Supplementary Materials

Algorithmic Fairness and Temporal Generalization in Early Childhood Risk Prediction: A Multi-Dimensional Audit Using the ECLS-K:2011 Longitudinal Study

[Author Name][1]

[1][Department], [University], [City, State]

February 2026

This document contains supplementary materials for the main manuscript. Section S1 presents supplementary figures not included in the main text. Section S2 provides supplementary tables with detailed results. Section S3 presents the full math outcome comparison referenced in the main paper.

S1. Supplementary Figures

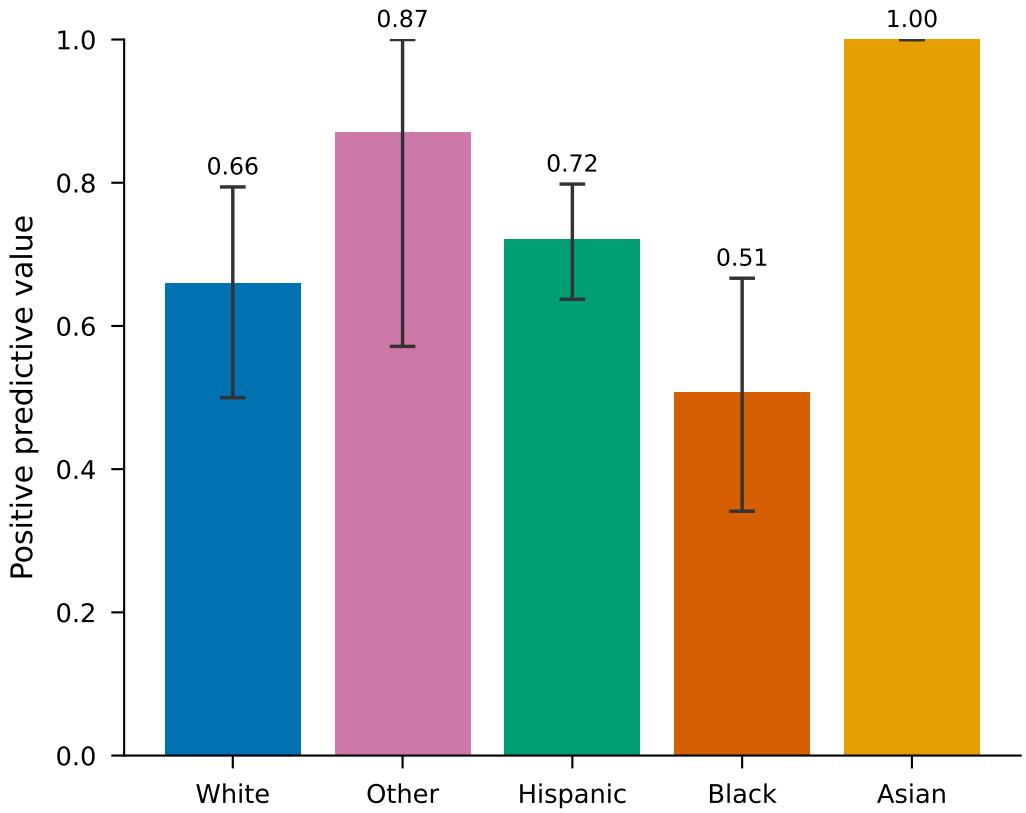

Figure S1. Positive Predictive Value (PPV) by Racial/Ethnic Group with 95% Bootstrap Confidence Intervals. PPV represents the proportion of students flagged as at-risk who are truly at-risk. The Other group shows the highest PPV (0.871) but with wide confidence intervals reflecting small sample size, while the Black group has the lowest PPV (0.508), indicating that nearly half of Black students flagged as at-risk are false positives.

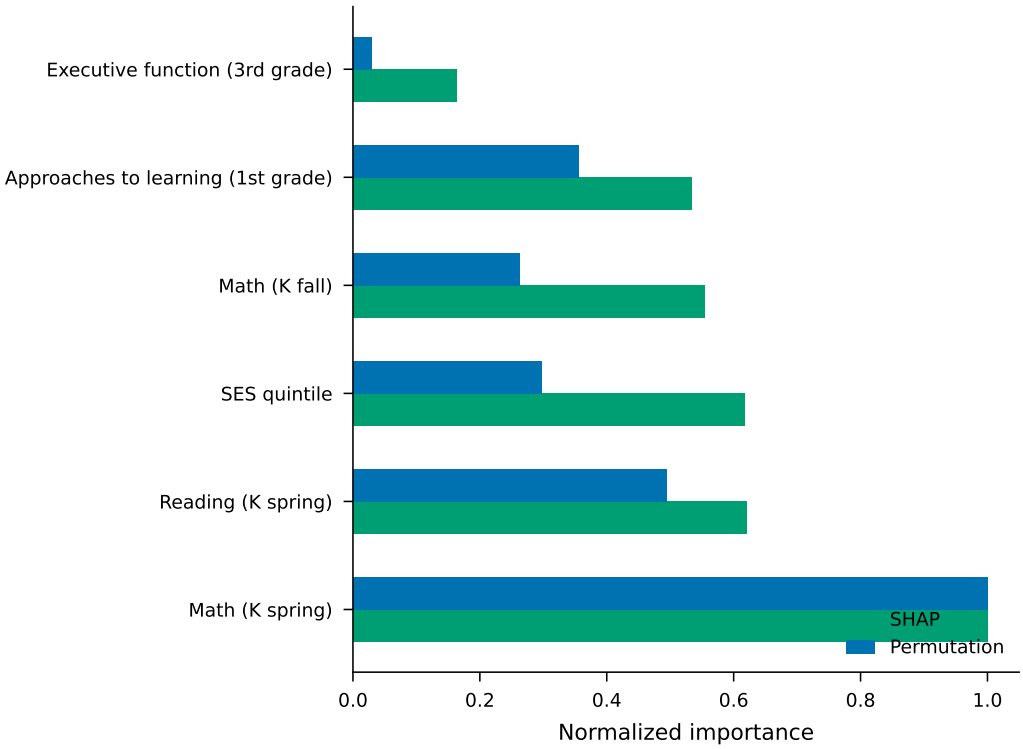

Figure S2. SHAP vs. Permutation Importance Comparison. Normalized importance scores from SHAP (mean absolute SHAP values) and permutation importance are compared across all features. Both methods identify the same top-5 predictors with high agreement (mean agreement = 0.87), with spring kindergarten math (X2MTHETK) as the dominant predictor. Features with zero importance under both methods (race, sex, language, early ATL) were effectively excluded by elastic net regularization.

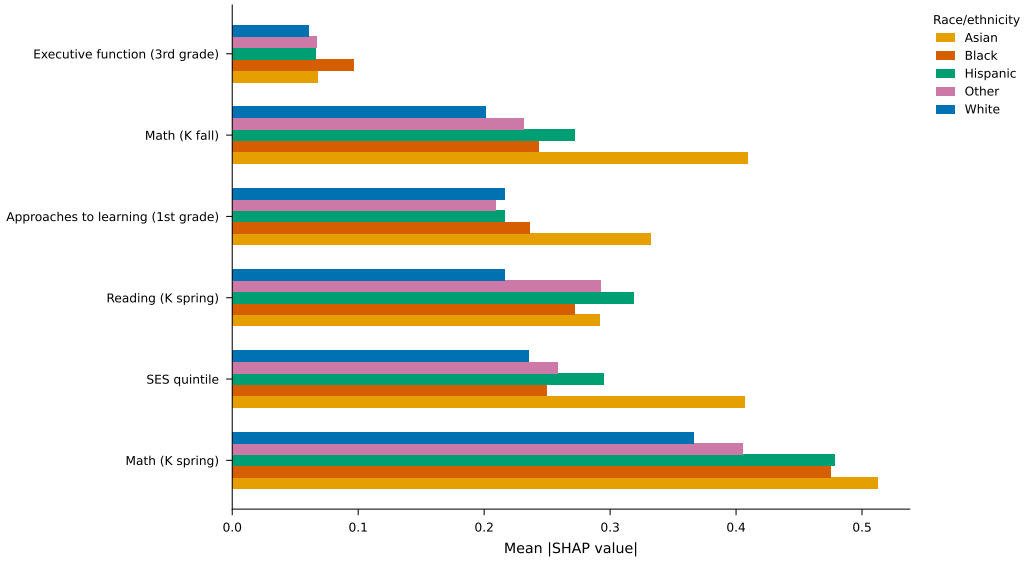

Figure S3. SHAP Feature Importance by Racial/Ethnic Group. Mean absolute SHAP values computed separately for each demographic group reveal whether the model relies on different features for different populations. Top-5 feature rankings are identical across groups, but the magnitude of math score importance is somewhat higher for Black and Hispanic students, suggesting cognitive scores carry relatively more predictive weight for minority students.

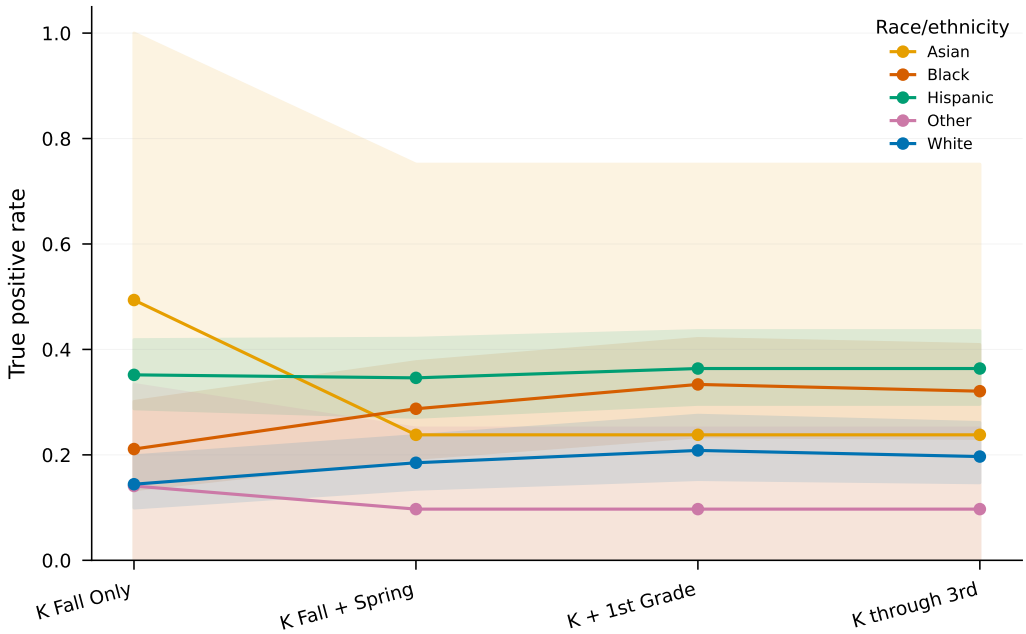

Figure S4. True Positive Rate Trends Across Temporal Scenarios by Racial/Ethnic Group. TPR is plotted for each group across four developmental windows (K Fall Only, K Fall + Spring, K + 1st Grade, K through 3rd). Hispanic students consistently show the highest TPR across all scenarios, while White students show the lowest. The Hispanic–White TPR gap remains substantial regardless of how much longitudinal data is available.

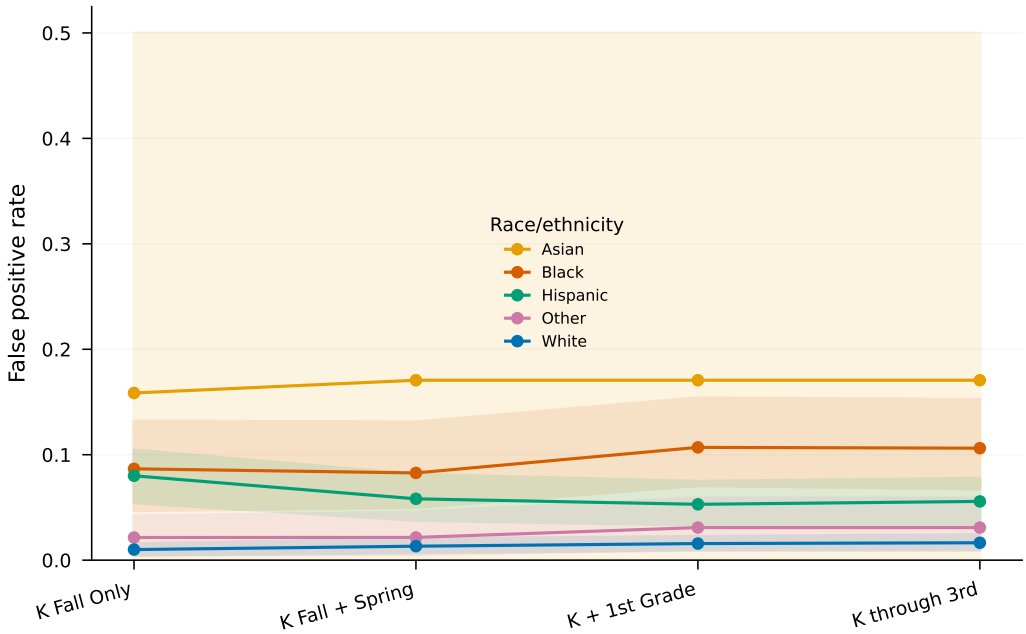

Figure S5. False Positive Rate Trends Across Temporal Scenarios by Racial/Ethnic Group. FPR is plotted for each group across four developmental windows. Black students consistently experience the highest FPR across all scenarios, approximately 8–10 times higher than the White FPR. Additional longitudinal data does not meaningfully reduce these FPR disparities.

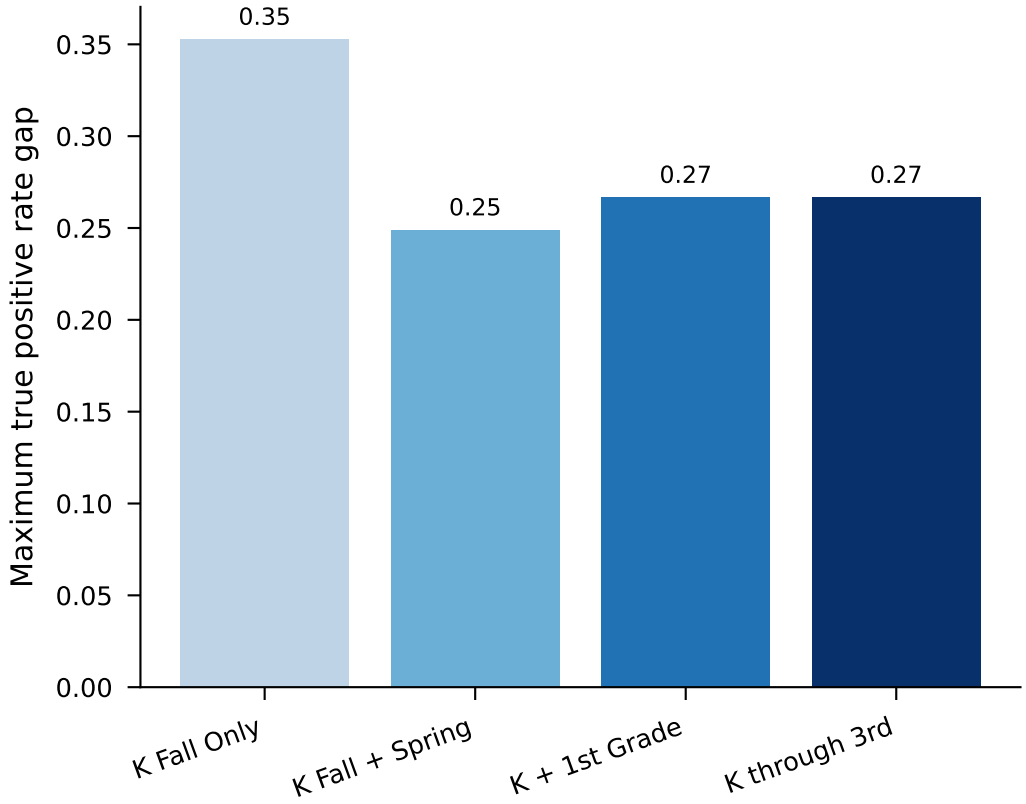

Figure S6. Temporal Fairness Gap. This figure summarizes how fairness disparities evolve across the four temporal prediction windows, illustrating the "temporal paradox" discussed in the main text: additional data improves overall accuracy but does not resolve—and may even exacerbate—fairness disparities across demographic groups.

S2. Supplementary Tables

Table S1. SHAP vs. Permutation Importance Comparison. Normalized importance scores, agreement between methods, and rank differences. Features are ordered by SHAP importance. Agreement is computed as $1 - |SHAP_{norm} - Perm_{norm}|$. Both methods identify X2MTHETK as the dominant predictor with perfect agreement at the top.

Feature	**SHAP (norm.)**	**Perm. (norm.)**	**Agree.**	**SHAP Rk.**	**Perm. Rk.**
X2MTHETK	1.000	1.000	1.000	1	1
X2RTHETK	0.621	0.494	0.873	2	2
X1SESQ5	0.618	0.297	0.680	3	4
X1MTHETK	0.554	0.263	0.709	4	5
X4TCHAPP	0.533	0.356	0.823	5	3
X6DCCSSCR	0.163	0.029	0.866	6	6
X1RTHETK	0.000	0.000	1.000	9.5	9.5
X12LANGST	0.000	0.000	1.000	9.5	9.5
X2TCHAPP	0.000	0.000	1.000	9.5	9.5
X1TCHAPP	0.000	0.000	1.000	9.5	9.5
X_CHSEX_R	0.000	0.000	1.000	9.5	9.5
X_RACETH_R	0.000	0.000	1.000	9.5	9.5

Table S2. Permutation Importance with Bootstrap 95% Confidence Intervals. Importance was computed using 50 bootstrap iterations. Features with importance indistinguishable from zero (CI includes 0) are not significantly predictive. Only six features have non-trivial importance; elastic net regularization excluded all others.

Feature	**Mean Imp.**	**Std.**	**CI 2.5%**	**CI 97.5%**
X2MTHETK	0.046	0.004	0.042	0.054
X2RTHETK	0.023	0.003	0.018	0.027
X4TCHAPP	0.016	0.003	0.012	0.022
X1SESQ5	0.013	0.003	0.007	0.019
X1MTHETK	0.012	0.002	0.007	0.017
X6DCCSSCR	0.001	0.001	−0.001	0.003
X1RTHETK	0.000	0.000	0.000	0.000
X1TCHAPP	0.000	0.000	0.000	0.000
X2TCHAPP	0.000	0.000	0.000	0.000
X_CHSEX_R	0.000	0.000	0.000	0.000
X_RACETH_R	0.000	0.000	0.000	0.000
X12LANGST	0.000	0.000	0.000	0.000

Table S3. Calibration Error Across Temporal Scenarios. Expected Calibration Error (ECE), Maximum Calibration Error (MCE), and Brier score for each demographic group under each temporal prediction scenario. ECE Ratio is computed relative to the White reference group within each scenario. Black students consistently show the highest ECE ratios (5–7× White), and this disparity worsens with additional longitudinal data.

Scenario	Group	N	ECE	MCE	Brier	ECE Ratio
K Fall Only	White	1,458	0.015	0.337	0.088	1.00
	Hispanic	629	0.027	0.089	0.167	1.73
	Other	213	0.021	0.301	0.084	1.37
	Black	301	0.098	0.579	0.195	6.36
K Fall + Spring	White	1,458	0.016	0.246	0.083	1.00
	Hispanic	629	0.039	0.124	0.158	2.47
	Other	213	0.035	0.715	0.084	2.17
	Black	301	0.077	0.262	0.182	4.85
K + 1st Grade	White	1,458	0.017	0.181	0.082	1.00
	Hispanic	629	0.044	0.091	0.156	2.57
	Other	213	0.035	0.742	0.088	2.04
	Black	301	0.099	0.220	0.183	5.79
K through 3rd	White	1,458	0.015	0.142	0.082	1.00
	Hispanic	629	0.041	0.119	0.156	2.71
	Other	213	0.033	0.742	0.086	2.18
	Black	301	0.100	0.224	0.184	6.60

Table S4. Detailed Sensitivity Analysis: Fairness Metrics by At-Risk Threshold and Demographic Group (with 95% Bootstrap Confidence Intervals). This table supplements Table 9 in the main text by providing the group-level TPR, FPR, and PPV values underlying each threshold's fairness criteria assessment.

%ile	Group	N	TPR [95% CI]	FPR [95% CI]	PPV [95% CI]
	White	1,505	0.050 [0.000, 0.106]	0.002 [0.000, 0.005]	0.490 [0.000, 1.000]
	Hispanic	574	0.261 [0.171, 0.353]	0.012 [0.002, 0.022]	0.734 [0.562, 0.947]
10th	Black	302	0.222 [0.108, 0.343]	0.003 [0.000, 0.011]	0.906 [0.714, 1.000]
	Other	223	0.000 [0.000, 0.000]	0.000 [0.000, 0.000]	—
	Asian	11	0.000 [0.000, 0.000]	0.000 [0.000, 0.000]	—
	White	1,420	0.164 [0.098, 0.225]	0.009 [0.003, 0.013]	0.651 [0.500, 0.834]
	Hispanic	653	0.407 [0.324, 0.484]	0.049 [0.030, 0.068]	0.704 [0.612, 0.800]
20th	Black	308	0.296 [0.186, 0.417]	0.118 [0.077, 0.164]	0.420 [0.269, 0.568]
	Other	217	0.063 [0.000, 0.222]	0.000 [0.000, 0.000]	1.000 [1.000, 1.000]
	Asian	12	0.521 [0.000, 1.000]	0.255 [0.000, 0.625]	0.513 [0.000, 1.000]
	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]
25th	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]
	White	1,431	0.258 [0.201, 0.317]	0.029 [0.019, 0.037]	0.620 [0.536, 0.721]
	Hispanic	624	0.499 [0.433, 0.565]	0.132 [0.101, 0.163]	0.681 [0.623, 0.745]
30th	Black	310	0.442 [0.351, 0.543]	0.178 [0.134, 0.229]	0.543 [0.446, 0.647]
	Other	223	0.109 [0.000, 0.226]	0.038 [0.010, 0.067]	0.287 [0.000, 0.626]
	Asian	8	0.661 [0.000, 1.000]	0.185 [0.000, 0.500]	0.694 [0.000, 1.000]

Table S5. Temporal Fairness: Group-Level Metrics Across Developmental Windows. TPR, FPR, and PPV with 95% bootstrap confidence intervals for each demographic group under four temporal prediction scenarios. Fairness disparities persist across all scenarios, confirming the temporal paradox discussed in the main text.

Scenario	Group	N	TPR [95% CI]	FPR [95% CI]	PPV [95% CI]
K Fall Only	White	1,458	0.144 [0.099, 0.198]	0.010 [0.005, 0.016]	0.661 [0.500, 0.833]
	Hispanic	629	0.352 [0.288, 0.418]	0.080 [0.055, 0.104]	0.639 [0.552, 0.738]
	Black	301	0.211 [0.134, 0.301]	0.087 [0.047, 0.132]	0.497 [0.342, 0.675]
	Other	213	0.141 [0.000, 0.333]	0.021 [0.005, 0.042]	0.423 [0.000, 0.751]
K Fall + Spring	White	1,458	0.185 [0.135, 0.236]	0.013 [0.007, 0.020]	0.655 [0.512, 0.787]
	Hispanic	629	0.346 [0.271, 0.421]	0.058 [0.038, 0.081]	0.705 [0.602, 0.794]
	Black	301	0.287 [0.193, 0.377]	0.083 [0.050, 0.131]	0.584 [0.452, 0.732]
	Other	213	0.097 [0.000, 0.251]	0.022 [0.005, 0.045]	0.330 [0.000, 0.715]
K + 1st Grade	White	1,458	0.208 [0.153, 0.275]	0.016 [0.009, 0.022]	0.641 [0.524, 0.767]
	Hispanic	629	0.364 [0.296, 0.436]	0.053 [0.033, 0.075]	0.734 [0.638, 0.822]
	Black	301	0.334 [0.235, 0.421]	0.107 [0.071, 0.154]	0.557 [0.435, 0.685]
	Other	213	0.097 [0.000, 0.251]	0.031 [0.010, 0.059]	0.261 [0.000, 0.573]
K through 3rd	White	1,458	0.197 [0.147, 0.261]	0.016 [0.009, 0.024]	0.617 [0.500, 0.755]
	Hispanic	629	0.364 [0.296, 0.436]	0.056 [0.034, 0.078]	0.724 [0.629, 0.815]
	Black	301	0.321 [0.231, 0.409]	0.106 [0.068, 0.152]	0.549 [0.415, 0.680]
	Other	213	0.097 [0.000, 0.251]	0.031 [0.010, 0.059]	0.261 [0.000, 0.573]

S3. Math Outcome Comparison

The main manuscript focuses on reading as the primary outcome, with math results summarized briefly. This section provides the full math outcome analysis.

S3.1 Performance Comparison

Table S6 compares overall model performance between reading and math prediction tasks.

Table S6. Model Performance Comparison: Reading vs. Math Outcomes. Math prediction (AUC = 0.867) substantially outperformed reading prediction (AUC = 0.848). The best-performing algorithm differed by outcome: elastic net for reading and logistic regression for math.

Outcome	Best Model	AUC	Accuracy	F1	Recall
Reading	Elastic Net	0.848	0.851	0.399	0.283
Math	Logistic Regression	0.867	0.847	0.459	0.366

S3.2 Fairness Metrics by Outcome

Table S7 presents the group-level fairness metrics for both outcomes side by side.

Table S7. Fairness Metrics by Racial/Ethnic Group for Reading and Math Outcomes (with 95% Bootstrap Confidence Intervals). Both outcomes show higher TPR for Hispanic and Black students relative to White students. Math shows notably higher FPR for Black students (0.161) than reading (0.095).

Outcome	Group	N	TPR [95% CI]	FPR [95% CI]	PPV [95% CI]
Reading	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]
Math	White	1,475	0.259 [0.196, 0.327]	0.024 [0.016, 0.032]	0.578 [0.478, 0.677]
	Hispanic	622	0.530 [0.459, 0.610]	0.093 [0.069, 0.118]	0.695 [0.625, 0.764]
	Black	288	0.326 [0.242, 0.429]	0.161 [0.113, 0.210]	0.526 [0.393, 0.641]
	Other	225	0.061 [0.000, 0.200]	0.024 [0.005, 0.042]	0.161 [0.000, 0.504]
	Asian	12	1.000 [1.000, 1.000]	0.113 [0.000, 0.333]	0.755 [0.333, 1.000]

S3.3 Disparity Comparison

Table S8 presents formal disparity metrics for both outcomes relative to the White reference group.

Table S8. Disparity Metrics by Outcome (Reference: White). Disparate impact is flagged when the TPR ratio falls below 0.80 (four-fifths rule). Math prediction uniquely triggers disparate impact for the Other group (TPR ratio = 0.24), indicating domain-specific fairness concerns not present in reading prediction.

Outcome	Group	TPR Rat.	TPR Diff	FPR Rat.	FPR Diff	Disp. Imp.
Reading	Asian	4.661	+0.589	0.000	−0.011	No
	Hispanic	2.445	+0.233	5.855	+0.053	No
	Black	1.823	+0.132	8.587	+0.082	No
	Other	1.611	+0.098	0.465	−0.006	No
Math	Asian	3.841	+0.740	4.681	+0.087	No
	Hispanic	2.028	+0.268	3.890	+0.069	No
	Black	1.243	+0.063	6.795	+0.138	No
	Other	**0.240**	−0.198	1.008	+0.000	**Yes**

The most notable difference between reading and math outcomes is the disparate impact flagged for the Other racial group in math prediction (TPR ratio = 0.24). While all minority groups have *higher* TPR than White students for reading, the Other group has dramatically *lower* TPR for math (0.061 vs. 0.259 for White). This finding indicates that fairness properties are domain-specific: a model that passes fairness criteria for one academic outcome may fail for another, reinforcing the importance of outcome-specific fairness auditing.

Additionally, math prediction shows substantially higher FPR for Black students (0.161 vs. 0.095 for reading), meaning that non-at-risk Black students are even more likely to be incorrectly flagged in math than in reading. The Asian group achieves perfect recall in math (TPR = 1.000), though the very small sample size (N = 12) limits interpretation.