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

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 critical assumption underlying these early warning systems (EWS) has gone largely untested—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 others (Mehrabi et al. 2021), and that fairness itself is multi-dimensional: 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. Yet few studies have examined whether EWS perform equitably in longitudinal contexts, across multiple fairness dimensions simultaneously, or using the modern toolkit of calibration analysis, intersectional auditing, and uncertainty quantification.

This study fills that gap through a comprehensive policy 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 post-hoc bias mitigation.

Our goal is not to introduce novel algorithms but to conduct a rigorous *policy audit* demonstrating that fairness failures are robust across seven algorithmic approaches. By showing that the source of inequity is structural rather than algorithmic, 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 comprehensive, multi-dimensional fairness audits of longitudinal educational prediction models using nationally representative data, examining group fairness, calibration fairness, and intersectional fairness simultaneously.
- We demonstrate that the convergence of seven ML algorithms—including three state-of-the-art gradient boosting methods—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 × SES) analysis to educational prediction, revealing suggestive patterns of under-identification of high-SES minority students that warrant replication with larger samples.
- We identify a *temporal fairness paradox*: additional longitudinal data from kindergarten through 3rd grade improves accuracy but fails to resolve—and in some cases worsens—fairness disparities. To our knowledge, this is the first empirical demonstration that more data does not naturally produce fairer educational predictions.
- 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.

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.

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.

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 fail catastrophically for specific race-by-SES subgroups (Kearns et al. 2018).

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.

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.

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.

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 IRT-based ability estimate that allows for longitudinal comparisons. In our sample, 15.7% of children 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.

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)

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.

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

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 $n_{estimators} \in \{100, 200\}$, $max_{depth} \in \{3, 5, 7\}$, $learning_{rate} \in \{0.01, 0.1\}$.
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 LightGBM, 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.

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

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 (four-fifths rule).

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

Statistical Parity: Satisfied if positive rate ratios exceed 0.80.

We computed 95% bootstrap confidence intervals for all group-level fairness metrics using 500 bootstrap iterations, 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.

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:

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

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.

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).

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.

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.

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.

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. Outcome attrition was substantial: 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. Missing data rates ranged from 5.3% (spring kindergarten reading) to 37.0% (5th-grade outcome), with executive function (24.1%) and 1st-grade approaches to learning (25.9%) showing intermediate rates. This level of attrition raises the possibility that our primary complete-case results are affected by selection bias if missingness is related to outcomes or protected attributes.

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.

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 legal and ethical concerns analogous to those surrounding affirmative action in other domains. 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). We present threshold optimization as a *diagnostic tool* to illustrate the theoretical bounds of post-hoc equalization, not as a recommended intervention for deployment.

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 18,151 children meeting

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
Missing	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
Missing	2,063	11.4
Sex		
Male	9,273	51.1
Female	8,840	48.7
5th Grade Reading Risk		
At-Risk (¡25th %ile)	2,857	15.7
Not At-Risk	15,294	84.3

inclusion criteria, 15.7% were classified as at-risk in reading by 5th grade, providing adequate prevalence for model training.

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 state-of-the-art 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 highest AUC.

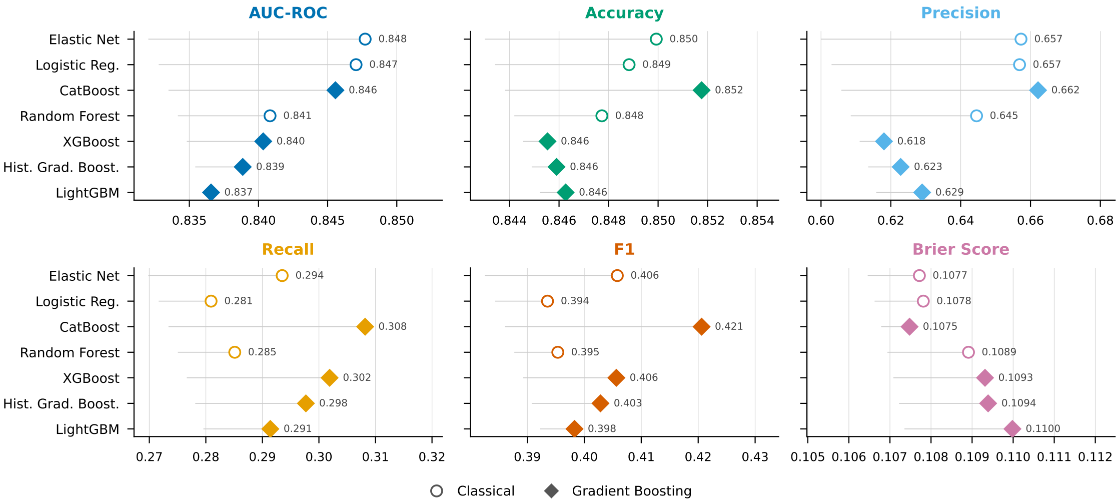
Table 3 presents feature importance from three complementary methods: SHAP values, elastic net coefficients, and permutation importance with bootstrap confidence intervals.

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.

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.

Table 4 presents performance metrics by racial/ethnic group with bootstrap 95% confidence intervals.

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



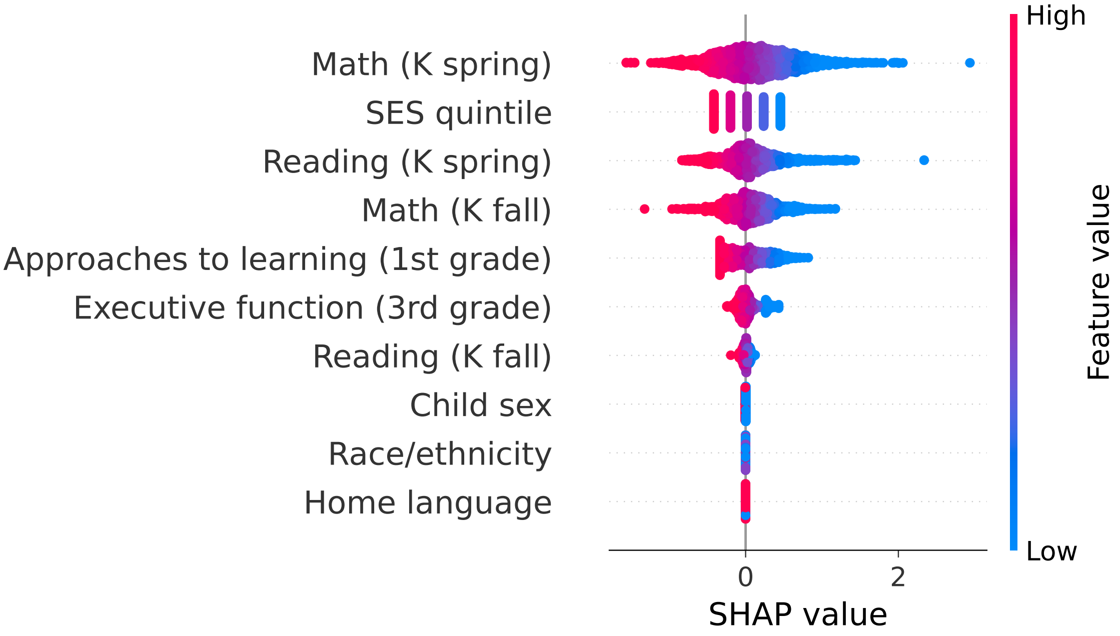
Feature	Mean |SHAP|	EN Coef.	Perm. Imp. [95% CI]
Spring K Math (X2MTHETK)	0.410	0.501	0.046 [0.042, 0.054]
Spring K Reading (X2RTHETK)	0.254	0.350	0.023 [0.018, 0.027]
SES Quintile (X1SESQ5)	0.253	0.303	0.013 [0.007, 0.019]
Fall K Math (X1MTHETK)	0.227	0.288	0.012 [0.007, 0.017]
Approaches to Learning, 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

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.

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

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.

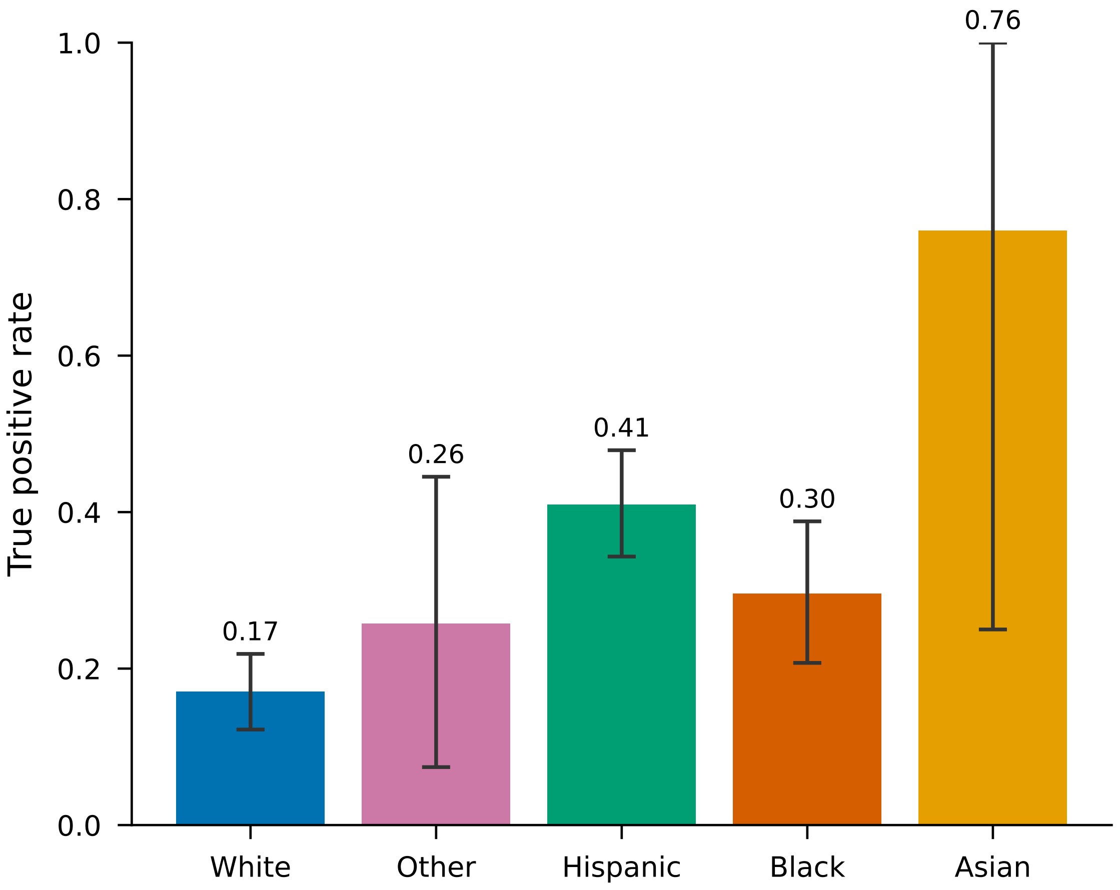
Table 6 presents calibration metrics by demographic group.

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 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.

Table 7 presents fairness metrics for selected race × SES subgroups, revealing patterns invisible in single-attribute analysis.

A suggestive finding was the model's failure to identify any at-risk Black students in the 4th SES quintile (TPR = 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),



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

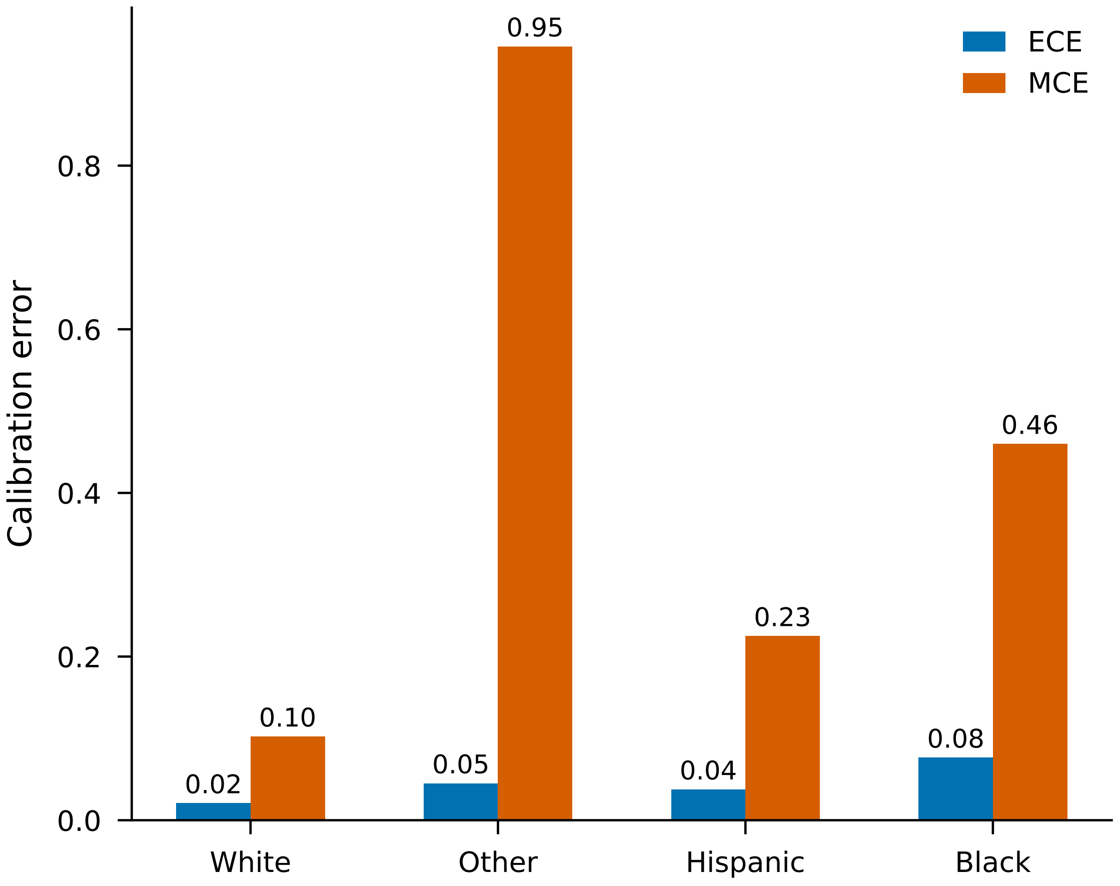
Group	N	ECE	MCE	Brier	ECE Ratio
White	1,462	0.022	0.112	0.082	1.00
Hispanic	623	0.036	0.115	0.155	1.65
Other	225	0.051	0.940	0.090	2.31
Black	300	0.074	0.456	0.162	**3.35**

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.

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. 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

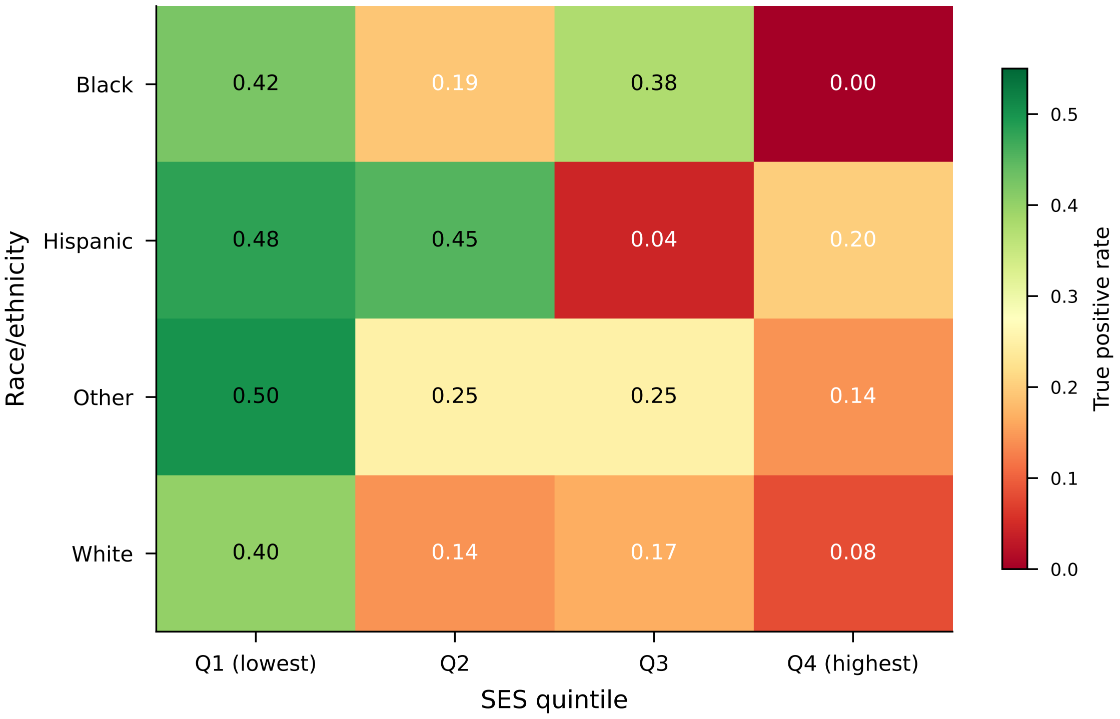


×

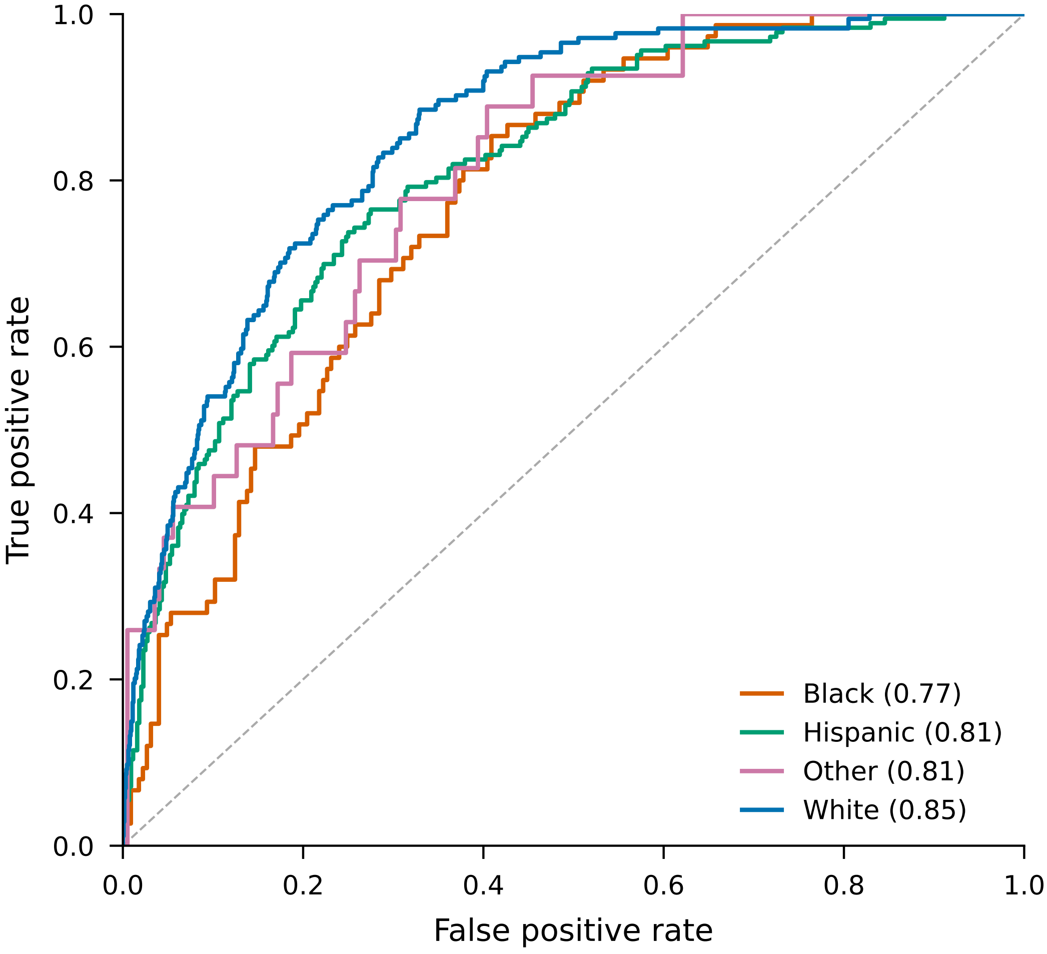
Subgroup	N	Prevalence	TPR	FPR	Accuracy
Low SES (Q1)					
Hispanic Q1	269	38.7%	0.481	0.121	0.725
Black Q1	80	32.5%	0.423	0.204	0.675
White Q1	107	28.0%	0.400	0.117	0.748
Mid SES (Q2–Q3)					
Hispanic Q2	145	29.0%	0.452	0.058	0.800
Black Q3	60	26.7%	0.375	0.068	0.783
White Q3	314	13.4%	0.167	0.004	0.885
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

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

across all scenarios, and calibration disparities persisted. We term this the *temporal fairness paradox*: additional longitudinal data improves accuracy without resolving—and in some cases worsening—fairness disparities. To our knowledge, this is the first empirical demonstration that the widely held assumption that “more data produces fairer predictions” does not hold in longitudinal educational settings. The paradox arises because additional developmental data reinforces the

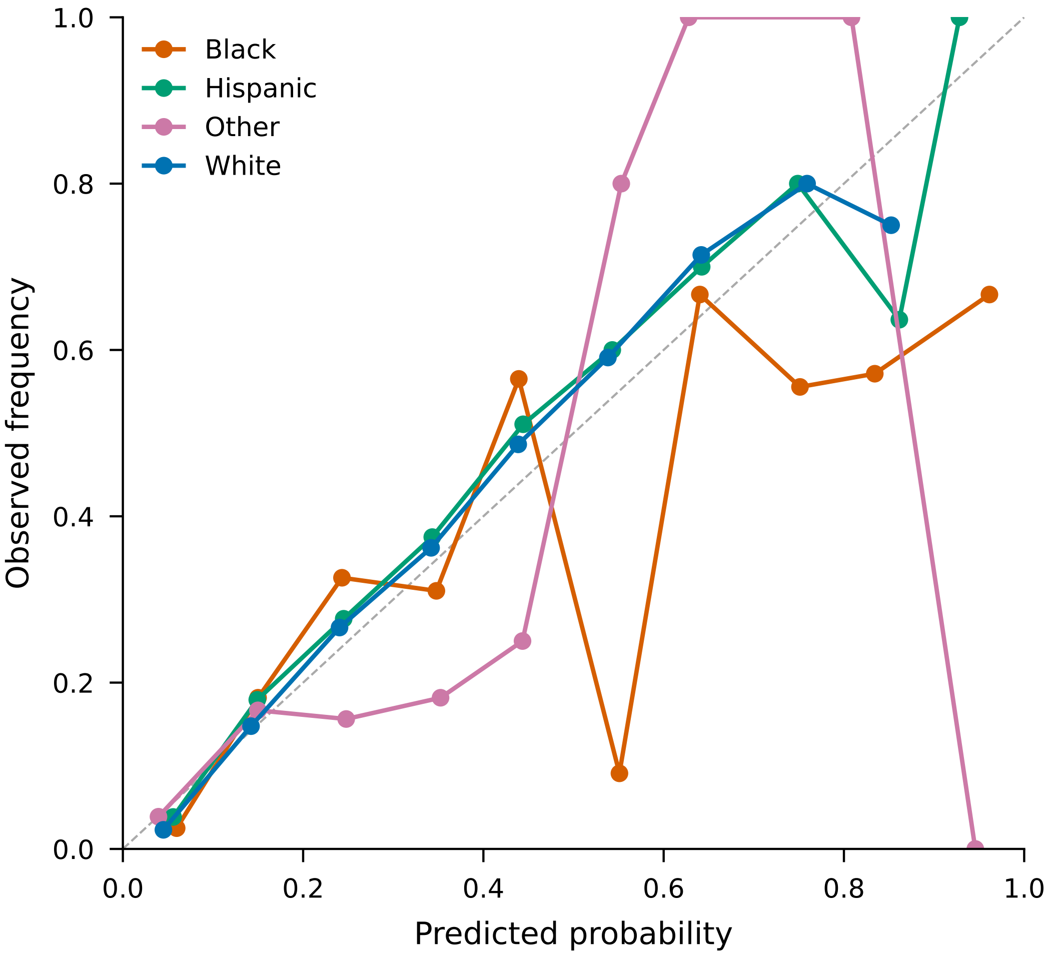


$\times N = 42 \approx p = 0.41 N < 50$



same socioeconomic signals that drive differential performance, rather than introducing orthogonal information that could equalize predictions across groups.

A critical question is whether the fairness findings above are artifacts of the specific threshold used to define "at-risk."



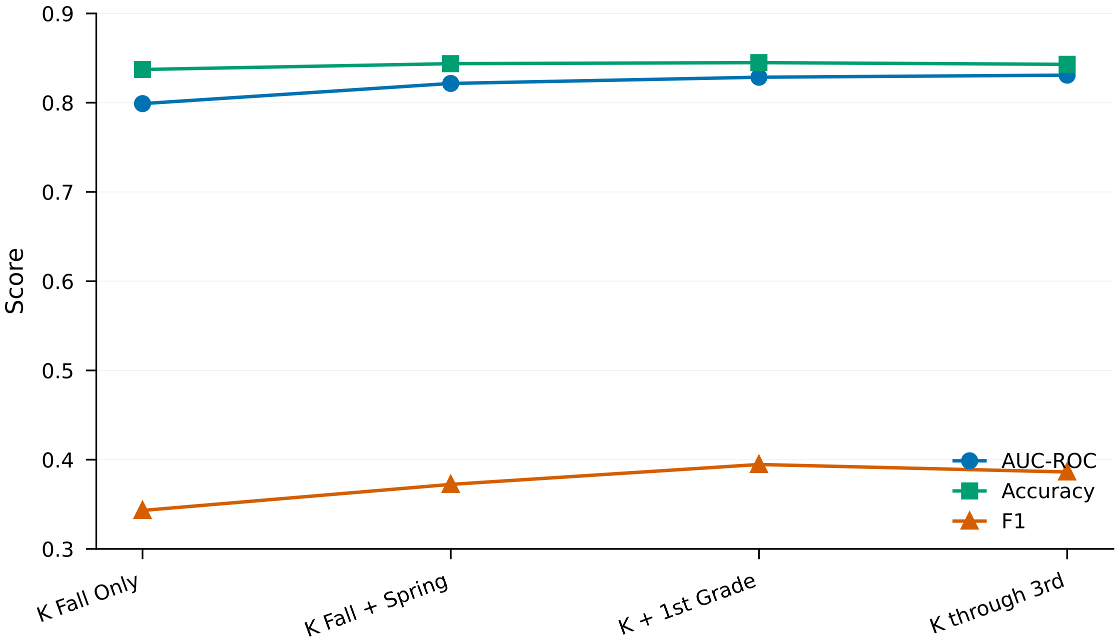
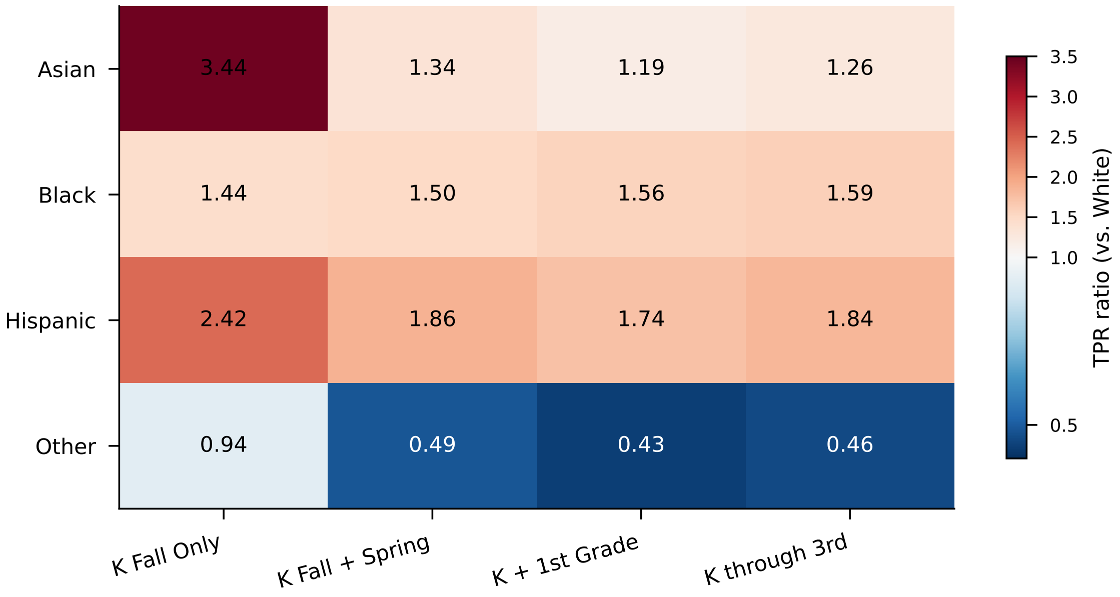


Table 9 presents fairness criteria compliance across four at-risk threshold definitions.

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



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

fragility underscores that fairness assessments cannot be divorced from the operationalization of the outcome, and that claims of fairness compliance are contingent on specific analytical choices.

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.

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 0.29 to 0.53 (+0.24); Hispanic from 0.41 to 0.55 (+0.14); White from 0.17 to 0.37 (+0.20)—indicating that attrition disproportionately removed detectable at-risk students from the analytic sample. TPR ratios moved toward parity (Hispanic-White: 2.46 → 1.48; Black-White: 1.83 → 1.42), though the absolute Black-White TPR gap increased slightly (0.16 vs. 0.13 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.

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. This convergence between IPW and unweighted results provides additional evidence that the fairness findings are robust to selection bias adjustment.

As a secondary analysis, we compared reading and math outcomes. Math prediction (AUC = 0.867, best model: Logistic Regression) substantially outperformed reading prediction (AUC = 0.848, best model: Elastic Net). Fairness patterns were broadly similar across domains, with Hispanic and Black students showing higher TPR than White students for both outcomes. However, math prediction uniquely exhibited disparate impact for the Other racial group (TPR ratio = 0.24 vs. White), a pattern not observed in reading. This suggests that the domain of the outcome can affect which groups experience adverse fairness impacts, reinforcing the importance of outcome-specific fairness auditing. Full results are presented in Appendix B.

We applied threshold optimization to equalize TPR across groups, targeting the overall model TPR of 0.283. Table 10 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 catastrophic 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.

The central finding of this study is that fairness failures in early childhood risk prediction are structural, not algorithmic. 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 architecture (Grinsztajn et al. 2022).

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 *temporal fairness paradox*—that additional longitudinal data improved accuracy (AUC 0.799 to 0.831) without resolving fairness disparities—undermines the expectation that more information naturally produces 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 post hoc 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. This outcome, a direct consequence of the impossibility results of Chouldechova (2017), underscores that group-specific thresholds raise both practical and legal concerns as a form of race-conscious classification. Finally, MICE sensitivity

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

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.

The fairness disparities documented above demand explanation. Why does a model that excludes race as a predictor nonetheless produce racially differential outcomes? 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 |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 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).

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. Similarly, the temporal fairness paradox 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.

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.

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: 0.29 → 0.53; Hispanic: 0.41 → 0.55; White: 0.17 → 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.
- **Post-hoc mitigation only:** We examined only post-hoc threshold adjustment. In-processing methods (e.g., adversarial debiasing, fairness constraints during training) might achieve better accuracy-fairness trade-offs.
- **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.

Several directions for future research emerge from this study:

- **In-processing fairness methods:** Future work should evaluate constraint-based methods that incorporate fairness during model training, potentially achieving better accuracy-fairness trade-offs than post-hoc threshold adjustment.
- **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.
- **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.

When seven algorithms converge on nearly identical performance and exhibit the same pattern of fairness failures, the conclusion is clear: the source of inequity in educational risk prediction is structural, not algorithmic. No amount of algorithmic sophistication will resolve disparities that are embedded in the features, outcome definitions, and social stratification that shape the training data. The *temporal fairness paradox*—that more data improves accuracy without resolving fairness—reinforces this conclusion.

These findings carry a practical imperative. As predictive analytics become standard in K-12 settings, multi-dimensional 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 catastrophically. 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.

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). 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.

The ECLS-K:2011 public-use data file is freely available from the National Center for Education Statistics at . All analysis code, configuration files, and scripts to reproduce the results reported in this paper are available at [GitHub repository URL]. A permanent archival copy of the code is deposited at [Zenodo DOI]. The repository includes a complete pipeline () that reproduces all results, figures, and tables from the raw ECLS data file.

[Removed for double-blind review.]

The authors declare no competing interests.

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.

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.

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.

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.

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.

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.

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.

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

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.

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

- scikit-learn $\geq$ 1.4.0 (including HistGradientBoostingClassifier)
- 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
- 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.

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.

Table 11 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.

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).

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%