Fairness Implications of Unmodeled Reading Skills in Adaptive Learning Systems

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

Clara Belitz
College of the Holy Cross
Worcester, MA, USA
cbelitz@holycross.edu

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

Husni Almoubayyed
Carnegie Learning, Inc.
Pittsburgh, PA, USA
halmoubayyed@carnegielearning.com

Jaclyn Ocumpaugh
University of Houston
Houston, Texas, USA
jlocumpaugh@gmail.com

HaeJin Lee
University of Sydney
Darlington, NSW, Australia
haejin.lee@sydney.edu.au

Nidhi Nasiar
University of Pennsylvania
Philadelphia, PA, USA
nasiar@upenn.edu

Steve Ritter
Carnegie Learning, Inc.
Pittsburgh, PA, USA
sritter@carnegielearning.com

Ryan S. Baker
Adelaide University
Adelaide, South Australia, AU
ryan.baker@unisa.edu.au

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

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). 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 95% CI = [0.075, 0.095]), indicating that emerging reader status is predictive of mastery. We then conducted an intersectional fairness analysis of emerging reader status and traditional demographic groups. We found that students of different demographics, when considering emerging reader status, have varying rates of mastery of up to 10%. We 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 (Martin et al., 2020). While the benefits of such software are clear (e.g., Martin et al. 2020; Liu et al. 2017; Hur et al. 2022), there is a growing concern regarding different individual learning outcomes from interactions with the software (Liu et al., 2017; Almoubayyed et al., 2023a; Almoubayyed et al., 2023b). Previous research has shown that reading ability is a strong indicator of math success (Hübner et al., 2022; Grimm, 2008), but less work has examined this relationship in digital learning environments. This study builds on recent work from Almoubayyed et al. (2023a) and Stinar et al. (2025) that has sought to address this gap, examining how differences in student reading ability relate to outcomes in mathematics-focused adaptive learning software. We extend this previous research by analyzing whether demographics, combined with emerging reader status, could also explain students' differential experiences in a mathematics-focused Bayesian knowledge tracing system. Specifically, we additionally provide an intersectional fairness analysis with traditional demographic information and emerging reader membership.

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 (Corbett and Anderson, 1995). 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 (Corbett and Anderson, 1995). 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 (Doroudi and Brunskill, 2019).

Ensuring fairness in educational tools is crucial, as biased predictions can disproportionately impact students from historically underrepresented backgrounds, potentially exacerbating existing systemic educational inequalities (Kizilcec and Lee, 2022; Baker and Hawn, 2022). Inequalities can also come from other factors that affect how different students interact with learning systems, such as disability status, parental education levels, familiarity with technology, and more. In this study, we focus on students' reading ability as one such dimension, studying how this variable and its intersection with traditional demographics, such as race and gender, could influence 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), 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 (Richey et al., 2020; Fuentes, 1998; Vilenius-Tuohimaa et al., 2008), and it

has been used to predict long-term math understanding (Grimm, 2008). Ensuring fairness across student reading abilities could 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 related term "emergent literacy" was introduced to emphasize the linguistic skills and assets that should be drawn upon to help very young children acquire reading skills (Sulzby and Teale, 1991), and it has since expanded to include adult learners as well as learners who have learning disabilities (Klenk, 1994; Kurvers, 2015). In this study, we operationalize emerging reader status algorithmically, using an existing model that predicts whether students are emerging English language learners using interactions in the introductory phases of the learning session (Almoubayyed et al., 2023a). 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 (Ritter et al., 2007), as either *emerging readers* or *non-emerging readers*.

In tandem with reading comprehension research, studies have evaluated differences among race/ethnicity and gender groups in student math performance and reading comprehension. Historically, both the race/ethnicity and the gender of learners are correlated with achievement gaps in math performance and reading comprehension (Walkington et al., 2018; Reardon and Galindo, 2009; Hixson and McGlinchey, 2004; Breda et al., 2023). We further examine these inequalities in the context of BKT and how these groups intersect with emerging reader status.

This paper focuses on three main research questions: **RQ1**: Do emerging and non-emerging readers have different outcomes within a mathematics-focused adaptive learning software? **RQ2**: Is the MATHia BKT model (un)biased with respect to the students who are emerging readers? **RQ3**: Is the MATHia BKT model (un)biased for emerging readers within demographic groups?

We examine bias towards emerging readers by testing BKT model performance in two distinct but related processes. The first process analyzes the bias within each of the most common skills in MATHia (i.e., the skills with the largest number of student interactions). MATHia also has over 600 groupings of skills called "workspaces." The second process analyzes the aggregate BKT model performance on the most common workspaces in MATHia. We performed a two-pronged approach to measure the bias present within the learning session as a whole and the 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) (Barocas et al., 2023).

Thus, this 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 treat emerging reader status as an individual difference that can be combined with traditional demographics to identify areas of unfairness in BKT. The results show that emerging readers interact with MATHia in significantly different ways than other students—suggesting that there is an unmodeled factor (i.e., reading comprehension) in the curriculum that affects students' ability to achieve mastery. The results 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. The analysis scripts are at: https://osf.io/fvczr/.

2. Background

Traditionally, modeling math learning processes has focused on students' knowledge of math concepts (Knisley, 2015). Richey et al. (2020) call for a comprehensive view of modeling math learners that would include modeling non-math factors as well. However, if these non-math factors are less effective for students with diverse backgrounds, adding additional modeling could potentially widen the opportunity gap for students who have historically been underserved. Researchers have established areas where students could be underserved, including that current intelligent tutoring systems can provide unequal learning gains (Pape, 2018; Thomas et al., 2024). Furthermore, Ogan et al. (2015) found that inequalities could be due to cultural differences in learners, thus showing that educational models might not be generalizable to diverse student populations. Gašević et al. (2016) also discuss that the effectiveness of learning analytics is related to the diverse ways that learning tools are deployed in different classrooms and contexts. Inequalities in outcomes from learning tools could come from numerous places. This paper provides an option for understanding and handling unfairness that could emerge from applying learning analytics systems to a diverse student population (Gardner et al., 2019; Holt, 2024). In this section, we discuss related work in emerging readers, adaptive learning, and the measurement and mitigation of algorithmic unfairness.

2.1. Emerging Readers and Achievement Differences

Research has long shown that non-math factors can play a key role in the success of math learning (Carnevale et al., 1999; Lee, 2002; Cribbs et al., 2015), as math proficiency often requires some level of reading comprehension (Fuentes, 1998). Students' reading ability can predict their ability to solve math word problems and their understanding of the symbols in traditional math (Vilenius-Tuohimaa et al., 2008). Additionally, reading comprehension scores predict mastery across multiple math topics (Grimm, 2008).

Researchers have found links between struggling readers and protected classes of learners. Scammacca et al. (2015) surveyed interventions for struggling readers across three decades and found over 80 studies with a focus on improving reading comprehension for students. Alongside trying to help struggling readers, other researchers have found connections between reading ability and protected groups in education. For example, Catone and Brady (2005) found that individual education programs (IEPs) were specifically ineffective for students who had word-level reading difficulties. Also, Klingner et al. (2006) found a complex relationship between students who are English language learners, struggling readers, and students who have learning disabilities.

This study joins a growing body of research examining the connection between reading comprehension and math learning within adaptive learning software (Richey et al., 2020). 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 (Almoubayyed et al., 2023a). Emerging readers likely span a large range of demographics including, for example, groups marginalized in STEM or students from low socioeconomic

status households.

Furthermore, researchers have devoted decades to analyzing and understanding achievement gaps in mathematics and reading in schools (Willis, 2015; Grimm, 2008; Hübner et al., 2022). Mathematics achievement gaps have been demonstrated to be related to socioeconomic status, language proficiency, race, and gender, among other factors (Walkington et al., 2018; Tate, 1997). These gaps have been examined by researchers who theorize that girls' advantage in reading can be correlated to the achievement gap in genders for mathematics (Breda et al., 2023; Breda and Napp, 2019). We extend this line of research by evaluating how reading comprehension, at the intersection of traditional demographics, relates to outcomes and model predictions in real-world mathematics learning software.

2.2. MATHia Computer-based Learning Environment

MATHia is an intelligent tutoring system that is used for math mastery learning (Ritter et al., 2007). 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 (Corbett and Anderson, 1995).

Because skills address different mathematics topics and concepts, each one may potentially require different amounts of English language proficiency. In other words, there may be various English skills (e.g., comprehension) which, 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 (Mehrabi et al., 2021). For example, Friedman and Nissenbaum (1996) outline 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, and each of these has been found in educational contexts (Baker and Hawn, 2022). Also, Doroudi and Brunskill (2019) analyzed biases within knowledge tracing, and Zambrano et al. (2024) investigated bias in BKT and carelessness detectors. Similarly, measuring bias in academic exams related to student demographics has long been a topic in educational research (Hutchinson and Mitchell, 2019). More recently, educational data mining research has also focused on specific data biases (Baker and Hawn, 2022) as well as broad questions around ethics in AI-driven education (Holmes et al., 2022). These data biases can become an issue of unfairness because of the types of harm that come from biased systems; in the worst case, data bias can lead to avenues for educators to make decisions based on prejudice (Pangrazio, 2024). Furthermore, certain researchers have found specific groups of students who are impacted by bias. Yanagiura et al. (2023) 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), and Li et al. (2024) found bias against students with disabilities when predicting standardized

test scores in early childhood education settings. Additionally, Finkelstein et al. (2013) found that third-grade Black students learned more in science using their computer-based system when African-American Vernacular English was used, versus the English typically used in educational software. Baker and Hawn (2022) raised the concern that we do not fully know which groups are impacted by algorithmic bias, and emerging readers are not one of the groups they found in their review. Also, researchers have found that a lack of diverse representations of student behaviors could lead to model bias (Cock et al., 2024). 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. (2023) list 19 different fairness metrics created between 1971 and 2017. Specifically, measurements of fairness are split into either procedural or statistical guidelines (Green and Hu, 2018). Procedural fairness measures are geared towards making sure that statistically similar data points are treated similarly (Dwork et al., 2012). Alternatively, other definitions of fairness involve ensuring that all students are given the pedagogical strategies that they specifically need. For example, the *equal opportunity* measure in Hardt et al. (2016) tests if the true positive rate is balanced across groups in order to prevent unwanted discrimination. Other measures analyze the probabilities of error rates (Zafar et al., 2017), the difference in selection rates between groups (Barocas et al., 2023), or even domain-specific concerns, such as measuring model discrimination independent of model performance (Verger et al., 2023).

Alongside measurement, methods for mitigating the unfairness found have also been proposed. The methods are generally grouped into preprocessing, inprocessing (Zhang et al., 2018; Celis et al., 2019), and postprocessing (Hardt et al., 2016) based on where the mitigation strategy is applied in the machine learning pipeline. These methods have been successfully applied in education. Kizilcec and Lee (2022) reference different areas within an educational algorithm where bias can be mitigated. Furthermore, Jiang et al. (2024) compared reweighting, the disparate impact remover preprocessing method, and the equalized odds postprocessing method to mitigate unfairness in synthetic educational data. Hu and Rangwala (2020) compared the learning fair representations technique to their own domain-specific inprocessing technique for reducing bias in identifying at-risk students. Also, Stinar and Bosch (2022) analyzed how unfairness mitigation methods distort education data in unintended ways. These outlined unfairness methods still require adaptations or explorations within BKT (similar to how Xu et al. (2024) adapted methods for Markov modeling). In our own previous research, we examined the reweighting preprocessing method (Kamiran and Calders, 2012), and found minimal impact on the results (Stinar et al., 2025).

3. METHODS

In the following subsections, we describe our data collection and methodology in detail. In short, we obtained BKT data of students across the entire duration of two school years, estimated the probability of students being emerging readers, then used these predictions as the group categories for fairness analyses of the BKT models. Finally, we used the emerging reader categories alongside student gender and race/ethnicity groups within an intersectional analysis to further analyze the fairness of 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 with the most student interactions for each dataset (Badrinath et al., 2021).

3.2. Emerging Reader Estimator

The student category *emerging reader* is defined by an existing, validated reading comprehension model (Almoubayyed et al., 2023a). 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). The introductory workspace consists of the first interactions a student has within MATHia. While the problems in this workspace are not immediately related to mathematics mastery, the problems have students engage with different features of MATHia and process the adaptive features of the software. As a means of characterizing the approximate demands on reading ability students face in this context, we computed the Gunning fog index of readability for the introductory workspace (Gunning, 1969). The Gunning fog index was 7.5, which corresponds approximately to grade levels and thus indicates that students who comprehend the introductory workspace well are at least at a seventh-grade reading level.

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 the students are classified as non-emerging readers. Thus, our estimation of emerging reader status could conflate English language learner status with struggling reader status. Nevertheless, since the emerging reader estimator is predicting English Language Arts standardized test outcomes, ELL status could be inferable from our estimations, but we caution against using the estimations past relatively low-stakes adaptive supports. Stronger inferences could be made as future research better understands the relationship between estimated emerging reader status and ELL or 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 demographic 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. Additional counts of intersectional groups are shown in Table 1. While the original datasets had seven race/ethnicity groups, we focused on the groups that had large enough sample sizes to conduct a statistical analysis, as the others were quite small (some with fewer than 10 students). Demographic information for these students was provided by the school district under an IRB. For the purposes of our research, we will use categories as they were defined by the school district. We refer to the categories of male and female as "gender," though these terms are more frequently associated with sex, while acknowledging that gender is not a binary variable (Lindqvist et al., 2021). We will also combine race and ethnicity to reflect the school district's use of "Hispanic" as a race category, though it is more commonly described as an ethnicity encompassing multiple races (Benthall and Haynes, 2019).

3.4. Unfairness Analysis

For the main analyses, we compared the base rates, predicted rates, errors, and fairness measures for emerging and non-emerging readers at the intersection of traditional demographics. We compared aggregate and per-knowledge-component performance measurements and fairness metrics to identify bias in the BKT models. Furthermore, we calculated bootstrap confidence intervals on the outputs to determine if there were statistically significant differences between BKT estimates for emerging and non-emerging readers and traditional demographic groups.

Initially, we compared performance 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 four established measures of algorithmic fairness: statistical parity (Barocas et al., 2023), disparate impact (Zafar et al., 2017), average odds (Bellamy et al., 2019), and model calibration def-

Table 1: Intersectional demographic group counts for 2021–2022 and 2022–2023 academic years. *Emerge* and *nonE* refer to emerging and non-emerging readers. Gender groups are denoted as M (male) and F (female). Race/ethnicity groups are W (white), AA (African-American), and H (Hispanic).

Year	Group	Count
Gender intersections		
21–22	Emerge + M	656
	Emerge + F	460
	nonE + M	1,791
	nonE + F	1,820
22–23	Emerge + M	495
	Emerge + F	426
	nonE + M	1,463
	nonE + F	1,430
Race/ethnicity intersections		
21–22	Emerge + W	115
	Emerge + AA	717
	Emerge + H	213
	nonE + W	589
	nonE + AA	2,100
	nonE + H	643
22–23	Emerge + W	83
	Emerge + AA	553
	Emerge + H	234
	nonE + W	486
	nonE + AA	1,681
	nonE + H	514

initions of fairness (Kleinberg et al., 2017). We define statistical parity as the 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 version of 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 (Caruana and Niculescu-Mizil, 2004). 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 (alongside additional demographic information) as a measure of unfairness. For calibration, a value of 0 indicates perfect fairness.

We applied the previously described fairness definitions to provide multifaceted fairness

Table 2: Aggregate measurements of the BKT models trained on the two original datasets for differences in emerging and non-emerging readers.

Dataset	AUC	RMSE	Statistical Parity	Disparate Impact	Average Odds
2021–2022	0.696	0.386	-0.077	0.885	-0.052
2022–2023	0.694	0.393	-0.066	0.878	-0.047

analyses of BKT predicted outcomes for emerging and non-emerging readers. Overall, we cover the non-discrimination criteria of independence, separation, and sufficiency (Barocas et al., 2023).

Since the metrics and statistics we are using for fairness do not have a closed-form solution for significance like traditional methods of calculating p-values, we used statistical bootstrapping to estimate our confidence intervals. Specifically, we performed bootstrapping to determine if any differences between emerging and non-emerging readers within and between demographic groups 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 whether the difference in both base rates (i.e., the data) and model predictions was 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). Figure 1 displays the bootstrapped confidence intervals for the differences between emerging and non-emerging readers (RQ1). Table 2 summarizes the results of our aggregate BKT models trained on the two original datasets with respect to emerging reader membership (RQ2), and Table 4 describes the differences of BKT models trained on different skills in MATHia with respect to emerging reader membership (RQ2). Finally, we compared intersectional test statistic differences for emerging reader groups in intersectional groups of student gender and race/ethnicity (RQ3). These results are displayed in Figure 2, Table 5, and Table 6.

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 two different bootstrapped statistics for our two 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 three out of the four tests. The results of these analyses are in Table 3.

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

Table 3: Tabulated mean differences and confidence intervals of each bootstrapped statistic. The measurements of differences in base rates, predictions, and calibration were all significant.

Data	Base Rates	Predictions	RMSE	Calibration
21–22	.089 [.0747, .0950]	.044 [.0366, .0503]	-.0031 [-.0079, .0018]	.0046 [.0009, .0086]
22–23	.075 [.0722, .0932]	.054 [.0317, .0452]	-.0006 [-.0067, .0062]	.0054 [.0012, .0097]

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 reader 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 readers 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.

In addition to the base rates, we wanted to know if the distribution of mastery for emerging readers was statistically different from the distribution of mastery for non-emerging readers. We tested for significance in the difference in 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 the 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 concerning attempt-level correctness. Figure 1 illustrates the results of our bootstrapped RMSE comparison for the two original datasets. Unlike the previous two statistical tests, we did not find significance in the difference in RMSE between groups for either of the two 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 significantly 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 for emerging and non-emerging readers. For the 2021–2022 dataset, emerging and non-emerging readers had a difference in

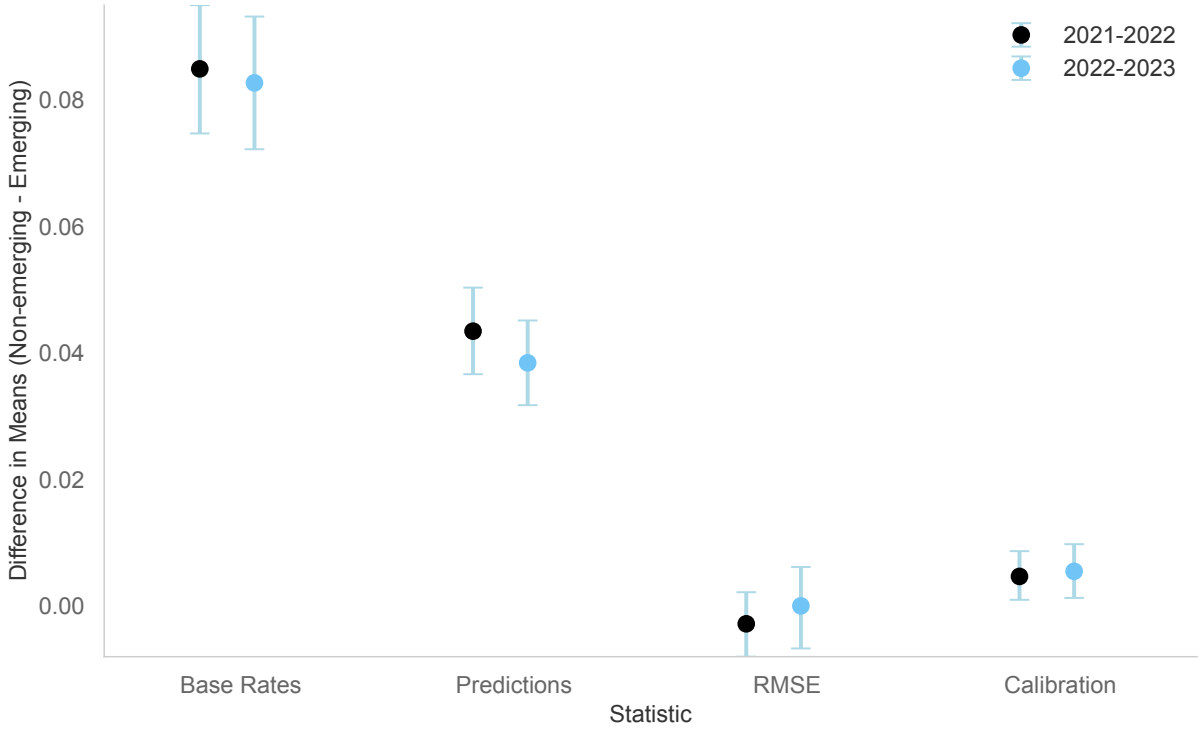


Figure 1: Mean differences and confidence intervals of each bootstrapped statistic. The measurements of differences in base rates, predictions, and calibration were all significant.

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 (un)biased with respect to the students who are emerging readers?

The fairness metrics displayed in Table 2 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, a disparate impact (i.e., a ratio where perfect fairness implies the value is 1) of 0.885 also indicates there was a bias towards predicting lower knowledge for emerging readers. Furthermore, an average odds of -0.052 indicates that the false positive and true positive rates between the groups are not equivalent. This indicates that emerging readers were predicted to gain less 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 2).

Table 4: Fine-grained fairness analyses of specific BKT models trained on different skills. We show the knowledge-component-specific BKT models trained on each year of data. The worst, best, and median BKT model in terms of statistical parity trained on specific skills are presented.

Skills	AUC	RMSE	Stat. Parity	Disp. Impact	Avg. Odds
geometric transform	0.512	0.285	-0.009	0.991	-0.007
identifying units	0.809	0.420	-0.175	0.736	-0.089
apply exponent	0.771	0.439	-0.262	0.609	-0.181

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 original datasets due to computational limitations, we included multiple examples of models trained on the skills present in the datasets. Specifically, Table 4 presents the worst, best, and median BKT model in terms of statistical parity trained on specific skills on the data.

The BKT model trained on the "geometric transform" skill displayed in Table 4 is a skill-specific model with less bias than the aggregate BKT. The statistical parity, disparate impact, and average odds fairness metrics all indicate less bias against emerging readers than the aggregate model. Though some skills had fairer models than the aggregate model, others had more biased models. For example, the models trained on the "identifying units" and "apply exponent" skills in Table 4 both exhibited more extreme biases than the aggregate model. The BKT models trained on the "apply exponent" skill had a difference of 26% in 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 using 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: Is the MATHia BKT model (un)biased for emerging readers within demographic groups?

RQ1–2 results established that the emerging reader status of a student relates to different rates of mastery in MATHia and significantly affects BKT. To better understand how emerging reader status effects on BKT might relate to existing achievement gaps, we considered four questions: (i) For a group of emerging or non-emerging readers, what difference does gender explain? (ii) For a given gender, what difference does emerging reader status explain? (iii) For a group of emerging or non-emerging readers, what difference does race/ethnicity explain? (iv) For a given race/ethnicity, what difference does emerging reader status explain? For each of these questions, we performed similar analyses as we did in RQ1 to determine if there are significant differences between the base rates, predicted rates of mastery, model errors, and model calibration; however, RQ3 incorporated intersectional student group differences. The results for questions (i) and (iii) are in Figure 2. The detailed results for all four questions are in Table 5 and Table 6, and the non-intersectional group differences are included in Table 6.

(i) For a group of emerging or non-emerging readers, what difference does gender explain? The differences between the test statistic and the genders of students within emerging and non-emerging reader groups are represented as M−F in Figure 2. We separated the data into emerging and non-emerging reader groups and then compared gender differences within

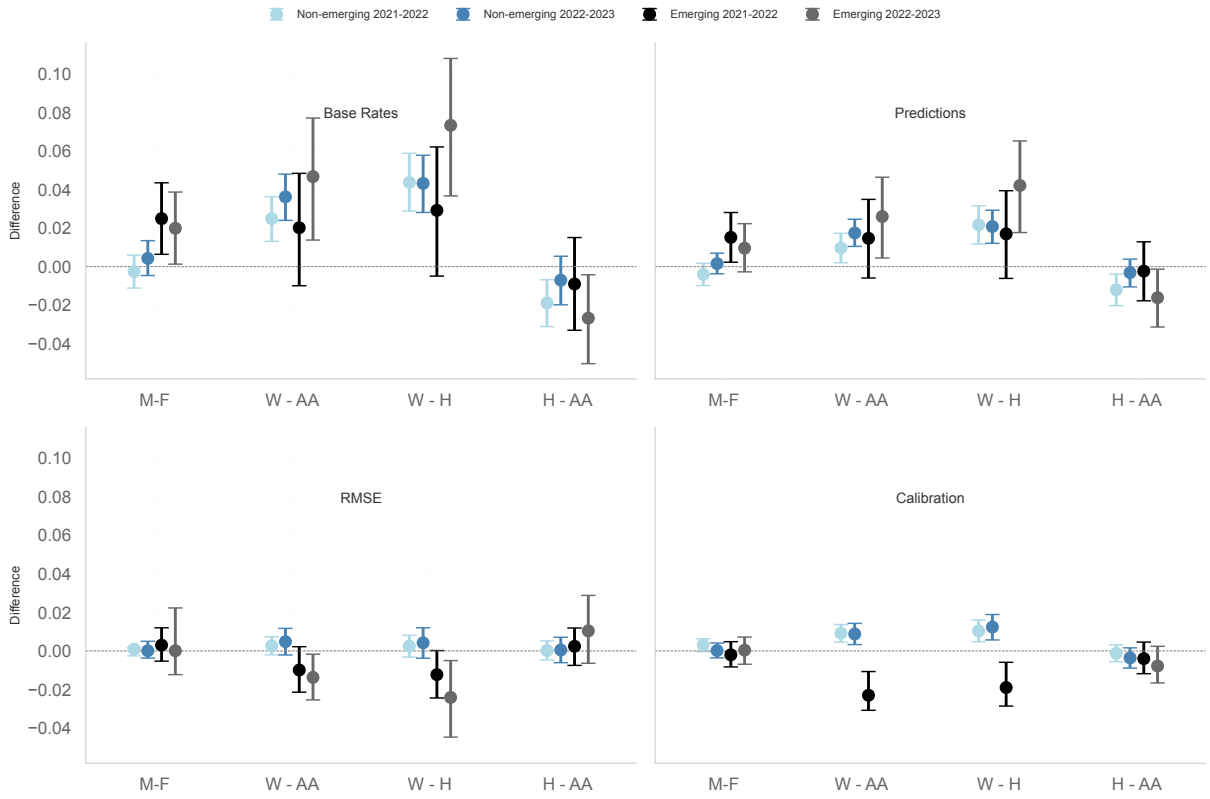


Figure 2: Differences in emerging reader status within and between gender demographics and race/ethnicity demographic groups across the four test statistics and both datasets. The genders are represented as M (male) and F (female). The race/ethnicity groups are W (white), AA (African-American), and H (Hispanic).

the reader groups. For the 2021–2022 data, we found that gender differences only significantly explained differences between the base and predicted rates of mastery for students who were emerging readers. Specifically, the base rates differed by .0249 ($CI_{2021-2022} = [.0064, .0435]$), and the predicted rates of mastery differed by .0152 ($CI_{2021-2022} = [.0023, .0281]$) when comparing differences between emerging readers who are male and female. In the 2022–2023 dataset, the base rates also significantly differed for the emerging readers when gender was considered (.0199, $CI_{2022-2023} = [.0013, .0387]$), but not for the predicted rates of mastery (.0096, $CI_{2022-2023} = [-.0027, .0223]$). Thus, across both datasets, we only found significant gender differences for emerging and non-emerging readers in base rates and predicted rates of mastery. Similar to the analysis of differences in just the emerging and non-emerging reader groups (RQ1), we found no significant difference here in the model errors. The significant difference in model calibration is not in these results that was previously found without considering the intersection of emerging reader status with gender (RQ1).

(ii) For a given gender, what difference does emerging reader status explain? To answer this question, we compared the four test statistics (base rates, predictions, RMSE, and calibration) for the differences in emerging and non-emerging groups within each of the two gender categories. The results for this question are shown in Table 5. We separated the data into gender groups, then compared emerging reader differences within the gender groups. For

Table 5: Differences in the test statistics for emerging reader status (*emerge* and *nonE*) within and between male (M) and female (F) genders for the 2021–2022 and 2022–2023 datasets.

Data	Comparison	Base Rates	Pred	RMSE	Calib
(i) For a group of emerging or non-emerging readers, what difference does gender explain?					
21–22	Emerge (M−F)	.0249 [.0064, .0435]*	.0152 [.0023, .0281]*	.0030 [-.0053, .012]	-.0020 [-.0083, .0048]
	nonE (M−F)	-.0026 [-.0111, .006]	-.0040 [-.0098, .0017]	.0009 [-.0026, .0017]	.0031 [-.0003, .0063]
22–23	Emerge (M−F)	.0199 [.0013, .0387]*	.0096 [-.0027, .0223]	.0001 [-.0123, .0223]	.0004 [-.0069, .0072]
	nonE (M−F)	.0043 [-.0046, .0135]	.0016 [-.0037, .0070]	.0001 [-.0037, .0050]	.0002 [-.0036, .0041]
(ii) For a given gender, what difference does emerging reader status explain?					
21–22	M (nonE−Emerge)	.0731 [.0596, .0872]*	.0350 [.0257, .0441]*	-.0038 [-.0115, .0031]	.0095 [.0042, .0143]*
	F (nonE−Emerge)	.1007 [.0857, .1161]*	.0542 [.0436, .0648]*	-.0017 [-.0075, .004]	.0044 [-.0007, .0097]
22–23	M (nonE−Emerge)	.0756 [.0608, .0902]*	.0347 [.0253, .0439]*	-.0002 [-.0083, .0073]	.0072 [.0018, .0128]*
	F (nonE−Emerge)	.0912 [.0747, .1061]*	.0427 [.033, .0531]*	-.0002 [-.0114, .0089]	.0073 [.0014, .0131]*

the 2021–2022 dataset, we found significant differences between the base rates and predicted rates of mastery for both the male group split on emerging reader status and the female group split on emerging reader status. Furthermore, we found a significant difference in the model calibration for the male group. Specifically, the non-emerging reader male group versus the emerging reader male group base rates had a difference of .0731 ($CI_{2021-2022} = [.0596, .0872]$), a difference in predictions of .0350 ($CI_{2021-2022} = [.0257, .0441]$), and a difference in calibration of .0095 ($CI_{2021-2022} = [.0042, .0143]$). For the non-emerging reader female group versus the emerging reader female group, the significant differences were the base rate difference of .1007 ($CI_{2021-2022} = [.0857, .1161]$) and the predicted rate of mastery difference of .0542 ($CI_{2021-2022} = [.0436, .0648]$). Analyses of the 2022–2023 dataset revealed similar results. The same test statistics were significantly different for the emerging and non-emerging reader gender groups, but model calibration was also significant for the female group. For non-emerging reader males versus emerging reader males, the base rate difference was .0756 ($CI_{2022-2023} = [.0608, .0902]$), the predicted rate of mastery difference was .0347 ($CI_{2022-2023} = [.0253, .0439]$), and the model calibration difference was .0072 ($CI_{2022-2023} = [.0018, .0128]$). For the female group, the base rate difference was .0912 ($CI_{2022-2023} = [.0747, .1061]$), the predicted rate of mastery difference was .0427 ($CI_{2022-2023} = [.0330, .0531]$), and the model calibration difference was .0073 ($CI_{2022-2023} = [.0014, .0131]$).

(iii) For a group of emerging or non-emerging readers, what difference does race/ethnicity explain? Similar to the previous questions, we answered this question by comparing the differences of the four test statistics for the emerging and non-emerging reader groups between different race/ethnicity groups in the data. For both the 2021–2022 and 2022–2023 datasets, we analyzed the white (W), African American (AA), and Hispanic (H) student groups (for the 2022–2023 dataset, due to a small population of emerging readers who were also white (n = 83), model calibration differences could not be calculated for this intersectional group). That is, we split the data by emerging reader groups, then compared race/ethnicity differences within the emerging reader groups. The results of our analyses can be seen in both Table 6 and the W−AA, W−H, and H−AA sections of Figure 2.

We observed that there were differences in the base rates and model predictions for non-emerging readers of different race/ethnicities. Out of the 12 bootstrapped test statistics analyzed in the 2021–2022 dataset for non-emerging reader groups splitting on race/ethnicity, 8 were significant. In contrast, only 1 of the 12 differences was significant for the emerging reader group splitting on race/ethnicity. Shown in the upper left facet of Figure 2, the base rate differ-

ences in the non-emerging reader group with white students and the non-emerging reader group with African-American students (W−AA) was .0249 ($CI_{2021-2022} = [.0131, .0363]$), the non-emerging reader group with white students and non-emerging reader group with Hispanic students (W−H) was .0437 ($CI_{2021-2022} = [.0288, .0588]$), and the non-emerging reader group with Hispanic students and the non-emerging reader group with African-American students (H−AA) was -.0188 ($CI_{2021-2022} = [-.0311, -.0068]$). Then, in the upper right facet of Figure 2, the differences in predicted rates of mastery are shown with W−AA for non-emerging readers being .0097 ($CI_{2021-2022} = [.0020, .0173]$), W−H being .0217 ($CI_{2021-2022} = [.0118, .0316]$), and H−AA being -.012 ($CI_{2021-2022} = [-.0202, -.0039]$).

For the 2022–2023 dataset, we similarly found that there were differences in the base rates and model predictions for non-emerging readers of different race/ethnicity groups. 8 of the 10 statistics for non-emerging reader intersections were significant. The base rate differences and predicted rates of mastery differences for each of the intersectional groups were significant for non-emerging readers. In contrast, for W−AA (-.0137, $CI_{2022-2023} = [-.0254, -.0017]$) and W−H (-.0241, $CI_{2022-2023} = [-.0447, -.0050]$) the model errors were also significantly different. For non-emerging reader differences in base rates for the race/ethnicity groups, W−AA (.0362, $CI_{2022-2023} = [.0240, .0480]$) and W−H (.0432, $CI_{2022-2023} = [.0281, .0578]$) were significant. For the emerging reader differences in base rates for the race/ethnicity groups, W−AA (.0467, $CI_{2022-2023} = [.0138, .0771]$), W−H (.0733, $CI_{2022-2023} = [.0367, .1080]$), and H−AA (-.0267, $CI_{2022-2023} = [-.0503, -.0042]$) were significantly different. The other significant differences in predicted rates of mastery and calibration, along with confidence intervals, can be found in Table 6.

In contrast to the 2021–2022 dataset, the 2022–2023 dataset had significant differences in the base rates and the predicted rates of mastery for the emerging reader groups of different race/ethnicity groups. That is, the difference in emerging reader base rates for W−AA was .0467 ($CI_{2022-2023} = [.0138, .0771]$) and the difference in emerging reader predicted rates of mastery for W−AA was .0260 ($CI_{2022-2023} = [.0045, .0464]$). We found similar trends in the emerging reader W−H differences and emerging reader H−AA differences. For emerging reader W−H, the base rate difference was .0733 ($CI_{2022-2023} = [.0367, .1080]$) and the predicted rate of mastery difference was .0421 ($CI_{2022-2023} = [.0177, .0652]$). Furthermore, for emerging reader H−AA, the difference in base rates was -.0267 ($CI_{2022-2023} = [-.0503, -.0042]$) and the difference in predicted rates of mastery was -.0161 ($CI_{2022-2023} = [-.0313, -.0113]$). The rest of the results can be found in both Figure 2 and Table 6.

(iv) For a given race/ethnicity, what difference does emerging reader status explain? To answer this question, we split the data by race/ethnicity groups, then compared emerging reader differences for each race/ethnicity group. As shown in Table 6, 7 of the 12 test statistic differences were significant for the 2021–2022 dataset. The difference in emerging reader base rates for W (.0868, $CI_{2021-2022} = [.0583, .1157]$) and H (.0728, $CI_{2021-2022} = [.0493, .0959]$) students when accounting for emerging reader status were significant, as were the differences in predicted mastery for W (.0391, $CI_{2021-2022} = [.0195, .0600]$), AA (.0440, $CI_{2021-2022} = [.0351, .0530]$), and H (.0355, $CI_{2021-2022} = [.0195, .0494]$). Also, the calibration differences when accounting for emerging reader status for the W (.0367, $CI_{2021-2022} = [.0245, .0439]$) and AA (.0045, $CI_{2021-2022} = [.0688, .0951]$) students were significant.

We observed that emerging reader status explained base rate differences and prediction differences for each race/ethnicity group in the 2022–2023 dataset. Furthermore, a different set of 7 of the 11 test statistics were significant. Each of the race/ethnicity groups had significant differ-

Table 6: Differences in the test statistics for emerging reader status (emerge and nonE) within and between race/ethnicity groups for the 2021–2022 and 2022–2023 datasets. The white group is represented by W, the African-American group is represented by AA, and the Hispanic group is represented by H.

Data	Comparison	Base Rates	Pred	RMSE	Calibration
(iii) For a group of emerging or non-emerging readers, what difference does race/ethnicity explain?					
21–22	Emerge (W−AA)	.0202 [-.0099, .0484]	.0147 [-.0059, .0349]	-.0099 [-.0214, .0022]	-.023 [-.0308, -.0107]*
	nonE (W−AA)	.0249 [.0131, .0363]*	.0097 [.0020, .0173]*	.0027 [-.0020, .0073]	.0092 [.0046, .0137]*
	Emerge (W−H)	.0292 [-.0049, .0621]	.0170 [-.0061, .0394]	-.0123 [-.0244, .0002]	-.019 [-.0286, -.0059]*
	nonE (W−H)	.0437 [.0288, .0588]*	.0217 [.0118, .0316]*	.0025 [-.0032, .0081]	.0104 [.0048, .0160]*
	Emerge (H−AA)	-.0090 [-.0330, .0151]	-.0023 [-.0177, .0129]	.0024 [-.0075, .0119]	-.0040 [-.0118, .0046]
	nonE (H−AA)	-.0188 [-.0311, -.0068]*	-.0120 [-.0202, -.0039]*	.0002 [-.0047, .0052]	-.0012 [-.0056, .0032]
22–23	Emerge (W−AA)	.0467 [.0138, .0771]*	.026 [.0045, .0464]*	-.0137 [-.0254, -.0017]*	–
	nonE (W−AA)	.0362 [.024, .048]*	.0175 [.0105, .0246]*	.0048 [-.0021, .0117]	.0088 [.0033, .0143]*
	Emerge (W−H)	.0733 [.0367, .108]*	.0421 [.0177, .0652]*	-.0241 [-.0447, -.005]*	–
	nonE (W−H)	.0432 [.0281, .0578]*	.0208 [.0121, .0293]*	.0042 [-.0038, .012]	.0124 [.0057, .0189]*
	Emerge (H−AA)	-.0267 [-.0503, -.0042]*	-.0161 [-.0313, -.0013]*	.0104 [-.0064, .0288]	-.0078 [-.0166, .0024]
	nonE (H−AA)	-.007 [-.0198, .0054]	-.0031 [-.0105, .0039]	.0005 [-.0061, .0071]	-.0036 [-.0089, .0016]*
(iv) For a given race/ethnicity, what difference does emerging reader status explain?					
21–22	W (nonE−Emerge)	.0868 [.0583, .1157]*	.0391 [.0195, .0600]*	.0077 [-.0029, .0177]	.0367 [.0245, .0439]*
	AA (nonE-Emerge)	.0045 [-.0005, .0093]	.0440 [.0351, .0530]*	-.0048 [-.0121, .0018]	.0045 [.0688, .0951]*
	H (nonE−Emerge)	.0728 [.0493, .0959]*	.0355 [.0195, .0494]*	-.0070 [-.0152, .0017]	.0072 [-.0012, .0149]
22–23	W (nonE−Emerge)	.0624 [.0323, .0942]*	.0252 [.0059, .0466]*	.0177 [.0053, .0297]*	–
	AA (nonE-Emerge)	.0729 [.0594, .0867]*	.0337 [.0249, .0424]*	-.0008 [-.0071, .0054]	.0020 [-.0031, .0075]
	H (nonE−Emerge)	.0926 [.0699, .1158]*	.0465 [.0329, .0606]*	-.0107 [-.0295, .0061]	.0062 [-.0041, .0149]
Non-intersectional differences					
21–22	W−AA	.0316 [.0208, .0428]*	.0146 [.0072, .0220]*	.0004 [-.0041, .0046]	.0055 [.0012, .0095]*
	W−H	.0475 [.0337, .0614]*	.0239 [.0147, .0331]*	-.0005 [-.0057, .0047]	.0044 [-.0007, .0096]
	H−AA	-.0159 [-.0273, -.0048]*	-.0093 [-.0167, -.0018]*	.0008 [-.0036, .0051]	.0011 [-.0030, .0052]
22–23	W−AA	.0451 [.0338, .0565]*	.0222 [.0154, .0209]*	.0023 [-.0038, .0083]	.0046 [-.0003, .0094]
	W−H	.0630 [.0488, .0770]*	.0317 [.0230, .0400]*	-.0016 [-.0104, .0069]	.0078 [.0020, .0139]*
	H−AA	-.0179 [-.0294, -.0064]*	-.0095 [-.0164, -.0026]*	.0038 [-.0035, .0117]	-.0033 [-.0082, .0013]

ences in base rates when accounting for emerging reader status. Specifically, W had a difference between emerging reader groups of .0624 ($CI_{2022-2023} = [.0323, .0942]$), AA had a difference between emerging reader groups of .0729 ($CI_{2022-2023} = [.0594, .0867]$), and H had a difference between emerging reader groups of .0926 ($CI_{2022-2023} = [.0699, .1158]$). The predicted rates of mastery were significantly different as well when accounting for emerging reader status for W (.0252, $CI_{2022-2023} = [.0059, .0466]$), AA (.0337, $CI_{2022-2023} = [.0249, .0424]$), and H (.0465, $CI_{2022-2023} = [.0329, .0606]$). The final significant difference was .0177 ($CI_{2022-2023} = [.0053, .0297]$) for the W group's model error.

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 emerging 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 for emerging readers when compared to the base rates across all groups. We also found consis-

tent differences in rates of mastery and model-predicted rates when considering the intersection of emerging reader status with gender and race/ethnicity.

The 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, as shown in Table 4, 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 outcomes of emerging and non-emerging readers were statistically different, and that these groups of students had different learning experiences within MATHia.

The findings imply both that reading ability is an important aspect in how students interact with mathematics learning software and, by extension, that there are ways to improve current learning software to better support learners with different levels of reading ability. The significance of the difference in 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 the analyses, in terms of how the groups experience mastery, how the groups are estimated to have achieved mastery by the software, and how they experience specific skills within the learning software. These findings are supported by previous research that students experience unequal learning gains while using the same educational system (Thomas et al., 2024).

In answering the first research question, we also began to examine the possible bias in MATHia that RQ2 addresses (*RQ2: Is the MATHia BKT model (un)biased with respect to the students who are emerging readers?*). Holistically, as seen in Table 2, the MATHia BKT model is, in some respects, fair. Both datasets have an absolute statistical parity measure of $< .08$ and an absolute average odds measure of $< .06$. As a reminder, both of these measures are satisfied when they are close to 0. The low statistical parity tells us that reading ability has minimal impact on the mastery prediction. The low average odds indicate that the difference in false positives 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 that does exist 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 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 (Friedler et al., 2021). Related to adaptive learning, Prihar et al. (2022) 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, 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.

The comparison of differences in skills (i.e., Table 4) 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. That is, as seen in Table 4, skills that require more reading comprehension likely exhibit more unfairness against emerging readers. For example, the more graphically represented "geometric transform" skill has less unfairness as opposed to the other two skills shown. 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. Aside from modeling a new skill, better understanding the impact that emerging reader status has on mastering math topics could justify modifying existing parameters commonly used in student modeling. For example, if emerging readers often understand the math topic but make mistakes due to the semantics of the math problem, the probability that a student incorrectly answers a problem despite mastering the skill (i.e., "slipping") could be weighted by the chance of a student being an emerging reader. 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).

When reducing unfairness for emerging readers within BKT, other biases could be introduced into the model. For example, modeling a "reading" skill alongside topic mastery could bias the model to overly rely on reading when recommending and providing a sequence towards mastery. Examining how a modeled reading skill or other options for unfairness mitigation impact BKT would further illustrate how researchers ought to mitigate the emerging reader unfairness.

Finally, since we found that emerging reader status does lead to different base rates and predicted rates of mastery, we analyzed if this difference can be further explained by additional demographics of the students (i.e., *RQ3: Is the MATHia BKT model (un)biased for emerging readers within demographic groups?*). Given previous research on achievement gaps in both mathematics and reading between student genders (Breda et al., 2023; Breda and Napp, 2019; Tate, 1997) and race/ethnicity groups (Hixson and McGlinchey, 2004; Willis, 2015; Reardon and Galindo, 2009; Hopewell and Escamilla, 2014), we hypothesized that some of the differences could be explained by previously researched demographic differences.

In terms of gender and emerging reader status, we found that emerging reader status explained large differences in base rates of mastery and predicted rates of mastery for both male and female students across both datasets. The difference in base mastery was nearly 10% for female students in the 2021–2022 dataset. This difference could be partially explained through previous work by Breda et al. (2023), which found that a high intent to learn and achieve in math is correlated with increased reading abilities in girls. Thus, since emerging reader status could be related to reading comprehension skills, our results are consistent with previous research around achievement gaps in math and reading. However, we find that both female *and* male students have different rates of mastery that are explained by emerging reader status. This implies that, albeit unequally, both groups are impacted by their membership in the emerging reader category. We further see this in the rows of Table 5 related to question (i). Across both years of data, we found that gender is not a significant explainer for non-emerging readers, yet

it was for emerging readers.

Consistent across 2021–2022 and 2022–2023, race/ethnicity partially explained base rates and predicted rates of mastery for non-emerging readers. Regarding the 2021–2022 dataset, this result implies emerging reader membership is correlated with mastery rates for students. Seemingly, all emerging reading students struggle across all race/ethnicity groups. This is in contrast to the significant differences between race/ethnicity groups when controlling for those who are non-emerging readers. However, when analyzing the 2022–2023 dataset, there were substantial differences for both the non-emerging readers and for the emerging readers when accounting for differences across race/ethnicity groups. Thus, in 2022–2023, race/ethnicity groupings are able to explain differences in mastery across the emerging reader groups. While further research needs to be done to better understand the relationship between emerging reader status and race/ethnicity groups of students, we find that, at a minimum, students with different demographic characteristics are differentially impacted by their emerging reader status when accounting for math mastery. Furthermore, these results support previous research that has evaluated achievement gaps between race/ethnicity groups of learners (Willis, 2015; Hixson and McGlinchey, 2004; Walkington et al., 2018). The analyses in this paper do, however, contain numerous comparisons that are somewhat exploratory in nature, increasing the possibility of type I error. Despite the possible error, using bootstrapping allowed us to measure significant interaction differences across learners in the BKT model, which would not be possible due to fairness measures not having a closed-form solution. Future confirmatory research in more contexts is needed to determine the extent to which findings in this context generalize.

Each of the contributions in this paper helps promote the design of more equitable learning software. In tandem with a 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 and Future Work

Although we present a large-scale study using a diverse group of students, there are many details about both 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 Northeastern 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. Similarly, the data provided to us by the school district was limited in its breadth of gender and race/ethnicity categories.

Likewise, within this data, there may be variation that is not yet accounted for. As previously discussed, there are many reasons that a student could be classified as an emerging reader. Students with learning disabilities may differ in their behaviors and learning versus 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 socioeconomic 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 become available.

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.

In our study, students are estimated to be emerging readers based on behavior during the introductory workspaces in MATHia. Thus, our emerging reader estimator is specific to the MATHia intelligent tutoring system. Future work could introduce ways to estimate emerging reader status during early interactions in other intelligent tutoring systems. Also, since our analysis is focused on MATHia, we only modeled differences with the BKT model. Future research could examine these differences within other modeling frameworks.

Since MATHia covers a wide range of math topics, some topics involve more word problems, while other topics are more graphical or numerical in nature. Now that we have established differences in outcomes for emerging readers overall, in future work, more fine-grained analyses can determine how problems or concepts with different levels of reading comprehension could relate to different levels of bias with regard to emerging readers. These results could provide implications for fields outside math as well, which are more focused on certain types of problems (e.g., physics or chemistry).

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 could 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).

While our previous study found minimal changes when using the reweighting unfairness mitigation algorithm for the emerging reader and non-emerging reader differences (Stinar et al., 2025), future work still needs to be done to evaluate whether unfairness mitigation strategies could still reduce bias in the intersectional demographic and emerging reader groups. Furthermore, future work should also test other unfairness mitigation techniques. Given the Markov model nature of BKT-based models, other preprocessing techniques (Zafar et al., 2017; Calmon et al., 2017) and postprocessing techniques (Hardt et al., 2016; Bellamy et al., 2019) may be strong candidates. Furthermore, possible inprocessing unfairness mitigation techniques could be created to specifically ensure skill-level fairness in BKT models. 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 (Barrett et al., 2024; Abdelrahman et al., 2023).

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 intelligent tutoring system. These disparities suggest that 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). Furthermore, we found that students of different genders and races/ethnicities have different rates of mastery and model interactions based on their emerging reader membership (RQ3). Thus, if we are able to mitigate unfairness due to the emerging reader status of a student, we might be able to also mitigate unfairness between traditional demographic groups. The results 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 found in specific skills. In sum, the results advance research about fairness in adaptive learning environments by illustrating the importance of considering emerging readers as a critical category for fairness assessments.

8. Acknowledgments

We thank Kole Norberg for his assistance with the analyses. 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. Declaration of Generative AI Software Tools in the writing process

During the preparation of this work, the authors used Qwen 3.6-27B in all sections as a final proofreading tool. After using this tool, the authors reviewed and edited the content as needed and take full responsibility for the content of the publication.

References

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

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

Almoubayyed, H., Fancsali, S. E., and Ritter, S. 2023b. Instruction-embedded assessment for reading ability in adaptive mathematics software. In *LAK23: 13th International Learning Analytics and Knowledge Conference*. ACM, Arlington TX USA, 366–377.

BADRINATH, A., WANG, F., AND PARDOS, Z. 2021. pyBKT: An Accessible Python Library of Bayesian Knowledge Tracing Models. In *Proceedings of the 14th International Conference on Educational Data Mining*. International Educational Data Mining Society, Virtual, 1–7.

BAKER, R. S. AND HAWN, A. 2022. Algorithmic Bias in Education. *International Journal of Artificial Intelligence in Education 32,* 4 (Dec.), 1052–1092.

BAROCAS, S., HARDT, M., AND NARAYANAN, A. 2023. *Fairness and Machine Learning: Limitations and Opportunities*. MIT Press, Cambridge, Massachusetts.

BARRETT, J., DAY, A., AND GAL, K. 2024. Improving Model Fairness with Time-Augmented Bayesian Knowledge Tracing. In *Proceedings of the 14th Learning Analytics and Knowledge Conference*. ACM, Kyoto Japan, 46–54.

BELLAMY, R. K. E., MOJSILOVIC, A., NAGAR, S., RAMAMURTHY, K. N., RICHARDS, J., SAHA, D., SATTIGERI, P., SINGH, M., VARSHNEY, K. R., ZHANG, Y., DEY, K., HIND, M., HOFFMAN, S. C., HOUDE, S., KANNAN, K., LOHIA, P., MARTINO, J., AND MEHTA, S. 2019. AI Fairness 360: An extensible toolkit for detecting and mitigating algorithmic bias. *IBM Journal of Research and Development 63,* 4/5 (July), 4:1–4:15.

BENTHALL, S. AND HAYNES, B. D. 2019. Racial categories in machine learning. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*. FAT* '19. Association for Computing Machinery, New York, NY, USA, 289–298.

BREDA, T., JOUINI, E., AND NAPP, C. 2023. Gender differences in the intention to study math increase with math performance. *Nature Communications 14,* 1 (June), 3664.

BREDA, T. AND NAPP, C. 2019. Girls' comparative advantage in reading can largely explain the gender gap in math-related fields. *Proceedings of the National Academy of Sciences 116,* 31 (July), 15435–15440. Publisher: Proceedings of the National Academy of Sciences.

CALMON, F., WEI, D., VINZAMURI, B., NATESAN RAMAMURTHY, K., AND VARSHNEY, K. R. 2017. Optimized Pre-Processing for Discrimination Prevention. In *Advances in Neural Information Processing Systems*, I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, Eds. Vol. 30. Curran Associates, Inc., Los Angeles, California, 3995–4004.

CARNEVALE, A. P., EDUCATIONAL TESTING SERVICE, AND HISPANIC ASSOCIATION OF COLLEGES AND UNIVERSITIES. 1999. *Education = success: empowering Hispanic youth and adults*. ETS leadership 2000 series. [Educational Testing Service], [Princeton, NJ].

CARUANA, R. AND NICULESCU-MIZIL, A. 2004. 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*. ACM, Seattle WA USA, 69–78.

CATONE, W. V. AND BRADY, S. A. 2005. The inadequacy of individual educational program (IEP) goals for high school students with word-level reading difficulties. *Annals of Dyslexia 55,* 1 (June), 53–78.

CELIS, L. E., HUANG, L., KESWANI, V., AND VISHNOI, N. K. 2019. Classification with Fairness Constraints: A Meta-Algorithm with Provable Guarantees. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*. FAT* '19. Association for Computing Machinery, New York, NY, USA, 319–328. event-place: Atlanta, GA, USA.

COCK, J. M., SALTINI, H., SHENG, H., RANJAN, R., DAVIS, R., AND KÄSER, T. 2024. Investigation of behavioral Differences: Uncovering Behavioral Sources of Demographic Bias in Educational Algorithms. In *Proceedings of the 17th International Conference on Educational Data Mining*. International Educational Data Mining Society, Atlanta, Georgia, 443–451.

CORBETT, A. T. AND ANDERSON, J. R. 1995. Knowledge tracing: Modeling the acquisition of procedural knowledge. *User Modelling and User-Adapted Interaction 4,* 4, 253–278.

CRIBBS, J. D., HAZARI, Z., SONNERT, G., AND SADLER, P. M. 2015. Establishing an explanatory model for mathematics identity. *Child Development 86,* 4, 1048–1062.

DOROUDI, S. AND BRUNSKILL, E. 2019. Fairer but Not Fair Enough On the Equitability of Knowledge Tracing. In *Proceedings of the 9th International Conference on Learning Analytics & Knowledge*. ACM, Tempe AZ USA, 335–339.

DWORK, C., HARDT, M., PITASSI, T., REINGOLD, O., AND ZEMEL, R. 2012. Fairness through awareness. In *Proceedings of the 3rd Innovations in Theoretical Computer Science Conference*. ACM, Cambridge Massachusetts, 214–226.

FINKELSTEIN, S., YARZEBINSKI, E., VAUGHN, C., OGAN, A., AND CASSELL, J. 2013. The Effects of Culturally Congruent Educational Technologies on Student Achievement. In *Artificial Intelligence in Education*, 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, Eds. Vol. 7926. Springer Berlin Heidelberg, Berlin, Heidelberg, 493–502. Series Title: Lecture Notes in Computer Science.

FRIEDLER, S. A., SCHEIDEGGER, C., AND VENKATASUBRAMANIAN, S. 2021. The (im)possibility of fairness: different value systems require different mechanisms for fair decision making. *Commun. ACM 64,* 4 (mar), 136–143.

FRIEDMAN, B. AND NISSENBAUM, H. 1996. Bias in computer systems. *ACM Transactions on Information Systems 14,* 3, 330–347.

FUENTES, P. 1998. Reading comprehension in mathematics. *The Clearing House 72,* 2, 81–88.

GARDNER, J., BROOKS, C., AND BAKER, R. 2019. Evaluating the Fairness of Predictive Student Models Through Slicing Analysis. In *Proceedings of the 9th International Conference on Learning Analytics & Knowledge*. LAK19. Association for Computing Machinery, New York, NY, USA, 225–234.

GAŠEVIĆ, D., DAWSON, S., ROGERS, T., AND GASEVIC, D. 2016. Learning analytics should not promote one size fits all: The effects of instructional conditions in predicting academic success. *The Internet and Higher Education 28*, 68–84.

GREEN, B. AND HU, L. 2018. 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)*. International Machine Learning Society, Stockholm, Sweden, 5.

GRIMM, K. J. 2008. Longitudinal associations between reading and mathematics achievement. *Developmental neuropsychology 33,* 3, 410–426. Place: England.

GUNNING, R. 1969. The Fog Index After Twenty Years. *Journal of Business Communication 6*, 13 – 3.

HARDT, M., PRICE, E., PRICE, E., AND SREBRO, N. 2016. Equality of Opportunity in Supervised Learning. *Advances in Neural Information Processing Systems 30*, 1–9.

HIXSON, M. D. AND MCGLINCHEY, M. T. 2004. The Relationship between Race, Income, and Oral Reading Fluency and Performance on Two Reading Comprehension Measures. *Journal of Psychoeducational Assessment 22,* 4 (Dec.), 351–364. Publisher: SAGE Publications Inc.

HOLMES, W., PORAYSKA-POMSTA, K., HOLSTEIN, K., SUTHERLAND, E., BAKER, T., SHUM, S. B., SANTOS, O. C., RODRIGO, M. T., CUKUROVA, M., BITTENCOURT, I. I., AND KOEDINGER, K. R. 2022. Ethics of AI in Education: Towards a Community-Wide Framework. *International Journal of Artificial Intelligence in Education 32,* 3 (Sept.), 504–526.

HOLT, L. 2024. Online mathematics programs may benefit most the kids who need it least. *Education Next 24,* 4, 26–31.

HOPEWELL, S. AND ESCAMILLA, K. 2014. Struggling Reader or Emerging Biliterate Student? Reevaluating the Criteria for Labeling Emerging Bilingual Students as Low Achieving. *Journal of Literacy Research 46,* 1 (Mar.), 68–89. Publisher: SAGE Publications Inc.

HU, Q. AND RANGWALA, H. 2020. Towards fair educational data mining: A case study on detecting at-risk students. In *Proceedings of the 13th International Conference on Educational Data Mining*. International Educational Data Mining Society, Virtual, 431–438.

HUR, P., LEE, H., BHAT, S., AND BOSCH, N. 2022. Using Machine Learning Explainability Methods to Personalize Interventions for Students. In *Proceedings of the 15th International Conference on Educational Data Mining*. International Educational Data Mining Society, Durham, United Kingdom, 438–445.

HUTCHINSON, B. AND MITCHELL, M. 2019. 50 years of test (un)fairness: Lessons for machine learning. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*. FAT* '19. Association for Computing Machinery, New York, NY, USA, 49–58.

HÜBNER, N., MERRELL, C., CRAMMAN, H., LITTLE, J., BOLDEN, D., AND NAGENGAST, B. 2022. Reading to learn? The co-development of mathematics and reading during primary school. *Child Development 93,* 6 (Nov.), 1760–1776. Publisher: John Wiley & Sons, Ltd.

JIANG, L., BELITZ, C., AND BOSCH, N. 2024. Synthetic Dataset Generation for Fairer Unfairness Research. In *Proceedings of the 14th Learning Analytics and Knowledge Conference*. ACM, Kyoto Japan, 200–209.

KAMIRAN, F. AND CALDERS, T. 2012. Data preprocessing techniques for classification without discrimination. *Knowledge and Information Systems 33,* 1 (Oct.), 1–33.

KIZILCEC, R. F. AND LEE, H. 2022. Algorithmic fairness in education. In *The Ethics of Artificial Intelligence in Education*, 1 ed., Routledge, Ed. W. Holmes & K. Porayska-Pomsta (Eds.), New York, 174–202.

KLEINBERG, J., MULLAINATHAN, S., AND RAGHAVAN, M. 2017. Inherent Trade-Offs in the Fair Determination of Risk Scores. In *8th Innovations in Theoretical Computer Science Conference (ITCS 2017)*, C. H. Papadimitriou, Ed. Leibniz International Proceedings in Informatics (LIPIcs), vol. 67. Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Dagstuhl, Germany, 43:1–43:23.

KLENK, L. 1994. Case study in reading disability: An emergent literacy perspective. *Learning Disability Quarterly 17,* 1, 33–56.

KLINGNER, J. K., ARTILES, A. J., AND BARLETTA, L. M. 2006. English Language Learners Who Struggle With Reading: Language Acquisition or LD? *Journal of Learning Disabilities 39,* 2 (Mar.), 108–128. Publisher: SAGE Publications Inc.

KNISLEY, J. 2015. A Four-Stage Model of Mathematical Learning. *THE MATHEMATICS EDUCATOR 12,* 1 (Apr.), 11–16.

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

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

LI, L., SRIVASTAVA, N., RONG, J., PIANTA, G., VARANASI, R., GAŠEVIĆ, D., AND CHEN, G. 2024. 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. Association for Computing Machinery, New York, NY, USA, 608–619.

LINDQVIST, A., SENDÉN, M. G., AND RENSTRÖM, E. A. 2021. What is gender, anyway: A review of the options for operationalising gender. *Psychology & Sexuality 12,* 4 (Oct.), 332–344.

LIU, M., KANG, J., ZOU, W., LEE, H., PAN, Z., AND CORLISS, S. 2017. Using Data to Understand How to Better Design Adaptive Learning. *Technology, Knowledge and Learning 22,* 3 (Oct.), 271–298.

LIU, M., MCKELROY, E., CORLISS, S. B., AND CARRIGAN, J. 2017. Investigating the effect of an adaptive learning intervention on students' learning. *Educational Technology Research and Development 65,* 6 (Dec.), 1605–1625.

MARTIN, F., CHEN, Y., MOORE, R. L., AND WESTINE, C. D. 2020. Systematic review of adaptive learning research designs, context, strategies, and technologies from 2009 to 2018. *Educational Technology Research and Development 68,* 4 (Aug.), 1903–1929.

MEHRABI, N., MORSTATTER, F., SAXENA, N., LERMAN, K., AND GALSTYAN, A. 2021. A survey on bias and fairness in machine learning. *ACM Comput. Surv. 54,* 6 (jul), 1–35.

OGAN, A., WALKER, E., BAKER, R., RODRIGO, M. M. T., SORIANO, J. C., AND CASTRO, M. J. 2015. Towards Understanding How to Assess Help-Seeking Behavior Across Cultures. *International Journal of Artificial Intelligence in Education 25,* 2 (June), 229–248.

PANGRAZIO, L. 2024. Data Harms: The Evidence Against Education Data. *Postdigital Science and Education 6,* 4 (Dec.), 1049–1054.

PAPE, B. 2018. Learner Variability Is the Rule, Not the Exception. *Digital Promise*, 1–6.

PRIHAR, E., SYED, M., OSTROW, K., SHAW, S., SALES, A., AND HEFFERNAN, N. 2022. Exploring common trends in online educational experiments. In *Proceedings of the 15th International Conference on Educational Data Mining*, A. Mitrovic and N. Bosch, Eds. International Educational Data Mining Society, Durham, United Kingdom, 27–38.

REARDON, S. F. AND GALINDO, C. 2009. The Hispanic-White Achievement Gap in Math and Reading in the Elementary Grades. *American Educational Research Journal 46,* 3 (Sept.), 853–891. Publisher: American Educational Research Association.

RICHEY, J. E., LOBCZOWSKI, N. G., CARVALHO, P. F., AND KOEDINGER, K. 2020. Comprehensive Views of Math Learners: A Case for Modeling and Supporting Non-math Factors in Adaptive Math Software. In *Artificial Intelligence in Education*, I. I. Bittencourt, M. Cukurova, K. Muldner, R. Luckin, and E. Millán, Eds. Vol. 12163. Springer International Publishing, Cham, 460–471. Series Title: Lecture Notes in Computer Science.

RITTER, S., ANDERSON, J. R., KOEDINGER, K. R., AND CORBETT, A. 2007. Cognitive Tutor: Applied research in mathematics education. *Psychonomic Bulletin & Review 14,* 2 (Apr.), 249–255.

SCAMMACCA, N. K., ROBERTS, G., VAUGHN, S., AND STUEBING, K. K. 2015. A Meta-Analysis of Interventions for Struggling Readers in Grades 4–12: 1980–2011. *Journal of Learning Disabilities 48,* 4 (July), 369–390. Publisher: SAGE Publications Inc.

STINAR, F. AND BOSCH, N. 2022. Algorithmic unfairness mitigation in student models: When fairer methods lead to unintended results. In *Proceedings of the 15th International Conference on Educational Data Mining*, A. Mitrovic and N. Bosch, Eds. International Educational Data Mining Society, Durham, United Kingdom, 606–611.

STINAR, F., LEE, H., BELITZ, C., NASIAR, N., FANCSALI, S., RITTER, S., ALMOUBAYYED, H., BAKER, R., OCUMPAUGH, J., AND BOSCH, N. 2025. Fairness of Bayesian Knowledge Tracing for Math Learners of Different Reading Ability. In *Proceedings of the 18th International Conference on Educational Data Mining*, C. Mills, G. Alexandron, D. Taibi, G. L. Bosco, and L. Paquette, Eds. International Educational Data Mining Society, Palermo, Italy, 170–181.

SULZBY, E. AND TEALE, W. 1991. Emergent literacy. *Handbook of reading research 2*, 727–757.

TATE, W. F. 1997. Race-Ethnicity, SES, Gender, and Lanuage Proficiency Trends in Mathematics Achievement: An Update. *Journal for Research in Mathematics Education 28,* 6 (Dec.), 652–679. Place: Reston VA, USA Publisher: National Council of Teachers of Mathematics.

THOMAS, D. R., GATZ, E., GUPTA, S., ALEVEN, V., AND KOEDINGER, K. R. 2024. The Neglected 15%: Positive Effects of Hybrid Human-AI Tutoring Among Students with Disabilities. In *Artificial Intelligence in Education*, A. M. Olney, I.-A. Chounta, Z. Liu, O. C. Santos, and I. I. Bittencourt, Eds. Springer Nature Switzerland, Cham, 409–423.

VERGER, M., LALLÉ, S., BOUCHET, F., AND LUENGO, V. 2023. Is your model "MADD"? A novel metric to evaluate algorithmic fairness for predictive student models. In *Proceedings of the 16th International Conference on Educational Data Mining*, M. Feng, T. Käser, and P. Talukdar, Eds. International Educational Data Mining Society, Bengaluru, India, 91–102.

VILENIUS-TUOHIMAA, P. M., AUNOLA, K., AND NURMI, J. 2008. The association between mathematical word problems and reading comprehension. *Educational Psychology 28,* 4, 409–426.

WALKINGTON, C., CLINTON, V., AND SHIVRAJ, P. 2018. How readability factors are differentially associated with performance for students of different backgrounds when solving mathematics word problems. *American Educational Research Journal 55,* 2 (Apr.), 362–414. Publisher: American Educational Research Association.

WILLIS, A. I. 2015. Literacy and Race: Access, Equity, and Freedom. *Literacy Research: Theory, Method, and Practice 64,* 1 (Nov.), 23–55. Publisher: SAGE Publications Inc.

XU, Y., DENG, C., SUN, Y., ZHENG, R., WANG, X., ZHAO, J., AND HUANG, F. 2024. 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, 54962–54982.

YANAGIURA, T., YANO, S., KIHIRA, M., AND OKADA, Y. 2023. Examining Algorithmic Fairness for First- Term College Grade Prediction Models Relying on Pre-matriculation Data. *Journal of Educational Data Mining 15,* 3 (Dec.), 1–25.

ZAFAR, M. B., VALERA, I., GOMEZ RODRIGUEZ, M., AND GUMMADI, K. P. 2017. Fairness Beyond Disparate Treatment & Disparate Impact: Learning Classification without Disparate Mistreatment. In *Proceedings of the 26th International Conference on World Wide Web*. International World Wide Web Conferences Steering Committee, Perth Australia, 1171–1180.

ZAMBRANO, A. F., ZHANG, J., AND BAKER, R. S. 2024. Investigating Algorithmic Bias on Bayesian Knowledge Tracing and Carelessness Detectors. In *Proceedings of the 14th Learning Analytics and Knowledge Conference*. ACM, Kyoto Japan, 349–359.

ZHANG, B. H., LEMOINE, B., AND MITCHELL, M. 2018. Mitigating Unwanted Biases with Adversarial Learning. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*. ACM, New Orleans LA USA, 335–340.