Measuring Feature-Level Changes in Preprocessing Unfairness Mitigation

Hanning Lin
University of California Berkley
hanning4@illinois.edu

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

ABSTRACT

Gender bias in machine learning models can perpetuate and amplify existing societal inequalities, especially in domains where the users are vulnerable such as in education. In this study, we measure the effectiveness of multiple preprocessing unfairness mitigation strategies to reduce gender bias in predictive models trained on student data. Furthermore, we investigate how these mitigation techniques influence the underlying feature representations of the data, yielding insights into the procedural mechanisms by which bias is reduced. Through a series of experiments using four education datasets, we improved fairness and found that some methods dramatically modify data more than others. Using a diverse set of fairness metrics and statistical divergence methods, such as Wasserstein distance, we observe that the level of data modification and the level of bias reduction are not strongly related. To further isolate the role of data transformation, we additionally analyze mitigation methods that induce comparable levels of feature-level data modification, enabling more controlled comparisons across techniques. These results provide an initial analysis into the complexity of the unfairness mitigation step for educational modeling.

Keywords

Algorithmic fairness, Unfairness mitigation, Data transformation, Grade prediction

1. INTRODUCTION

Researchers and educators have harnessed big data mining techniques as a powerful way to leverage increasing quantities of educational data to improve learning outcomes, personalize education, and support diverse groups of students [6, 10]. One common tool that leads to these benefits is predictive modeling, which has proven useful in tasks such as grade prediction, mastery prediction, course recommendation, and dropout prevention [2, 29, 20]. However, as predictive models become more influential in educational decision-making, concerns about the fairness of these models have grown. Bias in educational data can perpetuate inequalities and even undermine the potential benefits of educational data mining methods [33, 24, 7]. Addressing these biases is critical to ensuring that predictive models do not exacerbate existing disparities in educational systems.

Education researchers have already begun to mitigate and understand bias in machine learning-powered educational systems [43]. While there has been varying success in applying unfairness mitigation techniques, there has been limited exploration of how preprocessing-based unfairness mitigation methods transform data in unexpected or opaque ways [19, 13]. Understanding these data transformations is important, as they can have downstream effects on model performance, interpretability, and fairness.

In this paper, we evaluate multiple preprocessing unfairness mitigation strategies across four educational datasets using a standardized data-transformation pipeline. Beyond measuring overall fairness improvements, we analyze how these mitigation techniques influence feature-level data representations. Rather than relying on a single divergence measure, we assess mitigation effectiveness using multiple fairness metrics, allowing for a more comprehensive evaluation of bias reduction. In addition, we compare mitigation methods that induce similar levels of feature-level data modification, enabling more controlled comparisons that disentangle the effects of the mitigation mechanism from the extent of data transformation itself. Thus, we examine how different preprocessing unfairness mitigation techniques can alter educational data and how the transformations influence the fairness and utility of grade prediction models. Our analysis is shaped by the following research questions:

Research Question 1 (RQ1): Are preprocessing unfairness mitigation strategies applicable in reducing gender bias in grade prediction?

Research Question 2 (RQ2): How can preprocessing unfairness mitigation strategies differentially impact student data?

We answer RQ1 by analyzing how four preprocessing unfairness mitigation algorithms can reduce gender bias in four education datasets. To answer RQ2, we analyze if the amount of transformation is linked to the level of bias mitigated.

The preliminary results in this paper extend beyond tech-

nical aspects of applying unfairness mitigation to grade-prediction. Through uncovering how preprocessing techniques differentially transform student data, the research helps to better understand the tradeoffs between unfairness mitigation and data integrity. That is, modifying the data could result in procedural unfairness if the resulting data represents students in unexpected, inaccurate, or otherwise undesirable ways [33, 36]. Thus, we hope to help education researchers in deciding how to implement unfairness mitigation techniques into transparent educational systems.

2. BACKGROUND

Machine learning pipelines in educational data mining often result in some type of prediction or learner model with the goal of improving student learning. These models are often used to predict a student's risk of dropping out [39, 1, 25], whether a learner has mastered certain topics in learning sessions [16, 29], grades on exams [3, 40, 17, 30], to name a few applications. As these predictive models are applied into classroom and online settings, concerns about the biases within these models have become increasingly examined [5, 23, 35]. To address the biases in the models, researchers in education and related fields have developed techniques to mitigate and measure the biases in the development of the models [19, 38, 37]. These techniques are often grouped into categories based on which step in a machine learning pipeline that act on. Preprocessing techniques transform the dataset used to train machine learning models [15, 22, 44]. Inprocessing techniques add other steps and parameters to the training process of models [41, 31]. Postprocessing techniques edit the predictions of the models [32, 28].

Researchers in education have already begun to test the utility of different unfairness mitigation techniques. In fact, educators have cared about issues of bias and fairness for decades [21]. Recently, researchers have developed techniques specific for educational data [38, 15, 8], examined how these techniques impact the data themselves [33], and surveyed how these algorithmic fairness issues should be handled in education [4, 14]. In this paper, we contribute to this research by examining if certain preprocessing unfairness mitigation techniques are useful at mitigating gender bias in grade prediction. Furthermore, we examine if there is a relationship between the gender bias mitigated and the amount that preprocessing techniques distort educational data, which is essential to understand for situations where such data distortion is undesirable.

3. METHOD

In the following section, we detail the datasets and experimental design. For RQ1, we evaluate whether preprocessing unfairness mitigation strategies reduce gender bias in grade prediction by comparing models trained on original datasets to models trained on mitigated datasets. For RQ2, we quantify the magnitude of data transformation introduced by each mitigation technique and analyze whether greater data modification corresponds to greater bias reduction in the trained models.

3.1 Data

We use four student performance datasets covering diverse age groups for prediction tasks. For all datasets, we frame the problem as a binary classification task, with student gender as the sensitive attribute. Gender is treated as a binary sensitive feature in our analyses since we are following the terminology used in the datasets. Other potentially sensitive attributes are removed prior to analysis (we note that unfairness can also exist between intersectional groups, but for our preliminary analysis we focus on one demographic all the datasets share). Table 1 includes the base rates and demographics of the datasets.

Table 1: Base rates of gender across datasets

Dataset	Gender	No. Obs	Pass (%)
HAR	Man	301	21.93
	Woman	413	17.92
POR	Man	266	64.29
	Woman	383	73.37
MAT	Man	187	56.68
	Woman	208	49.52
HSP	Man	87	48.28
	Woman	58	25.86

Harvard Student Performance Data (HAR): The HAR dataset is used to predict final grades in a Physics course. HAR contains 21 categorical and numerical features for 714 students. The 21 features consist of academic information and individual differences.

Student Portuguese and Math Performance Data (POR, MAT): The POR dataset predicts final grades in a Portuguese language course. For standardization, we binarized the outcome into pass or fail. POR contains 30 categorical and numerical features across 649 students. Similar to HSP, features in POR consist of general academic information (e.g., grades, attendance, etc.) and and individual difference measures. The MAT dataset undergoes the same transformation but is used to predict math grades.

Higher Education Student Performance Data (HSP): The HSP dataset is used to predict end-of-term grades. For our binary classification task, we transform the final grades as pass or fail. HSP contains 31 categorical and numerical features across 145 students in total. Features in this dataset focus on academics alongside a few individual difference measures (e.g., parental information).

3.2 Experiments

To answer RQ1, we applied four preprocessing-based unfairness mitigation methods: reweighting (Rw), FAIR-SMOTE (FSm), SMOTE (Sm), and Disparate Impact Remover (DRe). These methods were selected because they modify the data at different levels: reweighting adjusts instance weights without altering feature values, whereas SMOTE and FAIR-SMOTE generate synthetic samples to rebalance class and group distributions.

Reweighting (Rw) is a preprocessing technique designed to mitigate bias related to sensitive attributes. It assigns different weights to training instances based on the joint distribution of the sensitive attribute and the class label, reducing the influence of biased correlations during model training [22].

Table 2: AUC and fairness results of random forest models trained on the original and unfairness mitigated datasets.

Dataset	Mitigation	AUC	SP	DI	AO
HAR	–	.63	.05	1.20	.05
	Rw	.70	.06	1.11	.06
	FSm	.71	.07	1.19	.07
	Sm	.68	.05	1.21	.05
	DRe	.89	.04	0.91	.02
POR	–	.63	-.09	0.91	-.10
	Rw	.63	-.01	0.99	-.01
	FSm	.58	-.10	0.89	-.11
	Sm	.65	-.14	0.85	-.14
	DRe	.55	.07	0.92	.06
MAT	–	.58	.27	3.47	.27
	Rw	.61	.04	1.11	.04
	FSm	.64	.24	3.28	.24
	Sm	.63	.32	2.48	.32
	DRe	.44	.02	0.85	.04
HSP	–	.77	.03	1.07	-.15
	Rw	.60	-.04	0.91	-.07
	FSm	.53	-.03	0.93	-.12
	Sm	.51	-.23	0.62	-.36
	DRe	.79	-.05	0.67	-.02

FAIR-SMOTE (FSm) extends SMOTE by incorporating fairness constraints, generating synthetic samples that balance both class labels and sensitive attribute distributions [11].

SMOTE (Sm) addresses class imbalance by generating synthetic samples for the minority class through interpolation between existing instances and their nearest neighbors [12].

Disparate Impact Remover (DRe) applies a feature repair step that adjusts values so group-wise distributions align more closely with a fairness target (e.g., the "80% rule"), rather than changing instance weights or the label column [18].

We evaluated the effectiveness of these mitigation strategies using both predictive performance and fairness metrics. Predictive performance was measured using the area under the ROC curve (AUC). Fairness was assessed using statistical parity difference (SP), disparate impact (DI), and average odds difference (AO) [7, 41, 9]. For all experiments, we trained a random forest classifier with default hyperparameters to model student performance. The models were implemented using the *Scikit-learn* Python library, and the unfairness mitigation algorithms were implemented using the AI Fairness 360 toolkit and DebiasEd [27, 9, 34].

To answer RQ2, we analyze how each method transforms feature distributions. For each dataset, we compare the distributions of each feature before and after preprocessing using the Wasserstein distance (a metric that measures how different two distributions are by calculating how much work it would take to transform one into the other) [26].

4. RESULTS

The results of the unfairness mitigation on AUC and fairness measures is in Table 2. Considering that each dataset had differences in base rates between groups, we expected that the baseline model for each dataset could have bias due to differences in predicted rates. There is some bias in the baseline models denoted as the first row for each of the datasets in Table 2. Perfect fairness for SP and AO is 0, and perfect fairness for DI is 1. We report means across 5-fold cross-validation for each dataset.

For RQ1, we compared the fairness metrics of the baseline model with the fairness metrics of the models trained on the unfairness mitigated data. For each dataset, at least one model trained using unfairness mitigated data was fairer in terms of at least one metric. For example, Rw resulted in better DI by .09 given the HAR dataset. The MAT baseline showed the strongest bias (SP = .27, DI = 3.47, AO = .27). All four unfairness mitigation techniques were able to reduce bias with three of them also improving AUC in tandem. Thus, we conclude that the chosen preprocessing methods are capable of reducing gender bias in these models, albeit at differing levels.

Given some gender bias can be mitigated, we analyzed how the preprocessing differentiall impacts student data (RQ2). We found the average Wasserstein distance of the features before and after preprocessing across the four techniques and four datasets. The results are in Figure 1.

Our preliminary analysis finds slight trends in the data transformation of the mitigation methods. For example, generally, Rw and Sm transformed data more than the FSm and DRe methods. Also, while Rw had the highest Wasserstein distance across all features for HAR, the transformation did not relate to high levels of unfairness mitigation or AUC improvement. In fact, DRe resulted in better performance and fairness despite also having a lower Wasserstein distance. In contrast, DRe did not have better performance for MAT (the dataset with the highest gender bias) while still having the lowest Wasserstein distance. Thus, even though these four methods transform data at similar rates, our preliminary analysis does not find a relationship between the levels of transformation and the unfairness mitigated. The results further demonstrate this with three unique unfairness mitigation methods resulting in the least gender bias and these methods being unrelated to the level of data transformation.

5. DISCUSSION AND FUTURE WORK

The experiments revealed that preprocessing unfairness mitigation strategies are effective in reducing gender bias in grade prediction (RQ1), and examined how different types of preprocessing distort student data (RQ2). The results indicate that the preprocessing methods do not equally mitigate gender bias in grade prediction, and that inequality is seemingly not related to the amount that a dataset is transformed.

There are cases where preprocessing unfairness mitigation strategies can reduce gender bias in grade prediction within models trained on each of the datasets. However, the preprocessing methods are not equally effective for each dataset. As expected from previous research, different unfairness mitigation methods are more effective at improving certain dimensions of fairness [41]. At the same time, these statistical fairness definitions encode specific normative assumptions and improving them does not automatically imply equitable

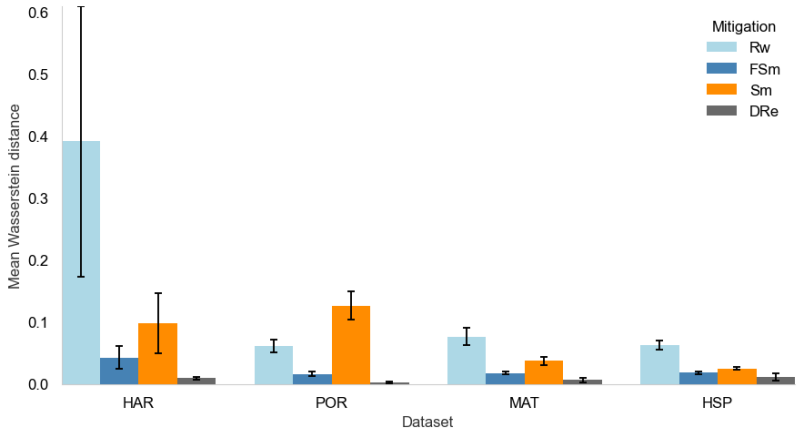

Figure 1: The average data divergence of the four unfairness mitigation methods on each of the datasets. On average, Rw distorts the data the most, followed by Sm, FSm, and DRe.

outcomes. However, notably, for the HAR and HSP datasets the model trained on a preprocessed version of the data using DRe resulted in a model that was fairer (in at least one dimension) *and* had better performance than the baseline. We also emphasize that mitigation is not uniformly beneficial: there are settings where fairness improves while AUC falls, where AUC improves while fairness worsens, or where one fairness metric improves while another degrades.

Preprocessing unfairness mitigation strategies having differential impacts on the data emphasizes the need for careful selection and adaptation of preprocessing unfairness mitigation methods in education. While DRe achieved fairness while minimally transforming the dataset, Rw achieved (sometimes better) fairness guarantees through distorting more data (in terms of the POR dataset). Furthermore, we found no relationship between the performance of a model, the unfairness of a model, and the amount in which the data was transformed. These findings are important for education researchers as they suggest that the most effective unfairness mitigation method is not necessarily the one that leads to the most transformation in the dataset. This flexibility in method is valuable in education as certain features being left untransformed could lead to greater explanatory power.

For practitioners, the correct mitigation to choose could be based on what must remain stable for the decision context: if preserving the literal feature values matters for communication and review, Rw may be preferable; if the priority is repairing dependence between protected attributes and outcomes in the input space, feature repair methods may be appropriate; if the issue is class imbalance, oversampling approaches may be relevant.

Therefore, future work can find mitigation strategies that improve fairness without sacrificing the transparency and interpretability of education models. Preprocessing techniques having seemingly no tradeoff between level of data distortion and unfairness mitigated offers opportunities for current preprocessing techniques to be improved and adapted for educational data. While DRe offers a possibility in this regard (through transforming only a few features), more research needs to be done to tailor this and other methods to education data. Moreover, Rw or Sm could be detrimental to the interpretability of models since they have greater data transformation. Furthermore, given that previous research has found that unfairness mitigation can transform certain features that could lead to procedural fairness harms [33], future work should move beyond our average estimates of Wasserstein distance. The next step would be to examine the specific features that are transformed by each preprocessing method. This could further provide insight into researchers as some methods could transform features that are important for model interpretability.

The study is limited by the sensitive attribute we chose to analyze, the unfairness mitigation techniques used, and the set of datasets. Since we focus on gender difference in grade prediction in a select set of tabular, categorical datasets, our results might not be generalizable to other types of data commonly used in educational data mining (e.g., knowledge tracing data [42]). Additionally, we only include a single (binary) sensitive attribute, which does not address intersectional fairness. Furthermore, we used only a few of an ever-increasing number of possible preprocessing unfairness mitigation techniques [19, 38]. Thus, our research could be further validated by testing the data distortion of other unfairness mitigation techniques.

The variability found in the relationship between dataset divergence, performance, and unfairness mitigation calls for multiple directions for future work. Our findings indicate that certain techniques could be promising in navigating this relationship, such as DRe, but future research needs to further test these hypotheses. Furthermore, future work can support the development of more education-specific preprocessing techniques with a focus on interpretability. Future extensions will include per-feature attribution of change (beyond dataset-level averages) paired with domain-informed features that must remain interpretable for instructors or advisors. Longer-term work could incorporate stakeholder-centered evaluation beyond offline metrics, and develop constraints common in education.

6. REFERENCES

[1] A. Alhothali, M. Albsisi, H. Assalahi, and T. Aldosemani. Predicting Student Outcomes in Online Courses Using Machine Learning Techniques: A Review. *Sustainability*, 14(10):1–14, 2022.

[2] K. E. Arnold and M. D. Pistilli. Course signals at Purdue: using learning analytics to increase student success. In *Proceedings of the 2nd International Conference on Learning Analytics and Knowledge*, LAK '12, pages 267–270, New York, NY, USA, 2012. Association for Computing Machinery.

[3] R. Asif, A. Merceron, S. A. Ali, and N. G. Haider. Analyzing undergraduate students' performance using educational data mining. *Computers & Education*, 113:177–194, Oct. 2017.

[4] R. S. Baker, L. Esbenshade, J. Vitale, and S. Karumbaiah. Using Demographic Data as Predictor Variables: a Questionable Choice. *Journal of Educational Data Mining*, 15(2):22–52, June 2023.

[5] R. S. Baker and A. Hawn. Algorithmic Bias in Education. *International Journal of Artificial Intelligence in Education*, 32(4):1052–1092, Dec. 2022.

[6] R. S. Baker and P. S. Inventado. Educational Data Mining and Learning Analytics. In J. A. Larusson and B. White, editors, *Learning Analytics: From Research to Practice*, pages 61–75. Springer New York, New York, NY, 2014.

[7] S. Barocas, M. Hardt, and A. Narayanan. *Fairness and Machine Learning: Limitations and Opportunities*. MIT Press, 2023.

[8] 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*, pages 379–389, Virtual Event USA, July 2021. ACM.

[9] R. K. E. Bellamy, A. Mojsilovic, S. Nagar, K. N. Ramamurthy, J. Richards, D. Saha, P. Sattigeri, M. Singh, K. R. Varshney, Y. Zhang, K. Dey, M. Hind, S. C. Hoffman, S. Houde, K. Kannan, P. Lohia, J. Martino, and S. Mehta. AI Fairness 360: An extensible toolkit for detecting and mitigating algorithmic bias. *IBM Journal of Research and Development*, 63(4/5):1–20, July 2019.

[10] C. Romero and S. Ventura. Educational Data Mining: A Review of the State of the Art. *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, 40(6):601–618, Nov. 2010.

[11] J. Chakraborty, S. Majumder, and T. Menzies. Bias in machine learning software: why? how? what to do? In *Proceedings of the 29th ACM Joint Meeting on European Software Engineering Conference and Symposium on the Foundations of Software Engineering*, ESEC/FSE 2021, pages 429–440, New York, NY, USA, 2021. Association for Computing Machinery.

[12] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer. SMOTE: Synthetic Minority Over-sampling Technique. *Journal of Artificial Intelligence Research*, 16:321–357, June 2002.

[13] Z. Chen, J. M. Zhang, F. Sarro, and M. Harman. A Comprehensive Empirical Study of Bias Mitigation Methods for Machine Learning Classifiers. *ACM Trans. Softw. Eng. Methodol.*, 32(4):1–30, May 2023.

[14] S. V. Chinta, Z. Wang, Z. Yin, N. Hoang, M. Gonzalez, T. L. Quy, and W. Zhang. FairAIED: Navigating Fairness, Bias, and Ethics in Educational AI Applications, July 2024.

[15] J. M. Cock, M. Bilal, R. Davis, M. Marras, and T. Kaser. Protected attributes tell us who, behavior tells us how: A comparison of demographic and behavioral oversampling for fair student success modeling. In *LAK23: 13th International Learning Analytics and Knowledge Conference*, pages 488–498, 2023.

[16] A. T. Corbett and J. R. Anderson. Knowledge tracing: Modeling the acquisition of procedural knowledge. *User Modelling and User-Adapted Interaction*, 4(4):253–278, 1995.

[17] P. Cortez and A. Silva. Using data mining to predict secondary school student performance. In J. L. Afonso, C. Cuoto, A. Lago Ferreiro, J. S. Martins, and A. Nogueiras Melèndez, editors, *Proceedings of 5th Annual Future Business Technology Conference*, volume 5, pages 5–12. EUROSIS-ETI, 2008.

[18] 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.

[19] M. Hort, Z. Chen, J. M. Zhang, M. Harman, and F. Sarro. Bias Mitigation for Machine Learning Classifiers: A Comprehensive Survey. *ACM J. Responsib. Comput.*, 1(2):1–52, June 2024.

[20] Q. Hu and H. Rangwala. Towards fair educational data mining: A case study on detecting at-risk students. In A. N. Rafferty, J. Whitehill, C. Romero, and V. Cavalli-Sforza, editors, *Proceedings of the 13th International Conference on Educational Data Mining, EDM 2020, Fully virtual conference, July 10-13, 2020*. International Educational Data Mining Society, 2020.

[21] B. Hutchinson and M. Mitchell. 50 Years of Test (Un)fairness: Lessons for Machine Learning. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, pages 49–58, Jan. 2019.

[22] F. Kamiran and T. Calders. Data preprocessing techniques for classification without discrimination. *Knowledge and Information Systems*, 33(1):1–33, Oct. 2012.

[23] R. F. Kizilcec and H. Lee. Algorithmic fairness in education. In *The Ethics of Artificial Intelligence in Education*, pages 174–202. Routledge, New York, 1 edition, Aug. 2022.

[24] N. Mehrabi, F. Morstatter, N. Saxena, K. Lerman, and A. Galstyan. A Survey on Bias and Fairness in Machine Learning. *ACM Computing Surveys*, 54(6):1–35, July 2022.

[25] C. Pan and Z. Zhang. Examining the algorithmic fairness in predicting high school dropouts. In B. PaaÃŸen and C. D. Epp, editors, *Proceedings of the 17th International Conference on Educational Data Mining*, pages 262–269, Atlanta, Georgia, USA, July 2024. International Educational Data Mining Society.

[26] V. M. Panaretos and Y. Zemel. Statistical Aspects of

Wasserstein Distances. *Annual Review of Statistics and Its Application*, 6(Volume 6, 2019):405–431, 2019. Type: Journal Article.

[27] 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(85):2825–2830, 2011.

[28] G. Pleiss, M. Raghavan, F. Wu, J. Kleinberg, and K. Q. Weinberger. On fairness and calibration. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, page 5684–5693, Red Hook, NY, USA, 2017. Curran Associates Inc.

[29] S. Qu, K. Li, B. Wu, X. Zhang, and K. Zhu. Predicting Student Performance and Deficiency in Mastering Knowledge Points in MOOCs Using Multi-Task Learning. *Entropy*, 21(12):12–16, Dec. 2019.

[30] R. J. Quinn and G. Gray. Prediction of student academic performance using Moodle data from a Further Education setting. *Irish Journal of Technology Enhanced Learning*, 5(1), Oct. 2019.

[31] N. Schreuder and E. Chzhen. Classification with abstention but without disparities. In C. de Campos and M. H. Maathuis, editors, *Proceedings of the Thirty-Seventh Conference on Uncertainty in Artificial Intelligence*, volume 161 of *Proceedings of Machine Learning Research*, pages 1227–1236. PMLR, July 2021.

[32] P. Snel and S. van Otterloo. Practical bias correction in neural networks: a credit default prediction case study. *Computers and Society Research Journal*, 3:1–12, 2022.

[33] F. Stinar and N. Bosch. Algorithmic unfairness mitigation in student models: When fairer methods lead to unintended results. In *Proceedings of the 15th International Conference on Educational Data Mining, EDM 2022*, Proceedings of the 15th International Conference on Educational Data Mining, EDM 2022. International Educational Data Mining Society, 2022.

[34] F. Stinar, J. Cock, R. Kizilcec, and T. Kaser. Applying DebiasEd: A Package for Mitigating Unfairness in Educational Data. In C. Mills, G. Alexandron, D. Taibi, G. L. Bosco, and L. Paquette, editors, *Proceedings of the 18th International Conference on Educational Data Mining*, pages 695–698, Palermo, Italy, July 2025. International Educational Data Mining Society.

[35] S. Tschiatschek, M. Knobelsdorf, and A. Singla. Equity and fairness of Bayesian knowledge tracing. In A. Mitrovic and N. Bosch, editors, *Proceedings of the 15th International Conference on Educational Data Mining*, pages 578–582, Durham, United Kingdom, July 2022. International Educational Data Mining Society.

[36] T. R. Tyler. What is Procedural Justice?: Criteria Used by Citizens to Assess the Fairness of Legal Procedures. *Law & Society Review*, 22(1):103–135, 1988.

[37] V. A. Vabenska, M. Verger, M. M. T. Rodrigo, C. J. G. Monterozo, R. S. Baker, M. Z. N. L. Saavedra, S. Lalla, and A. Shimada. Evaluating algorithmic bias in models for predicting academic performance of filipino students. In B. PaaÃŸen and C. D. Epp, editors, *Proceedings of the 17th International Conference on Educational Data Mining*, pages 744–751, Atlanta, Georgia, USA, July 2024. International Educational Data Mining Society.

[38] M. Verger, S. Lallé, F. Bouchet, V, and a. Luengo. Is Your Model "MADD"? A Novel Metric to Evaluate Algorithmic Fairness for Predictive Student Models. In *Proceedings of the 16th International Conference on Educational Data Mining*, pages 91–102. International Educational Data Mining Society, July 2023.

[39] W. Wang, Y. Zhao, Y. J. Wu, and M. Goh. Factors of dropout from MOOCs: a bibliometric review. *Library Hi Tech*, 41(2):432–453, June 2023.

[40] T. Yanagiura, S. Yano, M. Kihira, and Y. Okada. Examining Algorithmic Fairness for First- Term College Grade Prediction Models Relying on Pre-matriculation Data. *Journal of Educational Data Mining*, 15(3):1–25, Dec. 2023.

[41] 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, Perth Australia, Apr. 2017. International World Wide Web Conferences Steering Committee.

[42] A. F. Zambrano and R. S. Baker. Long-Term Prediction from Topic-Level Knowledge and Engagement in Mathematics Learning. In *Proceedings of the 14th Learning Analytics and Knowledge Conference*, LAK '24, pages 66–77, New York, NY, USA, 2024. Association for Computing Machinery.

[43] A. F. Zambrano, J. Zhang, and R. S. Baker. Investigating Algorithmic Bias on Bayesian Knowledge Tracing and Carelessness Detectors. In *Proceedings of the 14th Learning Analytics and Knowledge Conference*, pages 349–359, Kyoto Japan, Mar. 2024. ACM.

[44] R. Zemel, Y. Wu, K. Swersky, T. Pitassi, and C. Dwork. Learning Fair Representations. In S. Dasgupta and D. McAllester, editors, *Proceedings of the 30th International Conference on Machine Learning*, volume 28 of *Proceedings of Machine Learning Research*, pages 325–333, Atlanta, Georgia, USA, June 2013. PMLR.