Fairness of Bayesian Knowledge Tracing for Math Learners of Different Reading Ability

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

HaeJin Lee
University of Illinois Urbana–Champaign
haejin2@illinois.edu

Clara Belitz
University of Illinois Urbana–Champaign
cbelitz2@illinois.edu

Nidhi Nasiar
University of Pennsylvania
nasiar@upenn.edu

Stephen E. Fancsali
Carnegie Learning, Inc.
sfancsali@carnegielearning.com

Steve Ritter
Carnegie Learning, Inc.
sritter@carnegielearning.com

Husni Almoubayyed
Carnegie Learning, Inc.
halmoubayyed@carnegielearning.com

Ryan S. Baker
University of Pennsylvania
ryanshaunbaker@gmail.com

Jaclyn Ocumpaugh
University of Pennsylvania
ojaclyn@upenn.edu

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

ABSTRACT

Students' reading ability affects their outcomes in learning software even outside of reading education, such as in math education, which can result in unexpected and inequitable outcomes. We analyze an adaptive learning software using Bayesian Knowledge Tracing (BKT) to understand how the fairness of the software is impacted when reading ability is not modeled. We tested BKT model fairness by comparing two years of data from 8,549 students who were classified as either 'emerging' or 'non-emerging' readers (i.e., a measure of reading ability). We found that while BKT was unbiased on average in terms of equal predictive accuracy across groups, specific skills within the adaptive learning software exhibited bias related to reading level. Additionally, there were differences between the first-answer mastery rates of the emerging and non-emerging readers (M=.687 and M=.776, difference CI=[0.075, 0.095]), indicating that emerging reader status is predictive of mastery. Our findings demonstrate significant group differences in BKT models regarding reading ability, exhibiting that it is important to consider—and perhaps even model—reading as a separate skill that differentially influences students' outcomes.

Keywords

Algorithmic fairness, Bayesian knowledge tracing, reading ability

1. INTRODUCTION

Adaptive learning software has become a crucial tool in delivering personalized educational experiences to large numbers of students [41]. While the benefits of such software are clear (e.g., [41, 40, 43, 28]), there is a growing concern regarding different individual learning outcomes from interactions with the software [39, 2, 3]. Previous research has shown that reading ability is a strong indicator of math success [30, 24], but less work has examined this relationship in digital learning environments. This study builds on recent work from Almoubayyed et al. that has sought to address this gap, examining how differences in student reading ability relates to their outcomes in mathematics-focused adaptive learning software [2].

At the core of many adaptive learning systems is the Bayesian Knowledge Tracing (BKT) algorithm. BKT is a method for modeling and predicting student learning and knowledge [16]. Specifically, BKT leverages probabilistic inference to estimate a student's mastery of specific skills or topics over time based on their performance in different tasks that exercise a particular skill. BKT-driven systems can then provide more (or less) practice as needed to allow students to progress at the pace that is most appropriate for each individual [16]. In this paper, we refer to the possible areas of mastery as *skills* or *knowledge components*. While BKT has proven effective in providing personalized learning experiences and identifying knowledge gaps, concerns about its fairness and equity have emerged [17].

Ensuring fairness in educational tools in crucial, as biased predictions can disproportionately impact students from historically underrepresented backgrounds, potentially exacer-

bating existing educational inequalities [33, 6]. Inequalities can also come from other factors that effect how different students interact with learning systems, such as disability status, parental education levels, familiarity with technology, and more. In our study, we focus on students' reading ability as one such dimension that could relate to how students interact within an online learning system for mathematics. Because low reading ability can reflect several important groups of students who are often underserved in educational contexts (e.g., English language learners, students with learning disabilities, students whose parents have lower education levels, etc.), it is important to ensure that the underlying algorithms that predict when a student has mastered a skill are performing equally for these students.

Student reading ability is also important since research connects it to students ability to learn new math concepts [47, 22, 45], and it has be used to predict long-term math understanding [24]. Ensuring fairness across student reading abilities can improve current math-focused educational learning systems by accounting for the different interactions and outcomes that could be due to reading ability status.

Research on emerging readers now stretches back decades. The term was introduced to emphasize the linguistic skills and assets that should be drawn upon to help young children acquire reading skills [51], and it has since expanded to include adult learners as well as learners who have learning disabilities [34, 36]. In this study, we operationalize emerging reader status algorithmically, using an existing model that predicts if students are an emerging English language learner using interactions in the introductory phases of the learning session [2]. We then use BKT data to predict first-answer mastery across the emerging reader groups. Through the rest of the paper, we group students interacting with the math education software platform, Carnegie Learning's MATHia [48], as either *emerging readers* or *non-emerging readers*.

This paper focuses on three main research questions: **RQ1**: Do emerging and non-emerging readers have different outcomes with a mathematics-focused adaptive learning software? **RQ2**: Is the MATHia BKT model (un)biased with respect to the students who are emerging readers? **RQ3**: If there is bias in the BKT models, can we mitigate it through common unfairness mitigation strategies?

In larger adaptive learning systems, students can master many different, sometimes unrelated, skills. For example, MATHia has over 600 groupings of skills called "workspaces." We examine bias towards emerging readers by testing BKT model performance in two distinct but related processes. The first analyzes the aggregate BKT model performance on the most common workspaces in MATHia. The second analyzes the bias within each of the most common skills (i.e., the skills with the largest number of student interactions). We performed a two-pronged approach to measure the bias present within the learning session as a whole and bias present in specific MATHia skills. The approach is important, as even if no bias is present in MATHia as a whole, it could still be present in specific skills–particularly if some skills are more reliant on reading ability than others. To ensure that we are providing a comprehensive analysis of fairness, we examine how students' outcomes with MATHia are captured in terms of three anti-discrimination criteria: *independence* (i.e., equality of predicted outcomes), *separation* (i.e., equality of errors), and *sufficiency* (i.e., predictions reflect the same accuracy per group) [7]. In addition, we further apply an unfairness mitigation technique to serve as a part of a future adaptive process that suggests improvement to the models used.

Thus, our research has two main contributions. First, we combine a statistical bootstrapping analysis with classical fairness metrics to conduct a large-scale examination of differences in emerging and non-emerging readers' outcomes with adaptive learning software. Second, we analyze the effectiveness of reweighting, a bias mitigation method, on BKT models. These results show that emerging readers interact with MATHia in significantly different ways than other students–suggesting that there is an unmodeled factor (reading comprehension) in the curriculum that is affecting students' ability to achieve mastery. They also demonstrate that, while bias is low overall within MATHia, emerging readers may not be well-served by BKT algorithms for specific skills. In this way, we contribute to creating more equitable learning environments while also introducing a new category for consideration in this process, the emerging reader.

2. BACKGROUND

Richey et al. calls for a comprehensive view of modeling math learners that would include modeling non-math factors as well [47]. However, if these factors are less effective to students with diverse backgrounds, adding additional modeling could potentially widen the opportunity gap for students who have historically been under served. In this section, we discuss related work in emerging readers, adaptive learning and the measurement and mitigation of algorithmic unfairness.

2.1 Emerging Readers

Research has long shown that non-math factors can play a key role in the success of math learning [11, 37], as math proficiency often requires some level of reading comprehension [22]. Students' reading ability can predict their ability to solve math word problems and understanding the symbols in traditional math [45]. Additionally, reading comprehension scores predict mastery across multiple math topics [24].

Researchers have found links between struggling readers and protected classes of learners. Scammacca et al. surveyed interventions for struggling readers across three decades and found over 80 studies with a focus on improving reading comprehension for students [49]. Alongside trying to help struggling readers, other researchers have found connections between reading ability and protected groups in education. For example, Catone and Brady found that individual education programs (IEPs) were specifically ineffective for students who had word-level reading difficulties [13]. Also, Klinger et al. found a complex relationship between students whom are English language learners, struggling readers, and students who have learning disabilities [35].

This study joins a growing body of research examining the connection between reading comprehension and math learn-

ing within an adaptive learning software [47]. Here, we have operationalized the term *emerging readers* algorithmically, using a previously validated predictive model that forecasts who will be able to pass state standardized English language exams [2]. Emerging readers likely span a large range of demographics including, for example, groups marginalized in STEM or students from low socioeconomic status households.

2.2 MATHia Computer-based Learning Environment

MATHia is an intelligent tutoring system that is used for math mastery learning [48]. Specifically, MATHia is used in middle schools and high schools in the United States for over 500,000 students. Because large numbers of students use this system, and similar software like it, it is important that diverse populations of users can receive equal benefits from its implementation. Students interact with MATHia within different "workspaces" that consist of multiple skills that students can gain mastery in. The MATHia system provides students with repeated practice until it determines a student has gained mastery on all of the skills in a workspace. MATHia determines mastery using the BKT algorithm [16].

Because skills address different mathematics topics and concepts, each one addresses different math learning skills that potentially require different amounts of English language proficiency. In other words, there may be various English skills which (e.g., comprehension), unlike math skills, are typically not intentionally controlled or modeled during math learning, but which nevertheless systematically influence students' progress.

2.3 Measuring Algorithmic Unfairness

Bias can be present in algorithms at any step of the machine learning lifecycle [42]. For example, Friedman and College outline the multiple different kinds of biases as either coming from the preexisting world, technical aspects of the system, or some emergent property of the system being deployed [21], and each of these have been found in educational contexts [6, 21]. For example, Doroudi and Brunskill analyzed biases within knowledge tracing [17], and Zambrano et al. investigated bias in BKT and carelessness detectors [56]. Similarly, measuring bias in academic exams related to student demographics has long been a topic in educational research [29]. More recently, educational data mining research has also focused on specific data biases [6] as well as broad questions around ethics in AI-driven education [26]. These data biases can become an issue of unfairness because of the types of harm that come from biased systems; in a worst case, data bias can lead to avenues for educators to make decisions based on prejudice or disrespect [5]. Furthermore, certain researchers have found specific groups of students who are impacted by bias. Yanagiura at al. found that their collegiate early warning system had less predictive accuracy for an at-risk group of students (defined as belonging to two or more demographic or academic categories statistically associated with low GPA) [54] and Li et al. found bias against students with disabilities when predicting standardized test scores in early childhood education settings [38]. Furthermore, Finkelstein et al. found that third grade Black students learned more in science using their computer-based system when African-American Vernacular English was used [19], versus the English typically used in educational software. Baker and Hawn raised the concern that we do not fully know which groups are impacted by algorithmic bias [6], and the group studied in this paper is not one of the groups they found in their review. Also, researchers have found that lacking diverse representation of behaviors in students could lead to model bias [15]. Thus, possible biases in algorithms can result in unfairness for students.

To measure algorithmic biases, researchers in related fields have developed numerous ways to determine how biased the decisions of an algorithm are. Barocas et al. list 19 different fairness metrics created between 1971 and 2017 [7]. Specifically, measurements of fairness are split into either procedural or statistical guidelines [23]. Procedural fairness measures are geared towards making sure that statistically similar data points are treated similarly [18]. Alternatively, other definitions of fairness involve ensuring that all students are given the pedagogical strategies that they specifically need. For example, Hardt et al.'s *equal opportunity* measure tests if the true positive rate is balanced across groups [25] in order to prevent unwanted discrimination. Other measures analyze the probabilities of error rates [55], the difference in selection rates between groups [7], or even domain-specific concerns, such as measuring model discrimination independent of model performance [52]. Due to the many ways to measure fairness, many methods for mitigating unfairness have also been proposed.

2.4 Mitigating Unfairness

Unfairness mitigation strategies are grouped into preprocessing, inprocessing, and postprocesing methods based on where the strategy is applied in a machine learning pipeline. Preprocessing involves distorting the input data to remove correlations to sensitive features (e.g., reweighting [32], learning fair representations [57], and optimized preprocessing [10]); inprocessing adds constraints to the training or optimization processes (e.g., adversarial debiasing [58] and the meta-fair classifier [14]); postprocessing adjusts classifier outcomes (e.g., equalized odds postprocessing [25] and the reject option classification [32, 9]). In education, bias mitigation methods are also used. Kizilcek and Lee reference different areas within an educational algorithm where bias can be mitigated [33]. For example, Jiang et al. compared reweighting, the disparate impact remover preprocessing method, and the equalized odds postprocessing method to mitigate unfairness in synthetic educational data [31]. Hu and Rangwala compared the learning fair representations technique to their own domain-specific inprocessing technique for reducing bias in identifying at-risk students [27]. Furthermore, Stinar and Bosch analyzed how unfairness mitigation methods distort education data in unintended ways [50]. These outlined unfairness methods still require adaptations or explorations within BKT (similar to how Xu et al. adapted methods for Markov modeling [53])

3. METHODS

In the following subsections, we describe our data collection and methodology in detail. In short, Carnegie Learning provided us with BKT data of students across the entire duration of two school years, estimated the probability of stu-

dents being emerging readers, then used these predictions as the group categories for fairness analyses of the BKT models. Finally, we adapted a common unfairness mitigation strategy to attempt bias mitigation within the BKT models.

3.1 Bayesian Knowledge Tracing (BKT)

The BKT model represents a student's knowledge in terms of mastering specific skills. BKT assumes that the mastery process for each skill follows a hidden Markov process with the probability of mastery being updated as the student completes more problems related to the skill. The BKT model is based on four parameters: (i) the probability of a student already having a skill mastered, (ii) the probability the student masters the skill after a problem, (iii) the probability of guessing a problem correctly without mastery, (iv) the probability of incorrectly answering a problem despite mastering the skill (i.e., "slipping").

Using the pyBKT package for BKT implementation, we predicted the probability of mastery given first question correctness for each student across the 50 skills for each dataset with the most student interactions [4].

3.2 Emerging Reader Estimator

The student category *emerging reader* is defined by an existing, validated reading comprehension model [2]. The reading comprehension model estimates students' reading comprehension ability from the introductory workspace within MATHia (i.e., requiring only students' first few interactions with MATHia to estimate). Specifically, the estimator is trained on four categories of engineered features from the introductory workspace. The first is whether or not a student gets the problem step correct on the first attempt. The second is the total number of attempts per problem step. The third is the number of feedback prompts that the student has received per problem step. The final category tracks the number of hints that the student requests per problem step. The reading comprehension model was originally trained to predict English Language Arts standardized test outcomes as a probability; the lowest 25% of predicted probabilities per grade level are then classified as emerging readers, while the rest of students are classified as non-emerging readers. Thus, our estimation of emerging reader status could conflate English language learner status with struggling reader status. However, any early estimation of reading ability allows for diverse types of support to be given to students. Furthermore, it is often impossible to obtain demographic details from all systems. Thus, reading ability is helpful, especially in systems used at scale, where the developers or researchers using data from the system are unlikely to have access to test score or demographics data for all students but can still estimate their emerging reader status.

3.3 Data

We analyzed two datasets consisting of students' actions and outcomes in MATHia provided by Carnegie Learning. The two datasets are from the 2021–2022 and 2022–2023 school years in several schools from one school district in the Northeastern United States. We processed each dataset to consider only the student actions from the top 50 skills ranked by the number of students who completed each knowledge component, given that there is a long tail of infrequently used skills that are less relevant to characterizing BKT bias than the common components. During interactions, the students can either make an attempt at an answer or ask for a hint. Correct/not-correct labels given to BKT models are based on whether or not the student correctly answers on their first attempt with no hint, which is taken as one piece of evidence that the student has mastered that skill.

The 2021–2022 dataset consisted of 4,733 students interacting with the MATHia software. Of these students, 3,617 (76.4%) were considered to be non-emerging readers, and 1,116 (23.6%) students were estimated to be emerging readers; note that the percentages are not exactly 75%/25% due to ties in the probability scores used to measure reading. The 2022–2023 dataset contained interactions and outcomes of 3,816 students with 2,895 (75.8%) being non-emerging readers and 921 (24.2%) being emerging readers. This results in 8,549 students across both datasets. Since emerging readers are defined as the 25% of students with the lowest estimation from the emerging reader estimator with students grouped by year, our datasets, by definition, have representative samples of both emerging and non-emerging readers. These data were collected from students at the same grade level in each academic year.

3.4 Unfairness Mitigation

For our main analyses, we compared BKT models on the original, unmodified datasets with the datasets preprocessed using the reweighting unfairness mitigation algorithm [32]. To do this, we compared aggregate and per-knowledge component performance measurements and fairness metrics to identify bias in the BKT models. Furthermore, we performed bootstrapping on the outputs to determine if there were statistically significant differences between BKT estimates for emerging and non-emerging readers.

Initially, we compared aggregate and per-knowledge-component performance measurements and fairness metrics to identify possible bias in the BKT models. We trained the BKT models using five-fold cross-validation with the options of 0, 0.05, and 0.1 for slip probabilities and 0, 0.1, 0.2, and 0.3 for guess probabilities. We evaluated the BKT models using area under the curve (AUC) and root mean square error (RMSE) alongside the statistical parity [7], disparate impact [55], average odds difference [9], and equal calibration definitions of fairness [12]. We define statistical parity as difference in mastery rates between groups: specifically, $P(mastery|emerging) - P(mastery|nonemerging)$. We then define disparate impact as the ratio of selection rates between groups. That is, $\frac{P(mastery|emerging)}{P(mastery|nonemerging)}$. Average odds is a relaxed equalized odds that provides the average difference in false positive and true positive rates across groups. A value of 0 would indicate perfect fairness for statistical parity and average odds, whereas a value of 1 would indicate perfect fairness for disparate impact. Our definition of model calibration comes from Caruana and Niculescu-Mizil [12]. Lastly, we calculate model calibration by finding the true versus predicted probability across overlapping samples of 100 instances of the true and predicted values. We then computed the difference in calibration between predictions of emerging and non-emerging readers as a measure of unfairness. For calibration, a value of 0 indicates perfect

fairness.

We applied the previously described fairness definitions to provide a multifaceted fairness analyses of BKT predicted outcomes for emerging and non-emerging readers. Overall, we cover the non-discrimination criteria of independence, separation, and sufficiency [7].

Using the AI Fairness 360 toolkit (AIF360), and creating a wrapper for pyBKT to work as a scikit-learn classifier [44], we preprocessed our data using the reweighting unfairness mitigation algorithm [9]. reweighting is designed to address bias by adjusting weights of training instances to mitigate differences across groups [32]. The BKT models trained on the reweighed data are then compared to the models trained on the original data to examine the impact of unfairness mitigation on the BKT models for emerging readers.

Since the metrics and statistics we are using for fairness do not have a traditional closed-form solution like traditional methods of calculating p-value, we used statistical bootstrapping to estimate our confidence intervals. Specifically, we performed bootstrapping to determine if any differences between emerging and non-emerging readers were significant. With 10,000 iterations and measuring 95% confidence intervals, we bootstrapped the difference in means between emerging and non-emerging readers for the base rates, predicted values, model error at problem-solving steps, and model calibration. By comparing the difference in means, we examined if the difference in both base rates (i.e., the data) and model predictions were significant. Furthermore, by finding significance between the differences in model error (RMSE), we determined if BKT models provided similar errors to both groups. Finally, by bootstrapping the difference in calibration between the groups, we calculated if the model predictions were significantly different—and thus potentially unfair—for emerging readers.

4. RESULTS

Our results are organized by research question (RQ1–3). Table 1 summarizes the results of our aggregate BKT models trained on the two original datasets and our two datasets preprocessed using the reweighting unfairness mitigation algorithm. Table 2 describes the differences of BKT models trained on different skills in MATHia. Figure 1 displays the bootstrapped confidence intervals for the differences between emerging and non-emerging readers.

4.1 RQ1: Do emerging and non-emerging readers have different outcomes within a mathematics-focused adaptive learning software?

The results outlining bootstrapped 95% confidence intervals are in Figure 1. We compare four different bootstrapped statistics for our two original datasets and the two unfairness mitigated datasets. We statistically tested for differences in the overall base rates of mastery, predicted mastery, model errors, and model calibration between the two groups. We found significance in six out of the eight tests.

The emerging and non-emerging readers had differing base rates of mastery in both the 2021–2022 and 2022–2023 datasets. The base rate of mastery in the 2021–2022 dataset for emerging readers was .687 and for non-emerging readers it was .776. We bootstrapped the difference between these two means and the resulting confidence interval ($CI_{2021-2022} = [.0747, .0950]$) and mean difference can be seen in Figure 1. Since the difference in means is significant, we can conclude that the emerging and non-emerging readers groups had disparate rates of math mastery in MATHia. The base rates of mastery for emerging and non-emerging readers in the 2022–2023 dataset were .673 and .748 respectively. We performed the same bootstrapping analysis for significance on the 2022–2023 dataset and also found that the difference in means between the base rates of emerging and non-emerging reading was significant ($CI_{2022-2023} = [.0722, .0932]$). These results imply that across both years of data, emerging and non-emerging readers had significantly different rates of mastery.

Moving beyond the base rates, we wanted to know if the distribution of mastery emerging readers received was statistically different from the distribution of mastery that non-emerging readers received. We tested for significance in the difference of BKT predictions between groups. That is, we wanted to know if the distribution of mastery emerging readers received was statistically different from the distribution of mastery that non-emerging readers received. For the 2021–2022 dataset, emerging readers had an average mastery prediction of .708 and non-emerging readers had an average mastery prediction of .752. Similarly, for the 2022–2023 dataset, emerging readers had an average prediction of .679 and non-emerging readers had an average prediction of .733. Furthermore, we found that the difference between means of BKT model predictions for both datasets was significant ($CI_{2021-2022} = [0.0366, 0.0503]$, $CI_{2022-2023} = [0.0317, 0.0452]$). These results imply that the BKT models predicted mastery with significantly different distributions for the two groups (as they arguably should given the difference in base rates).

We also investigated whether the average RMSE for emerging readers was significantly different than the average RMSE for non-emerging readers in regards to attempt-level correctness. Figure 1 illustrates the results of our bootstrapped RMSE comparison for the two original and two unfairness mitigated datasets. Unlike the previous two statistical tests, we did not find significance in the difference in RMSE between groups for any of the four datasets. The 2021–2022 dataset had a mean difference in RMSE of -.0031 ($CI_{2021-2022} = [-0.0079, 0.0018]$), and the 2022–2023 dataset had a mean difference in RMSE of -.0006 ($CI_{2022-2023} = [-0.0067, 0.0062]$). These results show that the prediction errors were not substantially larger in one group than the other, since the confidence intervals include 0. This suggests two possible interpretations: (i) there is no meaningful difference in the model errors that each group experiences, and (ii) the model is splitting the difference between emerging and non-emerging readers, so that errors were similar but in opposite directions across groups. The overall RMSE would not capture that since RMSE does not indicate the direction of the error. The statistical parity results in RQ2 below further clarify these results, suggesting (ii) is the likely interpretation.

Lastly, we bootstrapped the difference in model calibration

Table 1: The table shows the aggregate measurements of the BKT models trained on the two original datasets and the reweighed versions of both original datasets. There is minimal to no difference between the aggregate measurements across the BKT models trained on the original and reweighed versions of the datasets. The mean difference between the AUC of our original datasets and the reweighting datasets is $\approx$ .002.

Dataset	AUC	RMSE	Statistical Parity	Disparate Impact	Average Odds
2021–2022	0.696	0.386	-0.077	0.885	-0.052
2021–2022 (RW)	0.699	0.386	-0.078	0.883	-0.054
2022–2023	0.694	0.393	-0.066	0.878	-0.047
2022–2023 (RW)	0.694	0.393	-0.066	0.878	-0.047

for emerging and non-emerging readers. For the 2021–2022 dataset, emerging and non-emerging readers had a difference in model calibration of .0046 ($CI_{2021-2022} = [.0009, .0086]$). For the 2022–2023 dataset, emerging and non-emerging readers had a difference in model calibration of .0054 ($CI_{2022-2023} = [.0012, .0097]$). Thus, both datasets showed a statistically significant—if quite small—difference in calibration between the two groups, such that BKT probabilities were slightly better calibrated for non-emerging readers.

4.2 RQ2: Is the MATHia BKT model unbiased with respect to the students who are emerging readers?

The fairness metrics displayed in Table 1 show bias in the aggregate BKT predictions for emerging readers. For the 2021–2022 original dataset, the -0.077 statistical parity value indicates that the aggregate model had a slight bias toward predicting lower knowledge for emerging readers. Similarly, disparate impact (i.e., a ratio where perfect fairness implies the value is 1) being 0.885 also indicates there was a bias towards predicting lower knowledge for emerging readers. Furthermore, average odds being -0.052 represents the false positive and true positive rates between the groups were not equivalent. This indicates that emerging readers were predicted to gain lower knowledge than non-emerging readers. Likewise, the 2022–2023 dataset exhibited the same trends in measurement with a -0.066 statistical parity, 0.878 disparate impact, and -0.047 average odds.

Despite the lack of perfect fairness, the levels of bias that the fairness metrics displayed were small for the aggregate models. For example, the -0.077 statistical parity for the 2021–2022 dataset implies a 7.7% difference in the predicted rate (Table 1. Table 1 also shows the impact of the BKT models' predictions being trained on the reweighted datasets. Notably, there was minimal difference across all fairness metric measurements once the data is reweighted, perhaps due to the already low level of bias in the predictions when aggregated across all topics.

To further understand the fairness of the BKT models, we examined the bias within skill-specific BKT models. While we could not include the analysis of all 50 skills in both of the original datasets and the two unfairness mitigated datasets due to computational limitations, we included multiple examples of models trained on the skills present in the datasets. Specifically, Table 2 presents the worst, best, and median BKT model in terms of statistical parity trained on specific skills on the original and unfairness mitigated data. Notably, the reweighting technique did not substantially impact any of the bias in the skill-level models.

The BKT model trained on the "geometric transform" skill displayed in Table 2 is a skill-specific model that had less bias than than the aggregate BKT. The statistical parity, disparate impact, and average odds fairness metrics all indicate less bias towards emerging readers than the aggregate model. Alongside some skills having fairer models, others had more biased models. For example, the models trained on the "identifying units" and "apply exponent" skills in Table 2 both exhibited more extreme biases than the aggregate model. The BKT models trained on the "apply exponent" skill had a difference of 26% mastery prediction for emerging and non-emerging readers (as presented by the statistical parity metric). Outside of the three skills presented, 36 of the 50 skills analyzed were biased against emerging readers in terms of statistical parity and disparate impact. 34 of the skills were biased against emerging readers considering the average odds metric. These results show that models for certain skills can contain large amounts of bias that the aggregate BKT representation does not display.

4.3 RQ3: If there is bias, can we mitigate it through common unfairness mitigation strategies?

Despite there being bias in specific skills, there is consistent minimal change when applying the reweighting unfairness mitigation algorithm to the data. First, in Table 1, we present that the reweighting method had minimal impact on the aggregate model predictions and had a mean difference of AUC from the models trained on the original dataset of $\approx$ .002, thus performance was minimally impacted by reweighting. Similarly, in Table 2, we found the reweighting preprocessed datasets yielded no apparent improved fairness for the knowledge component models due to there being minimal change in the fairness measures when reweighted. Finally, reweighting did not lead to the differences in emerging and non-emerging readers to be significant as further described in the previous section and seen in Figure 1. Table 1, Table 2, and Figure 1 each display how the reweighting unfairness mitigation algorithm was ineffective in mitigating the biases between emerging and non-emerging readers.

5. DISCUSSION

Our analyses resulted in a few main findings. The first is that BKT models, when considered in aggregate across all skills, had small but significant biases against emerging readers when examining common definitions of fairness. Despite the minimal bias in aggregate BKT, there was noticeable bias within certain skills. The second finding is that emerg-

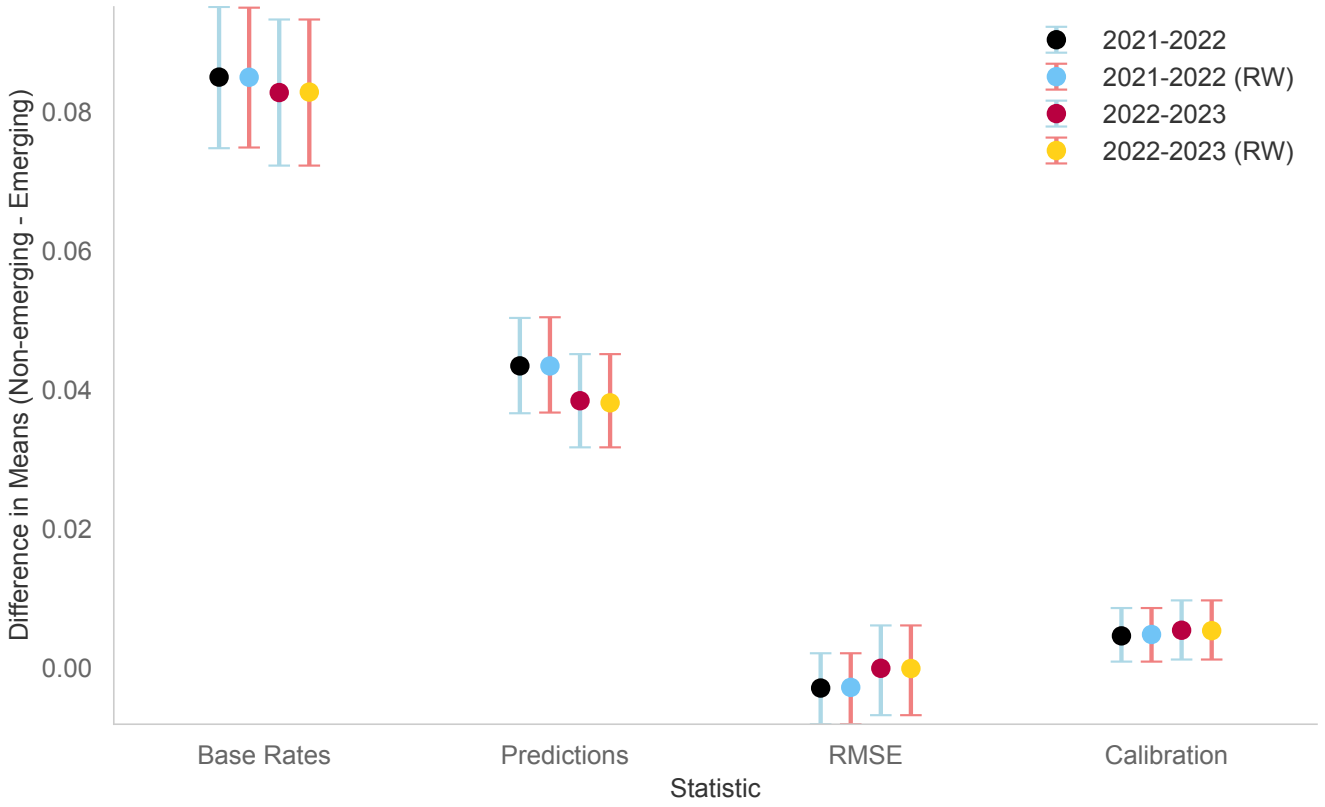

Figure 1: Presented are the mean differences and confidence intervals of each bootstrapped statistic. Across each statistic and for both of the datasets, the reweighting method had minimal impact on the results. Our measurements of difference in base rates, predictions, and calibration were all significant. The second and fourth intervals for each statistic represent the results from the datasets that were preprocessed with the reweighting (RW) algorithm.

ing readers had a significant difference in base rates and model predicted rates of answer correctness compared to the rest of the students. Furthermore, the BKT models did have a significant error in predicted rates when compared to the base rates across all groups.

Our analysis of student outcomes with MATHia showed several differences between emerging and non-emerging readers RQ1) Initially, when examining overall BKT model performance and the classical group fairness definitions (i.e., statistical parity, disparate impact, and average odds), we found minimal bias in the model against emerging readers. However, in Table 2 we found that specific skills in MATHia had varying levels of bias against emerging readers. Furthermore, by bootstrapping the differences in group base rates, predictions, model errors, and calibrations, we determined that emerging and non-emerging readers were not only statistically different, but also had different learning experiences within MATHia.

Our results imply both that reading ability is an important aspect in how students interact with mathematics learning software and that there are ways to improve current learning software to better support learners with different levels of reading ability. The significance in the difference of base rates and predictions between the two groups shows that the groups interact with the learning software in different ways. The difference in outcomes is shown throughout our analyses, in terms of how the groups experience mastery, how the groups are estimated mastery by the software, and how they experience specific skills within the learning software.

By answering the first research question, we also began to examine the possible bias in MATHia that RQ2 addresses (*RQ2: Is the MATHia BKT model unbiased with respect to the students who are emerging readers?*). Holistically, as seen in Table 1, the MATHia BKT model is in some respects fair. Both datasets have an absolute statistical parity measure of $< .08$ and absolute average odds measure of $< .06$. The low statistical parity tells us that reading ability has minimal impact on the mastery prediction. The low average odds represents that the difference in false positive and true positives is almost equal despite differing reading levels. Furthermore, Figure 1 illustrates that the mean model calibration for each group is similar. By analyzing each of the three anti-discrimination criteria (i.e., sufficiency, separation, and independence), we can conclude that the aggregate MATHia BKT model contains little unfairness towards emerging readers; however, the unfairness is still statistically significant (e.g., the difference in model calibrations between emerging and non-emerging readers is significant) given the large sample size. Since the base rates are different

Table 2: This table shows fine-grained fairness analyses of specific BKT models trained on different skills. We show both the knowledge component-specific BKT model trained on the original data and on the reweighed data. *RW* represents the knowledge component models trained on the reweighed datasets.

Skills	AUC	RMSE	Statistical Parity	Disparate Impact	Average Odds
geometric transform	0.512	0.285	-0.009	0.991	-0.007
geometric transform (RW)	0.512	0.284	-0.009	0.991	-0.007
identifying units	0.809	0.420	-0.175	0.736	-0.089
identifying units (RW)	0.809	0.420	-0.175	0.736	-0.089
apply exponent	0.771	0.439	-0.262	0.609	-0.181
apply exponent (RW)	0.771	0.439	-0.263	0.609	-0.181

for emerging and non-emerging readers, it is impossible for the model to be fair in every way. Thus, adaptive learning designers and researchers must decide if they want equally predicted rates of mastery or equally correct predictions, or a compromise between both goals. These results are similar to research in related fields that found not all definitions of fairness can be equally satisfied at the same time [20]. Related to adaptive learning, Prihar et al. [46] found significantly more learning when lowering the threshold for mastery because the sum of benefits (i.e., students get to experience more topics) outweighed the cost (i.e., students may not learn each topic as well). When resolving the discrepancy between base rate and predicted rate, then, it may be pedagogically favorable to increase the predicted mastery for non-emerging readers, prioritizing reducing differences in predicted mastery across groups at the cost of increasing miscalibration and potentially reducing accuracy. These results imply that when the unfairness is significant, even if small, there are still ways in which we can improve the BKT model and in turn other adaptive learning software to promote equity for emerging readers.

Our analysis of different skills (i.e., Table 2) gives the most straightforward way to promote equity in learning software for emerging readers, implementing different levels of mastery for different skills. The analysis also implies that certain skills rely more heavily on reading comprehension. Thus, BKT modeling for math curricula could potentially be improved by adding one or more "reading" skills into the steps that require higher levels of reading comprehension. Then, that reading skill could be estimated, like the math skills, throughout the adaptive learning software. This would allow the model to predict performance due to the reading skill, without conflating reading skills and math skills. Furthermore, the results show that skills in MATHia can be analyzed on a more fine-grained level to understand why some are biased against emerging readers (e.g., some math topics might require previous vocabulary that math courses do not teach).

Finally, since we observed bias within specific skills in MATHia, we tested if the reweighting unfairness mitigation algorithm was successful in mitigating that knowledge component-specific bias (i.e, *RQ3: If there is bias, can we mitigate it through common unfairness mitigation strategies?*). We used reweighting since it has been used in other educational tasks related to classification [50]. However, the reweighting unfairness mitigation algorithm had minimal, if any, changes to the BKT model results. Also, the reweighting method failed to reduce the unfairness present in any of the fine-grained skills. These results suggest that since reweighting works to ensure a definition of fairness within a whole dataset, and that the overall model is slightly unfair, the reweighting method also fails to notice (and correct for) the biases within specific skills and the significant differences in outcomes of emerging and non-emerging readers.

Each of our contributions helps promote the design of more equitable learning software. In tandem with better understanding of how unfairness mitigation methods interact with BKT models, the results can be captured in a few main contributions. We have quantitatively shown that emerging reading skill affects how students interact with math-based adaptive learning software, that meaningful non-math-based groupings can be found using initial outcomes in learning software, and that skill models with unmodeled difficulty factors (i.e., that do not model all of the underlying skills including non-math skills) can exhibit unfairness.

6. LIMITATIONS

Although we present a large-scale study using diverse students, there are many details about both our emerging and non-emerging students that we simply do not have access to. For example, the students in this study are primarily located within specific school districts in the Midwestern and North Eastern United States, and may represent different demographics (with different approaches to using online learning systems) than we might find in other parts of the world.

Likewise, within this data, there may be variation that is not yet accounted for. As we have discussed above, there are many reasons that a student could be classified as an emerging reason. Students with learning disabilities may differ in their behaviors and learning than students from English Language Learning (ELL) backgrounds. Likewise, ELLs are not monolithic groups, and they likely differ based on years of English exposure, reading fluency in their first language, and what language family their first language belongs to. Moreover, both emerging and non-emerging readers are likely to differ based on factors like socioecomic status (SES), as variables like these are known to affect students learning opportunities both inside and outside of the classroom.

Furthermore, we only analyzed the most commonly completed 50 skills in MATHia, ranked by the number of unique students who interacted with them, to focus on the most influential parts of the curriculum. It is possible that less frequently assigned content is non-random with respect to reliance on reading ability, and will merit further analysis as

data becomes available.

6.1 Future Work

The results of this study suggest directions for future work involving the relationship between reading ability and BKT estimates in other learning domains. In addition to better understanding how this relationship mediates learning more generally, work is needed to better understand how BKT will need to be adjusted to accurately predict the moment of learning for learners with different reading skills.

Although the relationship between reading and mathematics learning has been studied in other contexts, there is a need for more work on this issue in online learning systems, particularly those using adaptive algorithms. Further research can be done to understand more facets of emerging reader interactions and outcomes. For example, outside of traditional fairness analyses, the patterns of interactions in mathematics learning software of emerging versus non-emerging readers would help researchers better understand how the two groups differ in their interactions. That is, emerging readers might have different patterns while interacting with mathematics learning software over the long-term (e.g., patterns in requesting help or gaming tendencies).

In this study, we tested the reweighting algorithm's [32] ability to mitigate bias in the MATHia data. This algorithm was chosen due to its prevalence in related work and that the method does not modify ground truth labels [32]. Unfortunately, it was unable to substantially improve the fairness in this data, but future work should test other unfairness mitigation techniques. Given the Markov model nature of BKT-based models, other preprocessing techniques [55, 10] and postprocessing techniques [25, 32, 9] may be strong candidates. Furthermore, possible inprocessing unfairness mitigation techniques could be created to specifically ensure skill-level fairness in BKT models similar to Xu et al. [53]. For example, methods could be designed to introduce variable mastery thresholds for group-level mastery rates. These unfairness mitigation techniques can also be applied within different knowledge tracing frameworks [8, 1].

7. CONCLUSION

Our analysis shows how emerging and non-emerging readers interact with mathematics-focused adaptive learning software, which revealed insights into how adaptive learning systems—specifically, those using BKT—can be designed to support more equitable learning experiences. BKT models contained significant differences for emerging and non-emerging readers, indicating differences in how they interact with the ITS. These disparities suggest underlying reading-related biases in specific skills might affect student outcomes, or at least understanding of students' learning, if not accounted for.

Emerging readers do interact with the adaptive learning software differently (RQ2), and we were able to detect statistically significant prediction biases based on emerging reader status (RQ1). However, we were unable to successfully use reweighting to mitigate unfairness in specific BKT skills (RQ3), but our analyses do suggest promising opportunities for targeted interventions (e.g., for specific relevant reading skills) and opportunities for improving adaptive learning systems (e.g., by modeling reading skill) that decrease the bias imparted by specific skills. In sum, the results advance research about fairness in adaptive learning environments by illustrating the importance of emerging readers' as a critical category to consider in fairness assessments.

8. ACKNOWLEDGMENTS

This research was supported by NSF grant no. 2000638. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.

9. REFERENCES

[1] G. Abdelrahman, Q. Wang, and B. Nunes. Knowledge Tracing: A Survey. *ACM Computing Surveys*, 55(11):1–37, Nov. 2023.

[2] H. Almoubayyed, S. Fancsali, and S. Ritter. Generalizing predictive models of reading ability in adaptive mathematics software. In *Proceedings of the 16th International Conference on Educational Data Mining*, pages 207–216, Bengaluru, India, July 2023. International Educational Data Mining Society.

[3] H. Almoubayyed, S. E. Fancsali, and S. Ritter. Instruction-Embedded Assessment for Reading Ability in Adaptive Mathematics Software. In *LAK23: 13th International Learning Analytics and Knowledge Conference*, pages 366–377, Arlington TX USA, Mar. 2023. ACM.

[4] A. Badrinath, F. Wang, and Z. Pardos. pyBKT: An Accessible Python Library of Bayesian Knowledge Tracing Models. In *Proceedings of the 14th International Conference on Educational Data Mining*, pages 1–7, Virtual, 2021. International Educational Data Mining Society.

[5] 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.

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

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

[8] J. Barrett, A. Day, and K. Gal. Improving Model Fairness with Time-Augmented Bayesian Knowledge Tracing. In *Proceedings of the 14th Learning Analytics and Knowledge Conference*, pages 46–54, Kyoto Japan, Mar. 2024. 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):4:1–4:15, July 2019.

[10] F. Calmon, D. Wei, B. Vinzamuri, K. Natesan Ramamurthy, and K. R. Varshney. Optimized Pre-Processing for Discrimination Prevention. In I. Guyon, U. V. Luxburg, S. Bengio,

H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 30, page 3995–4004, Los Angeles, California, 2017. Curran Associates, Inc.

[11] A. P. Carnevale, Educational Testing Service, and Hispanic Association of Colleges and Universities. *Education = success: empowering Hispanic youth and adults*. ETS leadership 2000 series. [Educational Testing Service], [Princeton, NJ], 1999.

[12] R. Caruana and A. Niculescu-Mizil. Data mining in metric space: an empirical analysis of supervised learning performance criteria. In *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 69–78, Seattle WA USA, Aug. 2004. ACM.

[13] W. V. Catone and S. A. Brady. The inadequacy of individual educational program (IEP) goals for high school students with word-level reading difficulties. *Annals of Dyslexia*, 55(1):53–78, June 2005.

[14] L. E. Celis, L. Huang, V. Keswani, and N. K. Vishnoi. Classification with Fairness Constraints: A Meta-Algorithm with Provable Guarantees. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, FAT* '19, pages 319–328, New York, NY, USA, 2019. Association for Computing Machinery. event-place: Atlanta, GA, USA.

[15] J. M. Cock, H. Saltini, H. Sheng, R. Ranjan, R. Davis, and T. Käser. Investigation of behavioral Differences: Uncovering Behavioral Sources of Demographic Bias in Educational Algorithms. In *Proceedings of the 17th International Conference on Educational Data Mining*, pages 443–451, Atlanta, Georgia, July 2024. International Educational Data Mining Society.

[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] S. Doroudi and E. Brunskill. Fairer but Not Fair Enough On the Equitability of Knowledge Tracing. In *Proceedings of the 9th International Conference on Learning Analytics & Knowledge*, pages 335–339, Tempe AZ USA, Mar. 2019. ACM.

[18] C. Dwork, M. Hardt, T. Pitassi, O. Reingold, and R. Zemel. Fairness through awareness. In *Proceedings of the 3rd Innovations in Theoretical Computer Science Conference*, pages 214–226, Cambridge Massachusetts, Jan. 2012. ACM.

[19] S. Finkelstein, E. Yarzebinski, C. Vaughn, A. Ogan, and J. Cassell. The Effects of Culturally Congruent Educational Technologies on Student Achievement. In D. Hutchison, T. Kanade, J. Kittler, J. M. Kleinberg, F. Mattern, J. C. Mitchell, M. Naor, O. Nierstrasz, C. Pandu Rangan, B. Steffen, M. Sudan, D. Terzopoulos, D. Tygar, M. Y. Vardi, G. Weikum, H. C. Lane, K. Yacef, J. Mostow, and P. Pavlik, editors, *Artificial Intelligence in Education*, volume 7926, pages 493–502. Springer Berlin Heidelberg, Berlin, Heidelberg, 2013. Series Title: Lecture Notes in Computer Science.

[20] S. A. Friedler, C. Scheidegger, and S. Venkatasubramanian. The (im)possibility of fairness: different value systems require different mechanisms for fair decision making. *Commun. ACM*, 64(4):136–143, mar 2021.

[21] B. Friedman and C. College. Bias in computer systems. *ACM Transactions on Information Systems*, 14(3):330–347, 1996.

[22] P. Fuentes. Reading comprehension in mathematics. *The Clearing House*, 72(2):81–88, 1998.

[23] B. Green and L. Hu. The myth in the methodology: Towards a recontextualization of fairness in machine learning. In *Machine Learning: The Debates workshop at the 35th International Conference on Machine Learning (ICML)*, page 5, Stockholm, Sweden, 2018. International Machine Learning Society.

[24] K. J. Grimm. Longitudinal associations between reading and mathematics achievement. *Developmental neuropsychology*, 33(3):410–426, 2008. Place: England.

[25] M. Hardt, E. Price, E. Price, and N. Srebro. Equality of Opportunity in Supervised Learning. *Advances in Neural Information Processing Systems*, 30:1–9, 2016.

[26] W. Holmes, K. Porayska-Pomsta, K. Holstein, E. Sutherland, T. Baker, S. B. Shum, O. C. Santos, M. T. Rodrigo, M. Cukurova, I. I. Bittencourt, and K. R. Koedinger. Ethics of AI in Education: Towards a Community-Wide Framework. *International Journal of Artificial Intelligence in Education*, 32(3):504–526, Sept. 2022.

[27] 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*, pages 431–438, Virtual, July 2020. International Educational Data Mining Society.

[28] P. Hur, H. Lee, S. Bhat, and N. Bosch. Using Machine Learning Explainability Methods to Personalize Interventions for Students. In *Proceedings of the 15th International Conference on Educational Data Mining*, pages 438–445, Durham, United Kingdom, 2022. International Educational Data Mining Society.

[29] 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*, FAT* '19, page 49–58, New York, NY, USA, 2019. Association for Computing Machinery.

[30] N. Hübner, C. Merrell, H. Cramman, J. Little, D. Bolden, and B. Nagengast. Reading to learn? The co-development of mathematics and reading during primary school. *Child Development*, 93(6):1760–1776, Nov. 2022. Publisher: John Wiley & Sons, Ltd.

[31] L. Jiang, C. Belitz, and N. Bosch. Synthetic Dataset Generation for Fairer Unfairness Research. In *Proceedings of the 14th Learning Analytics and Knowledge Conference*, pages 200–209, Kyoto Japan, Mar. 2024. ACM.

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

[33] R. F. Kizilcec and H. Lee. Algorithmic fairness in education. In Routledge, editor, *The Ethics of Artificial Intelligence in Education*, pages 174–202. W.

Holmes & K. Porayska-Pomsta (Eds.), New York, 1 edition, Aug. 2022.

[34] L. Klenk. Case study in reading disability: An emergent literacy perspective. *Learning Disability Quarterly*, 17(1):33–56, 1994.

[35] J. K. Klingner, A. J. Artiles, and L. M. Barletta. English Language Learners Who Struggle With Reading: Language Acquisition or LD? *Journal of Learning Disabilities*, 39(2):108–128, Mar. 2006. Publisher: SAGE Publications Inc.

[36] J. Kurvers. Emerging literacy in adult second-language learners: A synthesis of research findings in the Netherlands. *Writing systems research*, 7(1):58–78, 2015.

[37] J. Lee. Racial and Ethnic Achievement Gap Trends: Reversing the Progress Toward Equity? *Educational Researcher*, 31(1):3–12, Jan. 2002. Publisher: American Educational Research Association.

[38] L. Li, N. Srivastava, J. Rong, G. Pianta, R. Varanasi, D. Gašević, and G. Chen. Unveiling goods and bads: A critical analysis of machine learning predictions of standardized test performance in early childhood education. In *Proceedings of the 14th Learning Analytics and Knowledge Conference*, LAK '24, page 608–619, New York, NY, USA, 2024. Association for Computing Machinery.

[39] M. Liu, J. Kang, W. Zou, H. Lee, Z. Pan, and S. Corliss. Using Data to Understand How to Better Design Adaptive Learning. *Technology, Knowledge and Learning*, 22(3):271–298, Oct. 2017.

[40] M. Liu, E. McKelroy, S. B. Corliss, and J. Carrigan. Investigating the effect of an adaptive learning intervention on students' learning. *Educational Technology Research and Development*, 65(6):1605–1625, Dec. 2017.

[41] F. Martin, Y. Chen, R. L. Moore, and C. D. Westine. Systematic review of adaptive learning research designs, context, strategies, and technologies from 2009 to 2018. *Educational Technology Research and Development*, 68(4):1903–1929, Aug. 2020.

[42] N. Mehrabi, F. Morstatter, N. Saxena, K. Lerman, and A. Galstyan. A survey on bias and fairness in machine learning. *ACM Comput. Surv.*, 54(6):1–35, jul 2021.

[43] Z. Pardos, Y. Bergner, D. Seaton, and D. Pritchard. Adapting Bayesian knowledge tracing to a massive open online course in edX. In S. D'Mello, R. Calvo, and A. Olney, editors, *Proceedings of the 6th International Conference on Educational Data Mining, EDM 2013*, Proceedings of the 6th International Conference on Educational Data Mining, EDM 2013, pages 137–145, Memphis, Tennessee, Jan. 2013. International Educational Data Mining Society.

[44] 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.

[45] K. A. Piia Maria Vilenius-Tuohimaa and J. Nurmi. The association between mathematical word problems and reading comprehension. *Educational Psychology*, 28(4):409–426, 2008.

[46] E. Prihar, M. Syed, K. Ostrow, S. Shaw, A. Sales, and N. Heffernan. Exploring common trends in online educational experiments. In A. Mitrovic and N. Bosch, editors, *Proceedings of the 15th International Conference on Educational Data Mining*, pages 27–38, Durham, United Kingdom, July 2022. International Educational Data Mining Society.

[47] J. E. Richey, N. G. Lobczowski, P. F. Carvalho, and K. Koedinger. Comprehensive Views of Math Learners: A Case for Modeling and Supporting Non-math Factors in Adaptive Math Software. In I. I. Bittencourt, M. Cukurova, K. Muldner, R. Luckin, and E. Millán, editors, *Artificial Intelligence in Education*, volume 12163, pages 460–471. Springer International Publishing, Cham, 2020. Series Title: Lecture Notes in Computer Science.

[48] S. Ritter, J. R. Anderson, K. R. Koedinger, and A. Corbett. Cognitive Tutor: Applied research in mathematics education. *Psychonomic Bulletin & Review*, 14(2):249–255, Apr. 2007.

[49] N. K. Scammacca, G. Roberts, S. Vaughn, and K. K. Stuebing. A Meta-Analysis of Interventions for Struggling Readers in Grades 4–12: 1980–2011. *Journal of Learning Disabilities*, 48(4):369–390, July 2015. Publisher: SAGE Publications Inc.

[50] F. Stinar and N. Bosch. Algorithmic unfairness mitigation in student models: When fairer methods lead to unintended results. In A. Mitrovic and N. Bosch, editors, *Proceedings of the 15th International Conference on Educational Data Mining*, pages 606–611, Durham, United Kingdom, July 2022. International Educational Data Mining Society.

[51] E. Sulzby and W. Teale. Emergent literacy. *Handbook of reading research*, 2:727–757, 1991.

[52] M. Verger, S. Lallé, F. Bouchet, and V. Luengo. Is your model "MADD"? A novel metric to evaluate algorithmic fairness for predictive student models. In M. Feng, T. Käser, and P. Talukdar, editors, *Proceedings of the 16th International Conference on Educational Data Mining*, pages 91–102, Bengaluru, India, July 2023. International Educational Data Mining Society.

[53] Y. Xu, C. Deng, Y. Sun, R. Zheng, X. Wang, J. Zhao, and F. Huang. Adapting static fairness to sequential decision-making: bias mitigation strategies towards equal long-term benefit rate. In *Proceedings of the 41st International Conference on Machine Learning*, ICML'24. JMLR.org, 2024.

[54] 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.

[55] 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.

[56] 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.

[57] 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. Issue: 3.

[58] B. H. Zhang, B. Lemoine, and M. Mitchell. Mitigating Unwanted Biases with Adversarial Learning. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, pages 335–340, New Orleans LA USA, Dec. 2018. ACM.