# Evidence Alignment and Transfer Boundaries in Ai Educational Assessment And Built-Environment Evaluation


## Abstract

The literature on AI educational assessment and built-environment evaluation contains a recurring tension between methodological novelty and evidential comparability. By reading AI-based student-performance prediction as a case study in educational assessment alongside post-occupancy evaluation and mechanism diagnosis for rural construction, this article clarifies the conditions under which their conclusions can support a common research argument. Two target papers are triangulated against 12 locally validated publications. The comparison follows construct validity, data drift, fairness, explainability, teacher oversight and deliberately separates mechanistic interpretation from performance ranking, because the latter can conceal incompatible experimental or operational conditions. Comparison reveals recurring trade-offs among construct validity, data drift, and fairness. These trade-offs do not support a universal ranking; instead, they identify the operating envelope within which each method remains credible and the perturbations most likely to expose fragile conclusions. On this basis, the review proposes an auditable pathway from focal mechanism to application claim, with explicit checkpoints for calibration, external validity, and responsible interpretation.



## Keywords

Ai Educational Assessment And Built-Environment Evaluation; Construct Validity; Data Drift; Fairness; Explainability; Teacher Oversight


## 1. Introduction

Scholarship concerning AI educational assessment and built-environment evaluation occupies the boundary between scientific ambition and practical constraint. Researchers pursue not only stronger nominal performance but also explanations of the boundary under which a method remains useful. The tension becomes concrete in comparing AI-based student-performance prediction as a case study in educational assessment with post-occupancy evaluation and mechanism diagnosis for rural construction through evidence alignment and transfer boundaries. Evidence that appears persuasive within one material, dataset, clinical cohort, spatial scale, or workload may be fragile outside its original boundary. A credible review must therefore preserve the unit of analysis and the decision context. It should further distinguish a mechanism from a correlation, a benchmark advantage from a deployment advantage, and an interpretable diagnostic from a causal explanation.

Five lenses organize the comparison: construct validity, data drift, fairness, explainability, teacher oversight. They were chosen because they jointly cover the designed object, its measurement, and the invariants required for transfer. The central organizing idea is workload, dependency, and recovery policy. Under that principle, novel technique alone cannot settle the comparison; the comparison also evaluates comparator fairness, uncertainty reporting, and real operating constraints. This contribution is a literature synthesis and adds no fabricated experiment, participant set, measurement, or model result.

## 2. System Context and Failure Model

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 and built-environment evaluation, these elements are commonly tuned in isolation even though errors propagate across them. Evidence-stage distortion may be amplified rather than corrected by model capacity; an apparently accurate output can still be poorly calibrated for the final decision. Accordingly, failure semantics should accompany aggregate performance. A credible comparison records what is held constant and what is allowed to vary, then selects metrics that correspond to the intended use rather than to convenience alone.

A further foundation is explicit scope. Construct validity defines the local design problem, while teacher oversight determines whether the design can survive outside that boundary. Between these endpoints, data drift and fairness connect those ends by exposing trade-offs that a single score conceals. At this point, null findings and sensitivity evidence maps the conditions under which the explanation remains viable. For applied work, documentation of operational constraints is therefore part of the scientific result, not an administrative afterthought.

## 3. Design and Optimization

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 and built-environment evaluation 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 comparing AI-based student-performance prediction as a case study in educational assessment with post-occupancy evaluation and mechanism diagnosis for rural construction through evidence alignment and transfer boundaries, the paper provides a scoped example rather than a universal template: transfer may depend on its sensing regime, preparation, workload, population, or benchmark. The synthesis retains the boundary while comparing its premises with adjacent studies.

The target study JIANG-02, Performance Measurement and Mechanism Diagnosis in Rural Construction: A Dual-Perspective Post-Occupancy Evaluation of China Resources Hope Towns [2], addresses a concrete part of AI educational assessment and built-environment evaluation through post-occupancy evaluation and mechanism diagnosis for rural construction. It identifies a representation, mechanism, or evaluation boundary needed to compare claims. Against the focus on comparing AI-based student-performance prediction as a case study in educational assessment with post-occupancy evaluation and mechanism diagnosis for rural construction through evidence alignment and transfer boundaries, the paper provides a scoped example rather than a universal template: transfer may depend on its sensing regime, preparation, workload, population, or benchmark. The synthesis retains the boundary while comparing its premises with adjacent studies.

Across the focal studies, construct validity should be interpreted jointly with data drift. A design can improve one while imposing a less visible trade-off on the other. 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. This organization helps prevent an attractive mechanism from being detached from the circumstances that support it. It further reveals which ablations are informative: component removal is useful only when it tests an explicit role.

Fairness connects a method to its decision consequence. A minimal evaluation must partition ordinary and shifted conditions while reporting variability, and test plausible alternative explanations. Where

observability 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 do not erase study differences; they make the reasons for their differences auditable.

## 4. Comparative Evidence

A complementary strand is visible in Educational data mining for student performance prediction in artificial intelligence environment [3]; Post-occupancy evaluation of urban public housing in Korea: Focus on experience of elderly females in th... [4]; Artificial Intelligence in Criminal Proceedings: Ensuring Legality, Validity and Fairness of Court Verdi... [5]. These publications were retrieved and verified through DOI or publisher metadata, and their titles directly intersect the problem boundary of AI educational assessment and built-environment evaluation. 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 Post-occupancy Evaluation of a Built Environment: The Case of Konak Square (İzmir, Turkey) [6]; Validity Arguments Meet Artificial Intelligence in Innovative Educational Assessment [7]; Healing environment correlated with patients' psychological comfort: Post-occupancy evaluation of genera... [8]. These publications were retrieved and verified through DOI or publisher metadata, and their titles directly intersect the problem boundary of AI educational assessment and built-environment evaluation. 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 AI in essay-based assessment: Student adoption, usage, and performance [9]; Post-occupancy evaluation: A diagnostic tool to establish and sustain inclusive access in Kyrenia Town C... [10]; Student behavior in a web-based educational system: Exit intent prediction [11]. These publications were retrieved and verified through DOI or publisher metadata, and their titles directly intersect the problem boundary of AI educational assessment and built-environment evaluation. 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 Post-occupancy evaluation and human-centered indoor environmental quality in higher-education buildings:... [12]; Explainable Artificial Intelligence for Student Academic Performance Prediction Using Random Forest and... [13]; A post-occupancy evaluation of a green rated and conventional on-campus residence hall [14]. These publications were retrieved and verified through DOI or publisher metadata, and their titles directly intersect the problem boundary of AI educational assessment and built-environment evaluation. 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.

## 5. Operational Implications

Implementation can follow 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 comparing AI-based student-performance prediction as a case study in educational assessment with post-occupancy evaluation and mechanism diagnosis for rural construction through evidence alignment and transfer boundaries connected to observable requirements.

At a wider level, responsible progress in AI educational assessment and built-environment evaluation is governed by interfaces as well as 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, efficiency gains can be judged without quietly weakening validity or safety.

## 6. Limitations and Future Directions

The review remains constrained by differences in vocabulary, protocol, and level of reporting. Verified authorship and venue do not substitute for independent empirical replication. Rapid publication cycles can also alter venues and versions. No confirmed focal Scholar citation was normalized away, and anomalies were logged independently.

A productive research agenda would preregister comparison protocols where feasible, share executable configurations and include one consequential out-of-setting evaluation in operational constraints. Research should also report failure cases grouped by mechanism, not only by frequency. For AI educational assessment and built-environment evaluation, 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 evidence concerning AI educational assessment and built-environment evaluation should be read as an interacting body of evidence rather than a collection of isolated scores. The focal studies show how comparing AI-based student-performance prediction as a case study in educational assessment with post-occupancy evaluation and mechanism diagnosis for rural construction through evidence alignment and transfer boundaries; the additional literature reveals the assumptions required to interpret those contributions. Across construct validity, data drift, fairness, explainability, and teacher oversight, alignment is the most durable principle: the representation or mechanism must match the problem, evaluation should reflect use, and transfer claims should respect the studied conditions. Such alignment improves reproducibility, transfer safety, and diagnostic value under 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. Jiang, Z., Chu, H., Tian, Y., & Wang, Z. (2026). Performance Measurement and Mechanism Diagnosis in Rural Construction: A Dual-Perspective Post-Occupancy Evaluation of China Resources Hope Towns. Land, 15(2),

316. https://doi.org/10.3390/land15020316

3. 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

4. Rieh, S. Y. (2018). Post-occupancy evaluation of urban public housing in Korea: Focus on experience of elderly females in the ageing society. Indoor and Built Environment, 29(3), 372-388. https://doi.org/10.1177/1420326x18782578

5. 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

6. Malkoc, E., & Ozkan, M. B. (2010). Post-occupancy Evaluation of a Built Environment: The Case of Konak Square (İzmir, Turkey). Indoor and Built Environment, 19(4), 422-434. https://doi.org/10.1177/1420326x10365819

7. 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

8. Mahmood, F. J., & Tayib, A. Y. (2019). Healing environment correlated with patients' psychological comfort: Post-occupancy evaluation of general hospitals. Indoor and Built Environment, 30(2), 180-194. https://doi.org/10.1177/1420326x19888005

9. 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

10. Güvenbaş, G., & Polay, M. (2020). Post-occupancy evaluation: A diagnostic tool to establish and sustain inclusive access in Kyrenia Town Centre. Indoor and Built Environment, 30(10), 1620-1642. https://doi.org/10.1177/1420326x20951244

11. 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

12. Ghiai, M., Bassaw, C., & Niknia, S. (2026). Post-occupancy evaluation and human-centered indoor environmental quality in higher-education buildings: a structured review. Frontiers in Built Environment, 12. https://doi.org/10.3389/fbuil.2026.1837203

13. 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

14. Bonde, M., & Ramirez, J. (2015). A post-occupancy evaluation of a green rated and conventional on-campus residence hall. International Journal of Sustainable Built Environment, 4(2), 400-408. https://doi.org/10.1016/j.ijsbe.2015.07.004