LSAR's Report (verbatim)

Paper Information

- **Title**: Engagement** **and Belonging Add Predictive Signal for High School** **Dropout
- **Target Venue**: JEDM (confidence: 1.00)
- **Page Count**: 37 pages
- **Review Date**: 2026-08-29

1. Paper Summary (150-250 words)

This paper uses HSLS:09 public-use data to test whether fall-of-ninth-grade engagement, school belonging, expectations, and school-context measures improve prediction of a later high school dropout episode beyond ninth-grade math achievement and SES. The outcome is X3EVERDROP, indicating any dropout episode by the 2013 follow-up, rather than permanent non-completion. The authors use a reconstructed school-cluster train/test split, compare five classifiers, report clustered bootstrap confidence intervals, and conduct block-incremental analyses. The full XGBoost model achieves an AUC of approximately 0.73, with the engagement/belonging/expectations/school-context block adding about 0.060 AUC points over an achievement-plus-SES baseline. The paper also reports precision-recall behavior at operational thresholds, calibration, subgroup performance by sex, race/ethnicity, and SES, and a sensitivity analysis excluding high-missingness predictors. The authors conclude that brief ninth-grade surveys can improve risk stratification, but that predicted probabilities must be recalibrated before threshold-based use. The paper presents three main contributions: a nationally representative incremental-signal estimate, threshold-based precision-recall reporting, and school-aware evaluation with clustered uncertainty.

2. Relevance Assessment

- **venue_fit**: Strong — The paper addresses a core EDM/learning-analytics problem—early identification of students at risk of dropout—and contributes a predictive-modeling study with fairness and reproducibility considerations. For JEDM, the work fits the journal's scope as an educational data mining application with policy implications.
- **track_fit**: The paper directly addresses several EDM topics: learner knowledge/performance modeling, learning analytics, early warning systems, equity/fairness/transparency, and reproducibility.
- **scope_concerns**: No out-of-scope flags. The paper does not emphasize the conference theme "across borders," but this is not disqualifying for JEDM. The main scope concern is that the claimed "nationally representative" inference should be interpreted carefully given the analytic sample and unweighted evaluation, as discussed below.

3. Strengths

- **Claim:** First nationally representative estimate of the incremental predictive value of ninth-grade engagement and belonging for dropout episodes above achievement and SES.
- **Claim:** Reporting precision-recall behavior at explicit early-warning thresholds instead of accuracy/AUC as the headline.
- **Claim:** School-aware train/test split, clustered bootstrap confidence intervals, subgroup analyses, and sensitivity analysis.

- **Clear, practical research question.** The block-incremental design directly asks whether engagement and belonging add signal beyond achievement and SES, which is the decision-relevant question for schools considering collecting survey data.
- **Attention to clustered data structure.** The paper uses a school-aware split and clustered bootstrap CIs, which is necessary for HSLS:09-like nested samples and often ignored in dropout-prediction work.
- **Threshold and calibration reporting.** The authors avoid treating AUC/accuracy as sufficient. They report precision, recall, F1/F2, confusion-matrix implications, Brier score, ECE, and calibration slope/intercept, and they clearly state that recalibration is needed before deployment.
- **Subgroup fairness analysis with explicit small-cell warnings.** The paper reports performance separately by sex, race/ethnicity, and SES quintiles and flags small subgroups as unreliable. This is an important step beyond aggregate performance.
- **Sensitivity analysis for missing data.** Dropping high-missingness predictors and showing the engagement/belonging signal is robust increases confidence in the main result.

4. Weaknesses

- **MAJOR — Possible preprocessing/imputation leakage.** The paper states that missing values were imputed with `IterativeImputer`, but it does not state that the imputer was fitted only on training folds. If imputation was performed on the full dataset before the school-aware split, test-set information could leak into training. The predictive claims and confidence intervals would then be optimistic. This is a central methodological concern under branch (c).
- **MAJOR — Reconstructed school clusters undermine the school-aware design.** School IDs are suppressed in the public-use file, so the authors reconstruct clusters by matching school-level profile variables. This yields 935 pseudo-clusters against the expected 944, but cannot be verified. The school-aware split and clustered bootstrap both depend on this reconstruction. If true school membership is imperfectly recovered, some schools likely appear in both training and test, weakening the "unseen school" claim.
- **MAJOR — Internal inconsistency in the threshold/precision discussion.** The text reports a balanced model precision of 0.576 and recall of 0.674, but later states the confusion matrix includes 1,550 false positives. If 341 true positives were captured and 1,550 non-dropouts were flagged, precision would be approximately 0.18, not 0.576. The flagged count is important for policy interpretation and must be reconciled.
- **MAJOR — Outcome measurement validity is not described.** The paper treats X3EVERDROP as a dropout episode, but does not explain how this variable was collected or validated. If it is based on student self-report, measurement error could affect both the base rate and the predictive relationships. The practical contribution depends on accurately identifying true dropout episodes.
- **MAJOR/MINOR — Pre-specified moderation analysis appears unreported.** The Methods describe an SES moderation analysis and tertile comparisons, but the Results section does not appear to report those results. This is an internal reproducibility gap.
- **MINOR — Inconsistent CI reporting.** The abstract reports AUC 0.73, 95% CI [0.71, 0.75]; the model comparison section reports [0.708, 0.751]; the school-level variance section reports a clustered CI [0.706, 0.753]. The paper should clearly distinguish row-level versus cluster-level bootstrap CIs and use one primary interval consistently.

5. Detailed Dimensional Assessment

Dimension	Score (1-10)	Justification
Relevance	8	Dropout early warning is a core JEDM topic, and the decision-relevant question of whether engagement and belonging add signal beyond achievement and SES has clear practical value.
Novelty	6	The claim of a first nationally representative estimate for dropout episodes is a useful empirical contribution, but the design is otherwise an incremental block-incremental prediction study.
Theoretical/Conceptual Grounding	6	The work draws on engagement and belonging literatures, but the excerpt does not demonstrate deep theoretical integration; grounding appears adequate but not exceptional.
Methodological Rigor	4	School-aware split and clustered bootstrap are strengths, but possible imputation leakage, unverifiable reconstructed school clusters, and an internal precision/false-positive inconsistency raise central validity concerns.
Empirical Support / Results	4	Reported precision-recall behavior is directly undermined by the apparent mismatch between stated precision and false positive counts; also, pre-specified moderation analyses appear unreported.

Dimension	Score (1-10)	Justification
Significance & Impact	5	If valid, the findings could inform national early-warning decisions, but the methodological and reporting problems must be resolved before the work can influence practice or policy.
Ethics, Fairness & Equity	6	Subgroup analyses and demographic attention are positive, but there is no explicit discussion of fairness, measurement equity, or bias implications for at-risk subgroups.
Clarity of Communication	4	The paper is directionally clear but has important internal inconsistencies and under-described outcome construction, which seriously impede a reader's ability to evaluate the results.

6. Comparison with Related Work

The paper positions itself against a broad dropout-prediction literature but misses some of the most directly relevant prior work.

- **Lovelace et al. (2017)** is a critical omission. That study explicitly examined cognitive and affective engagement in predicting dropout and on-time graduation beyond school records. The submitted paper should differentiate itself from Lovelace et al. more precisely: the possible distinction is the nationally representative HSLS:09 sample and the dropout-episode outcome, but this is not yet shown.
- **Sorensen (2018)** used administrative data from North Carolina and found that behavioral indicators improve dropout prediction beyond achievement and demographics. The submitted result is conceptually similar in spirit, though with self-report psychosocial measures rather than administrative behavior variables. The paper should engage this parallel more deeply.
- **Bernecker et al. (2018)** provides methodological precedent for asking whether self-report survey data add predictive value over administrative data. The paper would benefit from framing its block-incremental analysis as part of this broader "self-report increment" literature.
- **Williams (2020)** is relevant to the precision–recall/base-rate discussion. The paper's claims about the base-rate ceiling are correct, but Williams already formalizes this relationship. The contribution is applying that reasoning clearly to ninth-grade dropout episodes.
- **Marcolino et al. (2025)** and other LMS-based studies support the general predictive value of engagement, but they use platform traces rather than self-report school belonging. The present paper's distinction between behavioral engagement and felt belonging is useful.

The baseline models are appropriate: logistic regression, random forest, XGBoost, elastic net, and stacking provide a reasonable model battery. A simple majority-class or no-skill baseline would be useful to include explicitly for calibration and decision-curve context, especially given the 11% base rate.

Cited References

- Orooji, M., & Chen, J (2019). *2019 International Conference on Machine Learning and Applications (ICMLA).* , 100–105. https://doi.org/10.1109/ICMLA.2019.00085. [VERIFIED]
- M. Lovelace, Amy L. Reschly, James J. Appleton (2017). *Beyond School Records: The Value of Cognitive and Affective Engagement in Predicting Dropout and On-Time Graduation.* . DOI: 10.5330/1096-2409-21.1.70 [VERIFIED] — Relevance: 10.0/10
- Lucy C. Sorensen (2018). *"Big Data" in Educational Administration: An Application for Predicting School Dropout Risk.* Educational Administration Quarterly. DOI: 10.1177/0013161X18799439 [VERIFIED] — Relevance: 9.0/10
- Samantha L. Bernecker, A. Rosellini, M. Nock, W. Chiu, Peter M. Gutierrez, I. Hwang, T. Joiner, James A. Naifeh, N. Sampson, A. Zaslavsky, M. Stein, R. Ursano, R. Kessler (2018). *Improving risk prediction accuracy for new soldiers in the U.S. Army by adding self-report survey data to administrative data.* BMC Psychiatry. DOI: 10.1186/s12888-018-1656-4 [VERIFIED] — Relevance: 8.0/10
- Christopher K. I. Williams (2020). *The Effect of Class Imbalance on Precision-Recall Curves.* Neural Computation. DOI: 10.1162/neco_a_01362 [VERIFIED] — Relevance: 8.0/10
- Markson Rebelo Marcolino, Thiago Reis Porto, Tiago Thompsen Primo, Rafael Targino, Vinicius F. C. Ramos, Emanuel Marques Queiroga, Roberto Muñoz, C. Cechinel (2025). *Student dropout prediction through machine learning optimization: insights from moodle log data.* Scientific Reports. DOI: 10.1038/s41598-025-93918-1 [VERIFIED] — Relevance: 9.0/10

7. Questions for Authors

1. **Preprocessing leakage:** Were all imputation and scaling steps fitted only on training folds, including within the 5-fold grouped cross-validation? Please provide a pipeline description or pseudocode showing that test information could not enter training through preprocessing.
2. **Outcome validity:** How was X3EVERDROP measured in HSLS:09? Is it based on self-report, parent report, or administrative/transcript data? What evidence supports its validity as a dropout episode rather than permanent non-completion?
3. **Cluster reconstruction:** Can you report sensitivity analyses varying the cluster construction rule? For example, how do the main AUC and CI change using row-level bootstrap, coarser school-profile grouping, or a held-out sample with true school IDs if available?
4. **Threshold metrics inconsistency:** The text reports a balanced-model precision of 0.576 but also states 1,550 false positives for the confusion matrix. These cannot both be true if 341 true positives were identified. Which threshold was used, and what are the correct TP, FP, FN, and TN counts?

8. Suggestions for Improvement

1. **Fix preprocessing reproducibility.** Fit `IterativeImputer` and any scaling/encoding inside cross-validation and the held-out training set. Include a preprint or appendix link with the exact pipeline/code.
2. **Validate cluster reconstruction.** Provide sensitivity analyses showing robustness to pseudo-school misclassification. If possible, compare results with row-level splits and coarsened cluster definitions. State clearly that the cross-school claim is provisional.
3. **Reconcile threshold reporting.** Replace the narrative confusion-matrix paragraph with a table showing threshold, predicted-positive count, TP, FP, FN, TN, precision, recall, F1/F2, and cluster-bootstrap uncertainty. Ensure all numbers are mutually consistent.
4. **Address outcome measurement.** Add a subsection describing X3EVERDROP's source, coding, and limitations. If self-report, consider a sensitivity analysis or at least discussion of possible misclassification.
5. **Report the promised moderation analysis.** Either present the SES moderation results or remove the pre-specified analysis from the Methods. If retained, show whether the incremental value of engagement/belonging varies meaningfully across SES.
6. **Strengthen fairness analysis.** In addition to subgroup AUC, report subgroup precision, recall, false-positive rates, and calibration at a fixed operating threshold. This is more relevant to early-warning fairness than AUC alone.

9. Minor Issues

- The abstract reports AUC 0.73, 95% CI [0.71, 0.75], but the Results section gives [0.708, 0.751] and later a clustered CI [0.706, 0.753]. Clarify which is primary.
- Keywords are duplicated before and after the abstract: "_ dropout prediction, early warning systems, school engagement, school" at the end is likely a formatting artifact.
- The phrase "mean / SHAP / = 0.175" appears with malformed notation; should be mean |SHAP|.
- The confusion-matrix false-positive count appears incompatible with the precision value, as noted above.
- If future-dated references such as 2026 citations are real, they should be marked as in press or online-first; otherwise the reference list may include nonverifiable or fabricated sources.

10. Overall Recommendation

- **Score**: 5.3 / 10
- **Recommendation**: Borderline
- **Confidence**: 4.3 - 6.2
- **Calibration (JEDM, n=10 published anchors)**: P25 5.30, median 5.5. This score is at or above the 25th percentile of published JEDM papers.
- **Measurement noise**: repeat reviews of an unchanged paper have differed by up to 1.9 points overall (measured test-retest MAD), and individual dimension scores by a full point on identical input. Do not read this score as precise to one decimal place.

Disclaimer: This review was generated by LSAR (Learning Science Auto-Reviewer), an AI-assisted review system. It is intended to provide rapid, constructive feedback to help authors improve their work. It should NOT be used as a substitute for human peer review, nor should it be used in any way that violates the reviewing policies of the target venue. AI-generated reviews may contain errors, miss nuances, and have systematic biases (e.g., underweighting novelty, overweighting technical validity). Authors should use this feedback as one input among many. Conference reviewers should NOT use this tool to generate or supplement their official reviews, as this violates the policies of AIED, EDM, L@S, and LAK.