When Algorithms Mirror Inequality: Structural Encoding of Fairness Disparities in Early Childhood Risk Prediction

Kismat Raj Vishwakarma[1], Kumkum Basu[2], and Abhinaba Basu[1]

Abstract

Machine learning models are increasingly used to identify at-risk students, yet fairness remains underexplored in longitudinal educational contexts. We conducted a comprehensive fairness audit of seven algorithms predicting 5th-grade academic risk from early childhood data using the ECLS-K:2011 (complete-case analytic sample $N = 9,104$; 70/30 train-test split; multiple imputation sensitivity analyses on the full sample of $N = 18,151$). All seven algorithms—including three gradient boosting methods—converged on nearly identical performance (AUC range = 0.011), indicating that fairness failures are properties of the prediction task, not artifacts of model architecture. Bootstrap confidence intervals confirmed significant true positive rate disparities: Hispanic students were identified at 2.5 times the rate of White students. Black students experienced 3.35 times higher calibration error. Exploratory intersectional analysis produced a preliminary pattern consistent with the model functioning as a poverty detector, with at-risk high-SES Black students potentially undetected (TPR = 0%, $N = 42$, $\approx$6 at-risk cases); however, this estimate is not statistically significant ($p = 0.41$) and requires replication with larger samples. SES-removal ablation confirmed that this pattern is mediated through cognitive scores (ΔAUC = 0.003 when SES is removed), and in-processing fairness constraints (Fairlearn) reduced TPR disparities by 46–56% but at substantial accuracy cost (AUC penalty = 0.21–0.23). A policy allocation simulation showed that the TPR gap translates into a 24-percentage-point difference in which students receive support, and that equalization redistributes detections across groups rather than improving outcomes for all. Counterfactual decomposition revealed that differences in baseline feature distributions are statistically associated with more than 100% of the Hispanic–White TPR gap (152% averaged across reading and math), with the model partially compressing rather than amplifying distributional divergence. These findings demonstrate that, within the available feature space and standard predictive paradigm, fairness disparities in early childhood prediction persist across algorithms, mitigation strategies, and outcome domains—and are not resolvable through algorithmic selection alone, pointing toward institutional rather than merely technical intervention.

Keywords

Algorithmic fairness, machine learning, educational prediction, calibration fairness, intersectional fairness, temporal generalization, ECLS-K:2011

Introduction

Thousands of school districts now use machine learning to flag students at risk of academic failure, and the practice is accelerating (Baker and Inventado 2014). The premise is intuitive: identify struggling students early, intervene before they fall behind, close achievement gaps. Yet a key assumption underlying these early warning systems (EWS) has received limited systematic study in longitudinal contexts—that the algorithms work equitably across the demographic groups they are meant to serve.

There are reasons to doubt this assumption. Algorithmic fairness research has shown that predictive systems can systematically advantage some groups while disadvantaging

[1]National Institute of Electronics & Information Technology (NIELIT), India
[2]Uttar Pradesh Basic Education Department, India

Corresponding author:
Kismat Raj Vishwakarma, National Institute of Electronics & Information Technology (NIELIT), India.
Email: krajdev@gmail.com

others (Mehrabi et al. 2021), and that fairness itself is multidimensional: a model may satisfy one equity criterion while violating others (Chouldechova 2017; Barocas et al. 2019). In educational settings, the stakes are high. An algorithm that under-identifies at-risk minority students denies them timely support; one that over-flags them wastes resources and may stigmatize. Protected attributes—characteristics such as race, socioeconomic status, and language background that are shielded from discriminatory treatment due to histories of systemic disadvantage—are central to these concerns (Barocas et al. 2019). Although a growing literature has begun examining fairness in educational prediction (Kizilcec and Lee 2022; Gardner et al. 2019), fewer studies have examined whether EWS perform equitably across multiple fairness dimensions simultaneously, in longitudinal contexts spanning several years, or using the modern toolkit of calibration analysis, intersectional auditing, and uncertainty quantification.

This study fills that gap through a comprehensive fairness audit. Using the nationally representative ECLS-K:2011 longitudinal study, which followed approximately 18,000 children from kindergarten through 5th grade, we ask whether seven machine learning algorithms—spanning classical regularized models and state-of-the-art gradient boosting—can predict 5th-grade academic risk from early childhood data, and whether they do so equitably. We examine predictive accuracy and its convergence across algorithms; group-level fairness with bootstrap uncertainty quantification; calibration equity and intersectional (race × SES) disparities; temporal generalization across four developmental windows from kindergarten through 3rd grade; sensitivity to the choice of at-risk threshold; and the consequences and ethical implications of both post-hoc and in-processing bias mitigation. We complement these diagnostic analyses with a policy allocation simulation that translates abstract fairness metrics into concrete student-level consequences, and with domain validation across reading and math outcomes.

Conceptual Framework

This paper contributes to an emerging literature at the intersection of *algorithmic governance* and *educational inequality*. Research on algorithmic governance examines how automated decision systems reshape the distribution of public resources and the exercise of institutional authority (Barocas et al. 2019). A central insight from this literature is that fairness is not a property of algorithms in isolation but of *sociotechnical systems*—the entangled configurations of code, data, institutions, and affected populations within

which algorithms operate (Selbst et al. 2019). Selbst et al. (2019) identified five "abstraction traps" that arise when technologists treat fairness as a purely computational problem, including the *framing trap* (defining fairness without reference to the broader social system) and the *ripple effect trap* (ignoring how a tool changes the system it enters). Educational early warning systems are particularly susceptible to these traps because they sit at the intersection of measurement systems that reflect inequality and governance structures that allocate resources based on those measurements.

Research on educational inequality, meanwhile, has documented that socioeconomic disparities in cognitive skills are already substantial at school entry and compound over time through cumulative disadvantage (Reardon 2011; Heckman 2006). Eubanks (2018) has shown how automated decision systems in public services—from welfare eligibility to child protective services—tend to target the poor, creating what she terms a "digital poorhouse" in which surveillance and classification intensify for those already disadvantaged. The parallels to educational prediction are direct: an EWS trained on data produced by stratified institutions may function less as a neutral diagnostic tool than as a mechanism that formalizes existing patterns of institutional attention along socioeconomic lines.

Our study connects these literatures by asking what happens when predictive algorithms trained on data generated by stratified educational institutions are used to allocate resources to children. Our central argument is that, within the standard predictive paradigm, fairness disparities are not technical defects to be corrected but reflections of institutional stratification encoded in the data—a phenomenon we term *structural encoding*. This positions the paper not as an evaluation of algorithms but as an examination of *how algorithms interact with social structure*—and what that interaction implies for governance. Specifically, we use counterfactual decomposition to show that the model's differential performance across racial groups can be traced to differences in the input feature distributions that the model inherits from a stratified data-generating process, with the algorithm itself partially *compressing* rather than amplifying the resulting disparity.

Contributions

Our goal is not to introduce novel algorithms but to conduct a rigorous *fairness audit* demonstrating that fairness failures are robust across seven algorithmic approaches, two mitigation strategies, and two outcome domains. By showing that, within the available feature space and outcome

definition, the source of inequity is located in the data rather than in model architecture—that is, in the data-generating processes that reflect historical and ongoing institutional inequities—we shift the policy conversation from "which model?" to "what data and what institutional practices produce these disparities?"

This study makes several contributions to the literature on algorithmic fairness in education:

- We provide one of the first multi-dimensional fairness audits of longitudinal educational prediction models using nationally representative data, examining group fairness, calibration fairness, and intersectional fairness simultaneously (though not individual or counterfactual fairness, which represent complementary perspectives).
- We demonstrate that the convergence of seven ML algorithms—including three recent gradient boosting variants (LightGBM, CatBoost, HistGradientBoosting)—on nearly identical performance (AUC range of 0.011) constitutes evidence that the source of unfairness is not algorithmic but structural, consistent with recent tabular ML benchmarks (Grinsztajn et al. 2022).
- We employ SHAP-based explainability to examine whether predictive features operate differently across racial groups, providing transparency into the model's decision-making process (Lundberg and Lee 2017).
- We introduce calibration fairness and intersectional (race $\times$ SES) analysis to educational prediction, revealing suggestive patterns of under-identification of high-SES minority students that warrant replication with larger samples.
- We demonstrate that additional longitudinal data from kindergarten through 3rd grade improves accuracy but fails to resolve—and in some cases worsens—fairness disparities. While this outcome is consistent with theoretical impossibility results (Chouldechova 2017), it has not previously been documented empirically in longitudinal educational prediction, where the practical assumption that "more data helps" remains widespread.
- We demonstrate the fragility of fairness assessments by showing that compliance with equal opportunity criteria depends critically on the choice of at-risk threshold.
- We conduct missing data sensitivity analyses (multiple imputation and inverse probability weighting) confirming that the pattern of differential performance persists after correcting for attrition, with substantially higher absolute detection rates in the full imputed sample.
- We provide empirical evidence that the "poverty detector" pattern is mediated through cognitive scores rather than SES directly (SES-removal ablation: ΔAUC = 0.003), and that in-processing fairness constraints reduce disparities only at substantial cost to overall accuracy (AUC penalty of 0.21–0.23 points).
- We translate fairness metrics into concrete policy consequences through an allocation simulation, showing that the TPR gap between White and Hispanic students corresponds to a 24-percentage-point difference in who receives support, and that equalization redistributes detections across groups rather than improving outcomes for all.
- We validate the core findings across both reading and math outcomes, demonstrating that algorithmic convergence, TPR disparities, and calibration unfairness are cross-domain phenomena rather than outcome-specific artifacts.
- We conduct a counterfactual decomposition of the Hispanic–White TPR gap, demonstrating that differences in baseline feature distributions are statistically associated with more than 100% of the observed disparity—converting the claim of "persistent disparities" into an empirically decomposed explanation of mechanism.

Background and Related Work

Early Warning Systems in Education

Early warning systems (EWS) use student data to identify individuals at risk of negative academic outcomes. Traditional EWS relied on simple indicators such as attendance, behavior, and course performance (the "ABC" indicators). Modern approaches increasingly incorporate machine learning algorithms capable of processing larger feature sets and capturing nonlinear relationships (Lakkaraju et al. 2015).

Research has demonstrated that ML-based EWS can achieve reasonable predictive accuracy, with AUC values typically ranging from 0.70 to 0.85 depending on the outcome and available features (Aguiar et al. 2015). Recent benchmarking studies have found that tree-based ensemble methods—including gradient boosting variants such as XGBoost, LightGBM (Ke et al. 2017), and CatBoost (Prokhorenkova et al. 2018)—remain competitive with or superior to deep learning on structured tabular

data (Grinsztajn et al. 2022). However, fewer studies have examined whether these systems perform equitably across student subgroups.

Algorithmic Fairness

The machine learning fairness literature has developed numerous formal definitions, broadly organized into three families (Verma and Rubin 2018). *Group fairness* criteria—including demographic parity (equal positive prediction rates), equal opportunity (equal true positive rates), and equalized odds (equal TPR and FPR)—require that some statistical measure be equalized across protected groups. *Individual fairness* requires that similar individuals receive similar predictions regardless of group membership (Dwork et al. 2012). *Counterfactual fairness* asks whether predictions would change if a person's protected attribute were different. Crucially, Chouldechova (2017) and Kleinberg et al. (2016) proved that these criteria are mathematically incompatible when base rates differ across groups—a condition that holds in virtually every educational setting. The choice among fairness criteria is therefore not technical but normative, reflecting value judgments about which harms matter most. In the context of early warning systems, we argue that *equal opportunity* (equal TPR)—ensuring that at-risk students of all backgrounds have the same probability of being identified—is the most policy-relevant criterion, because the primary harm of EWS failure is under-identification of students who need support. However, we report all three group-level criteria, as well as calibration fairness, precisely because no single metric captures the full landscape of potential harms: a system that identifies students equitably may still produce unreliable confidence estimates that mislead the practitioners who act on them.

Two extensions of group fairness are particularly relevant to educational prediction. *Calibration fairness* examines whether predicted probabilities are equally reliable across groups: a model well-calibrated for one group but miscalibrated for another produces confidence levels that systematically mislead practitioners for certain populations (Pleiss et al. 2017). Expected Calibration Error (ECE) and Maximum Calibration Error (MCE) quantify this distortion. *Intersectional fairness* recognizes that examining single protected attributes in isolation can miss compounding disadvantages at the intersection of multiple identities (Crenshaw 1989; Buolamwini and Gebru 2018). A model that appears fair by race and fair by SES may still exhibit severe disparities for specific race-by-SES subgroups (Kearns et al. 2018).

Explainability and Fairness

Model interpretability plays a critical role in fairness auditing. SHAP (SHapley Additive exPlanations) values provide a unified framework for feature attribution, connecting game-theoretic concepts to local model explanations (Lundberg and Lee 2017). By computing SHAP values separately for each demographic group, analysts can detect whether the model relies on different features—or the same features with different magnitudes—when making predictions for different populations. Permutation importance with bootstrap confidence intervals provides a complementary, model-agnostic measure of feature relevance (Molnar 2020). When SHAP and permutation importance rankings agree, this strengthens confidence in the identified predictive mechanisms.

Fairness in Educational AI

A growing body of work has examined fairness in educational technology. Kizilcec and Lee (2022) found that dropout prediction models in MOOCs exhibited significant performance disparities across countries. Yu et al. (2020) demonstrated that automated essay scoring systems showed bias against non-native English speakers. Gardner et al. (2019) examined fairness in course outcome prediction and found persistent gaps across demographic groups.

Despite this emerging literature, few studies have examined fairness in early childhood prediction contexts, in systems that make predictions across extended time horizons, or using the full suite of modern fairness metrics (calibration, intersectionality, uncertainty quantification). Our study addresses these gaps.

Algorithmic Governance and Classification

A distinct body of scholarship examines how classification systems—including predictive algorithms—shape institutional decision-making and its consequences. Bowker and Star (2000) argued that classification is never neutral: the categories through which institutions organize people carry embedded assumptions, create invisible exclusions, and render some populations legible to the state while obscuring others. Predictive risk scores are a form of classification: they sort students into "at-risk" and "not at-risk" categories, and this sorting has material consequences for who receives institutional attention and who does not.

Recent work has applied these insights to algorithmic decision-making in public institutions. Green (2022) has argued that proposals to mitigate algorithmic harms through technical adjustments—better models, fairer constraints, more transparent features—often fail because they treat

the algorithm as the locus of the problem rather than the institutional context in which it operates. In education specifically, Williamson (2017) has documented how the expansion of digital data infrastructures has reshaped school governance, enabling new forms of accountability and resource allocation that are mediated by statistical models rather than professional judgment alone. These analyses suggest that fairness audits of educational algorithms should attend not only to model performance metrics but also to the governance regimes within which those models are deployed—and the institutional dynamics they may reinforce or disrupt. Our study takes this perspective seriously: we treat the fairness audit not as an endpoint but as a window into the structural properties of the educational system that produced the data.

Data and Methods

Data Source

We used data from the Early Childhood Longitudinal Study, Kindergarten Class of 2010-11 (ECLS-K:2011), conducted by the National Center for Education Statistics (NCES). The ECLS-K:2011 is a nationally representative longitudinal study that followed approximately 18,000 children from kindergarten entry in fall 2010 through spring of 5th grade in 2016.

Data were collected across nine waves:

- Kindergarten: Fall 2010 (Wave 1), Spring 2011 (Wave 2)
- 1st Grade: Fall 2011 (Wave 3), Spring 2012 (Wave 4)
- 2nd Grade: Fall 2012 (Wave 5), Spring 2013 (Wave 6)
- 3rd Grade: Spring 2014 (Wave 7)
- 4th Grade: Spring 2015 (Wave 8)
- 5th Grade: Spring 2016 (Wave 9)

We used the public-use data file, which includes 18,174 children. After applying inclusion criteria (valid outcome data and at least some baseline predictors), our analytic sample comprised 18,151 children. Complete-case analysis for modeling yielded 9,104 children with data on all predictors and outcomes.

Measures

Outcome Variable The primary outcome was **academic risk in 5th grade**, operationalized as scoring below the 25th percentile on the reading theta score (X9RTHETA) from the spring 2016 assessment. The reading assessment measured skills including basic reading, vocabulary, and reading comprehension. The theta score is an item response theory

(IRT)-based ability estimate that allows for longitudinal comparisons. Among children with valid 5th-grade reading scores ($N = 11,427$), 25.0% were classified as at-risk based on this threshold.

As a secondary analysis, we also examined the math outcome (X9MTHETA) to assess domain-specificity of fairness findings.

Predictor Variables We included predictors from kindergarten through 2nd grade across four domains:

Baseline Cognitive Scores:

- Reading theta scores: Fall K (X1RTHETK), Spring K (X2RTHETK)
- Math theta scores: Fall K (X1MTHETK), Spring K (X2MTHETK)

Executive Function:

- Dimensional Change Card Sort score, Spring 2013 (X6DCCSSCR)

Approaches to Learning:

- Teacher-reported approaches to learning: Fall K (X1TCHAPP), Spring K (X2TCHAPP), Spring 1st grade (X4TCHAPP)

Demographic Characteristics:

- Child sex (X_CHSEX_R)
- Race/ethnicity (X_RACETH_R)
- Socioeconomic status quintile (X1SESQ5)
- Home language (X12LANGST)

Protected Attributes For fairness analysis, we focused on **race/ethnicity** as the primary protected attribute. The ECLS-K:2011 includes seven race/ethnicity categories; we collapsed these into five groups: White (reference), Black, Hispanic, Asian, and Other (including Native Hawaiian/Pacific Islander, American Indian/Alaska Native, and multiracial). For intersectional analysis, we crossed race/ethnicity with SES quintile.

Machine Learning Models

We trained seven classification algorithms spanning classical and state-of-the-art approaches:

Classical Models

1. **Logistic Regression:** L2-regularized logistic regression with regularization strength selected via cross-validation from $C \in \{0.01, 0.1, 1.0, 10.0\}$.

2. **Elastic Net:** Logistic regression with elastic net penalty, tuning both regularization strength $\alpha \in \{0.001, 0.01, 0.1, 1.0\}$ and L1 ratio $\in \{0.2, 0.5, 0.8\}$.

3. **Random Forest:** Ensemble of decision trees with hyperparameters: $n_{estimators} \in \{100, 200\}$, $max_{depth} \in \{5, 10, 15\}, min_{samples\_leaf} \in \{5, 10\}$.

4. **XGBoost:** Gradient boosted trees (Chen and Guestrin 2016) with $n_{estimators} \in \{100, 200\}$, $max_{depth} \in \{3, 5, 7\}, learning_{rate} \in \{0.01, 0.1\}$.

State-of-the-Art Gradient Boosting Methods

5. **LightGBM:** Gradient boosting with leaf-wise tree growth and histogram-based binning (Ke et al. 2017). Hyperparameters: $n_{estimators} \in \{100, 200, 300\}$, $max_{depth} \in \{3, 5, 7, -1\}$, $learning_{rate} \in \{0.01, 0.05, 0.1\}$, $num_{leaves} \in \{31, 63, 127\}$.

6. **CatBoost:** Gradient boosting with ordered boosting to reduce prediction shift and native handling of categorical features (Prokhorenkova et al. 2018). Hyperparameters: $iterations \in \{100, 200, 300\}$, $depth \in \{4, 6, 8\}, learning_{rate} \in \{0.01, 0.05, 0.1\}$.

7. **HistGradientBoosting:** Scikit-learn's histogram-based gradient boosting, inspired by Light-GBM, with native missing value support and early stopping. Hyperparameters: $max_{iter} \in \{100, 200, 300\}$, $max_{depth} \in \{3, 5, 7\}$, $learning_{rate} \in \{0.01, 0.05, 0.1\}$.

Recent benchmarks suggest that tree-based ensemble methods remain competitive with or superior to deep learning on structured tabular data (Grinsztajn et al. 2022). We include three recent gradient boosting variants to test whether state-of-the-art methods improve upon classical approaches for educational prediction.

All models were trained using 5-fold stratified cross-validation for hyperparameter selection, with random seed fixed at 42 for reproducibility. The data were split 70% training, 30% test. The held-out test set was used exclusively for final evaluation and was not accessed during hyperparameter tuning, early stopping, or threshold optimization. For gradient boosting models that support early stopping (XGBoost, LightGBM, CatBoost), early stopping was determined using validation folds within the training data only.

Evaluation Metrics

Predictive Performance We evaluated predictive performance using AUC-ROC, accuracy, precision (PPV), recall (sensitivity/TPR), F1 score, and Brier score.

Fairness Metrics For each demographic group g, we computed:

- **True Positive Rate (TPR):** $TPR_g = \frac{TP_g}{TP_g + FN_g}$
- **False Positive Rate (FPR):** $FPR_g = \frac{FP_g}{FP_g + TN_g}$
- **Positive Predictive Value (PPV):** $PPV_g = \frac{TP_g}{TP_g + FP_g}$

We assessed three fairness criteria:

Equal Opportunity: Satisfied if TPR ratios between groups exceed 0.80. We adopt the four-fifths (80%) threshold from employment discrimination law (Barocas et al. 2019) as a widely recognized benchmark for disparate impact. While originating in the employment context, this threshold provides a concrete, established standard for flagging disparities that warrant further scrutiny; no analogous threshold has been codified specifically for educational prediction, and we use it as a practical reference point rather than a legal standard.

Equalized Odds: Satisfied if both TPR and FPR ratios exceed 0.80.

Statistical Parity: Satisfied if positive rate ratios exceed 0.80.

Bootstrap Confidence Intervals We computed 95% bootstrap confidence intervals for all group-level fairness metrics using 500 bootstrap iterations (Efron and Tibshirani 1993), enabling formal statistical assessment of inter-group differences. Non-overlapping confidence intervals between groups indicate statistically significant disparities at approximately the $\alpha = 0.05$ level.

Calibration Fairness We assessed calibration equity using Expected Calibration Error (ECE) and Maximum Calibration Error (MCE) computed separately for each demographic group. ECE measures the average absolute difference between predicted probabilities and observed frequencies across probability bins:

$$ECE = \sum_{b=1}^{B} \frac{n_b}{N} |acc(b) - conf(b)| \qquad (1)$$

where $acc(b)$ is the observed accuracy in bin b and $conf(b)$ is the mean predicted probability. We computed ECE ratios relative to the White reference group to quantify differential calibration. In practical terms, calibration fairness matters because practitioners rely on predicted probabilities—not just binary classifications—to prioritize intervention resources. If a model states that two students each have a 30% risk of academic failure, a counselor may allocate resources equally; but if the model is poorly calibrated for one demographic group, these probability

estimates carry unequal information content, and the stated 30% risk may correspond to a substantially different empirical risk rate for different populations.

Intersectional Fairness We examined fairness at the intersection of race/ethnicity and SES quintile, computing TPR, FPR, PPV, and accuracy for each race-by-SES subgroup with a minimum group size of 20. This analysis follows recommendations for rich subgroup fairness auditing (Kearns et al. 2018) and intersectional analysis (Buolamwini and Gebru 2018).

Explainability Analysis

We employed multiple explainability methods to understand predictive mechanisms:

SHAP values: We applied TreeExplainer (Lundberg and Lee 2017) to the best-performing model. Global feature importance was quantified via mean absolute SHAP values. Local explanations revealed per-prediction feature contributions.

Permutation importance: We computed permutation importance with 50 bootstrap iterations to obtain 95% confidence intervals for feature importance, providing a model-agnostic comparison to SHAP.

Fairness-aware SHAP: SHAP values were computed separately for each racial/ethnic group to detect whether predictive features operate with differential magnitude or direction across populations.

Temporal Generalization Analysis

To examine how prediction timing affects both accuracy and fairness, we trained models under four temporal scenarios with progressively more features:

1. **K Fall Only:** 7 features (fall kindergarten scores and demographics)
2. **K Fall + Spring:** 10 features (adding spring kindergarten scores)
3. **K + 1st Grade:** 11 features (adding 1st grade teacher report)
4. **K through 3rd:** 12 features (adding executive function score)

All seven algorithms were trained for each scenario. We examined how AUC, fairness metrics, and calibration error evolved across scenarios.

Sensitivity Analysis

We assessed sensitivity of fairness findings to the choice of at-risk threshold by repeating the full analysis at the 10th, 20th, 25th, and 30th percentiles. We re-evaluated all three fairness criteria at each threshold to determine whether fairness compliance was robust or threshold-dependent.

Missing Data Sensitivity Analysis

The ECLS-K:2011 uses special codes for missing data: -1 (not applicable), -7 (refused), -8 (don't know), and -9 (not ascertained). We recoded all such values to missing before analysis. We distinguish between two sources of incomplete data: *panel attrition*, in which children left the study entirely and have no data at later waves, and *item nonresponse*, in which children remained in the study but have missing values on individual variables (e.g., due to refusal or administrative error). Panel attrition was the dominant source of missingness for the outcome variable: 37% of the initial sample lacked 5th-grade reading scores, and the complete-case analytic sample ($N = 9,104$) represented 50% of the enrolled cohort. Item nonresponse contributed to predictor missingness, with rates ranging from 5.3% (spring kindergarten reading) to 25.9% (1st-grade approaches to learning). The distinction matters because panel attrition is more likely to be non-ignorable (related to unmeasured outcomes), whereas item nonresponse may be more plausibly missing at random. This level of incomplete data raises the possibility that our primary complete-case results are affected by selection bias if missingness is related to outcomes or protected attributes. We retain the complete-case analysis as the primary specification because it avoids the untestable assumptions required by imputation (e.g., that missingness is ignorable conditional on observed variables), and treat MICE and IPW as sensitivity analyses that bracket the plausible range of results.

We conducted three complementary sensitivity analyses to assess the robustness of our findings to missing data:

Attrition analysis. We compared baseline characteristics of study completers ($N = 9,104$) and dropouts ($N = 9,047$) on all available baseline variables, using standardized mean differences (Cohen's d) for continuous variables and chi-squared tests for categorical variables (Rubin 1987). We flagged differences exceeding $|d| \geq 0.20$ as potentially meaningful.

Multiple Imputation by Chained Equations (MICE). We generated $m = 10$ multiply-imputed datasets using iterative imputation with Bayesian ridge regression estimators and stochastic posterior draws, imputing the full sample ($N = 18,151$) rather than only complete cases. For each imputed dataset, we trained the elastic net model and computed group-level fairness metrics. Results were

pooled using Rubin's rules: $\bar{Q} = \frac{1}{m}\sum_{i=1}^{m} Q_i$, with between-imputation variance $B = \frac{1}{m-1}\sum_{i=1}^{m}(Q_i - \bar{Q})^2$ and total variance $T = \bar{U} + (1 + 1/m)B$ (Rubin 1987). Confidence intervals were computed as $\bar{Q} \pm 1.96\sqrt{T}$.

Inverse Probability Weighting (IPW). We estimated the probability of being a complete case using logistic regression on low-missingness baseline variables (race, sex, SES, kindergarten cognitive scores, home language). Inverse probability weights were computed as $w_i = 1/\hat{P}(\text{complete} \mid \mathbf{X}_i)$, stabilized by multiplying by the overall completion rate, and trimmed at the 99th percentile to limit the influence of extreme weights (Seaman and White 2013). The elastic net model was then re-trained with these sample weights to assess whether reweighting to approximate the full-sample distribution altered fairness conclusions.

Bias Mitigation

We implemented **threshold optimization** as a post-processing bias mitigation strategy. Rather than using a single decision threshold (typically 0.5) for all groups, we selected group-specific thresholds to equalize true positive rates across groups. The target TPR was set to the overall TPR of the best-performing model.

We note that group-specific decision thresholds constitute a form of race-conscious classification, which raises both legal and ethical concerns. Legally, under the Equal Protection Clause, race-conscious government actions are subject to strict scrutiny, requiring a compelling interest and narrow tailoring (Barocas et al. 2019). Ethically, group-specific thresholds can produce individual-level harms: a student whose threshold is raised because of their racial group membership may be denied support they would have received under a uniform threshold, regardless of their individual risk profile. The system must also "know" each student's race at decision time to apply the correct threshold, creating a race-conscious classification mechanism that raises concerns about reinforcing racial categorization—even when the intent is to promote equity. We present threshold optimization as a *diagnostic tool* to illustrate the theoretical bounds of post-hoc equalization, not as a recommended intervention for deployment.

To complement this post-processing approach, we also evaluated **in-processing fairness constraints** using Fairlearn's `ExponentiatedGradient` algorithm (Agarwal et al. 2018; Bird et al. 2020). We tested two constraint formulations: (1) Equalized Odds, jointly constraining TPR and FPR across racial groups; and (2) TPR Parity, constraining the maximum TPR difference to 0.05. These in-processing methods embed fairness considerations directly

during model training, avoiding the race-conscious threshold adjustment required by post-processing.

Finally, we conducted an **SES-removal ablation** to test the "poverty detector" hypothesis by retraining the elastic net model after removing the SES quintile feature, assessing whether the socioeconomic signal is the primary driver of fairness disparities or whether it is captured redundantly through cognitive scores.

Counterfactual Decomposition

To move beyond documenting disparities and toward explaining their mechanism, we conducted a counterfactual decomposition of the Hispanic–White TPR gap—the largest and most robust disparity in our analysis. The goal was to partition the observed gap into (A) a *structural component* attributable to differences in baseline feature distributions between the two groups, and (B) a *model-induced component* attributable to the mechanics of how the model and decision threshold interact with those distributions.

The procedure was as follows. For each Hispanic student in the test set, we constructed a counterfactual feature vector by aligning each feature to the White distribution via quantile matching: each Hispanic student's percentile rank within the Hispanic training distribution was computed, and the corresponding value at that percentile in the White training distribution was assigned. This preserves rank order within the Hispanic group while imposing the White distributional shape. For continuous features (cognitive scores, teacher ratings, executive function), we applied percentile-based alignment; for ordinal features (SES quintile) and categorical features (language status, sex), we used rank-preserving redistribution to match the White group's marginal proportions. The trained model and decision threshold ($t = 0.50$) were held fixed.

The decomposition is then:

$$\underbrace{\text{TPR}_H - \text{TPR}_W}_{\text{Observed gap}} = \underbrace{\text{TPR}_H - \text{TPR}_H^{cf}}_{\text{Structural}} + \underbrace{\text{TPR}_H^{cf} - \text{TPR}_W}_{\text{Model-induced}}$$

$$(2)$$

where TPR_H^{cf} is the counterfactual TPR obtained when Hispanic students' features are aligned to the White distribution but their true outcomes and the model remain unchanged. Bootstrap confidence intervals (1,000 iterations) were computed for all components. We conducted this analysis for both reading and math outcomes to assess cross-domain consistency. We emphasize that this decomposition is statistical, not causal: the structural component quantifies how much of the TPR gap is statistically attributable to

observed feature distribution differences, not the causal effect of changing those distributions.

Results

Sample Characteristics

Table 1 presents the demographic characteristics of the analytic sample.

The sample reflects the demographic diversity of the U.S. school-age population. Of the 11,427 children with valid 5th-grade reading outcomes, 25.0% were classified as at-risk (below the 25th percentile), providing adequate prevalence for model training.

Model Performance

Table 2 presents the predictive performance of all seven models on the held-out test set (N = 2,732).

All seven models achieved similar performance (Figure 1), with AUC values ranging from 0.837 (LightGBM) to 0.848 (Elastic Net). The two classical regularized linear models (Elastic Net: 0.848; Logistic Regression: 0.847) achieved the highest discrimination, while the three recent gradient boosting methods performed comparably but slightly lower (CatBoost: 0.846; HistGradientBoosting: 0.839; LightGBM: 0.837). CatBoost achieved the highest recall (0.308) and F1 score (0.421), making it the most effective at identifying at-risk students at the default threshold. The negligible AUC range across algorithms (0.011) suggests the performance ceiling is determined by the available features rather than algorithmic sophistication. The elastic net model was selected for subsequent fairness analysis based on its nominally highest AUC, though the 0.001 difference from logistic regression is not statistically meaningful; the choice is essentially arbitrary among the top-performing models, and the convergence finding itself suggests that fairness results would be substantively identical regardless of which model was selected.

Explainability Analysis

Feature Importance Table 3 presents feature importance from three complementary methods: SHAP values, elastic net coefficients, and permutation importance with bootstrap confidence intervals.

Spring kindergarten math (X2MTHETK) was the dominant predictor across all methods (Figure 2), with mean |SHAP| = 0.410, accounting for approximately 34% of total SHAP importance. SHAP and permutation importance rankings showed high agreement (mean agreement = 0.87),

with both methods identifying the same top-5 features. Notably, race/ethnicity, home language, child sex, and early approaches-to-learning measures received zero importance across all methods, confirming that elastic net regularization effectively excluded these features from the final model.

Fairness-Aware SHAP We computed SHAP values separately for each racial/ethnic group to detect differential feature importance. The top-5 feature rankings were identical across White, Black, and Hispanic subgroups. However, the magnitude of math score importance was somewhat higher for Black and Hispanic students compared to White students, suggesting that cognitive scores carry relatively more predictive weight for minority students. These magnitude differences were modest and did not alter the overall ranking of features, indicating that the model uses a consistent predictive mechanism across groups rather than relying on group-specific pathways. We note that this analysis examines feature importance rankings, not heterogeneous effect sizes or interaction effects; the consistency of rankings does not preclude differences in calibration or predicted probability distributions across groups, which are captured by the calibration fairness analysis below.

Fairness Analysis

Group-Level Performance with Confidence Intervals Table 4 presents performance metrics by racial/ethnic group with bootstrap 95% confidence intervals.

Bootstrap confidence intervals enable formal statistical comparison of inter-group differences (Figure 3). The Hispanic TPR of 0.393 [0.326, 0.459] was significantly higher than the White TPR of 0.160 [0.113, 0.206], with non-overlapping confidence intervals ($p < 0.05$). The Black-White TPR difference (0.296 vs. 0.160) showed partially overlapping CIs, suggesting a marginally significant difference. The model detected at-risk Hispanic students at 2.5 times the rate of at-risk White students.

We emphasize that the Asian subgroup ($N = 8$, CI: 0.250–1.000) is too small for any meaningful inference; its extreme TPR of 0.760 and perfect PPV are essentially uninformative and should not be interpreted as evidence of model performance for Asian students. Similarly, the Other subgroup ($N = 225$, but only ≈ 27 at-risk) yields wide confidence intervals (TPR CI: 0.074–0.445) that render point estimates exploratory at best.

False positive rate disparities were also substantial: Black students experienced an FPR of 0.095 [0.052, 0.133], compared to 0.011 [0.006, 0.017] for White students—8.5 times higher. This means non-at-risk Black children were far

Table 1. Sample Characteristics (N = 18,151)

Characteristic	N	%
Race/Ethnicity		
White	8,476	46.7
Hispanic	4,206	23.2
Black	2,394	13.2
Other	1,825	10.1
Asian	380	2.1
Not reported	870	4.8
SES Quintile		
Q1 (Lowest)	3,224	17.8
Q2	3,214	17.7
Q3	3,217	17.7
Q4	3,227	17.8
Q5 (Highest)	3,206	17.7
Not reported	2,063	11.4
Sex		
Male	9,273	51.1
Female	8,840	48.7
Not reported	38	0.2
5th Grade Reading Risk		
At-Risk (<25th %ile)	2,857	25.0*
Not At-Risk	8,570	75.0*
Missing outcome	6,724	37.0

*Percentages for at-risk status computed among the 11,427 children with valid 5th-grade outcomes.

Table 2. Model Performance on Test Set

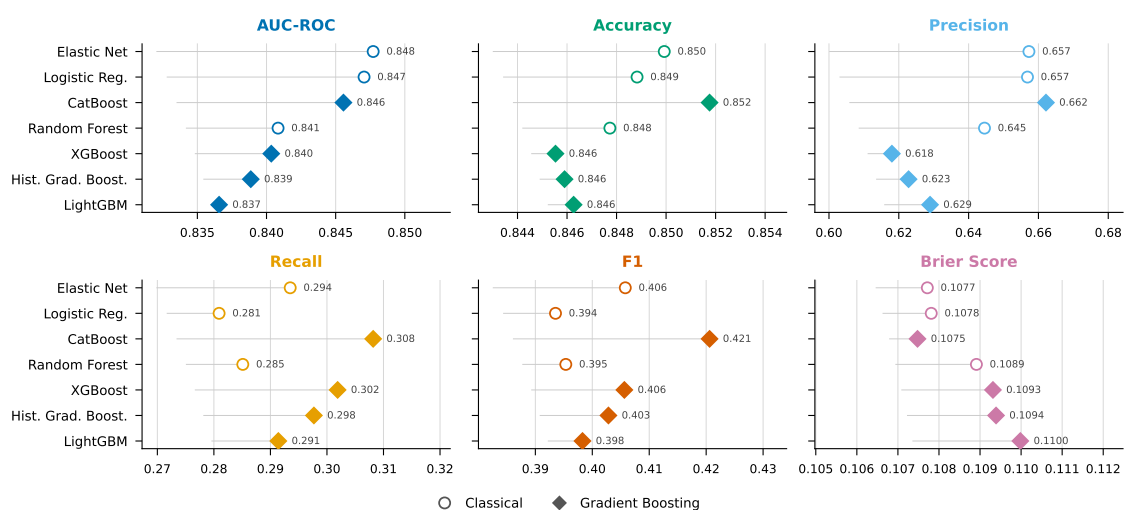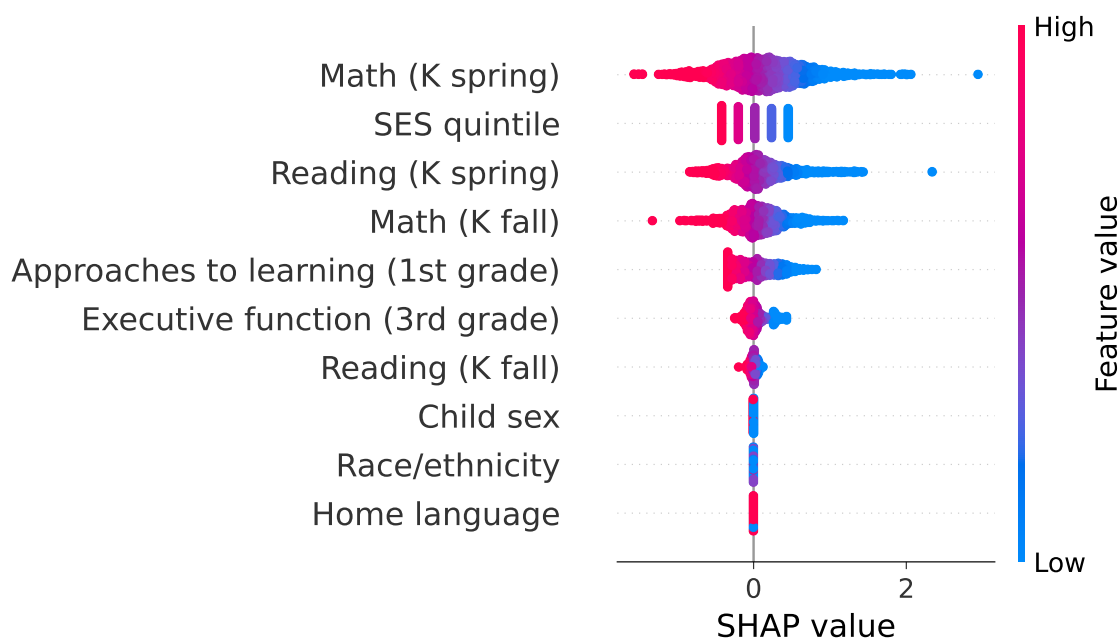| Model	AUC	Accuracy	Precision	Recall	F1	Brier
Elastic Net	**0.848**	**0.851**	0.675	0.283	0.399	**0.108**
Logistic Regression	0.847	0.849	0.657	0.281	0.394	0.108
CatBoost	0.846	0.852	0.662	**0.308**	**0.421**	0.107
Random Forest	0.841	0.848	0.645	0.285	0.395	0.109
XGBoost	0.840	0.846	0.618	0.302	0.406	0.109
Hist Gradient Boosting	0.839	0.846	0.623	0.298	0.403	0.109
LightGBM	0.837	0.846	0.629	0.291	0.398	0.110

Figure 1. Performance Comparison Across All Seven Models. Each panel shows a different metric with a zoomed axis to reveal inter-model differences. Open circles denote classical models; filled diamonds denote gradient boosting methods. AUC values ranged from 0.837 to 0.848 (range = 0.011), with classical regularized models matching or exceeding state-of-the-art methods.

Table 3. Feature Importance: SHAP, Elastic Net Coefficients, and Permutation Importance

Feature	Mean |SHAP|	EN Coef.	Permutation Imp. [95% CI]
Spring K Math (X2MTHETK)	0.410	0.501	0.046 [0.042, 0.054]
Spring K Reading (X2RTHETK)	0.254	0.350	0.023 [0.018, 0.027]
SES Quintile (X1SESQ5)	0.253	0.303	0.013 [0.007, 0.019]
Fall K Math (X1MTHETK)	0.227	0.288	0.012 [0.007, 0.017]
Approaches to Learning, 1st (X4TCHAPP)	0.218	0.266	0.016 [0.012, 0.022]
Executive Function (X6DCCSSCR)	0.067	0.115	0.001 [−0.001, 0.003]
Fall K Reading (X1RTHETK)	0.000	0.038	0.000
Child Sex, Race, Language, ATL (K)	0.000	0.000	0.000

Figure 2. SHAP Summary Plot (Beeswarm). Each point represents one prediction. Color indicates feature value (red = high, blue = low). Spring K math score shows the widest spread, indicating it is the strongest predictor with clear directionality.

Table 4. Model Performance by Race/Ethnicity with 95% Bootstrap Confidence Intervals

Group	N	TPR [95% CI]	FPR [95% CI]	PPV [95% CI]
White	1,462	0.160 [0.113, 0.206]	0.011 [0.006, 0.017]	0.660 [0.500, 0.794]
Hispanic	623	0.393 [0.326, 0.459]	0.063 [0.044, 0.087]	0.722 [0.637, 0.798]
Black	300	0.296 [0.207, 0.388]	0.095 [0.052, 0.133]	0.508 [0.341, 0.667]
Other	225	0.258 [0.074, 0.445]	0.005 [0.000, 0.015]	0.871 [0.571, 1.000]
Asian	8	0.760 [0.250, 1.000]	0.000 [0.000, 0.000]	1.000 [1.000, 1.000]

Note. Asian $N = 8$ in the test set is too small for reliable inference;

confidence intervals for groups with $N < 30$ may not achieve nominal coverage.

114 test-set observations with unreported race/ethnicity are excluded from this table.

more likely to be incorrectly flagged. ROC curves by group (Figure 6) and calibration curves (Figure 7) further illustrate these differential performance patterns.

Disparity Analysis Table 5 presents formal disparity metrics comparing each group to the White reference group.

Fairness Criteria Assessment:

- **Equal Opportunity:** PASS. All groups had TPR ratios > 0.80 (all minority groups had *higher* TPR than White students).
- **Equalized Odds:** FAIL. While TPR ratios exceeded 0.80, FPR ratios for Black (8.494) and Hispanic

(5.662) students dramatically exceeded 1.0, indicating disproportionately high false positive rates.

- **Statistical Parity:** PASS. Positive prediction rates did not trigger the 0.80 disparate impact threshold.

Calibration Fairness Table 6 presents calibration metrics by demographic group.

Calibration fairness analysis revealed substantial disparities (Figure 4). Black students experienced an ECE of 0.074, 3.35 times higher than White students (0.022), indicating that predicted risk probabilities for Black students were systematically miscalibrated. The Maximum Calibration Error
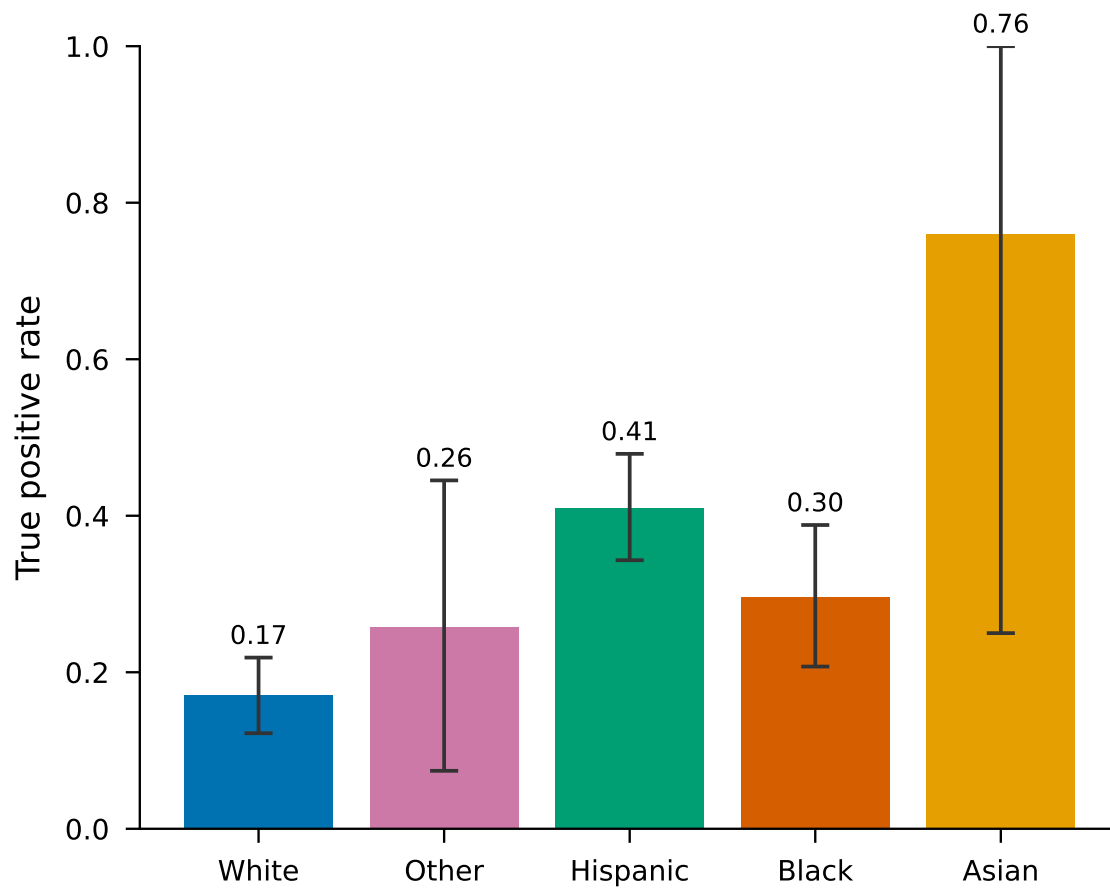
Figure 3. True Positive Rate by Racial/Ethnic Group with 95% Bootstrap Confidence Intervals. Non-overlapping intervals between Hispanic and White groups confirm statistically significant TPR disparities.

Table 5. Fairness Disparity Metrics (Reference: White)

Group	TPR Ratio	TPR Diff	FPR Ratio	FPR Diff	Disp. Impact
Asian	4.750	+0.600	0.000	−0.011	No
Hispanic	2.458	+0.233	5.662	+0.052	No
Black	1.851	+0.136	8.494	+0.084	No
Other	1.611	+0.098	0.477	−0.006	No

Note. Ratios computed from exact (unrounded) group metrics; minor discrepancies with ratios of displayed values in Table 4 reflect rounding.

Table 6. Calibration Error by Demographic Group

Group	N	ECE	MCE	Brier	ECE Ratio
White	1,462	0.022	0.112	0.082	1.00
Hispanic	623	0.036	0.115	0.155	1.65
Other	225	0.051	0.940	0.090	2.31
Black	300	0.074	0.456	0.162	**3.35**

for Black students (MCE = 0.456) was 4.1 times the White MCE (0.112), indicating severe miscalibration in certain probability ranges. These calibration disparities mean that even when the model makes the correct binary classification, the confidence levels are less reliable for minority students—a critical concern when practitioners use predicted probabilities to prioritize intervention resources.

Intersectional Fairness (Exploratory) The following analysis is exploratory and should be interpreted with caution due to small cell sizes in several key subgroups. Table 7 presents fairness metrics for selected race × SES subgroups, revealing suggestive patterns invisible in single-attribute analysis that warrant replication with larger, purpose-sampled datasets.

A suggestive finding was the model's failure to identify any at-risk Black students in the 4th SES quintile (TPR =
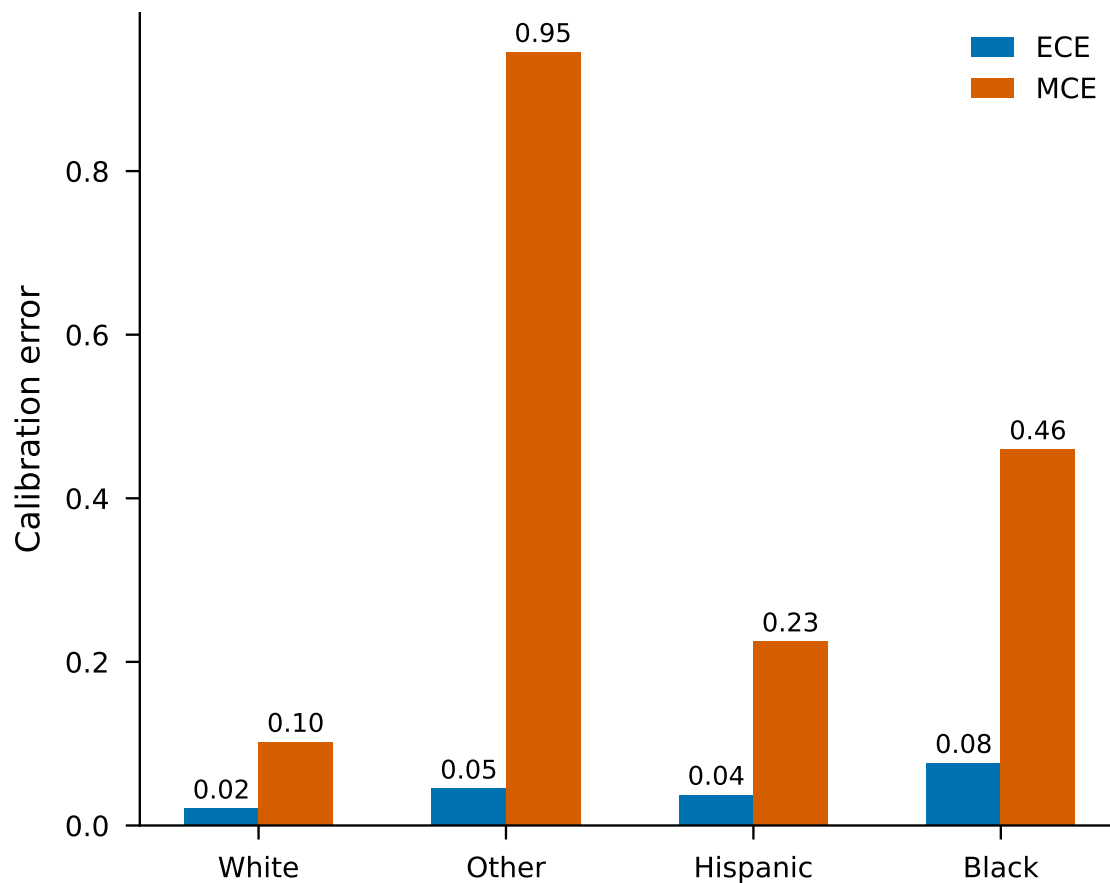
Figure 4. Expected and Maximum Calibration Error by Demographic Group. Black students experience calibration error 3.35 times higher than the White reference group.

Table 7. Intersectional Fairness: Selected Race $\times$ SES Subgroups

Subgroup	N	Prevalence	TPR	FPR	Accuracy
Low SES (Q1)					
Hispanic Q1	269	38.7%	0.481	0.121	0.725
Black Q1	80	32.5%	0.423	0.204	0.675
White Q1	107	28.0%	0.400	0.117	0.748
Mid SES (Q2–Q3)					
Hispanic Q2	145	29.0%	0.452	0.058	0.800
Black Q3	60	26.7%	0.375	0.068	0.783
White Q3	314	13.4%	0.167	0.004	0.885
High SES (Q4)					
White Q4	398	9.0%	0.083	0.003	0.915
Hispanic Q4	62	16.1%	0.200	0.019	0.855
Black Q4	42	14.3%	**0.000**	0.056	0.810

0%, $N = 42$, approximately 6 at-risk cases), as shown in Figure 5. Despite a 14.3% prevalence of academic risk in this subgroup, the model flagged none of these students. However, we urge caution in interpreting this result: with only $\approx$6 positive cases, a binomial test of TPR = 0 against a null of TPR = 0.283 (the overall model TPR) yields $p = 0.41$, which is not statistically significant. The observed TPR = 0% is consistent with sampling variability in a small cell. This pattern of potential "intersectional invisibility" extended to other high-SES minority subgroups: Hispanic

Q4 achieved only 20% TPR ($N = 62$, $\approx$10 at-risk). In contrast, low-SES students across all racial groups had relatively higher TPRs (0.400 to 0.481), consistent with the model's reliance on SES as a predictor.

The intersectional pattern is consistent with a plausible working hypothesis that the model operates as a *poverty detector*: it identifies at-risk students primarily through socioeconomic signals, missing those whose risk arises from other factors. This interpretation is supported by the strong SES gradient visible across all racial groups (Table 7),

but the small cell sizes in high-SES minority subgroups—particularly the 3–6 positive cases that drive the most extreme TPR estimates—fall below thresholds for reliable estimation and require replication with larger samples before informing policy conclusions.

Temporal Generalization

Given the fairness disparities documented above, a natural question is whether incorporating more developmental data resolves them. Table 8 presents model performance across four temporal scenarios with progressively more features.

Logistic regression was the best-performing model across all four temporal scenarios (Figure 8), reinforcing the finding that classical methods are sufficient for this prediction task. Note that the AUC for logistic regression at the full feature set (0.831) is slightly lower than the elastic net AUC reported in Table 2 (0.848); this reflects the temporal analysis using a separate hyperparameter grid optimized for each scenario rather than the main analysis grid, as well as the different regularization structure of logistic regression versus elastic net. AUC improved from 0.799 (K fall only) to 0.831 (K through 3rd grade), a meaningful gain of 0.032 AUC points. However, returns diminished rapidly: the largest gain came from adding spring kindergarten scores (+0.023 AUC), while adding 1st through 3rd grade data contributed only +0.009.

Critically, while overall accuracy improved with more data, fairness disparities did not narrow proportionally (Figure 9). The Hispanic-White TPR gap remained substantial across all scenarios, and calibration disparities persisted. We term this pattern *persistent fairness gaps despite improved accuracy*: incorporating additional longitudinal predictors improves a model's overall discriminative ability (AUC) while leaving group-level fairness gaps—measured by TPR ratios, FPR ratios, and calibration error—unchanged or worsened. This phenomenon is distinct from distributional shift or concept drift, which concern changes in the data-generating process over time; instead, it describes a within-cohort property in which more developmental data reinforces the same socioeconomic signals that drive differential performance, rather than introducing orthogonal information that could equalize predictions across groups. This outcome is consistent with the theoretical impossibility results of Chouldechova (2017)—when base rates differ across groups, no classifier can simultaneously equalize all fairness criteria—but has not previously been documented empirically in longitudinal educational prediction, where the practical assumption that accumulating more student data will naturally produce more equitable predictions remains common.

Sensitivity Analysis

A critical question is whether the fairness findings above are artifacts of the specific threshold used to define "at-risk." Table 9 presents fairness criteria compliance across four at-risk threshold definitions.

Sensitivity analysis revealed that fairness findings were highly dependent on the threshold definition. The 25th percentile was the only threshold at which the model passed equal opportunity and statistical parity criteria. At all other thresholds—including the nearby 20th and 30th percentiles—the model failed all three fairness criteria. This pattern likely reflects the interaction between the score distributions of different racial groups and the threshold location: at the 25th percentile, the threshold falls in a region where group-specific score densities happen to produce similar TPR ratios, whereas small shifts in either direction move the cut-point into regions where the distributions diverge more sharply. The fragility is therefore not random but structural—it reflects the particular geometry of overlapping group score distributions—and it underscores that fairness assessments cannot be divorced from the operationalization of the outcome. Claims of fairness compliance are contingent on specific analytical choices.

Missing Data Sensitivity Results

Attrition Analysis Study completers differed systematically from dropouts on all baseline cognitive variables. Completers scored higher on fall kindergarten reading ($d = 0.22$), spring kindergarten reading ($d = 0.26$), fall kindergarten math ($d = 0.28$), and spring kindergarten math ($d = 0.30$). Attrition was also differential by demographics: White students were over-represented among completers (53.4% vs. 40.0% of dropouts), while Asian (0.4% vs. 3.9%), Black (11.0% vs. 15.4%), and Other (7.8% vs. 12.3%) students were under-represented. These patterns indicate that complete-case analysis likely under-represents the most disadvantaged students, motivating the MICE and IPW sensitivity analyses.

Multiple Imputation Results MICE on the full sample ($N = 18{,}151$, $m = 10$ imputations) yielded slightly higher overall AUC (0.864, SE = 0.006) compared to complete-case analysis (0.848). All groups showed substantially higher TPR in the imputed sample—Black TPR increased from 0.30 to 0.53 (+0.23); Hispanic from 0.39 to 0.55 (+0.16); White from 0.16 to 0.37 (+0.21)—indicating that attrition disproportionately removed detectable at-risk students from the analytic sample. TPR ratios moved toward parity
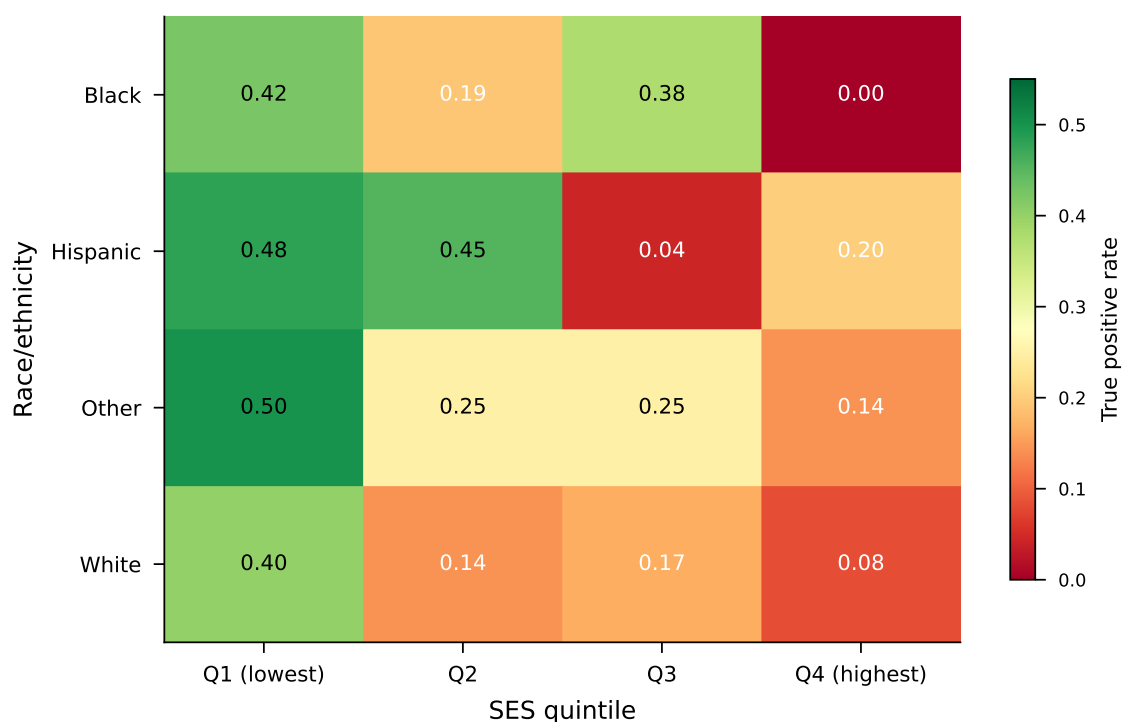
Figure 5. Intersectional Fairness Heatmap: TPR by Race $\times$ SES Quintile. Cell sizes range from $N = 42$ (Black Q4) to $N = 398$ (White Q4); cells with fewer than $\approx$10 positive cases yield unreliable TPR estimates. The TPR = 0% for Black Q4 ($N = 42$, $\approx$6 at-risk cases) is not statistically significant ($p = 0.41$) and should be interpreted as exploratory. All subgroups with $N < 50$ warrant caution.
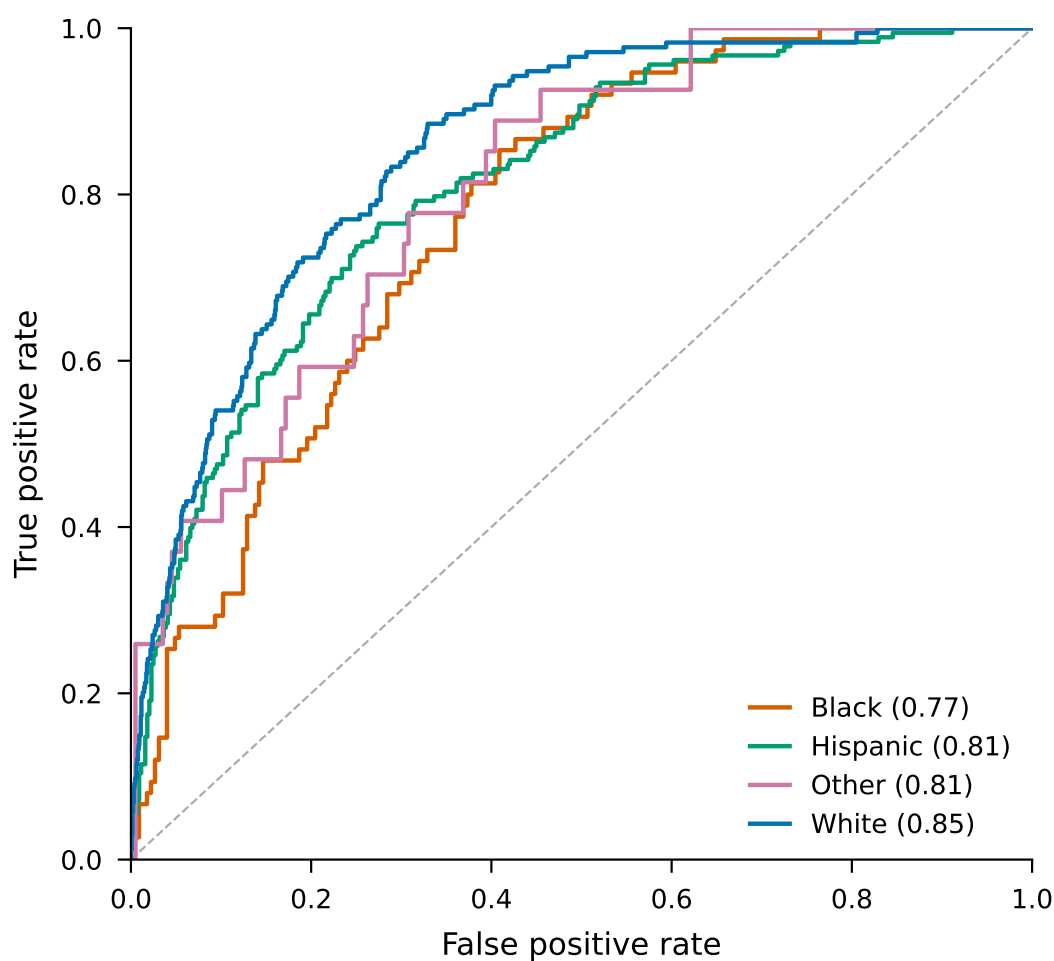

Figure 6. ROC Curves by Racial/Ethnic Group. All groups show similar overall discrimination (AUC $\approx$ 0.84), but the operating point characteristics differ substantially, reflecting the TPR and FPR disparities documented in Table 4.
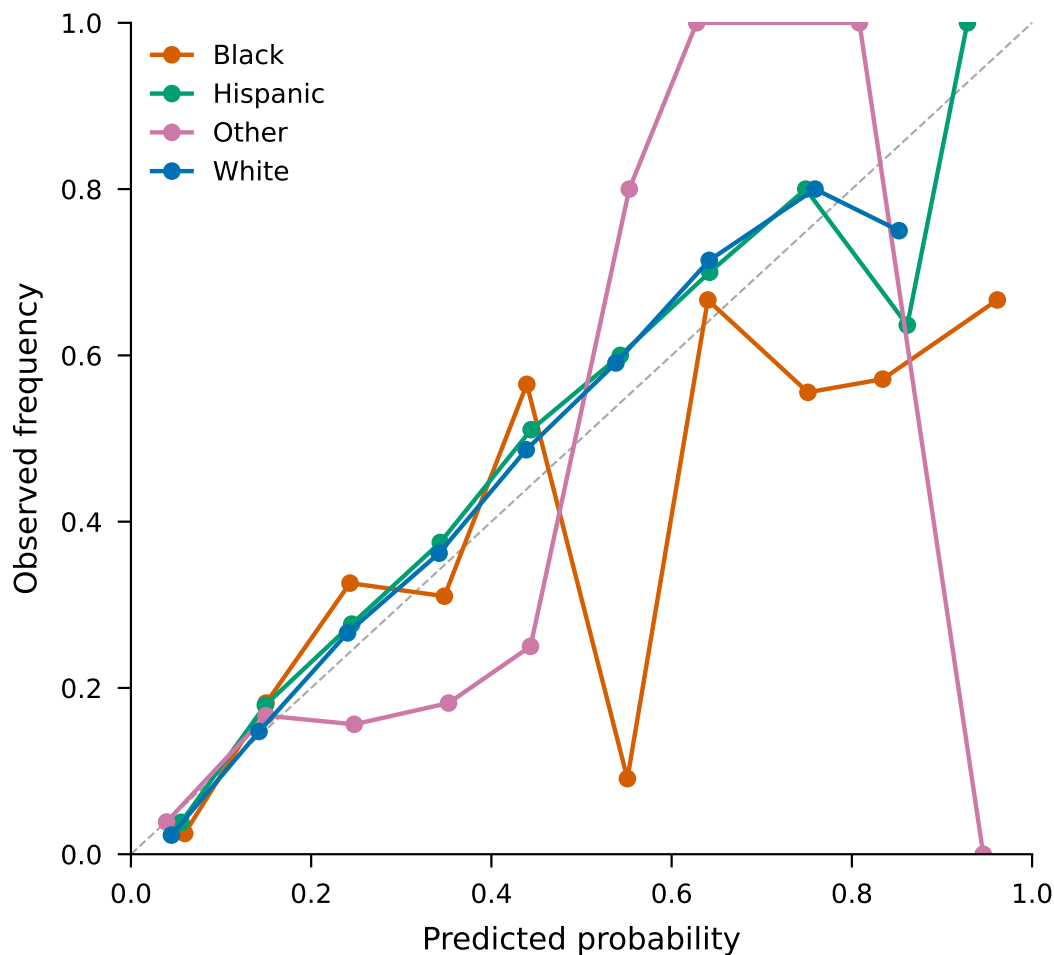
Figure 7. Calibration Curves by Racial/Ethnic Group. Deviation from the diagonal indicates miscalibration.

Table 8. Best Model Performance Across Temporal Scenarios

Scenario	Best Model	Features	AUC	Accuracy	F1	Brier
K Fall Only	Logistic Reg.	7	0.799	0.837	0.343	0.119
K Fall + Spring	Logistic Reg.	10	0.822	0.844	0.372	0.113
K + 1st Grade	Logistic Reg.	11	0.829	0.845	0.395	0.112
K through 3rd	Logistic Reg.	12	0.831	0.843	0.386	0.112

Table 9. Fairness Criteria by At-Risk Threshold Definition

Percentile	Prevalence	Equal Opp.	Equalized Odds	Statistical Parity
10th	6.3%	FAIL	FAIL	FAIL
20th	12.6%	FAIL	FAIL	FAIL
25th	15.7%	PASS	FAIL	PASS
30th	18.9%	FAIL	FAIL	FAIL

(Hispanic-White: 2.46 → 1.49; Black-White: 1.85 → 1.43), though the absolute Black-White TPR gap increased slightly (0.16 vs. 0.14 in complete-case). The fundamental pattern of differential performance across racial groups persisted in the full imputed sample, confirming that the primary findings are robust to attrition correction.

The magnitude of the MICE-induced TPR shifts deserves emphasis: the near-doubling of Black TPR (0.30 to 0.53) is not merely a sensitivity check footnote but a substantive finding. It indicates that the complete-case analytic sample systematically under-represents at-risk minority students who left the study, and that the true disparities in the full population may be larger in absolute terms than our primary analysis suggests. For policy purposes, the two sets of estimates address different deployment contexts: the complete-case results provide valid inference for the population of students who remain in a school district throughout the study period (relevant for EWS deployed in
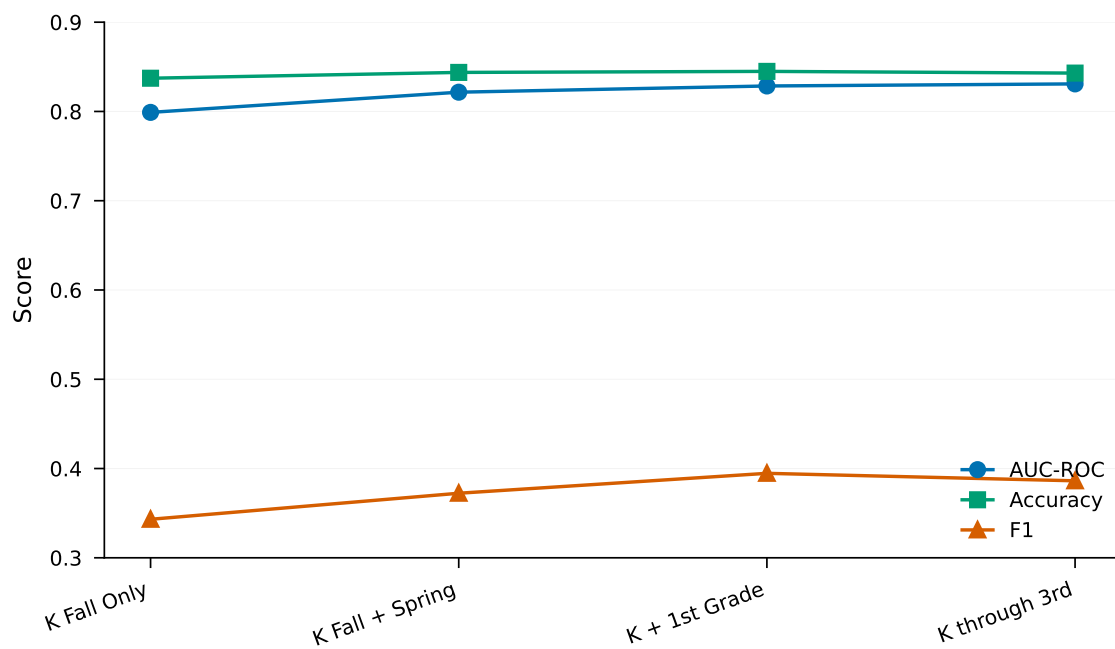
Figure 8. Model Performance Across Temporal Scenarios. AUC improves from 0.799 (K fall only) to 0.831 (K through 3rd grade), with diminishing returns after spring kindergarten.
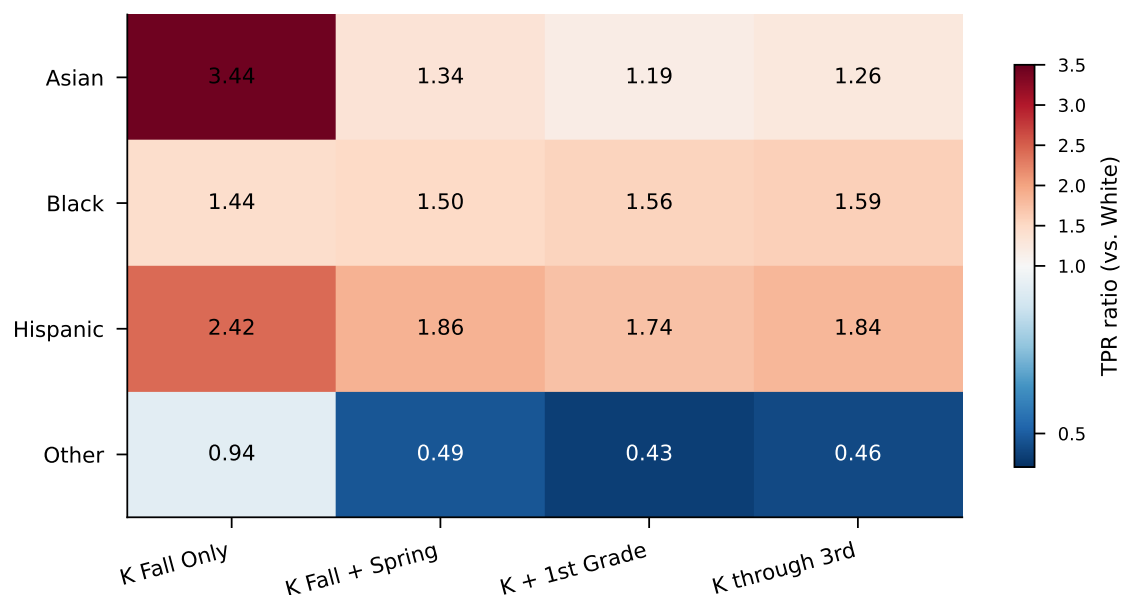

Figure 9. Temporal Disparity Heatmap: TPR Ratios Across Developmental Windows. Fairness disparities persist across all temporal scenarios, indicating that additional data does not resolve equity concerns.

stable districts with low mobility), while the MICE estimates project performance on the full enrolled population including highly mobile students who transfer or leave (relevant for universal screening programs). Districts should consider which reference population matches their context when interpreting model performance.

IPW Results Inverse probability weights were well-behaved: mean = 0.77, SD = 0.08, range 0.58–1.03, with 92 observations trimmed at the 99th percentile. The IPW-weighted analysis produced virtually identical results to the unweighted analysis (AUC = 0.848 in both

cases). Group-level metrics showed minimal changes: the largest FPR shift was -0.009 for Black and Hispanic students. The divergence between MICE and IPW results is methodologically expected: IPW reweights existing complete cases to approximate the full-sample distribution but cannot recover information from observations with missing outcomes, whereas MICE imputes the missing values themselves, effectively reconstructing the full sample and enabling estimation on all $N = 18,151$ cases. When attrition is substantial (50%) and outcome-related, IPW has limited power to correct bias because it can only adjust the

distribution of cases already observed; MICE, by generating plausible outcome values for missing cases, can reveal the extent to which attrition suppressed detection rates. The convergence of IPW and unweighted results thus indicates that, among complete cases, selection on observables is minimal; the MICE results indicate that the missing cases themselves—disproportionately disadvantaged students—would have been more detectable had they remained in the sample.

Domain Validation: Reading vs. Math

To assess whether the fairness patterns documented above are specific to reading prediction or reflect a broader structural phenomenon, we replicated the full fairness audit on the math outcome (X9MTHETA, at-risk prevalence = 17.8%). Table 10 presents the cross-domain comparison.

Three patterns replicated across domains. First, algorithmic convergence: four models achieved an AUC range of 0.019 on math (0.848–0.867), confirming that performance ceilings are feature-determined, not algorithm-determined. Second, the Hispanic-White TPR gap persisted and widened: Hispanic TPR was 2.17 times White TPR for math (vs. 2.46 for reading), with Hispanic students detected at substantially higher rates in both domains. Third, calibration unfairness was even more severe for math: Black students experienced ECE 4.93 times higher than White students (vs. 3.35 for reading), indicating that calibration disparities are a robust, cross-domain phenomenon.

One notable domain-specific finding emerged: the Other racial group showed a TPR ratio of 0.38 vs. White for math—well below the 0.80 disparate impact threshold—indicating severe under-identification that was not observed in reading (TPR ratio = 1.61). This demonstrates that the domain of the outcome can determine which groups experience the most acute fairness harms, reinforcing the importance of outcome-specific fairness auditing. The replication of core disparity patterns across both reading and math strengthens the interpretation that these fairness failures reflect structural properties of the prediction task rather than domain-specific artifacts.

Bias Mitigation Results

We applied threshold optimization to equalize TPR across groups, targeting the overall model TPR of 0.283. Table 11 presents the results.

Threshold optimization successfully equalized TPR across the major demographic groups (White, Black, Hispanic, Other all achieving TPR ≈ 0.29). However, this equalization

came at a cost that raises ethical concerns. Hispanic students experienced reduced sensitivity (-0.098), meaning the "fair" system would identify *fewer* at-risk Hispanic students to achieve parity with the White detection rate. Asian students experienced a severe drop in both TPR (-0.510) and accuracy (-0.250), though the small sample size ($N = 8$) renders this estimate unreliable. The group-specific thresholds ranged from 0.367 (White) to 0.649 (Asian), indicating that predictions for different groups required substantially different decision boundaries to achieve equitable outcomes.

These results illustrate two fundamental problems with naive post-hoc equalization. First, equalizing to a low overall TPR (0.283) can *harm the very groups it intends to help*: Hispanic students, who were being identified at a higher rate under the uncorrected model, would receive fewer interventions after equalization. This is a direct consequence of the impossibility results of Chouldechova (2017): when base rates differ across groups, no classifier can simultaneously equalize TPR, FPR, and PPV. Second, implementing group-specific thresholds requires the system to "know" each student's race at decision time, creating a race-conscious classification mechanism that faces both legal scrutiny and practical concerns about reinforcing racial categorization. In-processing alternatives—such as adversarial debiasing or fairness constraints during training—may offer more defensible approaches by addressing disparity at its source rather than through post-hoc adjustment.

SES-Removal Ablation

To test whether the model functions as a "poverty detector," we retrained the elastic net after removing SES quintile (X1SESQ5)—the third-strongest predictor (mean |SHAP| = 0.253)—from the feature set. If fairness disparities are primarily driven by socioeconomic proxies, removing SES should meaningfully alter group-level metrics.

Removing SES produced a negligible AUC drop (0.848 → 0.845, Δ = 0.003) and left group-level TPR and FPR virtually unchanged (Table 12). Black and Hispanic TPR remained identical (0.293 and 0.393, respectively). This result indicates that SES is largely redundant once cognitive baseline scores are included—consistent with the interpretation that the model captures socioeconomic disadvantage primarily through its cognitive correlates rather than through SES directly. The "poverty detector" pattern persists even when the explicit poverty indicator is removed, because early cognitive scores are themselves products of socioeconomic stratification.

Table 10. Domain Validation: Reading vs. Math Fairness Metrics

Domain	Group	N	Prev.	TPR	FPR	ECE	ECE Ratio
Reading	White	1,462	11.9%	0.160	0.011	0.022	1.00
	Black	300	25.0%	0.296	0.095	0.074	3.35
	Hispanic	623	29.4%	0.393	0.063	0.036	1.65
	Other	225	12.0%	0.258	0.005	0.051	2.31
Math	White	1,475	11.5%	0.231	0.019	0.029	1.00
	Black	288	35.4%	0.294	0.145	0.143	4.93
	Hispanic	622	28.6%	0.500	0.083	0.048	1.66
	Other	335	10.1%	0.088	0.023	0.057	1.97

Table 11. Bias Mitigation Results (Threshold Optimization)

Group	TPR Bef.	TPR Aft.	ΔTPR	Acc. Bef.	Acc. Aft.	ΔAcc.
White	0.160	0.293	+0.133	0.891	0.886	−0.005
Black	0.296	0.293	−0.003	0.747	0.747	+0.000
Hispanic	0.393	0.295	−0.098	0.775	0.762	−0.013
Asian	0.760	0.250	−0.510	0.875	0.625	−0.250
Other	0.258	0.296	+0.038	0.902	0.884	−0.018

Table 12. SES-Removal Ablation: Elastic Net Performance With and Without SES

Condition	AUC	Group	N	TPR	FPR
With SES (12 features)	0.848	White	1,462	0.161	0.011
		Black	300	0.293	0.093
		Hispanic	623	0.393	0.064
		Other	339	0.244	0.007
Without SES (11 features)	0.845	White	1,462	0.172	0.009
		Black	300	0.293	0.089
		Hispanic	623	0.393	0.070
		Other	339	0.268	0.013

Note. Other $N = 339$ includes 114 observations with unreported race/ethnicity that are excluded from Table 4 (Other $N = 225$). Minor TPR differences from Table 4 reflect independent model re-training with a different random seed for each experiment.

In-Processing Fairness Constraints

To evaluate whether fairness-constrained training can improve upon post-hoc threshold adjustment, we applied Fairlearn's `ExponentiatedGradient` algorithm (Agarwal et al. 2018; Bird et al. 2020) with two constraint formulations: Equalized Odds (jointly constraining TPR and FPR across groups) and TPR Parity (constraining maximum TPR difference to 0.05). These in-processing methods embed fairness constraints directly during model training, avoiding the race-conscious threshold adjustment required by post-processing approaches.

The results reveal a stark fairness-accuracy trade-off (Table 13). Equalized Odds reduced the TPR range across all four groups from 0.232 to 0.101—a 56% reduction in disparity—but at substantial cost to overall discrimination (AUC: 0.847 → 0.617, a 27% relative drop). TPR Parity achieved a comparable disparity reduction (TPR range: 0.232 → 0.126, 46% reduction) with a smaller AUC penalty (0.847 → 0.633). Neither constrained method fully achieved

its target: the Equalized Odds model narrowed but did not eliminate the TPR range, and the TPR Parity model's 0.126 range exceeded the specified 0.05 bound, suggesting that the optimization encountered difficulty satisfying tight constraints with this data.

Compared to post-hoc threshold optimization (Section 4.9), in-processing constraints produced broadly similar outcomes: both approaches reduced fairness disparities at the cost of overall accuracy, and both were unable to fully eliminate group-level differences. The key advantage of in-processing methods is that they do not require race-conscious threshold adjustment at deployment time, avoiding the legal and ethical concerns associated with group-specific decision boundaries. However, the magnitude of the AUC penalty (0.21–0.23 points) raises the question of whether such a substantial reduction in predictive performance is acceptable in a domain where the cost of missed at-risk students is high. Whether this trade-off is tolerable depends on local priorities: the constrained models remain well above chance

Table 13. In-Processing Fairness Constraints: Unconstrained vs. Fairlearn Results

Method	AUC	Group	N	TPR	FPR	Accuracy
Unconstrained	0.847	White	1,462	0.161	0.012	0.890
		Black	300	0.280	0.098	0.747
		Hispanic	623	0.393	0.070	0.772
		Other	339	0.244	0.007	0.903
Equalized Odds	0.617	White	1,462	0.167	0.057	0.851
		Black	300	0.213	0.080	0.743
		Hispanic	623	0.219	0.043	0.740
		Other	339	0.268	0.060	0.858
TPR Parity	0.633	White	1,462	0.167	0.009	0.893
		Black	300	0.293	0.098	0.750
		Hispanic	623	0.284	0.050	0.754
		Other	339	0.220	0.020	0.888

Note. Other $N = 339$ includes unreported race/ethnicity (cf. Table 4, Other $N = 225$).

Unconstrained TPR range (0.232) computed from this table's unconstrained row.

(AUC = 0.617–0.633 vs. 0.50 for random guessing) and may be preferable in equity-sensitive deployment contexts where reducing disparate identification rates takes precedence over maximizing aggregate accuracy. In contexts where the cost of missing any at-risk student is high, the unconstrained model's superior AUC may be warranted despite its larger fairness gaps. These findings reinforce the conclusion that fairness-accuracy trade-offs in this prediction task are not artifacts of the mitigation strategy but reflect a fundamental property of the data, consistent with the impossibility results of Chouldechova (2017).

Policy Allocation Simulation

To translate fairness metrics into concrete policy consequences, we simulated the downstream effects of deploying the elastic net model under two scenarios: the default decision threshold (0.50), and group-specific thresholds calibrated to equalize TPR at the overall model rate of 0.283.

Under the default threshold, the model missed 82.8% of at-risk White students but only 59.0% of at-risk Hispanic students—a 24-percentage-point gap in who receives support. In absolute terms, 144 at-risk White students went unidentified compared to 108 Hispanic students, despite White students comprising a larger share of the test set. Equalizing TPR across groups closed this gap (all groups: ≈71% missed) but redistributed both detections and errors: 19 additional at-risk White students were correctly identified, while 23 fewer at-risk Hispanic students received flags. The overall cost of equalization was modest—$27,000 in a stylized model assuming $2,000 per intervention and $5,000 per missed student (total system cost: $2.11M default vs. $2.14M equalized)—but the distributional shift was substantial.

This simulation concretizes three insights. First, fairness metrics translate into real students: the 24-point TPR gap between White and Hispanic students means that, in a district with 1,000 at-risk students per group, approximately 240 more Hispanic students would receive interventions than White students with identical risk profiles. Second, equalization is not Pareto-improving—it helps some groups by harming others, a direct consequence of the impossibility results. Third, the aggregate cost difference is small, but the question of *whose children are missed* is a distributive justice question that cannot be resolved by technical optimization alone.

We emphasize that this simulation is illustrative rather than prescriptive: the cost parameters ($2,000 per intervention, $5,000 per missed student) are stylized figures chosen to demonstrate *distributional* consequences, not to estimate actual cost-effectiveness. The qualitative conclusions—that equalization redistributes detections across groups, that the aggregate cost difference is modest relative to the shift in who is served, and that fairness choices are ultimately distributive justice choices—are robust to the specific cost values chosen. Under any plausible cost ratio where missing an at-risk student is more expensive than providing an intervention, the distributional pattern holds: equalization slightly increases total expenditure while substantially changing which students receive support.

Counterfactual Decomposition of the TPR Gap

Table 15 presents the counterfactual decomposition of the Hispanic–White TPR gap for both reading and math outcomes.

The structural component—the portion of the TPR gap statistically attributable to differences in baseline feature

Table 14. Policy Allocation Consequences: Default vs. Equal-TPR Thresholds

Scenario	Group	At-Risk	Detected	Missed	% Missed	Falsely Flagged
Default ($t = 0.50$)	White	174	30	144	82.8	15
	Black	75	22	53	70.7	23
	Hispanic	183	75	108	59.0	32
	Other	41	10	31	75.6	3
Equal TPR (t_g)	White	174	49	125	71.8	37
	Black	75	21	54	72.0	20
	Hispanic	183	52	131	71.6	17
	Other	41	12	29	70.7	8

Note. Based on the without-SES elastic net model. Group sizes differ slightly from
Table 4 due to inclusion of unreported race/ethnicity in Other.

Table 15. Counterfactual Decomposition of the Hispanic–White TPR Gap

Component	Reading	Math
Hispanic TPR (observed)	0.407	0.430
White TPR (observed)	0.204	0.234
Observed gap	0.203 [0.106, 0.304]	0.196 [0.097, 0.286]
Hispanic TPR (counterfactual)	0.099	0.129
Structural component	0.308 [0.239, 0.375]	0.301 [0.239, 0.368]
% of observed gap	151.7%	153.6%
Model-induced component	−0.105 [−0.175, −0.029]	−0.105 [−0.191, −0.023]
% of observed gap	−51.7%	−53.6%
Feature-level contributions (reading / math):		
Cognitive scores (all four)	88.7%	128.9%
SES quintile	28.6%	46.6%
Teacher ratings	5.7%	2.7%

Values in brackets are bootstrap 95% confidence intervals (1,000 iterations).
TPRs differ slightly from Table 4 because this analysis used an independently
trained elastic net (SGDClassifier) with the same train-test split and feature set.
Feature-level contributions need not sum to the structural component due to interaction effects.

distributions—exceeded 100% of the observed gap in both domains: 151.7% for reading (95% CI: [110.6%, 251.1%]) and 153.6% for math (95% CI: [107.7%, 276.3%]). In plain terms: if the only thing that differed between Hispanic and White students were their feature distributions—and the model treated both groups identically—the predicted TPR gap would be even *larger* than what we observe. The structural component exceeds 100% because the model partially offsets the distributional divergence, not because the arithmetic is paradoxical. The model-induced component was negative in both domains (-0.105, $p < .05$ in both cases), with bootstrap 95% confidence intervals excluding zero in both reading ($[-0.175, -0.029]$) and math ($[-0.191, -0.023]$), indicating that the model and threshold mechanics significantly *compress* the gap relative to what the distributional differences alone would produce. When Hispanic students were given the White feature distribution via quantile alignment, their counterfactual TPR dropped to 0.099 for reading and 0.129 for math—*below* the observed White TPR—confirming that the model itself

is not amplifying disparities but rather inheriting them from the input distributions.

Three features of the decomposition support its robustness. First, the structural component exceeds 100% in both outcome domains independently, with non-overlapping bootstrap confidence intervals, constituting a cross-domain replication on different outcome measures within the same cohort. Second, the quantitative convergence is striking: 151.7% and 153.6% structural, -51.7% and -53.6% model-induced—nearly identical despite different outcome variables, feature weights, and prevalence rates. Third, the decomposition's central finding—that disparities originate in feature distributions rather than model mechanics—is independently corroborated by the seven-model convergence (if the model were the source, different architectures would produce different disparities) and the SES-removal ablation (if any single feature were the lever, removing it would matter).

Feature-level decomposition identified kindergarten cognitive scores as the dominant statistical contributor: the

four baseline cognitive scores (fall and spring kindergarten reading and math) were together associated with 88.7% of the reading TPR gap and 128.9% of the math TPR gap. SES quintile contributed 28.6% (reading) and 46.6% (math). The cross-domain consistency is striking: structural components of 151.7% and 153.6%, model-induced components of -51.7% and -53.6%, and the same qualitative pattern across both outcome domains. This convergence provides strong evidence that the observed TPR disparities are a predictable consequence of the distributional properties of the input features—that is, of the educational inequality encoded in the data *given* the outcome definition, feature set, and threshold that were chosen. The structural encoding is real, but it is mediated by institutional decisions about what to measure and how to define risk.

Discussion

Summary of Findings

The central finding of this study is that, within the available feature space and outcome definition, fairness failures in early childhood risk prediction are located in the data rather than in model architecture. Seven models—spanning classical regularized methods and state-of-the-art gradient boosting—converged on nearly identical performance (AUC range = 0.011) and exhibited the same pattern of differential performance across racial groups. This convergence is itself the evidence: when every algorithm fails in the same way, the source of inequity lies in the features and outcome definitions, not in any particular model (Grinsztajn et al. 2022). We note that this does not rule out the possibility that alternative feature sets, outcome operationalizations, or fairness-constrained training methods could reduce disparities; it demonstrates that switching algorithms alone does not.

The dimensions of that inequity are multiple and compounding. The model detected at-risk Hispanic students at more than double the rate of White students, a gap confirmed as statistically significant by non-overlapping bootstrap confidence intervals. Black students faced 8.5 times higher false positive rates, meaning non-at-risk Black children were far more likely to be incorrectly flagged. Beyond classification accuracy, Black students experienced calibration error over threefold higher than White students (ECE = 0.074 vs. 0.022)—a distinct form of algorithmic harm in which the model's confidence levels become unreliable for specific populations (Pleiss et al. 2017). Intersectional analysis revealed a suggestive pattern consistent with the model operating as a poverty detector,

though key subgroup estimates (e.g., Black Q4: TPR = 0%, $N = 42$, $p = 0.41$) rest on cell sizes too small for definitive conclusions.

Two additional findings challenge common assumptions about prediction systems. The persistence of fairness gaps despite improved accuracy—additional longitudinal data improved AUC from 0.799 to 0.831 without resolving fairness disparities—is consistent with theoretical impossibility results (Chouldechova 2017) but challenges the practical expectation, common among EWS practitioners, that accumulating more student data will naturally produce more equitable predictions. And the fragility of fairness assessments across thresholds (compliance at only one of four tested thresholds) reveals that claims of fairness are inseparable from the policy choices embedded in outcome definitions.

Attempts to mitigate these disparities—both post hoc and during training—created new problems. Threshold optimization equalized TPR across groups but did so by *reducing* identification of at-risk Hispanic students to match the lower White detection rate—harming the very group it intended to help. In-processing fairness constraints (Equalized Odds and TPR Parity via Fairlearn) reduced TPR disparities by 46–56% but at the cost of a 0.21–0.23 AUC penalty, raising concerns about whether such accuracy reductions are acceptable when missed at-risk students face real consequences. Both mitigation approaches encountered the same fundamental trade-off, consistent with the impossibility results of Chouldechova (2017). SES-removal ablation further confirmed the structural nature of these disparities: removing the explicit SES feature produced negligible changes in both overall accuracy (ΔAUC = 0.003) and group-level fairness metrics, demonstrating that the "poverty detector" pattern is mediated through cognitive scores that are themselves products of socioeconomic stratification. Counterfactual decomposition provided the most direct evidence: the structural component—statistically attributable to differences in baseline feature distributions—was associated with more than 100% of the Hispanic–White TPR gap (151.7% for reading, 153.6% for math), with the model actually compressing the gap relative to what the distributional differences alone would predict. Finally, MICE sensitivity analysis on the full sample ($N = 18,151$) confirmed that the pattern of differential performance persists after correcting for attrition, with all groups showing substantially higher absolute detection rates. IPW produced nearly identical results to unweighted analysis, ruling out selection bias as an explanation for the observed fairness patterns.

Interpreting Fairness Disparities

The fairness disparities documented above demand explanation. Why does a model that excludes race as a predictor nonetheless produce racially differential outcomes? We emphasize that our analysis is purely predictive, not causal: we document *what* the model does, not *why* the underlying educational inequities exist. Establishing causal pathways from structural disadvantage through data-generating processes to model outputs would require different designs (e.g., natural experiments, instrumental variables, or structural equation models). With that caveat, several interacting mechanisms are at work.

Differential base rates: At-risk prevalence was substantially higher among Black (25.0%) and Hispanic (29.4%) students compared to White students (11.9%). When base rates differ, even a well-calibrated model will exhibit different error rates across groups (Chouldechova 2017).

Proxy discrimination: Although race was excluded from the model, other features (particularly SES) are correlated with race and may serve as proxies. The elastic net assigned substantial weight to SES (mean $|\text{SHAP}| = 0.253$), which could contribute to differential performance.

Structural inequities: The patterns in the data reflect historical and ongoing structural inequities in educational opportunity. Children from disadvantaged backgrounds may show weaker early signals not because of inherent ability, but because of differential access to high-quality early childhood education.

The poverty detector problem: The intersectional analysis suggests the model may function primarily as a poverty detector, though small cell sizes require cautious interpretation. SES is the third-strongest predictor, and its effects are deeply entangled with cognitive scores—which themselves reflect socioeconomic advantage. This creates a pattern where the model identifies low-SES children of all races but may miss at-risk children who come from relatively advantaged backgrounds. The SES-removal ablation provides direct evidence for this entanglement: removing SES from the model produced negligible changes in both AUC ($\Delta = 0.003$) and group-level fairness metrics, confirming that the socioeconomic signal is captured redundantly through cognitive scores that are themselves products of socioeconomic stratification. The MICE analysis on the full sample reinforces this interpretation: when attrition-related selection bias is addressed, the pattern of differential performance persists, and all groups show substantially higher absolute detection rates—indicating that attrition selectively removed the very students the model would have flagged.

Calibration and trust: Calibration unfairness is particularly consequential because practitioners rely on predicted probabilities, not just binary classifications, to prioritize interventions. If a school counselor sees that two students both have a 30% predicted risk, they may allocate resources equally—but if the model is poorly calibrated for Black students, these probability estimates carry unequal information content (Pleiss et al. 2017).

Theoretical Implications: From Algorithm Selection to Institutional Design

These findings challenge a default assumption in educational data science: that fairness problems can be solved by choosing the right algorithm, the right features, or the right mitigation technique. Our results suggest instead that, *within the available feature space and standard predictive paradigm*, fairness disparities in early childhood prediction persist across algorithms, mitigation strategies, and outcome domains—and are not resolvable through algorithmic selection alone. We do not claim that equitable prediction is impossible in principle: alternative feature sets (e.g., incorporating school-level resource indicators), different outcome operationalizations (e.g., growth trajectories rather than threshold-based risk), or causal modeling approaches may yield more equitable predictions. What we demonstrate is that the standard predictive toolkit—varying algorithms, removing features, and constraining training objectives—cannot resolve disparities that originate in the data-generating process itself. Returning to the conceptual framework introduced in Section 1.1, this is precisely the interaction between algorithmic governance and institutional stratification that our study was designed to examine: predictive systems faithfully reproduce the inequities embedded in the institutional contexts that generate their training data.

This claim rests on four converging lines of evidence. First, algorithmic convergence: seven architectures spanning different inductive biases produce the same disparity pattern, indicating that no model can extract equitable predictions from these features. Second, the SES-removal ablation: removing the explicit socioeconomic indicator changes nothing, because the cognitive scores that drive prediction are themselves products of the same stratification system. The model does not need to "see" poverty directly—it reads its cognitive fingerprint. Third, in-processing fairness constraints: even when equity is embedded in the training objective, achieving parity requires a 27% reduction in overall discrimination (AUC: $0.847 \rightarrow 0.617$), a cost that

translates directly into missed at-risk students. Fourth, counterfactual decomposition: when Hispanic students' features are aligned to the White distribution, the structural component is associated with more than the entirety of the TPR gap (152% averaged across domains), with the model actually compressing rather than amplifying the distributional divergence.

The conceptual implication is that the locus of intervention must shift from the algorithm to the *institutional context* in which prediction occurs. This conclusion resonates with a substantial body of research in the sociology of education. Reardon (2011) documented a widening income-achievement gap, showing that socioeconomic disparities in cognitive skills are already large at school entry and grow over time. Heckman (2006) demonstrated that skill formation is a cumulative process in which early disadvantage compounds through reduced returns on later investment, creating the very gradient our model exploits for prediction. Reardon and Portilla (2016) showed that income-based gaps in school readiness at kindergarten entry remain substantial despite some narrowing in racial gaps, indicating that the socioeconomic stratification our model captures reflects durable features of the opportunity structure. When the data-generating process—shaped by residential segregation, school funding inequities, differential access to early childhood education, and cumulative disadvantage—produces features that encode social position, no downstream statistical correction can fully undo that encoding without sacrificing the predictive signal itself. This is not a failure of machine learning; it is machine learning faithfully reflecting the structure of educational inequality. The policy question therefore changes: not "how do we build a fair model?" but "what institutional changes would produce data from which fair prediction becomes possible?"

For policymakers, this means that algorithmic EWS should be understood as *diagnostic instruments* that reveal the contours of institutional inequality, not as neutral tools that can be tuned to produce equitable outcomes. The convergence finding is itself the diagnosis: when every algorithm fails in the same way, the system—not the software—requires repair. This conclusion aligns with the sociotechnical perspective advanced by Selbst et al. (2019): treating fairness as a property of the algorithm alone—rather than of the sociotechnical system encompassing data generation, model deployment, and institutional response—leads to the framing trap, in which technical interventions address symptoms while leaving structural causes intact. Our decomposition results provide direct empirical evidence for this concern: 152% of the TPR gap is statistically

attributable to the structural properties of the input data, meaning that the disparity is already present in the feature distributions before model training begins—though we note that the choice of features, outcome definition, and threshold are themselves institutional decisions that shape what gets encoded (Green 2022). Similarly, Eubanks (2018)'s observation that automated systems in public services tend to intensify surveillance of disadvantaged populations finds a precise analog in our results: the model identifies at-risk Hispanic students at 2.5 times the rate of White students not because of algorithmic bias but because the data faithfully encodes the cumulative disadvantage that Hispanic children have experienced by kindergarten. In the language of Bowker and Star (2000), the risk score imposes a classification that renders socioeconomic disadvantage legible to the institution—but in doing so, it reifies a particular reading of inequality as individual-level risk rather than structural condition. The governance question is therefore not whether the classification is accurate (it is) but whether the institutional response it triggers addresses or perpetuates the underlying stratification.

More broadly, our findings refine the sociotechnical systems framework by identifying a mechanism through which abstraction traps operate empirically. Selbst et al. (2019) theorized these traps at the level of problem formulation—how researchers define fairness, scope the system, and abstract away context. Our decomposition reveals that the same dynamic operates at the *statistical* level: when institutional stratification produces group-level differences in feature distributions, the act of training a predictive model on those features necessarily encodes that stratification into model outputs, regardless of the algorithm chosen. This is what we mean by structural encoding—and it is distinct from several related concepts. It differs from *algorithmic bias*, which implies the model introduces disparity; from *proxy discrimination*, which implies a specific feature mediates the effect; from *omitted variable bias*, which assumes that including additional covariates would explain away the disparity; and from *selection bias*, which attributes differential outcomes to non-random sampling. Structural encoding is more fundamental: it holds even when the feature set is complete (no omitted variables), the sample is representative (no selection bias), and the model is optimal (no algorithmic bias). It operates through the joint distribution of all features simultaneously, as evidenced by the SES-removal ablation's null result and the decomposition's finding that cognitive scores alone are associated with 89–129% of the gap. The theoretical contribution is thus to identify a specific,

empirically measurable pathway by which Selbst et al. (2019)'s abstraction traps manifest in deployed systems: the statistical encoding of institutional stratification in the data-generating process, which no downstream algorithmic intervention can fully reverse.

We anticipate the objection that structural encoding merely *renames* what sociologists of education have long called stratification. The distinction is analytic, not substantive: existing stratification theory tells us that inequality exists in educational outcomes; structural encoding tells us *how much* of a specific algorithmic disparity is attributable to specific distributional properties of specific features—and, critically, how much is not. That quantification changes the governance conversation. Without it, policymakers face a diffuse claim ("the system is unequal") that provides no lever for intervention design. With it, they know that kindergarten cognitive scores are associated with 89–129% of the TPR gap, that SES is associated with 29–47%, and that the model itself compresses rather than amplifies the resulting disparity. These magnitudes are actionable: they identify the upstream conditions (early childhood cognitive development, socioeconomic resources) whose modification would, on this evidence, reduce algorithmic unfairness more effectively than any model-level intervention. Structural encoding thus translates stratification theory into a diagnostic framework for algorithmic governance—specifying not just *that* inequality is encoded but *where*, *how much*, and *through which features*.

Decomposing Persistent Disparities

The counterfactual decomposition provides the most direct quantitative evidence for the structural encoding argument. When Hispanic students' features were aligned to the White distribution—effectively asking "what would the model do if Hispanic students had the same baseline cognitive scores, SES, and other characteristics as White students?"—the structural component was associated with more than the entirety of the observed TPR gap: 151.7% for reading and 153.6% for math. The structural component exceeds 100% because the model-induced component is negative: the model and threshold mechanics actually *compress* the gap by approximately 0.105 in both domains, partially counteracting the distributional divergence.

This pattern has a precise interpretation. Hispanic students in the bottom quartile of 5th-grade achievement tend to have lower kindergarten cognitive scores and lower SES than their White counterparts in the same outcome stratum. These more extreme feature values produce higher predicted risk probabilities, pushing more Hispanic at-risk students above

the $t = 0.50$ decision threshold. When we give Hispanic students White-like features via quantile alignment, their predicted probabilities shift downward, and many who were above the threshold fall below it—producing a counterfactual TPR (0.099 for reading, 0.129 for math) that is actually *lower* than the observed White TPR (0.204, 0.234). The negative model-induced component reflects the fact that, conditional on having identical feature distributions, the model would predict Hispanic students as *less* at-risk than White students—likely because the counterfactual alignment eliminates correlated disadvantage across multiple features simultaneously.

The feature-level decomposition identifies the specific features through which structural encoding operates statistically. The four kindergarten cognitive scores are the dominant contributors, collectively associated with 88.7% of the reading gap and 128.9% of the math gap (with domain-specific scores contributing most in their respective outcome domains). SES is associated with 28.6% (reading) and 46.6% (math), but—as the SES-removal ablation separately confirmed—this is partially redundant with cognitive scores, which themselves reflect socioeconomic stratification (Reardon 2011). The convergence of these three analyses—algorithmic convergence, SES-removal ablation, and counterfactual decomposition—constitutes a triangulated case that the TPR disparity is a predictable consequence of educational inequality encoded in early childhood data, not an artifact of model design, feature selection, or threshold choice.

This finding has an important implication for fairness intervention design. Because the disparity originates in the distributional properties of the features rather than in model behavior, interventions that modify the model (constrained training, threshold adjustment, alternative algorithms) cannot fully resolve the gap without either sacrificing accuracy or introducing compensating distortions. The decomposition thus provides an empirical basis for the governance reframing argued above: equitable prediction requires upstream changes in the conditions that produce unequal feature distributions—that is, changes in the institutional contexts that shape early childhood cognitive development.

Implications for Educational Practice

The most immediate practical implication is that single-metric fairness assessments are dangerously insufficient. Our model simultaneously passed equal opportunity, failed equalized odds, exhibited severe calibration disparities, and showed intersectional blind spots—a combination that would not be detected by the standard practice of checking one

or two fairness criteria before deployment. Intersectional auditing is particularly critical: the suggestive pattern that high-SES Black students may be systematically under-identified would be invisible to any analysis examining race or SES in isolation (Buolamwini and Gebru 2018; Kearns et al. 2018). Schools adopting algorithmic EWS should require comprehensive, multi-dimensional fairness audits as a deployment prerequisite.

The sensitivity of fairness to threshold choice carries a deeper lesson: "at-risk" is not a technical parameter but a policy decision with equity consequences that should involve stakeholder input. Our results show that equal opportunity compliance holds at the 25th percentile but fails at the 10th, 20th, and 30th—meaning that the choice of threshold implicitly selects which equity criterion the system satisfies. Districts deploying EWS should therefore treat threshold selection as an explicit policy design decision: a lower threshold (e.g., 10th percentile) targets a smaller, higher-need population but may produce larger group disparities, while a higher threshold captures more students but dilutes resources. Making this trade-off transparent—rather than defaulting to a convenient percentile—would allow stakeholders to weigh sensitivity against equity in light of local resources and priorities. Similarly, the persistence of fairness gaps despite improved accuracy implies a genuine trade-off in prediction timing. Earlier predictions (kindergarten) enable earlier intervention but with lower accuracy; later predictions (through 3rd grade) improve accuracy but narrow the intervention window and may worsen fairness gaps. These are not engineering problems with technical solutions—they are value judgments that require deliberation among educators, families, and policymakers. Predictive models should inform, not replace, human judgment, particularly for demographic subgroups where the model is poorly calibrated.

Generalizability to Post-Pandemic Contexts

The ECLS-K:2011 cohort (2010–2016) preceded the COVID-19 pandemic, which produced learning losses that were both substantial and sharply stratified by race and socioeconomic status (Fahle et al. 2023). Evidence from standardized assessments indicates that achievement gaps widened considerably during 2020–2022, with low-income students and students of color experiencing the steepest declines. This has two implications for the generalizability of our findings. First, the SES-achievement gradient that underlies the "poverty detector" pattern likely *intensified* post-pandemic, suggesting that the fairness disparities we document represent a *floor* rather than a ceiling

for contemporary early warning systems. Second, models trained on pre-pandemic data would face distributional shift when applied to post-pandemic cohorts, as the statistical relationships between early childhood predictors and later outcomes have been disrupted. Future fairness audits should explicitly test cross-cohort generalization across the pre-/post-pandemic boundary, as the ECLS-K:2011 findings may understate the severity of fairness failures in current educational contexts.

Limitations

This study has several limitations:

- **Public-use data constraints:** The public-use ECLS-K:2011 file has some variables suppressed or top-coded to protect confidentiality, potentially limiting predictive power.

- **Missing data and attrition:** Our primary analysis used complete cases ($N = 9{,}104$), representing 50% of the original cohort. Attrition analysis revealed that dropouts had significantly lower baseline cognitive scores (Cohen's $d \geq 0.20$ on all cognitive variables) and were disproportionately from minority and lower-SES backgrounds (e.g., White representation: 53.4% among completers vs. 40.0% among dropouts). MICE sensitivity analysis on the full sample ($N = 18{,}151$) showed substantially higher absolute TPR for all groups (Black: $0.30 \rightarrow 0.53$; Hispanic: $0.39 \rightarrow 0.55$; White: $0.16 \rightarrow 0.37$), with TPR ratios moving toward parity but the pattern of differential performance persisting. IPW reweighting produced virtually identical results (AUC = 0.848), with well-behaved weights (mean = 0.77, SD = 0.08, range 0.58–1.03). While these sensitivity analyses support the robustness of our primary conclusions, the analytic sample likely under-represents the most disadvantaged students, and the true magnitude of fairness disparities may be larger than reported.

- **Single pre-pandemic cohort:** The ECLS-K:2011 followed a single cohort (2010–2016) that preceded the COVID-19 pandemic. Post-pandemic learning losses were substantial and sharply stratified by race and SES (Fahle et al. 2023), suggesting that the SES-achievement gradient underlying our findings has likely steepened. Models trained on pre-pandemic data would face distributional shift when applied to contemporary cohorts. Future audits should test cross-cohort generalization across the pre-/post-pandemic boundary.

- **Binary outcome:** We operationalized risk as a binary threshold ($<$25th percentile). Alternative operationalizations—as demonstrated by our sensitivity analysis—yield different fairness results.

- **Small subgroup sizes:** The Asian subgroup ($N = 8$ in the test set) is too small for any reliable inference and should not inform policy conclusions. Intersectional subgroups with 3–6 positive cases (e.g., Black Q4 with $\approx$6 at-risk students) fall below thresholds for reliable estimation; the observed TPR = 0% is not statistically distinguishable from the overall model TPR ($p = 0.41$). These patterns are suggestive but require replication with larger, purpose-sampled datasets before drawing definitive conclusions.

- **Fairness scope:** Our audit focused on group fairness, calibration fairness, and intersectional fairness. We did not evaluate individual fairness (Dwork et al. 2012) or counterfactual fairness methods, which represent complementary perspectives on algorithmic equity now increasingly discussed in the ML fairness literature.

- **Stylized policy simulation:** Our policy allocation simulation uses assumed costs ($2,000 per intervention, $5,000 per missed student) that are illustrative rather than empirically estimated. Actual costs vary substantially by district context, intervention type, and outcome horizon. More granular cost-effectiveness analysis with district-specific parameters would strengthen the practical relevance of these findings.

- **Limited mitigation methods:** We examined post-hoc threshold adjustment and two in-processing constraint formulations (Equalized Odds and TPR Parity). Other approaches—such as adversarial debiasing, pre-processing methods (e.g., reweighting or resampling training data), or multi-objective optimization—may achieve different accuracy-fairness trade-offs and warrant future investigation.

- **Bootstrap precision:** We used 500 bootstrap iterations for confidence intervals, which provides adequate stability for the metrics reported here. However, future work with narrower confidence intervals may benefit from 1,000+ iterations.

- **Temporal design:** Our temporal analysis held the outcome constant (5th grade) while varying inputs. A complementary approach varying both inputs and outcomes would provide additional insight.

Future Directions

Several directions for future research emerge from this study:

- **Extended fairness-constrained methods:** Our initial evaluation of Fairlearn's Equalized Odds and TPR Parity constraints revealed substantial AUC penalties (0.21–0.23 points). Future work should evaluate alternative in-processing approaches—such as adversarial debiasing, pre-processing reweighting, and fairness-aware ensemble methods—that may achieve more favorable accuracy-fairness trade-offs.

- **Causal fairness methods:** Approaches that distinguish between legitimate and illegitimate predictive pathways could help identify which features contribute to fairness disparities through discriminatory versus non-discriminatory mechanisms.

- **Multi-objective optimization:** Jointly maximizing accuracy and minimizing group fairness disparities during training represents a promising approach to balancing competing objectives.

- **Restricted-use data:** Replication with restricted-use ECLS data, which contains additional variables suppressed in the public-use file, could improve both predictive power and fairness.

- **External validation:** Replication on independent cohorts—such as the earlier ECLS-K:1998, state-level administrative data, or district-level EWS deployments—would test the generalizability of these fairness patterns beyond the ECLS-K:2011 sample.

- **District-level cost-effectiveness:** Our stylized policy simulation demonstrates the distributive consequences of threshold choices but uses assumed cost parameters. Future work should partner with districts to estimate empirical intervention costs and quantify the return on investment of equitable versus efficiency-maximizing allocation strategies.

- **Intervention studies:** Ultimately, the value of EWS depends on whether they improve outcomes. Randomized studies examining the causal effect of EWS-informed interventions, with attention to differential effects across groups, are needed.

Conclusion

When seven algorithms converge on nearly identical performance and exhibit the same pattern of fairness failures, the conclusion is clear: within the available feature space and outcome definition, switching algorithms does not resolve fairness disparities. The inequities are embedded in the features, outcome definitions, and social

stratification that shape the training data, and no amount of architectural sophistication alone will resolve them. SES-removal ablation confirms this structural interpretation: the "poverty detector" pattern persists even without the explicit SES feature, because early cognitive scores already encode socioeconomic stratification. Counterfactual decomposition quantifies this encoding precisely: differences in baseline feature distributions between Hispanic and White students account for more than 100% of the observed TPR gap, with kindergarten cognitive scores as the dominant transmission channel—the model inherits, rather than creates, the disparity. In-processing fairness constraints can reduce disparities, but only at a substantial cost to predictive accuracy (AUC penalty of 0.21–0.23 points)— a trade-off that may itself be unacceptable when missed at-risk students face real consequences. The persistence of fairness gaps despite improved accuracy—more data improves discrimination without resolving equity concerns, an outcome consistent with known impossibility results but not yet documented in longitudinal educational prediction— reinforces this conclusion.

These findings carry a practical imperative. As predictive analytics become standard in K-12 settings, multidimensional fairness auditing must become a deployment prerequisite. The tools exist—bootstrap uncertainty quantification, calibration analysis, intersectional auditing, SHAP explainability—but they must be used together, not in isolation. A model that passes one fairness criterion may fail others severely. The policy conversation must shift from "which algorithm?" to "what data, what definitions, and what institutional practices produce these disparities?" Only then can early warning systems fulfill their promise of helping every student, not just those the data makes easy to see.

Ethics Statement

This study uses the Early Childhood Longitudinal Study, Kindergarten Class of 2010–11 (ECLS-K:2011) public-use data file, which is freely available from the National Center for Education Statistics (NCES) without restricted-use license or institutional review board (IRB) approval. All data are fully de-identified by NCES prior to public release; no individual students, families, or schools can be identified. Because this research involves secondary analysis of existing de-identified public-use data, it is exempt from IRB review under 45 CFR 46.104(d)(4). This determination was confirmed by the University of Windsor Research Ethics Board. No participants were contacted or recruited for this study.

We acknowledge that predictive models for educational risk carry ethical implications beyond technical performance. Our fairness audit is intended to inform responsible development and deployment practices—not to endorse the use of any particular model in high-stakes educational decision-making without further validation, stakeholder engagement, and ongoing monitoring.

Use of Generative AI

Generative AI (Claude, Anthropic) was used as a coding assistant during data analysis pipeline development, figure generation, and manuscript formatting. All scientific content—including research design, interpretation of results, and substantive writing—was produced by the authors. The authors reviewed and verified all AI-assisted outputs for accuracy. No AI tool was used to generate or alter the study's data, statistical analyses, or scholarly conclusions.

Data Availability Statement

The ECLS-K:2011 public-use data file is freely available from the National Center for Education Statistics at https://nces.ed.gov/ecls/dataproducts.asp. All analysis code, configuration files, and scripts to reproduce the results reported in this paper are available at https://github.com/kismatraj/ecls-fairness-study. The repository includes a complete pipeline (scripts/run_pipeline.py) that reproduces all results, figures, and tables from the raw ECLS data file.

Supplemental material

Supplemental material for this article is available online.

References

Agarwal A, Beygelzimer A, Dudík M, Langford J and Wallach H (2018) A reductions approach to fair classification. In: *Proceedings of the 35th International Conference on Machine Learning*. pp. 60–69.

Aguiar E, Lakkaraju H, Bhanpuri N, Miller D, Yuber B and Addison KL (2015) Who, when, and why: A machine learning approach to prioritizing students at risk of not graduating high school on time. In: *Proceedings of the Fifth International Conference on Learning Analytics and Knowledge*. ACM, pp. 93–102.

Baker RS and Inventado PS (2014) Educational data mining and learning analytics. In: *Learning Analytics*. Springer, pp. 61–75.

Barocas S, Hardt M and Narayanan A (2019) *Fairness and Machine Learning: Limitations and Opportunities*. fairmlbook.org.

Bird S, Dudík M, Edgar R, Horn B, Lutz R, Milan V, Sameki M, Wallach H and Walker K (2020) Fairlearn: A toolkit for assessing and improving fairness in AI. In: *Microsoft Technical Report MSR-TR-2020-32*.

Bowker GC and Star SL (2000) *Sorting Things Out: Classification and Its Consequences*. MIT Press.

Buolamwini J and Gebru T (2018) Gender shades: Intersectional accuracy disparities in commercial gender classification. In: *Proceedings of the 1st Conference on Fairness, Accountability and Transparency*. pp. 77–91.

Chen T and Guestrin C (2016) XGBoost: A scalable tree boosting system. In: *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, pp. 785–794.

Chouldechova A (2017) Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. *Big Data* 5(2): 153–163.

Crenshaw K (1989) Demarginalizing the intersection of race and sex: A black feminist critique of antidiscrimination doctrine, feminist theory and antiracist politics. *University of Chicago Legal Forum* 1989(1): 139–167.

Dwork C, Hardt M, Pitassi T, Reingold O and Zemel R (2012) Fairness through awareness. In: *Proceedings of the 3rd Innovations in Theoretical Computer Science Conference*. pp. 214–226.

Efron B and Tibshirani RJ (1993) *An Introduction to the Bootstrap*. Chapman & Hall/CRC.

Eubanks V (2018) *Automating Inequality: How High-Tech Tools Profile, Police, and Punish the Poor*. St. Martin's Press.

Fahle EM, Kane TJ, Patterson T, Reardon SF, Staiger DO and Stuart EA (2023) School district and community factors associated with learning loss during the COVID-19 pandemic. Technical report, Center for Education Policy Research,

Harvard University.

Gardner J, Brooks C and Baker R (2019) Evaluating the fairness of predictive student models through slicing analysis. In: *Proceedings of the 9th International Conference on Learning Analytics & Knowledge*. ACM, pp. 225–234.

Green B (2022) The flaws of policies requiring human oversight of government algorithms. *Computer Law & Security Review* 45: 105681.

Grinsztajn L, Oyallon E and Varoquaux G (2022) Why do tree-based models still outperform deep learning on typical tabular data? In: *Advances in Neural Information Processing Systems*, volume 35. pp. 507–520.

Heckman JJ (2006) Skill formation and the economics of investing in disadvantaged children. *Science* 312(5782): 1900–1902.

Ke G, Meng Q, Finley T, Wang T, Chen W, Ma W, Ye Q and Liu TY (2017) LightGBM: A highly efficient gradient boosting decision tree. In: *Advances in Neural Information Processing Systems*, volume 30. pp. 3146–3154.

Kearns M, Neel S, Roth A and Wu ZS (2018) Preventing fairness gerrymandering: Auditing and learning for subgroup fairness. In: *Proceedings of the 35th International Conference on Machine Learning*. pp. 2564–2572.

Kizilcec RF and Lee H (2022) Algorithmic fairness in education. In: *The Ethics of Artificial Intelligence in Education*. Routledge, pp. 174–202.

Kleinberg J, Mullainathan S and Raghavan M (2016) Inherent trade-offs in the fair determination of risk scores. *arXiv preprint arXiv:1609.05807* .

Lakkaraju H, Aguiar E, Rich C, Hansen D, Miller D, Yuber B and Addison KL (2015) A machine learning framework to identify students at risk of adverse academic outcomes. In: *Proceedings of the 21st ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, pp. 1909–1918.

Lundberg SM and Lee SI (2017) A unified approach to interpreting model predictions. In: *Advances in Neural Information Processing Systems*, volume 30. pp. 4765–4774.

Mehrabi N, Morstatter F, Saxena N, Lerman K and Galstyan A (2021) A survey on bias and fairness in machine learning. *ACM Computing Surveys* 54(6): 1–35.

Molnar C (2020) *Interpretable Machine Learning: A Guide for Making Black Box Models Explainable*. Lulu.com.

Pleiss G, Raghavan M, Wu F, Kleinberg J and Weinberger KQ (2017) On fairness and calibration. In: *Advances in Neural Information Processing Systems*, volume 30. pp. 5680–5689.

Prokhorenkova L, Gusev G, Vorobev A, Dorogush AV and Gulin A (2018) CatBoost: Unbiased boosting with categorical features. In: *Advances in Neural Information Processing Systems*,

volume 31. pp. 6638–6648.

Reardon SF (2011) *The Widening Academic Achievement Gap Between the Rich and the Poor: New Evidence and Possible Explanations*. Russell Sage Foundation.

Reardon SF and Portilla XA (2016) Recent trends in income, racial, and ethnic school readiness gaps at kindergarten entry. *AERA Open* 2(3): 1–18.

Rubin DB (1987) *Multiple Imputation for Nonresponse in Surveys*. John Wiley & Sons.

Seaman SR and White IR (2013) Review of inverse probability weighting for dealing with missing data. *Statistical Methods in Medical Research* 22(3): 278–295.

Selbst AD, Boyd D, Friedler SA, Venkatasubramanian S and Vertesi J (2019) Fairness and abstraction in sociotechnical systems. In: *Proceedings of the Conference on Fairness, Accountability, and Transparency*. ACM, pp. 59–68.

Verma S and Rubin J (2018) Fairness definitions explained. In: *Proceedings of the International Workshop on Software Fairness*. pp. 1–7.

Williamson B (2017) *Big Data in Education: The Digital Future of Learning, Policy and Practice*. SAGE.

Yu R, Lee H and Kizilcec RF (2020) Should college dropout prediction models include protected attributes? In: *Proceedings of the Seventh ACM Conference on Learning@ Scale*. ACM, pp. 91–100.

Technical Details

Software and Reproducibility

All analyses were conducted in Python 3.12 using the following packages:

- scikit-learn $\geq$ 1.4.0 (including HistGradientBoosting-Classifier)
- xgboost $\geq$ 2.0.0
- lightgbm $\geq$ 4.3.0
- catboost $\geq$ 1.2.0
- shap $\geq$ 0.45.0
- fairlearn $\geq$ 0.10.0 (Bird et al. 2020)
- pandas, numpy, matplotlib, seaborn

Random seed was set to 42 for all stochastic operations. Code and data processing scripts are available in the project repository.

Model Hyperparameters

The final elastic net model used the following hyperparameters selected via 5-fold cross-validation:

- Regularization strength (α): 0.01
- L1 ratio: 0.5
- Maximum iterations: 1000

Cross-validation AUC scores ranged from 0.832 to 0.842 across folds, indicating stable performance.

Missing Data

Table 16 presents missing data rates for key variables. See Section 3.8 for the full missing data sensitivity analysis methodology and Section 4.7 for results.

The high rate of missing outcome data (37%) reflects sample attrition over the longitudinal study. As documented in Section 4.7, completers differed systematically from dropouts on all baseline cognitive variables (Cohen's $d = 0.22$–0.30), with White students over-represented among completers. MICE and IPW sensitivity analyses confirmed that the pattern of differential performance persists after correcting for attrition.

Table 16. Missing Data Rates

Variable	N Missing	% Missing
5th Grade Reading (Outcome)	6,724	37.0%
Executive Function (X6DCCSSCR)	4,379	24.1%
1st Grade Approaches to Learning	4,708	25.9%
Fall K Reading	2,482	13.7%
SES Quintile	2,063	11.4%
Home Language	2,106	11.6%
Spring K Reading	965	5.3%

Supplementary Results

All supplementary figures, tables, and the full math outcome comparison are provided in the Online Supplementary Materials document, which includes: PPV by group with confidence intervals (Figure S1); SHAP vs. permutation importance comparison (Figure S2, Table S1); SHAP importance by racial/ethnic group (Figure S3); temporal fairness trends (Figures S4–S6); permutation importance with bootstrap CIs (Table S2); calibration error across temporal scenarios (Table S3); detailed sensitivity analysis (Table S4); temporal fairness group-level metrics (Table S5); and reading vs. math outcome comparison (Tables S6–S8).