Algorithmic unfairness mitigation in student models: When fairer methods lead to unintended results

Frank Stinar
University of Illinois Urbana–Champaign
fstinar2@illinois.edu

Nigel Bosch
University of Illinois Urbana–Champaign
pnb@illinois.edu

ABSTRACT

Systematically unfair education systems lead to different levels of learning for students from different demographic groups, which, in the context of AI-driven education, has inspired work on mitigating unfairness in machine learning methods. However, unfairness mitigation methods may lead to unintended consequences for classrooms and students. We examined preprocessing and postprocessing unfairness mitigation algorithms in the context of a large dataset, the State of Texas Assessments of Academic Readiness (STAAR) outcome data, to investigate these issues. We evaluated each unfairness mitigation algorithm across multiple machine learning models using different definitions of fairness. We then evaluated how unfairness mitigation impacts classifications of students across different combinations of machine learning models, unfairness mitigation methods, and definitions of fairness. On average, unfairness mitigation methods led to a 22% improvement in fairness. When examining the impacts of unfairness mitigation methods on predictions, we found that these methods led to models that can and did overgeneralize groups. Consequently, predictions made by such models may not reach the intended audiences. We discuss the implications for AI-driven interventions and student support.

Keywords

Machine learning, Unfairness mitigation, Fairness, Data science applications in education

1. INTRODUCTION

Student assessment has been plagued with biases against students from different demographic groups [2]. Various fairness metrics have been designed to evaluate the severity of biases [27], including biases in automatic assessments powered by machine learning [13]. Researchers have explored methods for mitigating such biases [3, 16, 21], including for assessment-related tasks such as student grade prediction [18]. However, the implications of applying unfairness mitigation methods to educational datasets are currently not well understood. In this paper, we explore unfairness mitigation on the State of Texas Assessments of Academic Readiness (STAAR) dataset [25], including examples of what impacts might be felt by students when models are optimized for different definitions of fairness.

The influx of data stemming from education contexts has allowed computational methods to be used for understanding education [4]. Alongside the creation of computer-based assessment methods, researchers have also analyzed current assessment data with machine learning and other computational methods [17]. These forms of analysis allowed researchers to uncover new biases towards different demographic groups in education and assessment methods [3]. Moreover, newer methods do not necessarily fix the biases present in traditional assessments [21]. However, the newer methods are needed when dealing with complex (e.g., nonlinear, multimodal) educational datasets—as are new methods to address biases present in computational analyses [7, 8].

Biases in the field of education have differing definitions, differing impacts, and can come from different places within educational and social systems [3]. In the study presented in this paper we are unable to disentangle the systemic biases present in assessment systems, as is often the case with educational data that concerns specific contexts, learning environments, or topics. Instead, we focused on biases in statistical measurements and biases present in the machine learning models. In particular, we investigate how unfairness mitigation methods and machine learning models impact student representations in data (and predictions), and if the biases found in these methods are analogous to biases in assessments—which may come from the assessments themselves or from systemic biases. In this paper, we explore these issues by addressing two research questions:

Research Question 1 (RQ1): Do demographic differences in standardized assessment scores correspond to biases in machine learning models? Standardized tests often contain biases [21]. Machine learning models have been shown to exacerbate biases that are present in training data [5]. RQ1 therefore examines the link between test score differences (including biases) and the biases that are found in machine learning models trained on closely related data from the same students. By comparing the biases that machine learning models elicit and the biases present in standardized test data, we are able to disentangle some of the sources of bias

that occur in data-driven assessment, and thereby inform future work on mitigating such biases. In alignment with previous work [3], we expect that biases present in standardized assessments and machine learning models will not perfectly align. If the biases do not align, differences between them indicate that new biases may be introduced—or, ideally, that there are opportunities for reducing such biases.

Research Question 2 (RQ2): What are the implications of machine learning unfairness mitigation methods when applied to student test score prediction? Researching this provides insight into the real-world implications for applying fairness metric optimization to human-based datasets. These implications can provide insight into the types of interventions that could take place in learning environments, because applying machine learning (or even expert-created) models to the real world without substantively examining the possible impacts can lead to dire consequences [1, 5].

By answering these two RQs, we tie test score biases and machine learning biases together through unfairness mitigation strategies and examine the implications of applying these strategies in education. By comparing the outputs of unfairness-mitigated machine learning models with the actual biases present in standardized tests we found a relationship between the two. Then, we showed how learning environments and students are impacted from applying fairness processes. Next, we address related research to position our work in the space of educational data mining and AI fairness research.

2. RELATED WORK

We focus on work related to fair AI methods and fair AI used within educational data mining. For a more general overview, see Fischer et al. [12]; we discuss fair AI, specifically, in the remainder of this section.

We first discuss fairness research being done in AI, regardless of domain. Many different research approaches have differing definitions of fairness and what fair AI looks like. For example, Hu et al. increased fairness in part through comparing differing groups positive predictive rates [16]; others have examined several different statistical definitions of fairness [6], which can even be contradictory with each other [11]. Work has also been done to *explain* many of the definitions of fairness that are used in fair AI research [26], independent of domain.

There is an increasing body of work in the field of educational data mining that uses fair AI, including the impact caused by algorithmic bias in education systems [3, 20]. This research being done in educational data mining with fair AI methods shows the growing use of machine learning in education research, complementing existing work in related topics like fair assessment (e.g., [21, 2]). These publications represent how biases are seen within different data mining avenues.

Using similar connections made in previous work between fair AI methods and educational data mining, we harness machine learning models and unfairness mitigation methods to examine performance differences between demographic groups in standardized tests.

Table 1: Breakdown of Demographic Groups

Demographic Identifier	# Occurrences	% Sample
Female	672,545	17.7%
Male	670,664	17.6%
Economic disadvantage	648,716	17.1%
Hispanic	554,697	14.6%
White	440,972	11.6%
Special education	315,072	8.3%
African American	240,901	6.3%
Asian	125,308	3.3%
Two or more races	80,817	2.1%
Pacific Islander	1,273	0.03%
American Indian	1,050	0.02%

3. METHODS

We used the assessment database from the STAAR Texas Education Agency dataset. These data were collected from the Teaching Trust (a now defunct leadership development group with the goal of eliminating opportunity gaps for students) between 2012 and 2019 [25]. This dataset has information from over 5 million students, which is approximately 10% of the public school students in the United States. The data include at most one demographic identifier per student—e.g., a student's race or gender might be included, but not both. Table 1 contains the breakdown of demographic-related information in the dataset.

3.1 Machine Learning Models

The three machine learning models we used were the logistic regression, random forest, and extremely randomized trees models. We used 5-fold cross validation and tuned hyperparameters via grid search. We trained all models using `scikit-learn` [22]. The logistic regression used the default hyperparameters from `scikit-learn` (i.e., a small L_2 regularization penalty). The random forest and extremely randomized trees models underwent hyperparameter tuning for the maximum depth of trees and the proportion of features samples for each tree.

3.2 Model Evaluation

For each of the machine learning models we used four unfairness mitigation methods (described in more detail below): disparate impact preprocessing, reweighing, equalized odds postprocessing, and calibrated equalized odds postprocessing. We evaluated each of these model/method combinations in terms of accuracy measured via area under the receiver operating characteristic curve (AUC) and four unfairness metrics: statistical parity difference, disparate impact ratio, average odds difference, and equal opportunity difference (described in the Appendix).

3.3 Unfairness Mitigation Algorithms

We implemented unfairness mitigation algorithms with the `AIF360` Python library [7].

3.3.1 Disparate Impact Preprocessing

Disparate impact preprocessing compares the label (passing the STAAR test) base rate across groups. The algorithm

takes this rate and edits features of the original data so that it is impossible to tell which group an individual belongs to.

3.3.2 Reweighing

The reweighing algorithm adds weight to each example during model training based upon the proportion of students students in different demographic groups and outcome (e.g., positive vs. negative class) groups. The equation for the weight is given in Equation 1.

$$w_{positive/group1} = \frac{(N_{group1})(N_{positive})}{(N_{all})(N_{positive/group1})} \quad (1)$$

3.3.3 Equalized Odds Postprocessing

Equalized odds postprocessing works to optimize the equalized odds fairness metric by changing predicted labels as needed to satisfy the metric. Specifically, the algorithm solves a linear program for probabilities. From these probabilities, classification labels are given [15]. An equalized odds predictor is made for this program from predicting on equalized odds incentive measurements for all classes.

3.3.4 Calibrated Equalized Odds Postprocessing

Calibrated equalized odds postprocessing follows a similar process to equalized odds postprocessing; however, it optimizes for equalized odds over a calibrated model output. Calibrated output is found when probability predictions align with the actual probability of observing the predicted outcomes [23].

4. RESULTS

We describe our results with respect to the two research questions outlined in the Introduction.

4.1 RQ1: Machine Learning Bias compared to Assessment Bias

RQ1 asks if biases found in machine learning models are analogous to the biases present in the STAAR assessment. Table 2 contains the values of accuracy and fairness metrics found after applying unfairness mitigation algorithms.

On average we observed a 22% improvement in fairness metric evaluation. However, different unfairness mitigation methods led to different trends in results. We found the disparate impact preprocessing method had the smallest impact, on average, across all fairness metrics. However, disparate impact preprocessing yielded the highest accuracy for each model, perhaps because it impacted the models the least. For both the random forest and extremely randomized trees models, reweighing unfairness mitigation led to the fairest predictions across all fairness metrics. We found that for some unfairness mitigation methods, especially equalized odds and calibrated equalized odds, the predictions were so influenced that the accuracy was no better than chance level (i.e., AUC $\leq$ .500).

4.2 RQ2: Unfairness Mitigation Implications

We answered RQ2 by analyzing each unfairness mitigation algorithm's impact on the STAAR data or models being used. The disparate impact preprocessing algorithm removed distinctions between groups in the dataset itself. The disparate impact removal process preserves within-group ranking of singular data points; however, group membership of singular data points are changed so it is not possible to discern what groups individuals belong to. The reweighing algorithm adds weights to different data points based on frequency of group membership to remove bias. Equalized odds postprocessing and calibrated equalized odds postprocessing do not edit the original dataset. Instead, they change output labels of singular data points with the objective of maximizing the equalized odds of the classifications. Each algorithm thus impacts either the student data itself or the process used to classify students as passing or failing the assessment, the implications of which need to be understood to determine whether such methods are procedurally fair—that is, whether the process by which decisions are made is fair, not just the fairness of the decisions themselves [9, 6, 14].

We found that using these unfairness mitigation algorithms resulted in unintended consequences for how students were computationally represented. For example, after disparate impact preprocessing, students who originally belonged to one group were now considered part of a different group for training. In the case of STAAR data, if the data were bimodal, with each group having their own pattern of demographic identifiers, students would be represented in data with different demographic identifiers than they actually have.

The impacts of each unfairness mitigation method are seen in the manipulation of the STAAR data. For example, some individuals in the "Female" demographic group were now in the "Asian" demographic group after disparate impact preprocessing; the algorithm preserved the within-group ranking of individuals, but changed group membership in unintended ways. This may have created a more fair model in terms of predictions, but while ignoring or confusing the systemic education problem present in the data.

In contrast, the two equalized odds based algorithms effectively changed whether or not students in the training data passed or failed the assessment. This may be less drastic than shifting students' demographic identifiers; however, decisions made upon this unfairness mitigation are imperfect if they obfuscate the problem of unequal learning.

Finally, the reweighing algorithm introduces weights to students for fairer classification. This method does not change the students' features, and thus might be considered a more faithful representation of students. However, weighting students may cause unintended effects as demographic group and assessment score combinations become more impactful to classification of others. This is especially true for students from smaller groups, who may find their characteristics or behaviors become far more important to a model's decisions than is desirable.

5. DISCUSSION

Our first research question predicted that demographic differences in test scores do correspond to biases seen in machine learning models trained on those data. Our results show that there were similarities between the biases present; however, models can further propagate biases present in the

Table 2: Fairness metric calculations. The *No Model* row represents the metrics calculated on the original dataset (for metrics that could be calculated from outcome labels in the dataset) for comparison to the rest of the algorithms.

Model	Unfairness Mitigation	AUC	Stat Parity Diff	Disp Imp Ratio	Avg Odds Diff	EqOpp Diff
No Model	None	–	-.172	.766	–	–
LogReg	None	.579	.113	1.419	.124	.162
RandFor	None	.623	-.621	0.379	-.639	-.529
Extra-Trees	None	.623	-.692	0.373	-.614	-.517
LogReg	Disparate Impact	.524	.112	1.427	.122	.161
LogReg	Reweighing	.474	.286	2.071	.293	.348
LogReg	Equalized Odds	.384	.306	1.875	.323	.333
LogReg	Calibrated Equalized Odds	.483	.483	2.391	.503	.486
RandFor	Disparate Impact	.512	-.675	0.325	-.692	-.586
RandFor	Reweighing	.522	.114	1.226	.123	.176
RandFor	Equalized Odds	.471	-.760	0.208	-.769	-.687
RandFor	Calibrated Equalized Odds	.472	-.438	0.232	-.425	-.398
Extra-Trees	Disparate Impact	.528	-.686	0.258	-.698	-.606
Extra-Trees	Reweighing	.518	.081	1.176	.089	.145
Extra-Trees	Equalized Odds	.436	-.605	0.291	0.605	-.539
Extra-Trees	Calibrated Equalized Odds	.445	-.665	0.235	-.668	-.597

data when trained on unfairness-mitigated data. This aligns with other research on biases present in education [24].

Different types of biases can be measured in our machine learning models that can not be measured within STAAR. Models might include biases that are separate from biases present in the test scores, indicating that machine learning models may exacerbate biases already present in data or even introduce new biases. Additionally, putting assessment data into a machine learning pipeline with unfairness mitigation can lead to negative implications as theorized by RQ2.

Our expectation for RQ2 was that applying unfairness mitigation strategies would lead to unintended real-world consequences for students. Indeed, unfairness mitigation strategies manipulated data in ways that could be perceived as unfair, especially with respect to procedural fairness. We investigated this by examining how unfairness mitigation impacted the STAAR data and the classifications made. RQ2 results show if one is planning to administer data-driven interventions in education, applying unfairness mitigation will likely overgeneralize groups. Thus, students who require intervention may not receive it and students who do not need intervention may receive one, not because of model inaccuracies necessarily but because of unfairness mitigation strategies. For example, if students with differing amount of background knowledge are misrepresented, learning outcomes could be negatively impacted [19]. This imprecision of groups that comes with unfairness mitigation can lead to unintended consequences when applied to real-world applications. Thus, unfairness mitigation methods must be applied with caution.

5.1 Limitations and Future Work

The study in this paper explored one prediction task, which—though representative of a large proportion of U.S. students—is not representative of all assessments nor any of the other educational outcomes and constructs of possible interest. Similarly, we examined a few machine learning models with a selection of preprocessing and postprocessing unfairness mitigation algorithms. We focused on methods that are common in (or well suited to) education data contexts, but the space of possible models and unfairness mitigation algorithms is far larger. Thus, future work would benefit from working with other educational datasets, machine learning models, and unfairness mitigation algorithms to further examine how these methods can impact the representations of students in data analysis. Finally, a nearly insurmountable limitation of this work is that we are unable to disentangle the systemic biases that lead to different amounts of learning (e.g., structural racism and classism) versus the assessment biases (e.g., lack of cultural responsiveness) and algorithmic biases that contribute to biased measurement. We are unable to resolve this problem, but we do evaluate the alignment of these biases in this paper, and suggest that future work with quasi-experimental analyses may be one possible route to address this limitation.

5.2 Conclusion

Education is plagued with unfairness for differing demographic groups [10]. Unfairness mitigation methods have potential to reduce unfairness in data-driven assessment and student support, but when applied to educational datasets these methods may lead to unintended negative consequences. We explored machine learning pipelines with unfairness mitigation methods applied, and examined how these methods would affect the representations of individual students. We expect that these findings will guide the selection of unfairness mitigation methods in future work, and hope that with our findings in mind, when decisions are made from models based on educational data, less harm is done to students from a lack of caution when choosing models and algorithm.

6. REFERENCES

[1] J. Anders, C. Dilnot, L. Macmillan, and G. Wyness. Grade expectations: How well can we predict future grades based on past performance? CEPEO Working Paper Series 20-14, UCL Centre for Education Policy and Equalising Opportunities, 2020.

[2] K. Arbuthnot. *Filling in the blanks: Understanding standardized testing and the Black–White achievement gap.* IAP Information Age Publishing, Charlotte, NC, US, 2011.

[3] R. S. Baker and A. Hawn. Algorithmic bias in education. *International Journal of Artificial Intelligence in Education*, 2021.

[4] R. S. Baker and K. Yacef. The state of educational data mining in 2009: A review and future visions. *Journal of Educational Data Mining*, 1(1):3–17, Oct. 2009. The file is in PDF format. If your computer does not recognize it, simply download the file and then open it with your browser.

[5] M. Barenstein. ProPublica's COMPAS data revisited, 2019.

[6] C. Belitz, L. Jiang, and N. Bosch. Automating procedurally fair feature selection in machine learning. In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*, page 379–389, New York, NY, USA, 2021. Association for Computing Machinery.

[7] R. K. E. Bellamy, K. Dey, M. Hind, S. C. Hoffman, S. Houde, K. Kannan, P. Lohia, J. Martino, S. Mehta, A. Mojsilović, S. Nagar, K. N. Ramamurthy, J. Richards, D. Saha, P. Sattigeri, M. Singh, K. R. Varshney, and Y. Zhang. AI Fairness 360: An extensible toolkit for detecting and mitigating algorithmic bias. *IBM Journal of Research and Development*, 63(4/5):1–15, 2019.

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

[9] K. Burke and S. Leben. Procedural fairness: A key ingredient in public satisfaction. *Court Review*, 44(1-2):4–25, 2007.

[10] H.-Y. S. Cherng, P. F. Halpin, and L. A. Rodriguez. Teaching bias? Relations between teaching quality and classroom demographic composition. *American Journal of Education*, 128(2):000–000, 2022.

[11] I. Cojuharenco and D. Patient. Workplace fairness versus unfairness: Examining the differential salience of facets of organizational justice. *Journal of Occupational and Organizational Psychology*, 86:371–393, 2013.

[12] C. Fischer, Z. A. Pardos, R. S. Baker, J. J. Williams, P. Smyth, R. Yu, S. Slater, R. Baker, and M. Warschauer. Mining big data in education: Affordances and challenges. *Review of Research in Education*, 44(1):130–160, 2020.

[13] J. Gardner, M. O'Leary, and L. Yuan. Artificial intelligence in educational assessment: 'Breakthrough? Or buncombe and ballyhoo?'. *Journal of Computer Assisted Learning*, 37(5):1207–1216, 2021.

[14] N. Grgić-Hlača, M. B. Zafar, K. P. Gummadi, and A. Weller. Beyond distributive fairness in algorithmic decision making: Feature selection for procedurally fair learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32, pages 51–60, Apr. 2018.

[15] M. Hardt, E. Price, and N. Srebro. Equality of opportunity in supervised learning. volume abs/1610.02413, 2016.

[16] Q. Hu and H. Rangwala. Towards fair educational data mining: A case study on detecting at-risk students. In *Proceedings of the 13th International Conference on Educational Data Mining (EDM 2020)*, pages 431–437. International Educational Data Mining Society, 2020.

[17] S. Hussain, N. Abdulaziz Dahan, F. Ba-Alwib, and R. Najoua. Educational data mining and analysis of students' academic performance using WEKA. *Indonesian Journal of Electrical Engineering and Computer Science*, 9:447–459, 02 2018.

[18] 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*, page 608–617, New York, NY, USA, 2021. Association for Computing Machinery.

[19] S. Kalyuga. Expertise reversal effect and its implications for learner-tailored instruction. *Educational Psychology Review*, 19(4):509–539, Dec. 2007.

[20] R. F. Kizilcec and H. Lee. Algorithmic fairness in education. In W. Holmes and K. Porayska-Pomsta, editors, *Ethics in Artificial Intelligence in Education*. Taylor & Francis, in press.

[21] R. Mendoza-Denton. A social psychological perspective on the achievement gap in standardized test performance between white and minority students: Implications for assessment. *The Journal of Negro Education*, 83(4):465–484, 2014.

[22] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

[23] G. Pleiss, M. Raghavan, F. Wu, J. M. Kleinberg, and K. Q. Weinberger. On fairness and calibration. volume abs/1709.02012, 2017.

[24] J. G. Starck, T. Riddle, S. Sinclair, and N. Warikoo. Teachers are people too: Examining the racial bias of teachers compared to other american adults. *Educational Researcher*, 49(4):273–284, 2020.

[25] T. Texas Education Agency. STAAR. https://rptsvr1.tea.texas.gov/perfreport/account/2019/download/acctref.html, 2019.

[26] S. Verma and J. Rubin. Fairness definitions explained. In *Proceedings of the International Workshop on Software Fairness*, FairWare '18, page 1–7, New York, NY, USA, 2018. Association for Computing Machinery.

[27] M. B. Zafar, I. Valera, M. G. Rodriguez, and K. P. Gummadi. Fairness constraints: Mechanisms for fair classification. In *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics*,

pages 962–970, 2017.

APPENDIX

A. FAIRNESS METRICS

We measured fairness according to four quantitative definitions, detailed below. The combination of these metrics allowed us to holistically judge the impact of the unfairness mitigation algorithms on the machine learning models.

A.1 Statistical Parity Difference

The statistical parity difference metric compares the difference between the groups' probability of being predicted to pass the assessment. A value of 0 indicates both groups were predicted to pass the assessment with equal probability. A positive value means one group is predicted to pass the assessment more, a negative value means the other group is predicted to pass the assessment more.

$$SP = P(\hat{Y} = 1|D = group2) - P(\hat{Y} = 1|D = group1) \quad (2)$$

A.2 Disparate Impact Ratio

Disparate impact ratio is the ratio of the differing groups being predicted to pass the assessment. A value of 1 indicates that both groups are predicted to pass the assessment with equal probability. A value greater than 1 indicates that one group is predicted to pass the assessment more than the other group, while a value less than 1 indicates the opposite.

$$DI = \frac{P(\hat{Y} = 1|D = group2)}{P(\hat{Y} = 1|D = group1)} \quad (3)$$

A.3 Average Odds Difference

Average odds difference measures the average of the difference in the false positive rate and the true positive rate for the differing groups. A value of 0 indicates an equality of odds. A value of -1 or 1 indicates maximum possible inequality.

$$AO = \frac{\begin{array}{c}(FPR_{group2} - FPR_{group1})+\\(TPR_{group2} - TPR_{group1})\end{array}}{2} \quad (4)$$

A.4 Equal Opportunity Difference

The equal opportunity difference metric compares the difference in true positive rates between the two groups. A value of 0 indicates equality between groups. A value of -1 or 1 indicates high inequality.

$$EO = TPR_{group2} - TPR_{group1} \quad (5)$$