Investigating Algorithmic Bias on Bayesian Knowledge Tracing and Carelessness Detectors

Andres Felipe Zambrano*
Graduate School of Education,
University of Pennsylvania
afzambrano97@gmail.com

Jiayi Zhang
Graduate School of Education,
University of Pennsylvania
jzhang7718@gmail.com

Ryan S. Baker
Graduate School of Education,
University of Pennsylvania
ryanshaunbaker@gmail.com

ABSTRACT

In today's data-driven educational technologies, algorithms have a pivotal impact on student experiences and outcomes. Therefore, it is critical to take steps to minimize biases, to avoid perpetuating or exacerbating inequalities. In this paper, we investigate the degree to which algorithmic biases are present in two learning analytics models: knowledge estimates based on Bayesian Knowledge Tracing (BKT) and carelessness detectors. Using data from a learning platform used across the United States at scale, we explore algorithmic bias following three different approaches: 1) analyzing the performance of the models on every demographic group in the sample, 2) comparing performance across intersectional groups of these demographics, and 3) investigating whether the models trained using specific groups can be transferred to demographics that were not observed during the training process. Our experimental results show that the performance of these models is close to equal across all the demographic and intersectional groups. These findings establish the feasibility of validating educational algorithms for intersectional groups and indicate that these algorithms can be fairly used for diverse students at scale.

CCS CONCEPTS

• **Human-centered computing**; • **Social and professional topics**; • **Applied computing** → Education; Interactive learning environments;

KEYWORDS

Carelessness detection, Behavior detection, Bayesian knowledge tracing, Algorithmic bias, Fairness

ACM Reference Format:
Andres Felipe Zambrano, Jiayi Zhang, and Ryan S. Baker. 2024. Investigating Algorithmic Bias on Bayesian Knowledge Tracing and Carelessness Detectors. In *The 14th Learning Analytics and Knowledge Conference (LAK '24), March 18–22, 2024, Kyoto, Japan.* ACM, New York, NY, USA, 11 pages. https://doi.org/10.1145/3636555.3636890

*Corresponding author.

LAK '24, March 18–22, 2024, Kyoto, Japan

ACM ISBN 979-8-4007-1618-8/24/03…$15.00
https://doi.org/10.1145/3636555.3636890

1 INTRODUCTION

Increasingly, algorithms are used in education, but are they fair or are they biased? As defined in [17], algorithmic bias describes situations when a data-driven model makes "inequitable prediction across identity groups" (p228). In learning environments, inequitable predictions may potentially lead to unfairness in the allocation of resources and interventions, creating worse outcomes for often already historically disadvantaged groups [23].

Prior to 2020, model fairness was not commonly discussed in learning analytics or educational data [29], but where studied, evidence for algorithmic bias has frequently been found [6]. Models that detect a range of constructs have been found to achieve better predictive performance for some groups of students than others (e.g., [8, 20]). Since then, with the growing emphasis on developing trustworthy AI and fair student models [22], efforts to assess model fairness have increased (e.g. [40]). For example, studies have used slicing analysis (discussed below) to examine model performance for each demographic group in a sample, assessing whether the performance is comparable across these groups (e.g. [17, 43]). The same approach has been used to evaluate the fairness of models that predict students' likelihood of dropping out, passing, and academic success, using demographic categories such as race/ethnicity, gender, and socio-economic status (e.g. [14, 20, 23, 39, 42]).

Although there has been an increase in the amount of work dedicated to evaluating the fairness of predictive models, the criteria and methods for fairness evaluation have varied across studies [23], with most published papers emphasizing slicing analyses. However, recent articles in algorithmic bias more broadly (e.g. [6, 25]) have called for also looking at the intersectionality across demographic groups and transferability to new demographics.

Introduced by Crenshaw [12], intersectionality analysis examines the effects experienced by individuals who are associated with multiple group memberships. Crenshaw's work in examining discrimination against Black women revealed distinct experiences that differ from those of Black men or non-Black women. This discovery highlights the necessity of studying intersectional groups, since individuals belonging to two groups can have unique negative experiences not associated with either group individually, a possibility that can be overlooked when exclusively analyzing single-demographic groups. In recent years, increasing attention has been paid to intersectionality in algorithmic bias outside the learning analytics community [9], with initial work also beginning to explore this topic in learning analytics as well [32].

In addition, the generalizability of models may need to be assessed even beyond the current sample of groups available. In a certain sense, we can only truly assess model performance for the data we have, but some validation methods can shed more light

on out-of-sample performance than others. In specific, if we build and test our model on subsets of all groups, we can only make conclusions about new students from those groups, but if we build our model on some demographic groups and test on other unseen demographic groups, we can infer (to a degree) performance on demographic groups outside our current sample (see discussion in [3]). Analyzing transferability in this fashion is particularly pertinent when a model is set to be scaled for a broader audience that may include historically underrepresented learners who are not in the current sample. However, few learning analytics papers have analyzed generalizability between populations in this fashion (e.g., [27, 28, 35]). For instance, Ogan et al. [28] evaluated the transferability of a help-seeking model between students in three different countries.

Predictive models are used frequently in intelligent tutoring systems (ITS) to assess and reflect the state and the progress of learning in-real time, with the goal of informing adaptive decisions [21]. In the current study, we examined algorithmic bias in the context of an ITS, focusing on evaluating whether models were biased in three ways: across traditional single-demographic groups, the intersection between those groups, and model transferability when the target demographic is not considered when training the model. By exploring the possibility of algorithmic bias in these three distinct ways, we seek to obtain deeper understanding of the potential implications for different groups of learners when deploying learning analytics models in real settings.

In specific, we conduct the analysis for two types of models that have not been previously studied in terms of algorithmic bias: models that estimate students' knowledge of specific skills [30] and models that detect students' carelessness, a form of disengagement characteristic of high-performing students [16]. Specifically, we study algorithmic bias for Bayesian Knowledge Tracing (BKT; [13]), an older knowledge tracing algorithm that nonetheless is used at wide scale in several learning platforms. We also study algorithmic bias in detectors of carelessness built on top of BKT, using the approach from [4]. In this approach, a regression prediction model is trained to detect careless errors from behavior using carelessness estimates derived from BKT models as ground truth, an approach predictive of long-term student outcomes [1, 34]. As the second (machine learning-based carelessness detector) model relies on the results from the first (BKT) model, the fairness of BKT models must be tested as well. The primary objective of this study is therefore to assess potential biases in two different but interrelated models, evaluating their performance across intersectional demographic groups and transfer to new populations.

2 RELATED WORK

2.1 Algorithmic bias

In predictive modeling, to evaluate the overall performance of a model, the model is typically trained on a set of data and tested on a held-out testing dataset. However, the overall high performance of a model may hide the fact that the model may perform less well for individuals who are underrepresented in the data [23]. Slicing analysis as an approach addresses this issue by computing the performance of the model for each subgroup in the test set [38]. Specifically, slicing analysis evaluates a predictive model's performance by slicing the results of that model across different categories in the test set, providing a more granular examination of the predictive model and makes it possible to evaluate the relative performance or fairness of the model across subgroups [38].

Historically, most papers on algorithmic bias in learning analytics have looked at whether performance is comparable between groups. However, within machine learning more broadly, recent accounts have argued in addition to comparing the performance for each demographic group, algorithmic bias should also be evaluated for **intersectional groups** where learners are associated with multiple group memberships [9]. This is especially important given "biases can often be amplified in subgroups that combine membership from different categories, especially if such a subgroup is particularly underrepresented in historical platforms of opportunity" [19]. Yet, insufficient research has examined intersectionality in educational work on algorithmic bias (but see [32]). When evaluating intersectionality, guidelines have been proposed in [41] to address three practical concerns: which identifies to consider, how to handle very small groups, and how to evaluate a large number of subgroups.

As specific intersectional groups (or even entire non-intersectional groups) may be poorly represented in a specific sample, it is often desirable to assess potential performance for groups of learners outside the original sample – the model's **transferability**. This is particularly important when some groups of historically underrepresented learners will be in the eventual population of a learning analytics system, but are not in the current sample. For example, Ocumpaugh et al., [27] examined four affect detectors (boredom, confusion, engaged concentration, frustration) across three populations (urban, suburban, rural), and found that detectors were generally more accurate when applied to new students from the same population trained on than when they were applied to students from an unseen population. However, a model that predicts student reading proficiency, initially trained on student samples from a mid-western U.S. school district, was shown to demonstrate comparable performance when applied to a larger south-eastern district [2]. A carelessness model [35] and a help-seeking model [28] have also been evaluated for transferability for students in different countries.

2.2 Carelessness Detection and Bayesian Knowledge Tracing

Academic discussion of carelessness in classrooms dates back to the 1950s [15]. To empirically investigate this behavior, Clement [11] introduced a strategy of administering the same item in repeated assessments and treating an error as careless if the student could solve the item on some occasions but not others. This general intuition was extended to the broader case where multiple items involve the same skill by [4]. In this work, a two-step approach was proposed for detecting carelessness. In the first step, a contextual slip model was developed that estimates the probability of carelessness on a given question based on the performance on the next two questions. Using the estimates from bayesian knowledge tracing (BKT; [13]), the model contextually estimates the probability that the incorrect answer is the result of slipping (i.e., carelessness) as opposed to lack of knowledge by combining the student's prior knowledge estimate

and their performance on the next two questions containing the same skill. In essence, it is more likely that the first incorrect answer is a careless error if the student answers the next two questions containing the same skill correctly. Afterwards, in the second step, a machine-learned model is trained to predict these carelessness estimates from learner behavior (such as time to respond and hint use) without using data from the future.

Thus far, carelessness detection has been based on Bayesian Knowledge Tracing (BKT), a commonly used knowledge tracing model in real-world ITS and adaptive learning systems. BKT estimates a student's latent knowledge on a skill based on previous observable performance and predicts the probability of a student getting the next question correct [13]. Although more advanced knowledge tracing models that rely on deep-learning have been shown to better predict future student performance, both within-system (e.g. [18]) and beyond [37], BKT remains the most used knowledge tracing model in practice by many ITS, due to its predictability, stability, and ease of implementation. Additionally, BKT is comparable in performance to more advanced algorithms at estimating when a student has reached mastery [44], which remains the main application of knowledge tracing models.

In BKT, two learning parameters L_0 (initial probability of knowing each skill) and T (probability of learning the skill at the opportunity to work on a skill) and two performance parameters – G (probability of guessing) and S (probability of slipping) are used to calculate $P(L_n)$ for each skill, which reflects the mastery level of a skill for each student.

Using the BKT parameters, the contextual slip model in [4] infers carelessness by estimating the probability that an incorrect answer is a slip based on the BKT's model's prior estimates and the student's future performance (see equation below). In specific, the model uses future information, examining how well a student answers the subsequent two questions (i.e., $n+1$ and $n+2$) in order to infer the probability that a student's incorrectness at time n was due to not knowing the skill, or whether it is due to a slip. The probability that the student knew the skill at time n can be calculated, given information about the actions at time n+1 and n+2 (A_{n+1}, A_{n+2}), and the other parameters of the BKT as $P(A_n is\ a\ Slip|A_n\ is\ incorrect) = P(L_n|A_{n+1}A_{n+2})$.

The underlying logic is as follows: if the student's knowledge estimate is low and they continue to make errors in subsequent questions, it is more likely that the initial error was due to a lack of understanding. On the other hand, if the student's knowledge estimate was previously high and they do not make additional mistakes, the initial error is likely attributed to carelessness. For full mathematical details, readers are referred to [4].

This step of the approach yields inferences of carelessness, but those estimates rely on future information, which makes real-time detection impossible. To address this issue, Baker et al. [4] next use machine learning (ML) to predict carelessness in real-time using behavioral patterns. In this approach, features such as time taken and hint use are distilled and trained to predict the probability of carelessness, using machine learning algorithms (linear regression in the original work). This approach successfully predicted new carelessness labels for new students [4]. In addition, San Pedro et al. [35] examined how the carelessness model performs when it is trained and tested with student samples collected from different countries – Philippines and the United States. They found that models trained in one country correlated well to training labels for the other country.

These models were then used to study the relationship between carelessness and affect, engagement, and long-term outcomes. San Pedro et al. [35] found that students who frequently experienced engaged concentration were more likely to make careless errors. Similarly, Fancsali [16] showed that carelessness was positively correlated with engaged concentration but negatively associated with other disengaged behaviors, such as gaming the system and off-task. Carelessness during middle school has been found to be predictive of long-term outcomes, such as college enrollment [34], choice of STEM-related majors [36] and choice of STEM career post-college [1].

3 METHODS

3.1 Data

In this study, we examine data from 5,856 students enrolled in 12 middle and high schools in a district in a small city in the northeastern United States. These students engaged with Carnegie Learning's MATHia software [31] for mathematics instruction during the academic years 2021-2022. MATHia is an Intelligent Tutoring System (ITS) used by over 600,000 students in thousands of schools every year. The content within MATHia is structured into "workspaces," which consist of multi-step problems. Students advance by working through these pre-determined sequences of content. This system is closely aligned to the use of the BKT algorithm; BKT has been used within this system for well over two decades, and content has been adjusted to improve BKT fit (for example, by modifying items less well predicted by BKT).

The collected data includes students' interactions with the software (averaging 8124 actions over 433 problems and 1425 problem steps solved per student), as well as demographic data provided by the school district. The district demographic data covers standard categories such as age, gender, race/ethnicity (i.e., African American, Asian, Hispanic, White, Native American, Native Hawaiian and Pacific Islander, and Multi-race non-Hispanic), whether the student has special needs, is an English learner, or is economically disadvantaged. The demographic distributions are detailed in Table 1.

Among racial and ethnic groups, African Americans and Hispanics are the most represented groups in this district with 3,551 and 1,051 students, respectively, while Native Americans have the smallest representation with just 4 students. The data is close to balance in terms of gender, with 3,096 male and 2,753 female students. Data on non-binary students is limited, with only 7 identified, although this may be because many non-binary students have not declared their gender to the district. Additionally, the dataset includes 838 students with special needs, 1296 students who are English learners, and 4191 students with economic disadvantage.

For most combinations of race/ethnicity with other demographic groups, there are at least 50 students. However, intersections involving non-binary students and all races/ethnicities, and Native American students with all other demographics, each have fewer than 10 students. Beyond Native American and non-binary students,

Table 1: Distribution of each demographic. Only race/ethnicity and gender are mutually exclusive; therefore, the total column is not the sum of the elements of each row. Intersections with less than ten students are shown in bold.

Demographics	Male	Female	Non-Binary	Special Needs	English Learners	Economic Disadvantage	Total
African American	1905	1646	**0**	467	944	2639	3551
Asian	67	64	**1**	**5**	15	79	132
Hispanic	549	501	**1**	182	268	814	1051
White	435	391	**5**	131	54	454	831
Native American	**3**	**1**	**0**	**1**	**1**	**3**	**4**
Native Hawaiian and Pacific Islander	**4**	10	**0**	**2**	**1**	10	14
Multi race, non-hispanic	133	140	**0**	50	13	192	273
Total	3096	2753	**7**	838	1296	4191	5856

the dataset includes information about only 4 male Native Hawaiian and Pacific Islander students, 2 Native Hawaiian and Pacific Islander students with special needs, 1 Native Hawaiian student who is an English learner, and 5 Asian students with special needs. We highlight these particular demographics because the performance of detection algorithms may be poor and unreliable for these groups, not necessarily due to inherent biases but likely because of the small sample sizes available for training and validation within these demographic categories. Even a fair and high-performing model can obtain highly negative (or positive) performance for one specific student. This student, or a small group, might not represent the actual trends for that specific demographic group. For this reason, we excluded all intersectional groups with fewer than 10 students.

3.2 Bayesian Knowledge Tracing

For each of the demographics and intersections previously outlined, we calculate the knowledge estimates of each particular skill that students in this dataset practiced using Bayesian Knowledge Tracing (BKT; [13]). The parameters for the BKT models for each skill were fitted using brute-force grid search [5], as in that previous work. To avoid model degeneracy and ensure that the parameter values align with the model's conceptual meaning (such as a higher likelihood that students will correctly answer if they have already mastered the skill), we adopted the common practice of setting upper limits of 0.3 and 0.1 for the 'Guess' and 'Slip' parameters, respectively [4, 5]. The parameters of these BKT models are not designed to inherently favor or disfavor any demographic group. They are built solely using students' initial responses to each problem (correct or incorrect) without directly considering demographic characteristics [24]. However, due to uneven sample sizes across demographics, the parameters obtained through brute-force grid search (or any other optimization technique) may be more representative of demographics with a larger number of students. In addition, students in some demographic groups may have best-fitting parameters that are more different from the average of other groups, making a model fit across groups less accurate for them.

To evaluate this potential algorithmic bias of BKT models, we use each skill's BKT model parameters to estimate the probability of a student answering correctly. This estimate is then compared to the actual correctness rates. Specifically, we calculate the probability of a correct answer as the sum of two probabilities: the probability of not making a slip when the skill has been mastered, and the probability of guessing correctly when the skill is not yet mastered ($P(L_{n-1})P(\sim S) + P(\sim L_{n-1})P(G)$) [13]. We selected this task to validate the BKT model parameters because it enables a direct comparison between the model's predictions and the students' actual performance, which we already know to be either correct or incorrect. Moreover, this estimate represents what the BKT models are designed to predict. It incorporates not just the likelihood of skill mastery but also the probabilities of guessing correctly or making a slip.

To assess the effectiveness of these correctness estimations for each demographic, and consequently recognize any potential bias of the BKT model parameters, we conducted a 4-fold student level cross-validation that was stratified by demographics, ensuring that each fold maintained a demographic distribution similar to that of the overall dataset. We then evaluate model performance calculating the Area Under the Receiver Operating Characteristic Curve (AUC ROC; AUC for short) within each demographic group and intersection. This approach is commonly seen in learning analytics analyses of algorithmic bias [33, 43, 45]. Additionally, we also calculate max difference between AUC for the best and worst predicted group [23].

3.3 Carelessness detectors

Drawing on the parameters of each skill-specific BKT model [13], we estimate the likelihood that a student's incorrect answer stems from carelessness rather than a knowledge gap [4]. This estimation process takes into account the student's initial responses to the subsequent two problems that require the same skill, as well as the estimated level of the student's knowledge at the time of the error. These carelessness estimates derived from BKT models serve as the ground truth for training machine learning (ML) models of carelessness. These ML models aim to replicate BKT-derived ground truth for carelessness using features extracted exclusively from students' prior interactions with the educational software.

The feature set of these models encompasses elements such as the hints requested by the student, the nature of the response required for each problem, the total time and number of problems the

Table 2: Algorithmic Bias of BKT models for each demographic. Absolute differences between the performance of each demographic and the entire population are included.

Demographic	AUC	ΔAUC
All Students	0.782 (0.001)	-
Male	0.782 (0.002)	-0.001
Female	0.783 (0.001)	0.001
African American	0.783 (0.001)	0.001
Asian	0.769 (0.015)	-0.013
Hispanic	0.785 (0.003)	0.002
White	0.780 (0.002)	-0.002
Native Hawaiian and Pacific Islander	0.786 (0.013)	0.004
Multi race, non-hispanic	0.779 (0.007)	-0.004
Special Needs	0.788 (0.003)	0.006
Not Special Needs	0.781 (0.001)	-0.001
English Learners	0.790 (0.002)	0.008
Non-English Learners	0.780 (0.001)	-0.002
Economic Disadvantage	0.783 (0.001)	0.001
Not Economic Disadvantage	0.779 (0.002)	-0.003

student had previously engaged with to master the skill in focus, and the student's historical error count. These features draw inspiration from previous research that has employed similar approaches to train carelessness detectors [4, 35]. We employ Linear Regression (LR) in line with prior studies [4, 35], and introduce Random Forest Regressors (RF) as an alternative technique as a potentially higher-performance algorithm (but which still has relatively low risk of overfitting [7]). Both methods are implemented using the Scikit-learn library [26] using default parameters. As the labels are numerical probabilities, we use Pearson's r and Root Mean Squared Error (RMSE) as the metrics for the carelessness detectors. To address the imbalanced distribution of students across different demographic groups, we used Synthetic Minority Over-sampling Technique (SMOTE; [10]) and compared its results with models that do not use resampling.

To investigate potential algorithmic biases in these ML models among the defined demographic groups, we employ a similar methodology to the method used for the BKT estimates: a 4-fold student level cross-validation stratified by demographics and evaluating the model's performance for each specific demographic. In addition, we consider a "leave-one-demographic-out" strategy for both the training and testing phases. This approach allows us to examine the model's generalizability to demographic groups that were not part of the initial training set, a much less common approach in our field (but see [27]). This second step is important if our original training set does not contain all the groups that may eventually receive the detectors, which is frequently the case in models applied across a large geographical area.

4 RESULTS

4.1 Algorithmic Bias for BKT

Table 2 presents results confirming that our Bayesian Knowledge Tracing (BKT) models do not introduce any particular bias against any population. When evaluated using AUC, the correctness estimation derived from BKT models accurately distinguishes between correct and incorrect student answers 78.2% of the time (AUC= 0.782). A performance comparison across all demographic groups reveals a slight AUC difference of 0.017 between the best and worst predicted mutually exclusive groups (Native Hawaiian and Pacific Islander students and Asian students, respectively). The second highest difference in AUC (0.010) between non-overlapping groups is observed between English and non-English learners (AUC=0.790 and AUC=0.780, respectively). These modest differences showed that there is no evidence for bias in BKT estimates across single-demographic groups.

Similar results are observed when considering the intersections between demographics (Table 3). The AUC difference between the top and bottom predicted intersectional groups (Native Hawaiian and Pacific Islander students without special needs and Asian non-English learners, respectively) stands at 0.033. Although intersectional groups involving Asian students had some appearance of trend towards lower performance, the BKT models never had an AUC deviation greater than 0.015 from the overall student population for any such groups. Additionally, the model demonstrates a higher performance for Native American and Pacific Islander economically disadvantaged students (AUC of 0.792) and Hispanic and African American students with special needs or who are English learners (AUC of 0.790 for all of them). Nonetheless, the variations across these intersectional groups remain minor.

4.2 Algorithmic Bias on Carelessness detectors

After verifying that the estimated parameters from the BKT models for each skill (and therefore our ground truth for carelessness) do not introduce undue bias and offer a reliable approximation of actual student performance, we proceeded to train the machine learning (ML) models. These ML models aim to approximate the established ground truth without requiring future data, aligning with real-world applications. Table 4 shows the mean and standard deviation of the best-performing model for both the general population and each demographic group. The top-performing model was

Table 3: AUC ROC of BKT models for each intersection of demographics. Absolute differences between the performance of each demographic and the entire population are shown in parenthesis. Intersections with less than 10 students are not included.

	Male	Female	Special Needs	Not Special Needs	English Learners	Not English Learners	Economic Disadvantage	Not Economic Disadvantage
African American	0.783 (0.001)	0.783 (0.001)	0.790 (0.008)	0.782 (-0.001)	0.790 (0.008)	0.780 (-0.002)	0.783 (0.001)	0.783 (0.001)
Asian	0.770 (-0.012)	0.769 (-0.014)	-	0.768 (-0.015)	0.782 (-0.001)	0.767 (-0.016)	0.769 (-0.014)	0.769 (-0.014)
Hispanic	0.784 (0.002)	0.785 (0.002)	0.790 (0.008)	0.784 (0.002)	0.790 (0.007)	0.783 (0.001)	0.785 (0.003)	0.785 (0.003)
White	0.777 (-0.005)	0.782 (0.001)	0.783 (0.001)	0.779 (-0.003)	0.793 (0.010)	0.779 (-0.003)	0.780 (-0.003)	0.780 (-0.003)
Native Hawaiian and Pacific Islander	-	0.789 (0.007)	-	0.799 (0.017)	-	0.784 (0.002)	0.792 (0.010)	-
Multi race, non-hispanic	0.774 (-0.009)	0.784 (0.002)	0.773 (-0.010)	0.780 (-0.002)	0.794 (0.011)	0.778 (-0.004)	0.781 (-0.001)	0.774 (-0.008)

Table 4: Algorithmic fairness of Carelessness detectors for each demographic. Standard deviation of the cross-validation is shown in parenthesis.

Demographic	Pearson's r	Δr	RMSE	ΔRMSE
All Students (RF)	0.840 (0.001)	-	0.216 (0.018)	-
All Students (RF+SMOTE)	0.838 (0.001)	-0.002	0.217 (0.018)	0.001
All Students (LR)	0.706 (0.001)	-0.135	0.282 (0.021)	0.066
Male	0.838 (<0.001)	-0.002	0.216 (0.016)	0.001
Female	0.842 (0.001)	0.002	0.215 (0.021)	-0.001
African American	0.838 (0.002)	-0.002	0.217 (0.023)	0.001
Asian	0.836 (0.009)	-0.004	0.215 (0.042)	-0.001
Hispanic	0.843 (0.004)	0.002	0.215 (0.036)	-0.001
White	0.844 (0.004)	0.003	0.212 (0.033)	-0.003
Native Hawaiian and Pacific Islander	0.828 (0.038)	-0.013	0.215 (0.094)	-0.001
Multi race, non-hispanic	0.848 (0.006)	0.008	0.210 (0.039)	-0.006
Special Needs	0.834 (0.003)	-0.007	0.221 (0.026)	0.005
Not Special Needs	0.841 (0.001)	0.001	0.215 (0.019)	-0.001
English Learners	0.834 (0.003)	-0.006	0.223 (0.030)	0.007
Non-English Learners	0.842 (0.001)	0.002	0.214 (0.020)	-0.002
Economic Disadvantage	0.841 (0.001)	0.001	0.216 (0.020)	-0.001
Not Economic Disadvantage	0.837 (0.004)	-0.003	0.217 (0.064)	-0.001

a Random Forest Regressor (RF) with an r=0.840 and RMSE=0.216, a better result than was obtained for Linear Regression (LR), which performed more poorly than RF across all demographics and the general population (r=0.706 and RMSE= 0.282). We also experimented with SMOTE as an oversampling technique for underrepresented demographics. However, its performance for the RF models (r=0.838, RMSE=0.217) was slightly inferior than the performance without resampling. Consequently, we selected the RF model without resampling for the analyses presented in this paper.

For every group, the test-set correlation between the detector and ground truth exceeded 0.8, and the RMSE was below 0.23. This suggests that the ML model effectively mirrors the carelessness estimations produced by the BKT models. When assessed using Pearson's r, the least accurately predicted group was the Native Hawaiian and Pacific Islander students (r=0.828), while the most accurately predicted was the Multi-race, non-Hispanic students (r=0.848), a fairly minor 0.020 difference between correlations. RMSE reveals a similar trend. The highest difference between mutually exclusive groups was observed for the English and non-English learners (RMSE=0.223 and RMSE=0.214). In terms of race and ethnicity groups, the Multi-race, non-Hispanic students are again the best predicted group (RMSE=0.210), while African Americans rank as the least accurately predicted (RMSE=0.217). These results consistently demonstrate only minor discrepancies across different demographics, indicating limited evidence for bias for

Table 5: Pearson's r of carelessness detectors for each intersection of demographics. Percentages of difference between the performance of each demographic and the entire population are shown in parenthesis. Intersections with less than 10 students are not included.

	Male	Female	Special Needs	Not Special Needs	English Learners	Non-English Learners	Economic Disadvantage	Not Economic Disadvantage
African American	0.836 (-0.004)	0.840 (0.001)	0.831 (-0.009)	0.839 (-0.001)	0.831 (-0.010)	0.841 (0.001)	0.839 (-0.001)	0.835 (-0.005)
Asian	0.843 (0.002)	0.830 (-0.011)	-	0.837 (-0.003)	0.825 (-0.015)	0.837 (-0.003)	0.834 (-0.006)	0.839 (-0.001)
Hispanic	0.841 (0.001)	0.844 (0.004)	0.837 (-0.004)	0.844 (0.004)	0.841 (0.001)	0.844 (0.004)	0.843 (0.003)	0.843 (0.003)
White	0.841 (0.001)	0.846 (0.006)	0.834 (-0.006)	0.846 (0.006)	0.843 (0.003)	0.844 (0.004)	0.842 (0.002)	0.846 (0.006)
Native Hawaiian and Pacific Islander	-	0.825 (-0.016)	-	0.836 (-0.004)	-	0.828 (-0.012)	0.815 (-0.025)	-
Multi race, non-hispanic	0.846 (0.005)	0.850 (0.010)	0.849 (0.009)	0.848 (0.008)	0.844 (0.004)	0.848 (0.008)	0.848 (0.007)	0.848 (0.008)

Table 6: RMSE of carelessness detectors for each intersection of demographics. Percentages of difference between the performance of each demographic and the entire population are shown in parenthesis. Intersections with less than 10 students are not included.

	Male	Female	Special Needs	Not Special Needs	English Learners	Non-English Learners	Economic Disadvantage	Not Economic Disadvantage
African American	0.217 (0.002)	0.216 (0.001)	0.223 (0.007)	0.216 (0.001)	0.224 (0.009)	0.214 (-0.002)	0.217 (0.001)	0.217 (0.001)
Asian	0.209 (-0.006)	0.220 (0.004)	-	0.214 (-0.002)	0.223 (0.007)	0.214 (-0.002)	0.217 (0.001)	0.217 (0.001)
Hispanic	0.216 (0.001)	0.214 (-0.001)	0.219 (0.003)	0.214 (-0.002)	0.220 (0.004)	0.213 (-0.003)	0.215 (-0.001)	0.215 (-0.001)
White	0.213 (-0.002)	0.211 (-0.004)	0.221 (0.005)	0.210 (-0.006)	0.221 (0.005)	0.211 (-0.005)	0.214 (-0.002)	0.210 (-0.006)
Native Hawaiian and Pacific Islander	-	0.218 (0.003)	-	0.211 (-0.005)	-	0.215 (-0.001)	0.220 (0.005)	-
Multi race, non-hispanic	0.210 (-0.006)	0.209 (-0.006)	0.210 (-0.006)	0.210 (-0.006)	0.217 (0.001)	0.210 (-0.006)	0.210 (-0.006)	0.210 (-0.006)

the ML-based carelessness detector across individual demographic groups.

Tables 5 and 6 present the model's performance across several demographic intersections, considering r and RMSE, respectively. The worst-predicted intersectional group was the Native Hawaiian and Pacific Islander students with economic disadvantage (r=0.815), obtaining r 0.025 lower than the overall student population. This result contrasts with the third best-predicted group, Multi-race, non-Hispanic students, also with economic disadvantages (r=0.848). Similarly, female students from these two groups had fairly different performances from each other; the best-predicted group was an intersectional group involving female students (Multi race, non-hispanic; r=0.850), but so was the second-worst predicted group (Native Hawaiian and Pacific Islander; r=0.825). Although the difference between the predictions for Native Hawaiian and Pacific Islander students and Multi-race, non-hispanic students was observed when examining single demographics, this specific distinction for female students and students with economic disadvantage can only be observed through the intersectional analysis, underscoring the distinct findings produced by this type of analysis. While the differences are minor here (0.033 for students with economic disadvantage and 0.025 for female students between the two mentioned races), other contexts or constructs might reveal more pronounced disparities.

Differences were also observed between White students with and without special needs (r=0.834 and r=0.846, respectively), between Asian English and non-English learners (r=0.825 and r=0.837), and

Table 7: Algorithmic transferability of carelessness detectors for each demographic training with the other demographic groups.

Demographic	Pearson's r	Δr	RMSE	ΔRMSE
Male	0.835	-0.005	0.218	0.003
Female	0.837	-0.003	0.218	0.003
African American	0.832	-0.008	0.221	0.005
Asian	0.840	0.001	0.213	-0.003
Hispanic	0.843	0.003	0.215	-0.001
White	0.844	0.004	0.212	-0.004
Native Hawaiian and Pacific Islander	0.842	0.002	0.212	-0.004
Multi race, non-hispanic	0.850	0.010	0.208	-0.007
Special Needs	0.835	-0.005	0.220	0.005
Not Special Needs	0.821	-0.019	0.227	0.012
English Learners	0.832	-0.008	0.224	0.009
Non-English Learners	0.828	-0.012	0.221	0.006
Economic Disadvantage	0.831	-0.009	0.222	0.006
Not Economic Disadvantage	0.839	-0.001	0.215	-0.001

between Asian male and female students (r=0.843 and r=0.830). However, in all cases, these differences remain below 0.02. Similarly, the RMSE reveals a modest difference and the same trends between the intersectional groups described before, but all differences remain below 0.02. These results suggest that the ML carelessness detector consistently performs well across intersectional groups.

4.3 Transferability among different demographics

To evaluate the model's generalizability to demographic groups not included in its initial training, we utilized a "leave-one-demographic-out" strategy. Table 7 shows that Pearson's r does not decrease by more than 0.02 compared to the baseline performance of the entire population (r= 0.840, as detailed in Table 5), for any group. The worst transferability is observed for the students without special needs (when the model was trained with students with special needs; r=0.821), and for non-English learners (when the model was trained with English Learners; r=0.828), but those differences are modest. Similarly, the RMSE does not exceed an increase of 0.012 for any group compared to the broader student population's RMSE. Again, the worst-predicted group is the students without special needs when the model was trained on students with special needs (RMSE=0.227), but the difference is minor. Both findings suggest that, for every demographic group examined in this study, models trained using data from other demographics can achieve performance nearly equivalent to those trained with data from the demographic where the model is applied.

Using the same methodology for intersectional groups, we note a slight decrease in the model's transferability (Table 8 and 9), with the largest differences in performance appearing for the same intersectional groups observed in section 4.2. The largest decreases in the detector performance occur for Native Hawaiian and Pacific Islander female students (r=0.827, RMSE=0.219) and for Native Hawaiian and Pacific Islander students facing economic disadvantage (r=0.811, RMSE=0.226). These results contrast with the transferability observed for Multi-race, non-Hispanic female students (r=0.853, RMSE=0.208) and Multi-race, non-Hispanic students facing economic disadvantage (r=0.850, RMSE=0.209). This is the largest gap noted in all our bias evaluation methods but is still relatively small (less than 0.04 in Pearson's r and less than 0.02 in RMSE) compared to the levels of algorithmic bias reported in other work (see [4]). Notably, one of the groups with the least transferability also involved English learners (African American English learners, r=0.830, RMSE=0.225), showing again that the intersectional group can reveal results that cannot be observed from the single demographic groups analysis. However, once again, the difference still remains modest.

5 DISCUSSION AND CONCLUSION

In this research, we analyzed whether three forms of algorithmic bias were present for knowledge (BKT) and carelessness models. We found evidence that performance was close to equal across demographic groups, for these models, including intersectional categories, and tests where we held out entire demographic groups during model training (a test of model applicability to entirely new demographic groups), for carelessness. Overall, model performance was excellent across all groups – better than past detectors of these types in other data sets (e.g. [4, 18, 35]). We attribute this high overall performance to the use of a more contemporary algorithm for carelessness (Random Forest instead of Linear Regression) and the extensive past efforts made by the system developers to refine learning content so that the BKT algorithm would fit better. For all those cases, the detectors maintained successful performance, with the largest declines being less than 0.035 in AUC (for BKT), 0.040 in Pearson's r, and 0.020 in RMSE (for carelessness).

Several factors may explain the low degree of algorithmic bias. Algorithmic bias can often be due to limited sample size; although some groups were not well represented in the data, the dataset was large overall, comprising 5,856 students making 8124 actions and completing 433 problems spanning an entire academic year. Such a large sample size increases the representation of students with a

Table 8: Pearson's R of carelessness detectors for each intersection of demographics training with the other demographic groups. Intersections with less than 10 students are not included.

	Male	Female	Special Needs	Not Special Needs	English Learners	Non-English Learners	Economic Disadvantage	Not Economic Disadvantage
African American	0.836 (-0.004)	0.839 (-0.01)	0.833 (-0.007)	0.834 (-0.006)	0.830 (-0.010)	0.839 (-0.001)	0.836 (-0.004)	0.837 (-0.003)
Asian	0.846 (0.006)	0.835 (-0.005)	-	0.840 (-0.001)	0.831 (-0.09)	0.842 (0.002)	0.839 (-0.001)	0.842 (0.002)
Hispanic	0.843 (0.003)	0.846 (0.006)	0.839 (-0.001)	0.844 (0.004)	0.841 (0.001)	0.845 (0.005)	0.844 (0.004)	0.843 (0.003)
White	0.842 (0.002)	0.848 (0.008)	0.838 (-0.002)	0.846 (0.006)	0.847 (0.007)	0.844 (0.004)	0.844 (0.004)	0.846 (0.006)
Native Hawaiian and Pacific Islander	-	0.827 (-0.013)	-	0.854 (0.014)	-	0.842 (0.002)	0.811 (-0.029)	-
Multi race, non-hispanic	0.850 (0.010)	0.853 (0.013)	0.849 (0.009)	0.850 (0.010)	0.855 (0.015)	0.849 (0.009)	0.850 (0.010)	0.851 (0.011)

Table 9: RMSE of carelessness detectors for each intersection of demographics training with the other demographic groups. Intersections with less than 10 students are not included.

	Male	Female	Special Needs	Not Special Needs	English Learners	Non-English Learners	Economic Disadvantage	Not Economic Disadvantage
African American	0.218 (0.002)	0.217 (0.001)	0.221 (0.006)	0.219 (0.003)	0.225 (0.009)	0.215 (-0.001)	0.219 (0.003)	0.217 (0.001)
Asian	0.208 (-0.008)	0.217 (0.002)	-	0.212 (-0.004)	0.220 (0.005)	0.210 (-0.006)	0.214 (-0.001)	0.210 (-0.006)
Hispanic	0.215 (-0.001)	0.213 (-0.002)	0.218 (0.002)	0.214 (-0.002)	0.219 (0.004)	0.212 (-0.004)	0.215 (-0.001)	0.212 (-0.004)
White	0.212 (-0.003)	0.210 (-0.005)	0.219 (0.003)	0.210 (-0.006)	0.219 (0.003)	0.211 (-0.005)	0.213 (-0.003)	0.210 (-0.006)
Native Hawaiian and Pacific Islander	-	0.219 (0.003)	-	0.203 (-0.013)	-	0.212 (-0.004)	0.226 (0.010)	-
Multi race, non-hispanic	0.208 (-0.007)	0.208 (-0.008)	0.209 (-0.006)	0.208 (-0.008)	0.208 (-0.008)	0.209 (-0.007)	0.209 (-0.007)	0.208 (-0.008)

range of identities and attributes (within the specific district where the data was acquired), enhancing the likelihood of the model generalizing effectively for a new student. With this extensive dataset, there is a higher probability that the model has previously observed students bearing similarities to any new learner who studies in the same school district. Given this, future studies should investigate the impact of overall training set size on algorithmic bias, and perhaps consider sampling methods that can minimize bias in smaller samples. Second, all the students considered in this study are part of the same school district. This shared educational context, despite the students' varied demographics, likely provides a level of uniformity in their learning experiences. This uniformity makes it more feasible for the model to generalize across students within this shared school district. This may explain the results of the transferability analysis; even when trained on data from students of a different demographic, the model maintained its performance. Students from different demographic groups, due to their shared community, might exhibit behaviors parallel to their peers, regardless of demographic differences. This underlines the potential existence of patterns transcending these demographic variances, possibly influenced by contextual variables like the school environment, urban setting, or broader regional factors (such as state curricular standards and regional teacher training programs) that go beyond the traditional demographic categories considered in this study. Prior research on other constructs has found evidence for algorithmic bias when models are trained on students from a different geographical region [2, 27]. As such, further work is needed to understand which contextual differences impact the performance of models of different contexts.

Although the models employed in this study demonstrated high performance and reduced bias across multiple demographic groups,

it is uncertain whether these outcomes are transferable to regions with distinct cultural, socioeconomic, and linguistic backgrounds. Much of the current research on learning-related detectors is centered on students in the United States, usually from only one district or state, underscoring the need for similar studies in more diverse contexts. Future research must not only gather ample data to develop models that perform accurately but also confirm that these models are equally accurate for all demographic groups within the target population. This requirement could have additional challenges due to potentially reduced technological capabilities and stricter data collection regulations in different contexts. Additionally, content should also be tailored to the linguistic and cultural differences of these diverse and new populations before acquiring the data necessary to conduct this type of analysis. Despite these challenges, it is imperative to undertake these efforts and expand the research on the algorithmic bias of models across contexts and populations.

In summary, this study underpins how much work remains to be done before we, as a community, fully understand the scope and impact of algorithmic bias on learning analytics. Issues such as intersectionality and transfer to unseen populations have been understudied, and many types of models (even models frequently used in the real world, such as BKT) have not been studied in terms of algorithmic bias at all. Our findings here are fairly positive, but join a growing literature of more negative examples (e.g. [6]). Going forward, work should investigate a broader range of constructs, and aim to assess biases not only within conventional demographic categories but also at their intersections and beyond traditional categories like race, gender, and socioeconomic status. Only through such comprehensive evaluations can we ensure that algorithms used in practical settings are fair for all students.

ACKNOWLEDGMENTS

We thank the Carnegie Learning team for collecting and making available the data used in this practice. We also thank the National Science Foundation, award #DUE-2000405, for their support. We also thank Xiner Liu for conducting a software review to validate correctness and match to paper. Andres Felipe Zambrano thanks the Ministerio de Ciencia, Tecnología e Innovación and the Fulbright-Colombia commission for supporting his doctoral studies through the Fulbright-MinCiencias 2022 scholarship.

REFERENCES

[1] Ma. Victoria Almeda and Ryan S. Baker. 2020. Predicting Student Participation in STEM Careers: The Role of Affect and Engagement during Middle School. Journal of Educational Data Mining 12, 2 (2020), 33–47.
[2] Husni Almoubayyed, Stephen E. Fancsali, and Steve Ritter. 2023. Generalizing Predictive Models of Reading Ability in Adaptive Mathematics Software. In Proceedings of the 16th International Conference on Educational Data Mining.
[3] Ryan S. Baker. 2023. Big Data and Education. 7th Edition. Philadelphia, PA: University of Pennsylvania.
[4] Ryan S. J. d.Baker, Albert. T. Corbett, and Vincent Aleven. 2008. More accurate student modeling through contextual estimation of slip and guess probabilities in bayesian knowledge tracing. In Intelligent Tutoring Systems: 9th International Conference, ITS 2008, Montreal, Canada, June 23-27, 2008, Proceedings 9, 406-415. Springer Berlin Heidelberg.
[5] Ryan S. J. d. Baker, Albert T. Corbett, Sujith M Gowda, Angela Z Wagner, Benjamin A MacLaren, Linda R Kauffman, Aaron P Mitchell, and Stephen Giguere. 2010. Contextual slip and prediction of student performance after use of an intelligent tutor. In User Modeling, Adaptation, and Personalization: 18th International Conference, UMAP 2010, Big Island, HI, USA, June 20-24, 2010. Proceedings 18, Springer, 52–63.
[6] Ryan S. Baker and Aaron Hawn. 2022. Algorithmic bias in education. International Journal of Artificial Intelligence in Education (2021), 1–41.
[7] Leo Breiman. 2017. Classification and regression trees. Routledge.
[8] Brent Bridgeman, Catherine Trapani, and Yigal Attali. 2012. Comparison of human and machine scoring of essays: Differences by gender, ethnicity, and country. Applied Measurement in Education 25, 1 (2012), 27–40.
[9] Ángel Alexander Cabrera, Will Epperson, Fred Hohman, Minsuk Kahng, Jamie Morgenstern, and Duen Horng Chau. 2019. FairVis: Visual analytics for discovering intersectional bias in machine learning. In 2019 IEEE Conference on Visual Analytics Science and Technology (VAST), IEEE, 46–56.
[10] Nitesh V. Chawla, Kevin W. Bowyer, Lawrence O. Hall, and W. Philip Kegelmeyer. 2002. SMOTE: synthetic minority over-sampling technique. Journal of artificial intelligence research 16, (2002), 321–357.
[11] John Clement. 1982. Students' preconceptions in introductory mechanics. American Journal of physics 50, 1, 66–71.
[12] Kimberle Crenshaw. 1991. Race, gender, and sexual harassment. Sothern California Law Review 65.
[13] Albert T. Corbett and John R. Anderson. 1995. Knowledge tracing: Modeling the acquisition of procedural knowledge. User Modeling and User-Adapted Interaction 4, 4, 253-278.
[14] Oscar Blessed Deho, Chen Zhan, Jiuyong Li, Jixue Liu, Lin Liu, and Thuc Duy Le. 2022. How do the existing fairness metrics and unfairness mitigation algorithms contribute to ethical learning analytics? British Journal of Educational Technology 53, 4 (2022), 822–843.
[15] Merrill T. Eaton, Louis A. D'Amico, and Beeman N. Phillips. 1956. Problem behavior in school. Journal of Educational Psychology 47, 6, 350.
[16] Stephen Fancsali. 2015. Confounding Carelessness? Exploring Causal Relationships Between Carelessness, Affect, Behavior, and Learning in Cognitive Tutor Algebra Using Graphical Causal Models. In Educational Data Mining, 508–511.
[17] Josh Gardner, Christopher Brooks, and Ryan Baker. 2019. Evaluating the fairness of predictive student models through slicing analysis. In Proceedings of the 9th international conference on learning analytics & knowledge, 225–234.
[18] Theophile Gervet, Ken Koedinger, Jeff Schneider, Tom Mitchell, and others. 2020. When is deep learning the best approach to knowledge tracing? Journal of Educational Data Mining 12, 3 (2020), 31–54.
[19] Avijit Ghosh, Lea Genuit, and Mary Reagan. 2021. Characterizing intersectional group fairness with worst-case comparisons. In Artificial Intelligence Diversity, Belonging, Equity, and Inclusion, Proceedings of Machine Learning Research 14, 22–34.
[20] Qian Hu and Huzefa Rangwala. 2020. Towards Fair Educational Data Mining: A Case Study on Detecting At-Risk Students. International Educational Data Mining Society (2020).
[21] Yun Huang, Vincent Aleven, Elizabeth McLaughlin, and Kenneth Koedinger. 2020. A general multi-method approach to design-loop adaptivity in intelligent tutoring systems. In Artificial Intelligence in Education: 21st International Conference, AIED 2020, Ifrane, Morocco, July 6–10, 2020, Proceedings, Part II 21, Springer, 124–129.
[22] Mohammad Khalil, Paul Prinsloo, and Sharon Slade. 2023. Fairness, Trust, Transparency, Equity, and Responsibility in Learning Analytics. Journal of Learning Analytics 10, 1 (2023), 1–7.
[23] René F. Kizilcec and Hansol Lee. 2022. Algorithmic fairness in education. In The ethics of artificial intelligence in education. Routledge, 174–202.
[24] Matt J. Kusner, Joshua Loftus, Chris Russell, and Ricardo Silva. 2017. Counterfactual fairness. Advances in neural information processing systems 30.
[25] Lin Li, Lele Sha, Yuheng Li, Mladen Raković, Jia Rong, Srecko Joksimovic, Neil Selwyn, Dragan Gašević, and Guanliang Chen. 2023. Moral Machines or Tyranny of the Majority? A Systematic Review on Predictive Bias in Education. In LAK23: 13th International Learning Analytics and Knowledge Conference, 499–508.
[26] Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, and others. 2011. Scikit-learn: Machine learning in Python. Journal of machine Learning research 12, (2011), 2825–2830.
[27] Jaclyn Ocumpaugh, Ryan Baker, Sujith Gowda, Neil Heffernan, and Cristina Heffernan. 2014. Population validity for educational data mining models: A case study in affect detection. British Journal of Educational Technology 45, 3, 487–501.
[28] Amy Ogan, Erin Walker, Ryan Baker, Ma. Mercedes T. Rodrigo, Jose Carlo Soriano, and Maynor Jimenez Castro. 2015. Towards understanding how to assess help-seeking behavior across cultures. International Journal of Artificial Intelligence in Education 25, 229–248.
[29] Luc Paquette, Jaclyn Ocumpaugh, Ziyue Li, Alexandra Andres, and Ryan Baker. 2020. Who's Learning? Using Demographics in EDM Research. Journal of Educational Data Mining 12, 3 (2020), 1–30.
[30] Radek Pelánek. 2017. Bayesian knowledge tracing, logistic models, and beyond: an overview of learner modeling techniques. User Modeling and User-Adapted Interaction 27, 313–350.
[31] Steven Ritter, Jhon R. Anderson, Kenneth R. Koedinger, and Albert Corbett. 2007. Cognitive Tutor: Applied research in mathematics education. Psychonomic bulletin & review 14, 249-255.

[32] Nathalie Rzepka, Linda Fernsel, Hans-Georg Müller, Katrarina Simbeck, and Niels Pinkwart. 2023. Unbias me! Mitigating Algorithmic Bias for Less-studied Demographic Groups in the Context of Language Learning Technology. Computer-Based Learning in Context 6, 1, 1-23.

[33] Nathalie Rzepka, Katharina Simbeck, Hans-Georg Müller, and Niels Pinkwart. 2022. Fairness of In-session Dropout Prediction. In CSEDU (2), 316–326.

[34] Maria Ofelia San Pedro, Ryan Baker, Alex Bowers, and Neil Heffernan. 2013. Predicting college enrollment from student interaction with an intelligent tutoring system in middle school. In Educational Data Mining 2013.

[35] Maria Ofelia Z. San Pedro, Ryan S. J. d. Baker, and Ma. Mercedes T Rodrigo. 2014. Carelessness and affect in an intelligent tutoring system for mathematics. International Journal of Artificial Intelligence in Education 24, 189–210.

[36] Maria Ofelia San Pedro, Jaclyn Ocumpaugh, Ryan S Baker, and Neil T Heffernan. 2014. Predicting STEM and Non-STEM College Major Enrollment from Middle School Interaction with Mathematics Educational Software. In Educational Data Mining, 276–279.

[37] Richard Scruggs, Ryan S. Baker, and Bruce M. McLaren. 2020. Extending Deep Knowledge Tracing: Inferring Interpretable Knowledge and Predicting Post System Performance. In Proceedings of the 28th International Conference on Computers in Education.

[38] David Sculley, Jasper Snoek, Alex Wiltschko, and Ali Rahimi. 2018. Winner's curse? On pace, progress, and empirical rigor.

[39] Lele Sha, Mladen Raković, Angel Das, Dragan Gašević, and Guanliang Chen. 2022. Leveraging class balancing techniques to alleviate algorithmic bias for predictive tasks in education. IEEE Transactions on Learning Technologies 15, 4 (2022), 481–492.

[40] Jonathan Vasquez Verdugo, Xavier Gitiaux, Cesar Ortega, and Huzefa Rangwala. 2022. Faired: A systematic fairness analysis approach applied in a higher educational context. In LAK22: 12th International Learning Analytics and Knowledge Conference, 271–281.

[41] Angelina Wang, Vikram V. Ramaswamy, and Olga Russakovsky. 2022. Towards intersectionality in machine learning: Including more identities, handling underrepresentation, and performing evaluation. In Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency, 336–349.

[42] Renzhe Yu, Hansol Lee, and René F Kizilcec. 2021. Should college dropout prediction models include protected attributes? In Proceedings of the eighth ACM conference on learning@ scale, 91–100.

[43] Michael Yudelson, Steve Fancsali, Steve Ritter, Susan Berman, Tristan Nixon, and Ambarish Joshi. 2014. Better data beats big data. In Educational data mining 2014.

[44] Jiayi Zhang, Rohini Das, Ryan Baker, and Richard Scruggs. 2021. Knowledge tracing models' predictive performance when a student starts a skill. In Proceedings of the 14th International Conference on Educational Data Mining. EDM, Paris, France, 625–629.

[45] Jiayi Zhang, Juliana Ma. Alexandra L. Andres, Stephen Hutt, Ryan S. Baker, Jaclyn Ocumpaugh, Nidhi Nasiar, Caitlin Mills, Jamiella Brooks, Sheela Sethuaman, Tyron Young, and others. 2022. Using Machine Learning to Detect SMART Model Cognitive Operations in Mathematical Problem-Solving Process. Journal of Educational Data Mining 14, 3 (2022), 76–108.