Who Gets the Benefit of the Doubt? Racial Bias in Machine Learning Algorithms Applied to Secondary School Math Education

Haewon Jeong
UC Santa Barbara

Michael D. Wu
Harvard University

Nilanjana Dasgupta
UMass Amherst

Muriel Médard
MIT

Flavio Calmon
Harvard University

ABSTRACT

We present a large-scale study demonstrating how machine learning models trained on student data can perpetuate and—at worst—amplify racial discrimination and bias patterns that have existed in the American education system. We base our study on two datasets of middle and high school students collected in the United States. We find that standard machine learning models used to predict student math performance consistently give more benefit of the doubt to White and Asian students and are more pessimistic in their predictions to Black, Hispanic, and Native American students. Even more dangerously, these disparities are hidden by high accuracy numbers—the standard figure-of-merit used to evaluate machine learning performance. We also study the fairness implications of the racial composition of datasets used to train machine learning models that predict student performance. Our results show that changing the racial composition of the training dataset produces a surprising trade-off between false-positive and false-negative predictions between student groups. We discuss how we can leverage this effect as a tool for reducing racial gaps in prediction error patterns while preserving accuracy. We also benchmark several state-of-the-art fairness interventions on student data and report their performance. Our analyses provide guidelines for creating more racially just machine learning models in education.

Keywords

Racial justice, Machine learning, STEM education, Algorithmic discrimination, Fairness

1. INTRODUCTION

Machine learning (ML) algorithms are routinely used to support decisions that impact millions of students and their educational opportunities. In recent years, ML has been rapidly adopted in areas such as grading [25, 36], personalized learning [36, 48], and school admissions [36, 51]. One of the highest profile applications of algorithmic decision-making to education occurred in 2020 when the UK used a data-driven algorithm to assign 4.6 million grades for the A-level examinations due to Covid-19 restrictions on in-person test taking. This algorithm was found to systematically assign higher grades to students from historically high-performing schools in wealthier regions regardless of students' objective performance and sparked nationwide protests [1, 47].

ML algorithms are prone to discrimination in domains where racial inequalities are already pervasive. Data-driven algorithms can inherit and exacerbate human biases in applications such as criminal justice [6, 7], child welfare [8], and hiring [35], resulting in unfair decisions for historically underprivileged groups. In K-12 education—and STEM subjects in particular—racial disparities are widespread, with inequities existing in school funding, access to advanced placement classes, teacher perception, among many other areas [40, 41, 20]. The persistent racial disparities that exist in K-12 schools create a high-stakes minefield for ML algorithms. Nevertheless, the use of ML in education-related applications continues to increase at an unrestrained pace, with little to no guidelines and best practices to ensure that deployed algorithms are fair to students from diverse backgrounds.

In this work, we demonstrate concrete examples of how racial inequities emerge when ML algorithms are used to predict students' future math performance in secondary education. Our analysis is based on training popular ML models on large-scale datasets collected from middle schools and high schools across the United States. We find that the standard pipeline for training and deploying ML models—collecting a representative dataset, then fitting a ML model to maximize predictive accuracy—can systematically fail Black, Hispanic, and Native-American (BHN) students compared to White and Asian (WA) students when applied to predict future math performance.[1] Accuracy measures the rate of misclassification, but not all errors committed by a ML model are equal. False-positives give the benefit of the doubt to students and provide more opportunities, whereas false-

*This work was also presented at the NeurIPS 2021 MathAI4Ed Workshop.

[1]We categorize student race in to WA and BHN as White and Asian students are well-represented demographic groups in STEM education in the US [18].

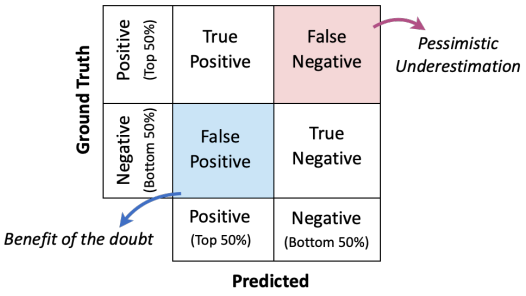

Figure 1: A confusion matrix for binary classification problems with the different errors that a ML model can make when predicting student performance (top or bottom 50% of their class). False positives in the blue bottom left box result in "benefit of the doubt" to students, providing more educational opportunities, and false negatives in the red top right box result in "pessimistic underestimations" which can undercut students' educational opportunities.

negatives undercut students' future potential (see Fig. 1). Our experiments show that standard ML algorithms achieve comparably high accuracy for both WA and BHN students when predicting future math performance, but the patterns of misclassifcation can be strikingly different between these two groups: WA students receive substantially more benefit of the doubt while BHN students receive more pessimistic predictions. The harm is silent: the usual procedure of optimizing ML models for accuracy may mask predictions that deprive BHN students of educational opportunities. Our findings suggest that, when predicting student future performance, false-positive rates and false-negative rates across student populations must be closely monitored.

We explore the impact of changing the racial composition of the training set and observe an intriguing trade-off between false positive rate (FPR) and false negative rate (FNR) as we vary the ratio of WA and BHN students. We show that by varying mixtures of BHN and WA students in the training set, the gap between error rates can be reduced with minimal impact on the model's overall accuracy. This result indicates that selecting the fairest demographic composition is not always straightforward. In fact, we show that the most counter-intuitive choice of using a homogeneous training set comprised only of one group can result in a model with the smallest disparity between the groups. Moreover, we report the performance of several fairness interventions. We observe that there is no clear winner among the five methods we tested. While state-of-the-art methods that do not utilize race as a feature can reduce FPR and FNR gaps substantially with a small sacrifice in accuracy, they cannot close the gaps completely. Only the methods that require student race information for the prediction can bring FPR and FNR differences close to zero.

Together, our findings are of direct value to data scientists, school administrators, and teachers who are considering using ML to support pedagogical decisions. The results presented next suggest three critical best practices: 1) monitor differences between false-positive and false-negative rates across student groups to ensure that all students receive comparable benefit of the doubt regardless of their racial background, and 2) judiciously vary the racial composition of training sets in order to close the gap between false-positive and false-negative rates, 3) apply different fairness interventions to further close this gap at a potential accuracy cost.

The main contributions of this paper are:

- We demonstrate that ML models that predict student performance can perpetuate biases that exist in the American education system. To the best of our knowledge, our results represent the first comprehensive and real-world study on the discrimination risks of deploying ML in education.
- Drawing from the literature on social psychology, we show how gaps in ML performance across different student populations may have detrimental downstream effects on educational opportunities. Specifically, we identify the *benefit of the doubt*, given be difference in false positive rates across student groups, as a meaningful metric for evaluating the disparity in ML performance in education.
- We show that the racial composition of the training dataset can impact ML bias and suggest best practices for data collection and balancing of student data.
- We benchmark several fairness interventions on real-world education data. Our results indicate that there is no single "best" fairness intervention. Depending on ethical constraints (e.g., if one can use sensitive attributes, how much accuracy one can sacrifice), a model developer should explore different options.
- Our findings expose several challenges for the responsible deployment of ML in education and serve as a guideline for data scientists working with student data.

2. RELATED WORK

Racial biases induced by machine learning systems have been reported in a number of areas including recidivism prediction [3], hiring [35, 53], child welfare [8], opioid use disorder detection [28], healthcare [37], and speech recognition [30]. For quantifying the discrimination of ML algorithms, various metrics were proposed. The ones that are most relevant to our work are equalized odds [21], disparate mistreatment [56], and error rate balance [7]. They examine the difference in FPR and FNR between majority and minority groups (e.g., male vs. female), and deem the model fair if FPR and FNR are equal in both groups. Many research works proposed a fair learning method that aims to achieve equalized odds [21, 56, 2, 5, 9, 52], and state-of-the-art methods were shown to reduce FPR and FNR substantially with a small loss in accuracy when evaluated on widely used datasets (e.g., COMPAS or Adult). We apply some of these methods to the education data and compare their performance in Section 5.

As the application of ML is explored in various areas of education (e.g., high school dropout prevention [11], MOOC dropout prediction [17], college admission [51]), several recent works study potential bias issues of ML applied to

	Schools	# Students	Features	Label
MSS	10 private middle schools	$\sim$ 3,000	Student/parent surveys, student demographic information, students' past academic performance	9-th grade math score
HSLS	940 public high schools	$\sim$ 20,000		

Table 1: Summary of the datasets used in the paper.

education. In [46, 22, 54], the authors consider training a ML model for detecting at-risk students at college (low-performing or early dropouts) and examine the fairness of the model across different subpopulations (e.g., gender or race). Predicting student success and future grades at college is studied in [55, 23]. For online learning systems, fairness in knowledge tracing [13] and MOOC dropout detection [19] have been studied.

This paper examines ML fairness applied in secondary education. We especially focus on math education in middle and high schools because decisions made at this period are critical to students' future STEM education and career due to the cumulative nature of math. Our study also relates to a realistic scenario where an algorithm is used for tracking and class placements as math is the most tracked class in the US [4]. A similar line of work includes [38, **?**]. Furthermore, the data we examine are distinct from previous works as it not only contains students' past performance and demographic information but also contains extensive survey answers from students and parents, including questions regarding their stereotypes toward STEM subjects. Furthermore, this is the first work that provides a comprehensive empirical analysis on applying state-of-the-art fair learning methods on education data.

Finally, we want to note that the balancing approach we discuss in Section 4 is conceptually related to resampling or reweighting techniques for achieving fairness when there exists class or label imbalance [26, 44, 31].

3. UNEQUAL BENEFIT OF THE DOUBT

We describe next the discrimination patterns that emerge when ML models used for predicting student math performance. We also discuss the implications of this bias in terms of limiting educational opportunities for BHN students. In the subsequent sections, we explore how this observed bias can be mitigated by changing the composition of the training set, as well as by applying existing fairness interventions. In all analyses we present in the paper, we use two datasets: the middle school study (MSS) dataset [10] and the public-use high school longitudinal study 2009 (HSLS) dataset [43]. The MSS dataset is contains roughly 3,000 entries collected from ten private middle schools (5 coeducational and 5 all-girl schools) in the US. The publicly available HSLS dataset is collected from 20,000+ students from 940 public and private high schools in the US across 50 different states and District of Columbia. Both datasets include student surveys, student demographic information, parent surveys, and students' math performance across several years.

Math is a foundational subject in STEM education. Accurate predictions of students' future math performance can enable better educational resource distribution to boost students with potential (e.g., advanced class placements or gifted program recommendations). We train several ML models (logistic regression, SVM, random forests, see Materials and Methods section for details) to predict if a student will be a top 50% performer in their future math class (positive prediction) or bottom 50% performer (negative prediction). Students' past performance is not enough to predict future success and persistence. In fact, with the HSLS dataset, we observe that prediction accuracy is 68.2 $\pm$ 0.1 % if we make predictions about students' math performance in the 9th grade based only on their past performance. Accuracy improves to 75.0 $\pm$ 0.1 %, by utilizing more features in the data such as student and parent's survey answers.

Being able to take advantage of more data for more accurate predictions with ML sounds promising. However, the deployment of ML may not benefit all racial groups equally. The models we trained achieve comparable accuracy across WA and BHN students in predicting future math performance. However, when examining metrics beyond accuracy, significant racial inequalities emerge. Even though a similar average accuracy indicates that there are roughly equal numbers of misclassified points among WA and BHN students, how they are misclassified is staggeringly different.

For each model, we examine four different metrics: accuracy, false positive rate (FPR), false negative rate (FNR), and predicted base rate (PBR). These metrics can be understood using *a confusion matrix* described in Fig. 1. False positive prediction refers to students who did not belong in the top 50% of math performers based on their actual grades but received a positive prediction from the ML model. In other words, they are given the *benefit of the doubt* from the ML prediction. On the flip side, false negative predictions are students who did belong in the top 50% in reality, but are given a negative prediction. This is a *pessimistic underestimation* of their future performance. By examining FPR and FNR, we discover that WA students are consistently given more benefit of the doubt, while BHN students are consistently underestimated in predicting their future performance despite similar accuracy numbers for both groups. This shows that narrowly focusing on accuracy can give an illusion of fairness when there is significant discriminatory impact on minority students.

Middle school dataset results.. We performed binary classification on whether a student will be a top 50% performer in their 9th grade math courses or a bottom 50% performer. We removed all features related to students' race such as their parents' place of birth. Since the MSS dataset is collected only from private schools, it has a relatively small number of BHN students—they make up only 26% of the data. We undersampled WA students to generate a balanced dataset with a roughly equal number of WA and BHN students. As different subsamples chosen from WA students in

		WA	BHN	Difference
MSS	Size	128	132	–
	PBR	**0.575**	0.488	0.087 (+15.13%)
	FPR	**0.319**	0.281	0.038 (+11.91%)
	FNR	0.205	**0.276**	-0.071 (-34.63%)
	Accuracy	0.740	0.721	0.019 (+2.57%)
HSLS	Size	2867	1486	–
	PBR	**0.580**	0.347	0.233 (+40.17%)
	FPR	**0.304**	0.176	0.128 (+42.11%)
	FNR	0.209	**0.371**	-0.162 (-77.51%)
	Accuracy	0.750	0.750	0.000 (+0.00%)

Table 2: Random forest results for math performance predictions on the Middle School Study dataset (MSS) and the High School Longitudinal Survey (HSLS) dataset. Size represents the number of samples in the test set. In both datasets, PBR and FPR are higher for WA students and FNR is higher in BHN students, while accuracy is similar between the groups. The higher FPR in WA students suggests that they are getting more "benefit of the doubt" predictions and the higher FNR in BHN students shows that they are receiving more "pessimistic underestimation" predictions.

the training set can lead to different models, we ran multiple iterations of data balancing and averaged their performance. We split the balanced dataset into a training set (70%) and a test set (30%). We ran 30 runs of data balancing, and with each balanced dataset, we ran 30 different train/test splits. In total, we get results from 900 models and compute the average and standard error by aggregating the performance metrics of each model. The result of training a random forest model is summarized in Table 2.

First, notice that the difference in accuracy between WA and BHN is small ($< 3\%$ relative difference). However, FNR was considerably smaller for WA students compared to BHN. The relative difference in FNR was up to 35%. At the same time, FPR is 12% higher for WA students. In other words, WA students are less prone to get an underestimated prediction and more likely to receive the benefit of the doubt from the trained ML model. We also observe that PBR is higher in WA students than in BHN students. This may reflect the difference in the ground truth data. The observed base rate was 0.52 for WA and 0.46 for BHN students. However, the PBR difference from the trained random forest models was about 0.09, indicating that the existing racial performance gap is exaggerated in the ML model's predictions.

High school dataset results.. We ran a similar experiment on the public-use HSLS dataset to replicate our findings on the MSS dataset. We again trained binary classification models to predict top and bottom 50% performers in the standardized test taken in the 9th grade. We train different models with 30 different train/test splits and obtain the average and standard error. We did not perform data balancing for the HSLS dataset since it was collected from both public and private schools from all states (the MSS dataset was collected only from private schools, from a few states) and its racial representation is close to national statistics. The results are summarized in Table 2.

We identify a very similar pattern of bias in the HSLS dataset. There is a negligible difference in accuracy between the racial groups. However, the difference in FPR and FNR is substantial. FPR is 42% higher for WA than BHN students in all models, and FNR is 78% lower for WA than BHN students. The gaps in FNR and FPR are even wider than the MSS dataset analysis. This may be due to a bigger gap in the ground truth base rate: 0.57 for WA and 0.38 for BHN (difference = 0.19). Yet again however, the existing racial performance gap is exaggerated in the ML predictions with a PBR of 0.23.

3.1 Connection to fairness metrics in the ML literature

The four metrics (PBR, FPR, FNR, and accuracy) we evaluate throughout the paper are related to some of the widely-used fairness metrics in ML: statistical parity [14] and equalized odds/opportunity [21]. Consider two population groups, a minority group (group 0) and a majority group (group 1). When an ML model has the same PBR for group 0 and group 1, it satisfies *statistical parity*. When the model has the same FNR and FPR (i.e., $\text{FPR}_0 = \text{FPR}_1$ and $\text{FNR}_0 = \text{FNR}_1$), it satisfies the *equalized odds* criterion. A relaxed version is *equalized opportunity*, which only requires equal FNRs: $\text{FNR}_0 = \text{FNR}_1$. These metrics have been heavily analyzed on datasets from domains such as criminal justice, income prediction, and healthcare [21, 31, 39, 21, 5, 12, 2], but to the best of our knowledge, such evaluation has not been reported on K-12 education data. Our results clearly illustrate that significant differences in equalized odds metrics can also arise when off-the-shelf ML algorithms are applied on secondary school student data.

3.2 Why do we observe unequal benefit of the doubt?

How does a ML model systematically underestimate BHN students' performance and overestimate WA students' performance despite not using any race-related features to make predictions? Well-calibrated classifiers are bound to have gaps in FPR and FNR between groups when they have different base rates [39]. Are the gaps observed when predicting student math performance simply caused by the difference in base rates? To examine this question, we performed experiments of selective subsampling BHN or WA students to equalize the base rate of the two groups. The result of subsampling underperforming BHN students to inflate their base rate is given in the first columns of Table 3. While the PBRs of both groups increase in this regime, the gaps in FPR and FNR still remain. We observe the same trend when we subsample high-performing WA students to lower the base rate of WA.

Another possible hypothesis is that there are features in the data that are covariates (i.e., proxies) for students' race. The trained ML model can then exploit these features to assign disparate predictions to different racial groups. To test this hypothesis, we designed the following experiment. First, we identify if any features reveal information about students' race by training an ML model that performs a new binary classification task of predicting whether a student is WA or BHN on the HSLS dataset. If it performs better than random guessing, we can conclude that other features in the

	Selective subsampling BHN			Removing implicit race features			Removing all parent features		
	WA	BHN	Difference	WA	BHN	Difference	WA	BHN	Difference
PBR	**0.655**	0.549	0.106 (+16.18%)	**0.558**	0.378	0.180 (+32.26%)	**0.524**	0.378	0.146 (+27.86%)
FPR	**0.402**	0.305	0.097 (+24.13%)	**0.279**	0.205	0.074 (+26.52%)	**0.289**	0.231	0.058 (+20.07%)
FNR	0.153	**0.267**	-0.114 (-74.51%)	0.229	**0.339**	-0.110 (-48.03%)	0.294	**0.380**	-0.086 (-29.25%)
Accuracy	0.739	0.716	0.023 (+3.11%)	0.749	0.744	0.005 (+0.67%)	0.708	0.712	-0.004 (-0.56%)

Table 3: Random forest results on the high school (HSLS) dataset with selective subsampling and removing a subset of features. To examine the sources of bias, we performed three experiments: subsampling BHN students in the bottom 50% to raise the base rate of the group, removing features that are most related to race, and removing all socioeconomic and parent variables. We omit standard errors due to space constraints. All three methods reduce FPR and FNR gaps, compared to the result in Table 2, but it is far from eradicating the gaps.

data predict students' race. In our case, baseline accuracy of random guessing is 0.66 as a model that predicts everyone is WA achieves the accuracy of 0.66 as 66% of the population is WA. A random forest model we trained achieved the accuracy of 71% accuracy. We then ranked the most relevant features used in the prediction to infer the most race-revealing features. The five most predictive features were: `S1LANG1ST` (student's first language), `P1MARSTAT` (parent 1's marital status), `X1FAMINCOME` (family income), `X1PAR2EDU` (parent 2's highest level of education), and `X1PAR2OCC2` (parent 2's current/most recent occupation). Then, we trained a new model without using these five features, i.e., with 47 features instead of 52. If the racial gap reduces by removing the most race-related features, it supports our hypothesis that implicitly race-related features were being used to assign different predictions to different races. The result of training a random forest model without the implicit race features in the second column of Table 3. We observe that gaps in PBR, FPR, and FNR all decrease substantially. To further investigate this issue, we also trained models after removing all socioeconomic variables as well as all parent survey variables. On average, this narrows the FPR and FNR gaps further by $\sim$0.02, but at the sacrifice of accuracy (a drop of $\sim$0.04).

These results demonstrate that salient features used in the prediction task have different distributions for WA and BHN students and that removing race information from the training data is not enough to prevent racially discriminatory performance. However, a careful feature selection can reduce performance gaps. This observation is congruent with reports of ML bias in other applications such as criminal recidivism prediction [24]. One can also employ preprocessing techniques that reduce race-related information in the data while maintaining the useful information for prediction [57, 16, 32].

3.3 Implications of FNR and FPR gaps in educational opportunities

When predicting student performance, unequal error rates have real-world consequences. Consider a scenario where we use a random forest model trained on the HSLS dataset for 9th grade math placement. Students who are predicted to be in top 50% will be placed in the advanced-level math class and students who receive bottom 50% prediction will be placed in the basic math class. The FPR of 0.30 for WA students (see Table 2) means that 30% of the students who would not perform well in the 9th grade will be placed in the advanced class. They are given the benefit of the doubt and the opportunity to learn more advanced math. On the other hand, only 18% of the BHN students get the same benefit of the doubt (FPR=0.18). The FNR of 0.21 in WA students indicates that 21% of WA students who would in fact perform well in the future will be placed in the basic class by the ML algorithm. For BHN students, a startlingly high 37% will be incorrectly placed in the basic class, their academic potential ignored by the algorithm.

The downstream effects of such misclassification is disproportionately detrimental to BHN students. Missing the opportunity to take foundational math classes such as Algebra 1 can prevent them from taking further advanced classes in the following years. Indeed, past research shows that middle school algebra is a strong early predictor of educational outcomes in high school and college [4, 33, 34, 49]. Moreover, they are at risk of losing interest in STEM subjects because of the pessimistic prediction by the algorithm. It was shown in prior research that low test scores or class placements to less-advanced classes discourage students from historically marginalized groups more because they elevate negative stereotypes [42, 45]. For these reasons, even if it has similar accuracy for all racial groups, a model that gives 42% more benefit of the doubt to WA students and 78% more pessimistic underestimation of BHN students' ability cannot be considered fair by any means.

4. RACIAL COMPOSITION OF THE TRAINING SET AND FAIRNESS

In the experiments we presented in the previous section, we balanced the MSS dataset to have roughly 50% of WA and 50% of BHN students, as the original dataset collected from private schools had a far fewer number of BHN students than the national average. For the HSLS dataset, which is collected from both public and private schools sampled according to national demographics, we did not balance racial groups as the given dataset is already a nationally representative sample. Was it *fair* to rebalance the MSS dataset to have the equal number of data points in each group, and was it *fair* to use the HSLS dataset as it is? When dealing with a dataset made up of different population groups, whether to balance the dataset and how to balance the dataset are unavoidable choices that a data scientist has to make in the

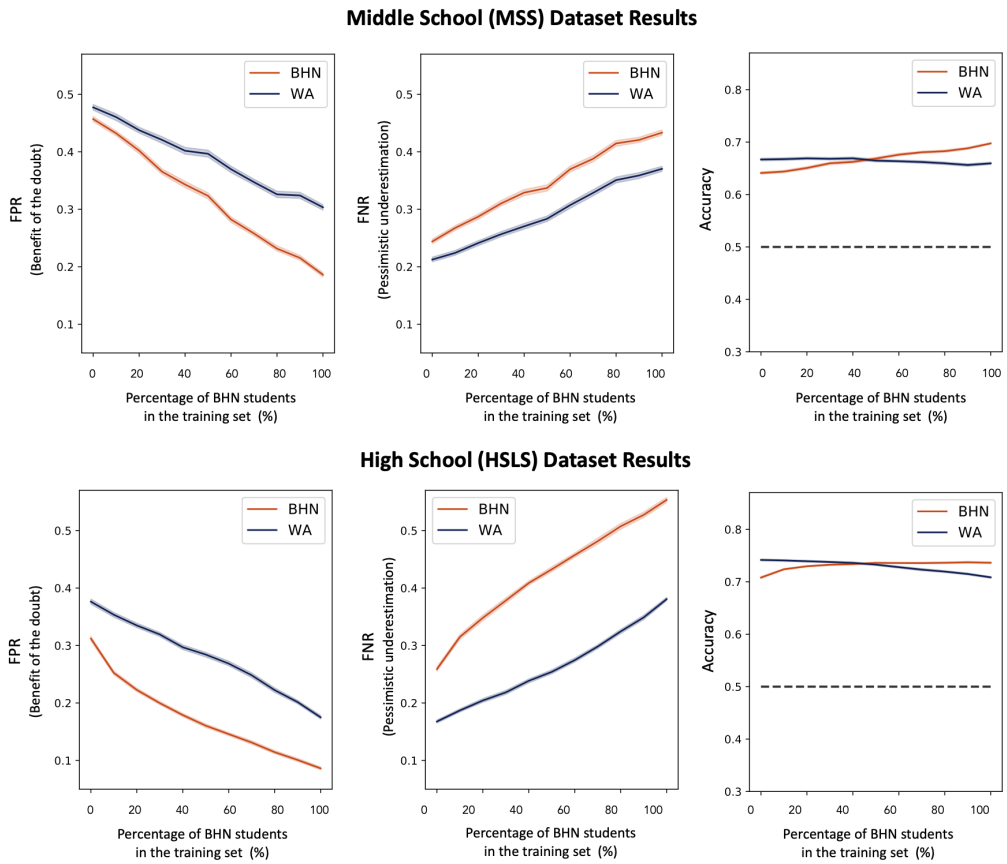

Figure 2: Results of changing racial composition of the training set on middle school (MSS) and high school (HSLS) datasets. We plot false positive rate (FPR), false-negative rate (FNR), and accuracy for each group, White/Asian (WA) and Black/Hispanic/Native American (BHN), as we change the percentage of BHN students in the training set (p) from 0 to 100%. For the MSS dataset, we ran 30 iterations of different train/test splits and for the HSLS dataset, we ran 10 iterations. The lines and the shaded regions represent the average and bootstrap confidence interval. In both datasets, we observe that FPR monotonically decreases and FNR monotonically increases as we increase p. The black dashed line in the accuracy plot represents the accuracy of random guessing. As we are using more BHN data points, the accuracy of BHN improves while the accuracy of WA decreases. However, the range of accuracy difference is substantially smaller than the range of FPR and FNR differences. This suggests that by changing the racial composition of the training set, we are essentially trading off false positive predictions with false negative predictions while maintaining similar accuracy.

data preprocessing stage in the ML pipeline. Despite its importance, past research has not rigorously investigated the question of what racial composition of training sets would produce the most fair model.

To investigate this question, we trained ML models with different racial mixtures in the training set. We discover that as we change the racial composition of the training set, from only WA students to only BHN students, FPR and FNR change significantly while overall accuracy remains close to constant. In essence, by changing the training set racial mixture, the model is trading off false positive predictions and false negative predictions.

In the experiments, we varied the proportion of BHN students in a training set, $p = \frac{(\#\text{BHN})}{(\#\text{WA}) + (\#\text{BHN})}$, from 0 to 1 in the interval of 0.1. For all p, we ensure that the entire training set size (i.e., number of all data points in the training set) is the same. To achieve different p, we subsample from each group and fit a random forest classifier for each subsampled dataset. The results are summarized in Fig. 2.

The results we observe are striking. Focusing solely on accuracy may lead to the incorrect conclusion that the effect of different racial compositions of a training set is minute: the accuracy for each group does not vary more than 0.05 as we change p from 0 to 1 (i.e., 0% to 100% BHN). However, FPR and FNR metrics change drastically with different racial compositions of the training set. In both the MSS and HSLS results, FPR monotonically decreases and FNR monotonically increases for both BHN and WA students as we increase p from 0% BHN to 100% BHN. The range of FPR and FNR changes is significant. In the MSS results, FPR moves from 0.45 to 0.25 and FNR moves from 0.2 to 0.4. In the HSLS results, the range of FPR difference is from $\sim$0.4 to 0.1 and FNR moves from 0.2 to 0.5. The gaps in FPR and FNR remain throughout different values of p. However, in the MSS dataset results, the gaps in both FPR and FNR tend to reduce as p gets closer to 0. Similarly, in the HSLS experiments, the FPR/FNR gaps reduce slightly

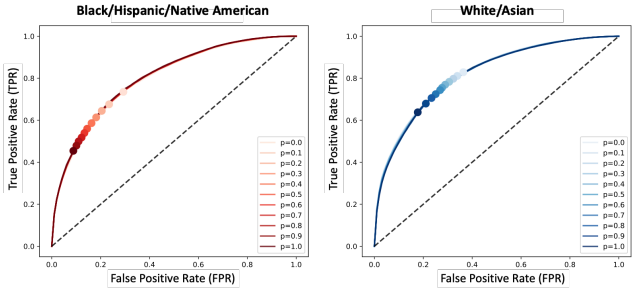

Figure 3: ROC curve analysis of training with different racial mixtures on the HSLS dataset. We plot receiver operating characteristic (ROC) curves for each different mixture (i.e., each different p). Lighter colors represent smaller p (i.e., fewer BHN students in the training set) and darker colors represent bigger p. Eleven different ROC curves almost overlap and are indistinguishable from each other. The circle markers denote the operating point of the trained model with different values of p. Notice that as p increases, the markers move down on the curve and have smaller FPR and TPR values. This implies that training with different racial mixtures mimics the behavior of choosing different decision thresholds of a fixed classifier.

around $p = 0$. Finally, even though the difference in accuracy is small, we observe that as we increase p, i.e., as we use more BHN students in the training set, the accuracy for BHN increases and the accuracy for WA decreases.

4.1 ROC curve analysis

There is a clear trade-off between FPR and FNR as we change the racial mixture of the training set. This behavior can be mapped to the receiver operating characteristic (ROC) curve of the classifiers used to predict student math performance. For binary classification, the ML models we considered produce a *score* S ($0 \leq S \leq 1$) for each input sample. Predictions are then made by thresholding the score as follows:

$$\text{Prediction} = \begin{cases} \text{Positive}, & \text{if } S \geq 0.5, \\ \text{Negative}, & \text{if } S < 0.5. \end{cases}$$

After training, a ML model can be viewed as a function that computes a score S from the input data. In most off-the shelf models, the default score threshold for predicting positive outcomes is 0.5, but this threshold can be adjusted. Increasing the threshold above 0.5 leads to fewer data points receiving positive predictions and, equivalently, a lower FPR and a higher FNR. When predicting student performance, this corresponds to fewer students being flagged as high performers and less benefit of the doubt overall. Conversely, lowering the threshold below 0.5 results in a lower FNR and a higher FPR. ROC curves show how true positive rate (TPR) and FPR change as we vary the threshold for a trained classifier (i.e., for a fixed function that computes S from the input), where TPR is defined as:

$$\text{TPR} = \frac{\text{True Positive}}{\text{True Positive} + \text{False Negative}} = 1 - \text{FNR}.$$

Notice that changing a score threshold for a trained classifier trades off FPR and FNR, similar to what we observed when changing the racial mixture of the training set. Is changing the racial composition of the training set equivalent to changing the score threshold? The goal of following analysis is to understand if models trained on different racial mixtures produce different FPR and FNR because they have different ROC curves or if they approximately learn the same scoring function (i.e., have the same ROC curves) but use different thresholds (i.e., correspond to different points on the ROC curve).

We plot the ROC curves for the trained models for each p in Fig. 3. Recall that p is the proportion of BHN data points in the training set. The lighter color lines and markers in the plot represent smaller p. We observe that the eleven ROC curves for different values of p mostly overlap with insignificant differences. The markers in the plot represent the operating point for the trained classifier for each p. As we increase p, these points move down on the curve. This is essentially equivalent to increasing the decision threshold of a fixed classifier to be more conservative in making positive predictions. In other words, the classifier decreases FPR and increases FNR.

We now explain this behavior with the base rate change induced by different racial mixture. Base rate (BR) is defined as:

$$\text{BR} = \frac{\text{True Positive} + \text{False Negative}}{\text{All}}.$$

As we increase p to include more BHN students, the overall BR of the training set becomes smaller. The classifier that maximizes accuracy on the ROC curve corresponds to the point that has the tangent of $\frac{1-\text{BR}}{\text{BR}}$. Since BR decreases with increasing p, the $\frac{1-\text{BR}}{\text{BR}}$ becomes larger. Hence, the operating point moves down on the ROC curve where the slope is steeper.

4.2 Fair racial mixture for training

How should one choose the racial mixture of training sets in order to produce a fair model? Ideally, a ML model should achieve similar performance across metrics (e.g., accuracy, PBR, FPR, and FNR) for different student racial groups. However, when the data from two groups do not follow the same statistical distribution, a model cannot achieve equality in all metrics [7, 29]. It is thus important to identify the impact of prioritizing different metrics when predicting student performance.

If we consider achieving equal accuracy as a fairness goal, the results in Fig. 2 suggest that balancing the dataset to have an equal number of BHN and WA students would be the right solution. However, our results echo that accuracy should not be the sole metric of focus as FPR and FNR gaps show severe discriminatory effects. If we choose a fairness criterion to be equal FPR and FNR, Fig. 2 indicates that simply changing the racial mixture of the training set cannot achieve this. If we relax the condition and choose a training set that has the smallest gap in FPR and FNR, the best p is 0.0, meaning the training set is 100% WA. This can benefit BHN students as it not only has the smallest FPR/FNR gaps, but also has the smallest FNR for all groups. Reducing false negative predictions can be especially beneficial for minority students, who are at greater risk of losing interest

in STEM subjects by receiving negative predictions. Contrary to the most intuitive choices of fair training set—one that contains an equal number of WA and BHN students or one that follows the national demographics—we show that using homogeneous samples of WA students can produce the fairest model in terms of the FPR and FNR gap.

5. FAIR INTERVENTIONS ON EDUCATION DATASETS

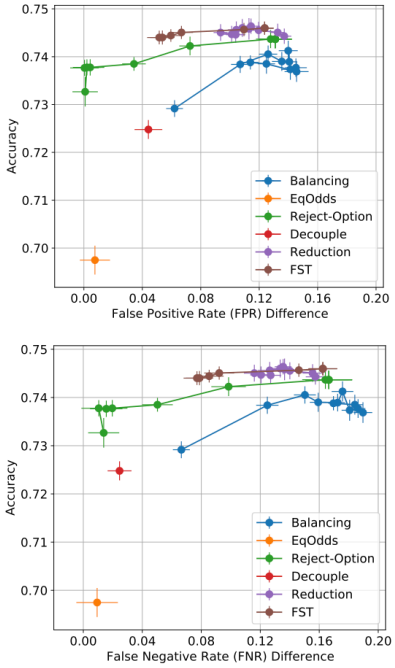

Figure 4: Comparison of different fairness interventions applied to the HSLS dataset. We plot accuracy over FPR and FNR differences for the balancing approach we introduced in Section 4 and five different existing fair intervention methods. Standard errors for both x and y axis are obtained from 10 runs of each method with different train-test splits.

In Section 3, we saw that WA students receive disproportionately more benefit of the doubt and BHN students receive more pessimistic predictions when off-the-shelf ML models are used to predict student math performance. We also reported that neither artificially equating the base rates between the groups by subsampling nor removing features that are most related to race can eradicate the racial gaps completely. In Section 4, we showed that different choices of data balancing can be a surprisingly effective for reducing FPR and FNR gaps with small loss in accuracy. In this section, we apply state-of-the-art fair learning methods and study how effective they are in mitigating bias in student performance prediction.

We compare five different methods: decoupling [15], equalized odds post-processing (EqOdds) [21], reject-option [27], reduction [2], and fair score transformer (FST) [52]. Decoupling [15, 50] is the simple method of training a separate model for each group, and can be an easy and effective method a data scientist can adopt if it is legal and ehtical to do so. The other four methods are designed to specifically reduce FPR and FNR gaps. However, there is a crucial distinction. EqOdds and reject-Option methods require sensitive attributes (i.e., student race information) at deployment. On the other hand, reduction and FST methods, as implemented here, only require sensitive attributes for training a model but not for making predictions.

The results of applying the five methods are illustrated in Figure 4. We first notice that EqOdds achieve FPR and FNR gaps closest to 0, albeit losing accuracy considerably. FST and reduction methods achieve the the best accuracy in the high FPR and FNR region. Evaluations on other datasets shown in the previous literature [2, 52] also demonstrated that FST and the reductions approaches often show the most competitive performance. However, as we exclude sensitive attributes at test time, these methods cannot achieve close-to-perfect fairness compared to methods that utilize sensitive attributes. The reject-option method exhibits competitive accuracy and it also extends to the region where FPR and FNR differences are very small by using the sensitive attribute. The balancing approach we introduced in the previous section (blue line) shows a trade-off curve comparable to the reject-option method even though it does not utilize sensitive attributes during training or testing time.

6. DISCUSSION

Our analyses on the middle school and high school datasets show that even when an algorithm gives significantly more benefit of the doubt to the privileged groups of students compared to historically marginalized groups, accuracy for the two groups can be almost equal. This result serves as a cautionary tale on the danger of blindly following the standard practice of choosing a model that achieves high accuracy. To build a fair algorithm, it is necessary to also examine FPR and FNR metrics. We frame these metrics as *benefit of the doubt* and *pessimistic underestimation* to make them relevant to a potential application scenario and understandable to educators and policymakers. When a data scientist observes unequal benefit of the doubt between racial groups in ML models, they can consider using intervention methods that reduce the FPR and FNR gaps. As mentioned earlier, equal FPR and FNR is referred to as *equalized odds* in the fair ML literature [21]. This notion of group fairness has been studied extensively and various methods have been proposed to reduce FPR/FNR disparities, ranging from preprocessing techniques [31], postprocessing techniques [39, 21], to adding FPR/FNR equality as an objective during the training process [5, 12, 2].

Our experiments of altering the demographic composition in training sets add a new dimension to fairness-ensuring interventions in ML. By using a training set with different ratios of population groups, we arrive at different models which can improve FPR/FNR disparities with little to no sacrifice in accuracy. By doing an analysis similar to Fig. 2, a data scientist can select a training set that improves the benefit of the doubt given to all groups. The main advantages of this method is that it does not require deploying different models for different groups (or using race as an input feature) nor any change to the data beyond varying the composition of the training set. This intervention can also easily be paired with other existing fair learning algorithms later in the ML pipeline. Theoretical analysis of this mechanism to support

our empirical study will be an interesting future direction.

7. REFERENCES

[1] K. Adam. The U.K. used an algorithm to estimate exam results. the calculations favored elites. https://www.washingtonpost.com/world/europe/the-uk-used-an-algorithm-to-estimate-exam-results-the-calculations-favored-elites/2020/08/17/2b116d48-e091-11ea-82d8-5e55d47e90ca_story.html, 2020.

[2] A. Agarwal, A. Beygelzimer, M. Dudík, J. Langford, and H. Wallach. A reductions approach to fair classification. In *International Conference on Machine Learning*, pages 60–69. PMLR, 2018.

[3] J. Angwin, J. Larson, S. Mattu, and L. Kirchner. Machine bias. propublica (2016). *URL: https://www. propublica. org/article/machine-bias-risk-asses sments-in-criminal-sentencing*, 2016.

[4] C. Berwick. Is it time to detrack math? https://www.edutopia.org/article/it-time-detrack-math, 2019.

[5] L. E. Celis, L. Huang, V. Keswani, and N. K. Vishnoi. Classification with fairness constraints: A meta-algorithm with provable guarantees. In *Proceedings of the conference on fairness, accountability, and transparency*, 2019.

[6] A. Chohlas-Wood. Big data's disparate impact. https://www.brookings.edu/research/understanding-risk-assessment-instruments-in-criminal-justice, 2020.

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

[8] A. Chouldechova, D. Benavides-Prado, O. Fialko, and R. Vaithianathan. A case study of algorithm-assisted decision making in child maltreatment hotline screening decisions. In *Conference on Fairness, Accountability and Transparency*, pages 134–148. PMLR, 2018.

[9] A. Cotter, H. Jiang, M. R. Gupta, S. Wang, T. Narayan, S. You, and K. Sridharan. Optimization with non-differentiable constraints with applications to fairness, recall, churn, and other goals. *J. Mach. Learn. Res.*, 20(172):1–59, 2019.

[10] N. Dasgupta, K. Thiem, A. Coyne, H. Laws, M. Barbieri, and R. Wells. The impact of communal learning contexts on adolescent self-concept and achievement: Similarities and differences across race and gender. *Journal of Personality and Social Psychology*, 2021.

[11] F. Del Bonifro, M. Gabbrielli, G. Lisanti, and S. P. Zingaro. Student dropout prediction. In *International Conference on Artificial Intelligence in Education*, pages 129–140. Springer, 2020.

[12] M. Donini, L. Oneto, S. Ben-David, J. Shawe-Taylor, and M. Pontil. Empirical risk minimization under fairness constraints. *arXiv preprint arXiv:1802.08626*, 2018.

[13] S. Doroudi and E. Brunskill. Fairer but not fair enough on the equitability of knowledge tracing. In *Proceedings of the 9th international conference on learning analytics & knowledge*, pages 335–339, 2019.

[14] C. Dwork, M. Hardt, T. Pitassi, O. Reingold, and R. Zemel. Fairness through awareness. In *Proceedings of the 3rd innovations in theoretical computer science conference*, pages 214–226, 2012.

[15] C. Dwork, N. Immorlica, A. T. Kalai, and M. Leiserson. Decoupled classifiers for group-fair and efficient machine learning. In *Conference on fairness, accountability and transparency*, pages 119–133. PMLR, 2018.

[16] M. Feldman, S. A. Friedler, J. Moeller, C. Scheidegger, and S. Venkatasubramanian. Certifying and removing disparate impact. In *proceedings of the 21th ACM SIGKDD international conference on knowledge discovery and data mining*, pages 259–268, 2015.

[17] W. Feng, J. Tang, and T. X. Liu. Understanding dropouts in moocs. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2019.

[18] R. Fry, B. Kennedy, and C. Funk. Stem jobs see uneven progress in increasing gender, racial and ethnic diversity. *Pew Research Center Science & Society*, 2021.

[19] J. Gardner, C. Brooks, and R. Baker. Evaluating the fairness of predictive student models through slicing analysis. In *Proceedings of the 9th international conference on learning analytics & knowledge*, pages 225–234, 2019.

[20] S. Gershenson, S. B. Holt, and N. W. Papageorge. Who believes in me? the effect of student–teacher demographic match on teacher expectations. *Economics of education review*, 52:209–224, 2016.

[21] M. Hardt, E. Price, and N. Srebro. Equality of opportunity in supervised learning. *Advances in neural information processing systems*, 2016.

[22] Q. Hu and H. Rangwala. Towards fair educational data mining: A case study on detecting at-risk students. *International Educational Data Mining Society*, 2020.

[23] W. Jiang and Z. A. Pardos. Towards equity and algorithmic fairness in student grade prediction. In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*, pages 608–617, 2021.

[24] J. E. Johndrow and K. Lum. An algorithm for removing sensitive information: application to race-independent recidivism prediction. *The Annals of Applied Statistics*, 13(1):189–220, 2019.

[25] J. L. Joyce and L. Harris. Artificial intelligence (AI) and education. *Focus, Congressional Research service, August.*, 2018.

[26] F. Kamiran and T. Calders. Data preprocessing techniques for classification without discrimination. *Knowledge and information systems*, 2012.

[27] F. Kamiran, A. Karim, and X. Zhang. Decision theory for discrimination-aware classification. In *2012 IEEE 12th International Conference on Data Mining*, pages 924–929. IEEE, 2012.

[28] A. E. Kilby. Algorithmic fairness in predicting opioid use disorder using machine learning. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, pages 272–272, 2021.

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

[30] A. Koenecke, A. Nam, E. Lake, J. Nudell, M. Quartey,

Z. Mengesha, C. Toups, J. R. Rickford, D. Jurafsky, and S. Goel. Racial disparities in automated speech recognition. *Proceedings of the National Academy of Sciences*, 117(14):7684–7689, 2020.

[31] E. Krasanakis, E. Spyromitros-Xioufis, S. Papadopoulos, and Y. Kompatsiaris. Adaptive sensitive reweighting to mitigate bias in fairness-aware classification. In *Proceedings of the 2018 World Wide Web Conference*, pages 853–862, 2018.

[32] C. Louizos, K. Swersky, Y. Li, M. Welling, and R. Zemel. The variational fair autoencoder. *arXiv preprint arXiv:1511.00830*, 2015.

[33] T. Loveless. The 2013 brown center report on american education: How well are american students learning? Technical report, The Brown Center on Education Policy at the Brookings Institution, 2013.

[34] A. V. Maltese and R. H. Tai. Pipeline persistence: Examining the association of educational experiences with earned degrees in stem among us students. *Science education*, 95(5):877–907, 2011.

[35] G. Mann and C. O'Neil. Hiring algorithms are not neutral. *Harvard Business Review*, 9, 2016.

[36] D. Newton. From admissions to teaching to grading, AI is infiltrating higher education. https://hechingerreport.org/from-admissions-to-teaching-to-grading-ai-is-infiltrating-higher-education/, 2021.

[37] Z. Obermeyer, B. Powers, C. Vogeli, and S. Mullainathan. Dissecting racial bias in an algorithm used to manage the health of populations. *Science*, 366(6464):447–453, 2019.

[38] J. Ocumpaugh, R. Baker, S. Gowda, N. Heffernan, and C. Heffernan. Population validity for educational data mining models: A case study in affect detection. *British Journal of Educational Technology*, 45(3):487–501, 2014.

[39] G. Pleiss, M. Raghavan, F. Wu, J. Kleinberg, and K. Q. Weinberger. On fairness and calibration. *arXiv preprint arXiv:1709.02012*, 2017.

[40] S. F. Reardon and X. A. Portilla. Recent trends in income, racial, and ethnic school readiness gaps at kindergarten entry. *Aera Open*, 2(3), 2016.

[41] R. Reeves, E. Rodrigue, and E. Kneebone. Five evils: Multidimensional poverty and race in america. *Economic Studies at Brookings Report*, 1:1–22, 2016.

[42] C. Riegle-Crumb, B. King, and Y. Irizarry. Does stem stand out? examining racial/ethnic gaps in persistence across postsecondary fields. *Educational Researcher*, 48(3):133–144, 2019.

[43] J. E. Rogers, E. Ritchie, and L. B. Fritch. Hsls:09 base year to second follow-up public-use data file. https://nces.ed.gov/pubsearch/pubsinfo.asp?pubid=2018142, June 2018.

[44] Y. Romano, S. Bates, and E. Candes. Achieving equalized odds by resampling sensitive attributes. *Advances in Neural Information Processing Systems*, 33:361–371, 2020.

[45] T. Sanabria and A. Penner. Weeded out? gendered responses to failing calculus. *Social Sciences*, 6(2):47, 2017.

[46] P. Sapiezynski, V. Kassarnig, and C. Wilson. Academic performance prediction in a gender-imbalanced environment. 2017.

[47] H. Smith. Algorithmic bias: should students pay the price? *AI & society*, 35(4):1077–1078, 2020.

[48] J. Snow. AI technology is disrupting the traditional classroom. here's a progress report. https://www.pbs.org/wgbh/nova/article/ai-technology-is-disrupting-the-traditional-classroom/, 2019.

[49] M. K. Stein, J. H. Kaufman, M. Sherman, and A. F. Hillen. Algebra: A challenge at the crossroads of policy and practice. *Review of Educational Research*, 81(4):453–492, 2011.

[50] H. Wang, H. Hsu, M. Diaz, and F. P. Calmon. To split or not to split: The impact of disparate treatment in classification. *IEEE Transactions on Information Theory*, 67(10):6733–6757, 2021.

[51] A. Waters and R. Miikkulainen. Grade: Machine learning support for graduate admissions. *Ai Magazine*, 35(1):64–64, 2014.

[52] D. Wei, K. N. Ramamurthy, and F. Calmon. Optimized score transformation for fair classification. In *Proceedings of the Twenty Third International Conference on Artificial Intelligence and Statistics*, volume 108, 26–28 Aug 2020.

[53] C. Wilson, A. Ghosh, S. Jiang, A. Mislove, L. Baker, J. Szary, K. Trindel, and F. Polli. Building and auditing fair algorithms: A case study in candidate screening. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, pages 666–677, 2021.

[54] R. Yu, H. Lee, and R. F. Kizilcec. Should college dropout prediction models include protected attributes? In *Proceedings of the Eighth ACM Conference on Learning@ Scale*, pages 91–100, 2021.

[55] R. Yu, Q. Li, C. Fischer, S. Doroudi, and D. Xu. Towards accurate and fair prediction of college success: Evaluating different sources of student data. *International Educational Data Mining Society*, 2020.

[56] M. B. Zafar, I. Valera, M. Gomez Rodriguez, and K. P. Gummadi. Fairness beyond disparate treatment & disparate impact: Learning classification without disparate mistreatment. In *Proceedings of the 26th international conference on world wide web*, pages 1171–1180, 2017.

[57] R. Zemel, Y. Wu, K. Swersky, T. Pitassi, and C. Dwork. Learning fair representations. In *International conference on machine learning*, 2013.