LSAR Review Report

Paper Information

- **Title**: **Do Ninth-Grade Math and Science Identity Predict STEM Degree Non-Completion? Evidence from HSLS:09 and Machine Learning**
- **Target Venue**: EDM (confidence: 1.00)
- **Page Count**: 12 pages
- **Review Date**: 2026-03-25

1. Paper Summary (150-250 words)

This study investigates whether 9th-grade non-cognitive constructs—specifically math and science identity, self-efficacy, school belonging, and engagement—predict STEM degree non-completion beyond traditional academic and socioeconomic controls. Using the nationally representative HSLS:09 dataset, the authors tracked 11,560 students from 9th grade (2009) through postsecondary enrollment (2016). Five machine learning models (Logistic Regression, Random Forest, XGBoost, ElasticNet, StackingEnsemble) were trained with school-aware cross-validation. The StackingEnsemble achieved an AUC of 0.744 (95% CI [0.720, 0.767]), though all models performed within a narrow 5-point AUC band, indicating robust predictive signal. SHAP analysis revealed that science identity (rank 3) and math identity (rank 4) together rivaled math achievement in predictive importance. Subgroup analysis uncovered a 5.1-point AUC gap by gender and an 8.3-point gap by race, suggesting that early warning systems built on these predictors may perform unevenly across demographic groups. The study concludes that non-cognitive constructs measured in 9th grade carry actionable signal for early identification of STEM non-completers, but equity-aware deployment strategies are needed.

2. Relevance Assessment

- **venue_fit**: **Strong** — The paper directly addresses EDM's methodological emphasis by applying ML with rigorous interpretability analysis (SHAP) to a nationally representative educational dataset. The early warning system framing aligns with EDM's applied focus on actionable educational insights.
- **track_fit**: The paper addresses multiple EDM topic areas: learner cognitive and behavior modeling, equity/fairness, early warning systems, and machine learning for educational data. The intersection of identity constructs and STEM outcomes is timely given EDM's growing interest in non-cognitive predictors.
- **scope_concerns**: None. The paper is squarely within EDM's scope.

3. Strengths

- **Methodologically rigorous design**: School-aware cross-validation (GroupShuffleSplit) preventing data leakage from school-level norms, cluster-level bootstrap confidence intervals,

and a five-model comparison battery demonstrate strong EDM methodological standards (Section 3.3–3.4).

- **Substantive equity analysis**: The explicit subgroup performance analysis by gender (5.1-point AUC gap) and race (8.3-point gap) is a genuine contribution to the fairness/equity literature in educational ML. Most prior studies do not interrogate differential model performance across demographic groups (Section 4.3).
- **Clear positioning of the 9th-grade intervention window**: The paper makes a compelling case for why 9th-grade predictors matter—the earliest actionable window before course pathway consolidation—addressing a documented gap in prior EDM literature (Section 1).
- **Robustness evidence**: Sensitivity analysis excluding high-missingness variables, near-identical clustered vs. standard bootstrap CIs, and model performance homogeneity across fundamentally different algorithms (linear, tree-based, ensemble) all strengthen confidence in the findings.
- **Actionable feature importance quantification**: The SHAP analysis provides concrete magnitudes (science identity SHAP = 0.241, math identity = 0.196) that early warning system designers can use for feature prioritization, moving beyond "statistically significant" to "practically important."

4. Weaknesses

- **MAJOR — Ambiguous outcome definition**: The outcome variable X4RFDGMJSTEM is described as "a binary indicator of whether the student had completed (or was on-track to complete) a STEM degree as of the 2016 update panel." The inclusion of "on-track" students introduces outcome heterogeneity: some students in the "non-completion" class may eventually complete a STEM degree, while the 2016 snapshot is not a true terminal outcome. This is not adequately acknowledged as a limitation and could bias effect estimates toward the null.
- **MAJOR — Restricted generalizability from sample selection**: The analytic sample excludes 50.8% of the original HSLS:09 cohort who lack valid STEM degree records (students who never enrolled in postsecondary education). The paper acknowledges this is "not random attrition" but does not sufficiently discuss the implications. Postsecondary enrollees represent a selected subsample that has already survived educational sorting; findings may not generalize to the full population of 9th graders. The discussion of the ICC attenuation (Section 4.4) partially addresses this but could be stronger.
- **MAJOR — Missing mediation/mechanism analysis**: The paper identifies identity as a strong predictor but does not test the theoretical mechanism through which it operates. Is identity mediating the effect of achievement on STEM outcomes? Do identity and achievement interact? Without testing SCCT's proposed mediation pathways, the paper identifies correlational patterns without theoretical deepening. The Related Work invokes SCCT extensively but the Results do not test SCCT predictions.
- **MINOR — StackingEnsemble excluded from SHAP but claimed as best model**: The StackingEnsemble achieved the highest AUC (0.744) but was excluded from SHAP interpretability analysis because it is a composite model. The XGBoost model (AUC = 0.742) was used for SHAP instead, but feature importance rankings could differ between the ensemble and the individual model. This is a minor inconsistency that could be addressed by applying SHAP to all base models.
- **MINOR — Incomplete exploration of HSLS:09 temporal structure**: HSLS:09 contains multiple follow-up waves with repeated measures of identity and self-efficacy. Using only the base-year predictors limits the ability to model growth trajectories (cf. Yeung, 2022, who explicitly modeled developmental trajectories). The paper's argument for early identification is strengthened by showing that base-year constructs are sufficient, but not weakened by the absence of trajectory analysis.

5. Detailed Dimensional Assessment

DimensionScore (1-10)Justification Relevance8Addresses a critical problem (STEM workforce diversity) at the earliest actionable intervention window (9th grade), directly fitting EDM's focus on educational prediction and early warning systems. The emphasis on early identification before course pathway consolidation addresses a documented gap. Novelty6The specific combination of 9th-grade psychological constructs with ML prediction is moderately novel, and the fairness subgroup analysis represents a genuine contribution. However, the core methodology (SHAP, model comparison, bootstrap CIs) is established in the field, and using HSLS:09 with ML is not itself new. Theoretical/ Conceptual Grounding5SCCT is extensively invoked in Related Work but the Results do not test SCCT's proposed mediation pathways, interaction effects, or theoretical predictions. The paper identifies correlational patterns without deepening theoretical understanding—identity's mechanism remains untested. Methodological Rigor9Exceptional methodological standards: school-aware GroupShuffleSplit prevents leakage, cluster-level bootstrap CIs, five-model comparison battery, SHAP interpretability, and robustness checks (sensitivity analysis, CI comparisons). This represents EDM best practices. Empirical Support / Results6Statistical evidence is sound with proper CIs and effect sizes (SHAP values), but two major threats to validity exist: (1) outcome heterogeneity from 'on-track' classification mixing terminal and non-terminal states, and (2) 50.8% sample exclusion of non-postsecondary enrollees creating selected subsample. Significance & Impact7High practical utility for early warning system designers with concrete SHAP magnitudes. Policy implications for 9th-grade intervention are significant, though the restricted sample limits population-level generalizability. Scalable framework for other datasets. Ethics, Fairness & Equity8The explicit demographic subgroup analysis (5.1-point gender AUC gap, 8.3-point race gap) is a genuine contribution to fairness in educational ML. Most prior work ignores differential model performance across groups; this paper explicitly interrogates it. Clarity of Communication7Well-structured paper with effective use of tables and figures, clear section organization, and accessible writing for an interdisciplinary audience. The methodological details are sufficiently thorough without obscuring the main narrative.

6. Comparison with Related Work

The paper appropriately cites Yeung (2022) on developmental trajectories and Dangur-Levy (2022) on self-efficacy mediation of gender gaps, positioning its own contribution as extending the temporal window to 9th grade and applying ML rather than regression-based mediation analysis. Wongvorachan et al. (2024), which also uses HSLS:09 for bias analysis, is relevant but not cited in relation to the fairness methodology. The comparison with Ibrahim (2022) on weighted ensemble frameworks is appropriate but could be deeper regarding the specific handling of subgroup performance.

Missing from reference list: The Alhadabi (2021) SEM-Tree study on science identity and achievement is highly relevant and could strengthen the theoretical discussion of identity-achievement relationships. The Glandorf et al. (2024) LAK paper on temporal and between-group variability in dropout prediction is directly relevant to the subgroup disparity analysis but is not cited.

Cited References

- T. Zhao, Lara Perez-Felkner (2022). *Perceived abilities or academic interests? Longitudinal high school science and mathematics effects on postsecondary STEM outcomes by gender and race*.

International Journal of STEM Education. DOI: 10.1186/s40594-022-00356-w [VERIFIED] — Relevance: 9.0/10

- Tarid Wongvorachan, Okan Bulut, Joyce Xinle Liu, Elisabetta Mazzullo (2024). *A Comparison of Bias Mitigation Techniques for Educational Classification Tasks Using Supervised Machine Learning*. Inf.. DOI: 10.3390/info15060326 [VERIFIED] — Relevance: 8.0/10
- Amal Alhadabi (2021). *Science Interest, Utility, Self-Efficacy, Identity, and Science Achievement Among High School Students: An Application of SEM Tree*. Frontiers in Psychology. DOI: 10.3389/fpsyg.2021.634120 [VERIFIED] — Relevance: 7.0/10
- Dominik Glandorf, Hye Rin Lee, G. Orona, Marina Pumptow, Renzhe Yu, Christian Fischer (2024). *Temporal and Between-Group Variability in College Dropout Prediction*. International Conference on Learning Analytics and Knowledge. DOI: 10.1145/3636555.3636906 [VERIFIED] — Relevance: 7.0/10

7. Questions for Authors

1. **Outcome definition**: You describe the outcome as "completed (or on-track to complete) a STEM degree as of 2016." Can you clarify what proportion of the "non-completion" class consists of on-track students versus confirmed non-completers? If on-track students eventually complete, this introduces outcome misclassification that could bias results.
2. **Mechanism testing**: The paper invokes SCCT extensively but does not test SCCT's mediation predictions. Did you consider testing whether identity mediates the effect of math achievement on STEM outcomes? This would strengthen the theoretical contribution beyond identifying correlates.
3. **Why only base-year predictors?** HSLS:09 includes follow-up waves measuring identity and self-efficacy at multiple time points. Using only base-year predictors supports the early identification argument, but was there a reason not to include trajectory measures (cf. Yeung, 2022) or at least test whether repeated measures improve prediction?
4. **StackingEnsemble interpretability**: Since the StackingEnsemble achieved the highest AUC, why was it excluded from SHAP analysis? Did you consider applying SHAP to the base model predictions within the stacking framework, or training a separate XGBoost model to verify that feature rankings are consistent?

8. Suggestions for Improvement

1. **Test SCCT mediation paths**: Add a mediation analysis (e.g., using the `mediation` R package or structural equation modeling) to test whether math/science identity mediates the relationship between math achievement and STEM non-completion. This would deepen the theoretical contribution beyond correlational description.
2. **Use repeated measures if available**: If HSLS:09 follow-up waves contain identity/self-efficacy measures, test whether trajectory of identity (growth curve) or change scores improve prediction. Even a sensitivity analysis comparing base-year-only to base-year-plus-11th-grade predictors would strengthen the "early window" argument.
3. **Clarify outcome classification**: Provide a table or description of the exact STEM degree coding in HSLS:09 and the proportions of "completed," "on-track," and "non-completers." Discuss how outcome misclassification would affect interpretation.
4. **Apply SHAP to all base models**: Report SHAP importance for Logistic Regression, Random Forest, and ElasticNet in addition to XGBoost to verify that the identity-achievement ranking is consistent across model families, not just XGBoost.
5. **Add interaction terms or subgroup SHAP**: Given the documented AUC disparities, consider computing separate SHAP importance rankings for male vs. female and White vs. Black

subgroups to identify which predictors drive the disparity. This would provide actionable guidance for equity-aware intervention design.

9. Minor Issues

- **Reference formatting**: Several references in the text (Karade, Kubsch, Ibrahim) lack complete bibliographic entries. The Dangur-Levy citation in the text body does not appear in the reference list.
- **Figure numbering**: Figures 8 and 11 both appear to be confusion matrices for the XGBoost model, creating potential confusion. Clarify the distinction.
- **Calibration curve interpretation**: The calibration curve (Figure 10) shows reasonable calibration, but the text does not specify the number of bins or the calibration method used (e.g., isotonic regression, Platt scaling).
- **Abstract specificity**: The abstract states AUC of 0.744 with 95% CI [0.720, 0.767] but does not specify that these are cluster-level bootstrap CIs. Specifying this in the abstract would improve precision.

10. Overall Recommendation

- **Score**: 7.0 / 10
- **Recommendation**: Weak Accept
- **Confidence**: 6.1 - 7.8
