LSAR Review Report

Paper Information

- **Title**: Fairness-Through-Unawareness Does Not Produce Equitable Predictions: Evidence from HSLS:09 on Subgroup Disparities in College Attendance Prediction
- **Target Venue**: EDM (confidence: 1.00)
- **Page Count**: 8 pages
- **Review Date**: 2026-03-23

1. Paper Summary (150-250 words)

This study investigates whether fairness-through-unawareness (FTU)—excluding sensitive demographic attributes from predictive models—produces equitable subgroup predictions for postsecondary college attendance. Using the HSLS:09 dataset (17,335 students, nationally representative), the authors trained six ML models (Logistic Regression, ElasticNet, Random Forest, XGBoost, MLP, StackingEnsemble) to predict college attendance using 15 features spanning academic achievement, motivational constructs, and demographic background. School-aware evaluation used GroupShuffleSplit with 139 pseudo-school clusters held out for testing. SHAP analysis identified baseline math scores (SHAP=1.61) as dominant, with educational expectations (SHAP=0.52) and SES (SHAP=0.45) as next most important. Despite excluding race, sex, and SES from models, racial subgroup AUCs ranged from 0.465 to 0.690 (gap=0.225), far exceeding the 5% disparity threshold. The core contribution is demonstrating that FTU fails as a fairness intervention because predictive signal merely migrates to correlated proxies. The paper recommends routine subgroup fairness auditing over variable exclusion as a fairness heuristic. A critical limitation is that the full-versus-reduced model comparison was not executed, limiting the direct empirical evidence for the FTU claim.

2. Relevance Assessment

- **venue_fit**: Strong — The paper directly addresses EDM's core concerns: algorithmic fairness, equity in educational prediction, and methodological standards for bias auditing. The topic of fairness-through-unawareness is underexplored in EDM specifically.
- **track_fit**: Addresses multiple EDM topics: Equity, privacy, transparency, fairness; Learner performance modeling; Learning analytics; Human factors, transparency, explainability.
- **scope_concerns**: None. The paper is well within scope for EDM.

3. Strengths

- **Important and timely research question**: The paper addresses a critical gap—whether fairness-through-unawareness (FTU) actually produces equitable predictions—which is widely deployed in practice but underexplored empirically. This directly addresses EDM's growing interest in fairness and transparency.
- **Strong methodological foundation**: The study uses a large, nationally representative longitudinal dataset (HSLS:09) with temporal ordering (9th-grade predictors → 7-year outcomes), applies school-aware evaluation with GroupShuffleSplit maintaining complete school separation, uses clustered bootstrap CIs (1,000 resamples) to account for within-school correlation, and employs SHAP for model interpretability.
- **Comprehensive model comparison**: Training six diverse classifiers (linear models, tree-based, neural network, stacking ensemble) with proper hyperparameter tuning via RandomizedSearchCV provides a thorough evaluation of modeling approaches.
- **Subgroup fairness analysis as primary evaluation**: Computing AUC separately for race (8 categories), sex (2 categories), and SES (planned but not executed) is methodologically appropriate and necessary for fairness auditing. The finding that racial AUC gaps exceed 5% threshold is compelling evidence.
- **Policy-relevant findings with clear implications**: The paper's recommendation that stakeholders invest in fairness-aware model development rather than variable exclusion provides actionable guidance for educational AI practitioners.

4. Weaknesses

- **FATAL: Core FTU comparison not executed** — The paper's central empirical claim—that fairness-through-unawareness fails—requires direct comparison of full versus reduced models on subgroup fairness metrics. The paper explicitly states that "metric_change was null" and the reduced model AUC and subgroup fairness comparison were not generated. The SHAP evidence (SES as proxy) is indirect and insufficient. This is a fundamental methodological gap that undermines the paper's core contribution.
- **FATAL: SES quintile subgroup analysis not generated** — The paper explicitly lists X1SESQ5 in research_spec.subgroup_analyses but the results are absent. Since SES is the paper's primary example of a protected attribute proxy (third most important feature), its absence from the subgroup fairness results severely limits the completeness of the fairness analysis.
- **MAJOR: Survey weights not incorporated** — HSLS:09 uses complex stratified multi-stage probability sampling. The paper acknowledges that "the machine learning models were trained and evaluated without survey weights," meaning reported metrics may not generalize to the national population. This is a significant methodological limitation for a study using nationally representative data.

- **MAJOR: School cluster reconstruction is imperfect** — The paper reconstructed 692 pseudo-school clusters versus 944 expected schools (26.7% discrepancy). While clustered CIs were computed, the imperfect reconstruction affects ICC estimation and may not fully account for school-level variance. The paper acknowledges this limitation but does not adequately quantify its impact on results.
- **MAJOR: Negative direction for math achievement unexplained** — The finding that higher baseline math scores are associated with *lower* predicted college attendance probability (Section 4.3) contradicts established literature and is explained only superficially. This counterintuitive result warrants deeper investigation and theoretical justification, as it may indicate model misspecification or data issues rather than a genuine educational phenomenon.

5. Detailed Dimensional Assessment

Dimension	Score (1-10)	Justification
Relevance	8	The paper addresses a critical and timely gap in educational data mining—evaluating whether fairness-through-unawareness actually delivers equitable predictions. This directly aligns with EDM's growing focus on fairness and transparency in predictive systems, making it highly relevant for practitioners deploying models in high-stakes educational contexts.
Novelty	5	While the research question about FTU effectiveness is valuable, the core empirical contribution is undermined by not executing the central comparison (full vs. reduced model). The methodological approach follows standard fairness auditing practices rather than introducing novel methods, limiting the incremental contribution to the field.
Theoretical/Conceptual Grounding	7	The paper is well-grounded in fairness theory, appropriately referencing fairness-through-unawareness as a deployed but underexplored approach. Literature coverage includes relevant educational equity work and SHAP-based interpretability. The theoretical framing of proxy discrimination is sound, though the incomplete execution weakens the conceptual contribution.
Methodological Rigor	4	Despite several strong elements (clustered bootstrap CIs, school-aware GroupShuffleSplit, proper hyperparameter tuning), there are fatal flaws: the core FTU comparison was never generated, SES subgroup analysis is missing from results, survey weights were ignored, and school cluster reconstruction has a 26.7% discrepancy. These methodological gaps fundamentally undermine the paper's central claims.

Dimension	Score (1-10)	Justification
Empirical Support / Results	4	The finding of racial AUC gaps exceeding 5% is compelling evidence, but the absence of the primary comparison (full vs. reduced model) and missing SES analysis severely limits empirical support. The unexplained negative direction of math achievement on college attendance is a red flag that suggests potential model misspecification rather than genuine educational phenomenon.
Significance & Impact	6	The policy recommendation (invest in fairness-aware modeling rather than variable exclusion) is actionable and meaningful for practice. However, the methodological limitations reduce the paper's potential influence on policy and practice, as the core empirical claim remains unproven. The findings are suggestive but not conclusive.
Ethics, Fairness & Equity	7	The paper's explicit focus on subgroup fairness analysis across race, sex, and SES demonstrates strong ethical awareness. The study appropriately examines demographic disparities and acknowledges methodological limitations. However, the incomplete execution of the primary fairness analysis (FTU comparison) represents a missed opportunity for rigorous equity evaluation.
Clarity of Communication	7	The paper is well-written with clear organization and effective use of tables and figures for subgroup fairness results. The abstract appropriately summarizes the contribution. The writing is accessible for an interdisciplinary audience, though the critical methodological limitations (non-execution of core comparison) could be more prominently flagged.

6. Comparison with Related Work

Gandara et al. (2023) — *"Inside the Black Box: Detecting and Mitigating Algorithmic Bias Across Racialized Groups in College Student-Success Prediction."* AERA Open. DOI: 10.1177/23328584241258741 — This is the most directly relevant prior work, auditing bias in models that *include* protected attributes. The present paper extends this by testing whether *excluding* protected attributes addresses disparities, which is a logical next step. However, Gandara et al. also use ELS:02 rather than HSLS:09, and this paper does not cite or compare to their bias mitigation findings.

Kearns et al. (2018) — *"An Empirical Study of Rich Subgroup Fairness for Machine Learning."* FAT* — Provides the theoretical foundation for why subgroup fairness analysis is necessary (marginal fairness constraints may not ensure fairness on complex subgroups). The present paper's finding that racial AUC varies substantially across groups (0.465–0.690) aligns with Kearns et al.'s concern about rich subgroup fairness.

Ingels et al. (2011) — HSLS:09 documentation is properly cited as the data source, establishing methodological credibility for the dataset choice.

Baker et al. (2023) — *"Using Demographic Data as Predictor Variables: a Questionable Choice."* — This EDM-adjacent work is relevant but not cited. The Baker et al. paper directly addresses whether demographic variables should be included in educational prediction models, which is closely related to the FTU question.

Missing citations: The paper does not cite the fairness-aware policy gradient paper by Solheim & De Santis ([1] in references) despite it being listed in references. Additionally, the paper could benefit from citing Ashurst & Weller (2023) on fairness without demographic data, which is closely related to the FTU question.

Baselines: The paper uses standard ML models as baselines (Logistic Regression, Random Forest, XGBoost) which is appropriate for EDM. However, it does not compare against prior college attendance prediction models, making it difficult to assess whether the AUC=0.813 performance is competitive with state-of-the-art.

Cited References

- Denisa G'andara, Hadis Anahideh, Matthew P. Ison, Anuja Tayal (2023). *Inside the Black Box: Detecting and Mitigating Algorithmic Bias Across Racialized Groups in College Student-Success Prediction*. AERA Open. DOI: 10.1177/23328584241258741 [VERIFIED] — Relevance: 9.0/10

- Michael Kearns, Seth Neel, Aaron Roth, Zhiwei Steven Wu (2018). *An Empirical Study of Rich Subgroup Fairness for Machine Learning*. FAT. DOI: 10.1145/3287560.3287592 [VERIFIED] — Relevance: 8.0/10

- S. Ingels, B. Dalton, T. Holder, Erich Lauff, Laura J. Burns (2011). *The High School Longitudinal Study of 2009 (HSLS:09): A First Look at Fall 2009 Ninth-Graders. NCES 2011-327.. .* https://www.semanticscholar.org/paper/74d9ca7a4b7924736525a9118e402d3bd0b0be83 [VERIFIED] — Relevance: 9.0/10

- R. S. Baker, Lief Esbenshade, Jonathan Vitale, Shamya Karumbaiah (2023). *Using Demographic Data as Predictor Variables: a Questionable Choice. .* https://www.semanticscholar.org/paper/7dccc97532f1691892c4a017e01e3bbf02e166a5 [VERIFIED] — Relevance: 8.0/10

- Carolyn Ashurst, Adrian Weller (2023). *Fairness Without Demographic Data: A Survey of Approaches*. Conference on Equity and Access in Algorithms, Mechanisms, and Optimization. DOI: 10.1145/3617694.3623234 [VERIFIED] — Relevance: 8.0/10

7. Questions for Authors

1. **The reduced model comparison was not executed** — The paper states "metric_change was null" for the sensitivity analysis. Can you clarify: (a) Was this a technical failure in the EDM-ARS

pipeline? (b) What would the reduced model AUC and subgroup AUCs have been? Without this comparison, the central claim that FTU fails relies only on indirect SHAP evidence.

2. **The SES quintile subgroup analysis is listed in research_spec but missing from results** — Why was this analysis not generated? Given that SES is identified as the primary protected attribute proxy (third most important feature), its absence from subgroup fairness results is a critical gap.
3. **The negative direction for math achievement (SHAP=-1.61)** contradicts established literature showing math achievement positively predicts college attendance. Can you provide additional analysis or theoretical justification for this counterintuitive finding? Is it possible this reflects model misspecification, data quality issues, or a genuine phenomenon specific to HSLS:09?
4. **How do the AUC=0.813 results compare to prior college attendance prediction models** in the literature? Without a comparison to existing benchmarks, it is difficult to assess whether this performance represents state-of-the-art or merely adequate prediction.

8. Suggestions for Improvement

1. **Execute the full-versus-reduced model comparison**: This is the paper's most critical gap. Retrain models excluding X1SES, X1RACE, and X1SEX, and report reduced model AUC alongside full model AUC. Compute subgroup AUCs for reduced models to directly demonstrate that disparities persist. The SHAP evidence alone is insufficient to support the FTU claim.
2. **Generate the SES quintile subgroup analysis**: Given that SES is the paper's primary example of a proxy variable, its absence from subgroup fairness results is a critical omission. This analysis should be prioritized to provide a complete fairness picture.
3. **Incorporate survey weights or explain the impact of not using them**: If survey weights cannot be incorporated into the scikit-learn pipeline, consider using weighted bootstrap procedures or switching to R's survey package for key analyses. Alternatively, conduct a sensitivity analysis comparing weighted versus unweighted estimates to assess the magnitude of potential bias.
4. **Investigate the negative math achievement direction**: Conduct additional analysis to understand why higher baseline math scores predict lower college attendance. Examine whether this reflects differential institution type (elite four-year vs. community college), delayed enrollment, SES confounding, or data quality issues. This finding needs theoretical justification before publication.
5. **Add model comparison to prior benchmarks**: Include a brief comparison with published college attendance prediction models (e.g., using different datasets or methods) to contextualize the AUC=0.813 performance.

9. Minor Issues

- **Reference [1] is cited in text but appears incomplete**: "Solheim and Marco De Santis" should be "Solheim and De Santis" (initials omitted).
- **Figure labels are generic**: Figures are labeled as "Figure 1", "Figure 2", etc. Descriptive titles (e.g., "Figure 2: ROC curves for all six models") appear only in the caption text.

- **The calibration curve and confusion matrix (Figures 7-8)** are reported but not discussed in the Results section. These could support or complicate the interpretation of model reliability for fairness-sensitive deployment.
- **Small subgroup sample sizes**: American Indian/Alaska Native (n=13) and Hispanic/no race specified (n=60) have unstable AUC estimates. Consider reporting confidence

10. Overall Recommendation

- **Score**: 5.8 / 10
- **Recommendation**: Borderline
- **Confidence**: 4.8 - 6.8

> **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.