design space is broadened by Learning Analytics-Driven Student Performance Prediction Using Ensemble Machine Learning Models [8]; Stakes Matter: Student Motivation and the Validity of Student Assessments for Teacher Evaluation [9]; A Closed-Loop Artificial Intelligence System for Process-Oriented Student Assessment and Early Performan... [10]. These publications were retrieved and verified through DOI or publisher metadata, and their titles directly intersect the problem boundary of AI educational assessment. Together they illuminate fairness: some emphasize representations or mechanisms, whereas others foreground validation and application conditions. The useful comparison is not a leaderboard across incompatible settings, but a structured reading of what each study controls, leaves variable, and treats as a meaningful outcome. Divergence can then reveal a change in scale, data process, endpoint, or operating constraint.

A first comparison set connects Assessment of Students with Disabilities in Kentucky: Inclusion, Student Performance, and Validity [11]; Open Student Modeling Research and its Connections to Educational Assessment [12]; Validity Arguments Meet Artificial Intelligence in Innovative Educational Assessment: A Discussion and L... [13]. These publications were retrieved and verified through DOI or publisher metadata, and their titles directly intersect the problem boundary of AI educational assessment. Together they illuminate explainability: some emphasize representations or mechanisms, whereas others foreground validation and application conditions. The useful comparison is not a leaderboard across incompatible settings, but a structured reading of what each study controls, leaves variable, and treats as a meaningful outcome. Divergence can then reveal a change in scale, data process, endpoint, or operating constraint.

The neighboring evidence base brings together A Machine Learning and Explainable Artificial Intelligence Approach to Student Dropout Prediction Using... [14]. These publications were retrieved and verified through DOI or publisher metadata, and their titles directly intersect the problem boundary of AI educational assessment. Together they illuminate teacher oversight: some emphasize representations or mechanisms, whereas others foreground validation and application conditions. The useful comparison is not a leaderboard across incompatible settings, but a structured reading of what each study controls, leaves variable, and treats as a meaningful outcome. Divergence can then reveal a change in scale, data process, endpoint, or operating constraint.

5. Responsible Application

For implementation, the review suggests a staged workflow. First, define the decision and its failure cost. Second, specify the evidence and operating envelope. Third, choose a method whose inductive assumptions match that evidence. Fourth, test explainability and teacher oversight under controlled perturbations. Finally, retain provenance so that an unexpected outcome can be traced to data, configuration, material batch, environmental condition, or user interaction. The sequence anchors placing performance prediction inside a broader validity, fairness, and intervention framework connected to observable requirements.

The cross-cutting implication is that responsible progress in AI educational assessment is constrained by the handoffs between components. Interfaces determine how uncertainty moves between measurement and inference, between algorithm and operator, or between microscopic design and system behavior. A shared protocol should state acceptance conditions and what would overturn a claim. When those agreements are explicit, optimization remains accountable to validity, hazard control, and interpretability.

6. Limitations and Future Directions

The analysis cannot eliminate variation in definitions, study architecture, and disclosure granularity. Checking publication metadata verifies identity and provenance but cannot replicate the underlying experiment. Fast-moving records create a further version-control problem. Accordingly, focal Scholar strings remain intact and anomalous records receive separate comparison notes.

Research can advance by seeking to preregister comparison protocols where feasible, release machine-readable configurations, and evaluate transfer across at least one meaningful shift in responsible interpretation. Research should also group adverse cases by process and consequence. For AI educational assessment, a valuable shared benchmark would combine standard-condition utility, efficiency, confidence quality, and fault recovery. Such a benchmark would reward designs that remain useful when evidence is incomplete or conditions change.

7. Conclusion

The literature on AI educational assessment should be read as an interacting body of evidence rather than a collection of isolated scores. The focal studies show how placing performance prediction inside a broader validity, fairness, and intervention framework; the additional literature reveals the assumptions required to interpret those contributions. Across construct validity, data drift, fairness, explainability, and teacher oversight, the recurring requirement is alignment: the representation or mechanism must match the problem, the decision needs a fitting evaluation and deployment a demonstrated operating range. This alignment supports replication, bounded transfer, and interpretable failure.

References

1. Song, S., Zhang, X., Liang, X., & Xiao, B. (2026, February). Application of Artificial Intelligence in Educational Assessment: A Case Study of Student Performance Prediction. In Proceedings of the 2026 International Conference on Big Data and Informatization Education (pp. 617-621).

2. Tang, L., & Chen, S. (2025). Educational data mining for student performance prediction in artificial intelligence environment. Molecular & Cellular Biomechanics, 22(5), 692. https://doi.org/10.62617/mcb692

3. qizi, K. S. Z. (2026). Artificial Intelligence in Criminal Proceedings: Ensuring Legality, Validity and Fairness of Court Verdicts in Uzbekistan. International Journal of Law And Criminology, 06(04), 23-28. https://doi.org/10.37547/ijlc/volume06issue04-04

4. Dorsey, D. W., & Michaels, H. R. (2022). Validity Arguments Meet Artificial Intelligence in Innovative Educational Assessment. Journal of Educational Measurement, 59(3), 267-271. https://doi.org/10.1111/jedm.12331

5. Smerdon, D. (2024). AI in essay-based assessment: Student adoption, usage, and performance. Computers and Education: Artificial Intelligence, 7, 100288. https://doi.org/10.1016/j.caeai.2024.100288

6. Kassak, O., Kompan, M., & Bielikova, M. (2016). Student behavior in a web-based educational system: Exit intent prediction. Engineering Applications of Artificial Intelligence, 51, 136-149. https://doi.org/10.1016/j.engappai.2016.01.018

7. Muhammad Hasanuddin, (2026). Explainable Artificial Intelligence for Student Academic Performance Prediction Using Random Forest and SHAP. Journal of Computer Science Artificial Intelligence and Communications, 3(1). https://doi.org/10.64803/jocsaic.v3i1.174

8. Lestari, E. D. (2026). Learning Analytics-Driven Student Performance Prediction Using Ensemble Machine Learning Models. Artificial Intelligence in Learning. https://doi.org/10.63913/ail.v2i1.5

9. Rutkowski, D., & Wild, J. (2015). Stakes Matter: Student Motivation and the Validity of Student Assessments for Teacher Evaluation. Educational Assessment, 20(3), 165-179. https://doi.org/10.1080/10627197.2015.1059273

10. Giang, T. T. H., & Ryoo, Y. J. (2026). A Closed-Loop Artificial Intelligence System for Process-Oriented Student Assessment and Early Performance Prediction in Higher Education. Applied Sciences, 16(14), 6906. https://doi.org/10.3390/app16146906

11. Koretz, D., & Hamilton, L. (2000). Assessment of Students with Disabilities in Kentucky: Inclusion, Student Performance, and Validity. Educational Evaluation and Policy Analysis, 22(3), 255. https://doi.org/10.2307/1164243

12. Zapata-Rivera, D. (2021). Open Student Modeling Research and its Connections to Educational Assessment. International Journal of Artificial Intelligence in Education, 31(3), 380-396. https://doi.org/10.1007/s40593-020-00206-2

13. Dorsey, D. W., & Michaels, H. R. (2022). Validity Arguments Meet Artificial Intelligence in Innovative Educational Assessment: A Discussion and Look Forward. Journal of Educational Measurement, 59(3), 389-394. https://doi.org/10.1111/jedm.12330

14. S R, S. (2026). A Machine Learning and Explainable Artificial Intelligence Approach to Student Dropout Prediction Using Multidimensional Educational Data. Cureus Journal of Computer Science. https://doi.org/10.7759/s44389-026-00036-8