XAI Reveals the Causes of Attention Deficit Hyperactivity Disorder (ADHD) Bias in Student Performance Prediction

HaeJin Lee
School of Information Sciences
University of Illinois
Champaign, IL, USA
haejin2@illinois.edu

Clara Belitz
School of Information Sciences
University of Illinois
Champaign, IL, USA
cbelitz2@illinois.edu

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

Nigel Bosch
School of Information Sciences
University of Illinois
Champaign, IL, USA
Department of Educational Psychology
University of Illinois
Champaign, IL, USA
pnb@illinois.edu

Abstract

Uncovering algorithmic bias related to sensitive attributes is crucial. However, understanding the underlying causes of bias is even more important to ensure fairer outcomes. This study investigates bias associated with Attention Deficit Hyperactivity Disorder (ADHD) in a machine learning model predicting students' test scores. While fairness metrics did not reveal significant bias, potential subtle bias indicated by variations in model performance for students with ADHD was observed. To uncover causes of this potential bias, we correlated SHapley Additive exPlanations (SHAP) values with the model's prediction errors, identifying the features most strongly associated with increasing prediction errors. Behavioral and self-reported survey features designed to measure students' use of effective learning strategies were identified as potential causes of the model underestimating test grades for students with ADHD. Behavioral features had a stronger correlation between absolute SHAP values and prediction errors (up to $r = .354$, $p = .013$) for students with ADHD than for those without ADHD. Students with ADHD often use unique yet effective approaches to studying in online learning environments—approaches that may not be fully captured by traditional measures of typical student behaviors. These insights suggest adjusting feature design to better account for students with ADHD and mitigate bias.

CCS Concepts

• **Applied computing** → **E-learning**; • **Human-centered computing** → *Empirical studies in HCI*; • **Computing methodologies** → **Artificial intelligence**.

LAK 2025, Dublin, Ireland

ACM ISBN 979-8-4007-0701-8/25/03
https://doi.org/10.1145/3706468.3706521

Keywords

Explainable AI, Algorithmic bias, Machine Learning, Self-regulated Learning, Attention Deficit Hyperactivity Disorder

ACM Reference Format:
HaeJin Lee, Clara Belitz, Nidhi Nasiar, and Nigel Bosch. 2025. XAI Reveals the Causes of Attention Deficit Hyperactivity Disorder (ADHD) Bias in Student Performance Prediction. In *LAK25: The 15th International Learning Analytics and Knowledge Conference (LAK 2025), March 03–07, 2025, Dublin, Ireland*. ACM, New York, NY, USA, 11 pages. https://doi.org/10.1145/3706468.3706521

1 Introduction

Self-regulated Learning (SRL) is a broad category of learning activities in which students learn largely on their own, which requires students' active use of various metacognitive strategies, such as goal setting, self-reflection, and self-evaluation [26, 51, 75]. The effective use of SRL skills has been shown to positively correlate with student academic achievement [34, 54, 69]. However, students' levels of SRL skill usage and even their approaches to engaging in SRL can vary [15, 47]. These differences in SRL strategies are particularly relevant for students with Attention Deficit Hyperactivity Disorder (ADHD), a prevalent neurodevelopmental disorder characterized by hyperactivity, inattention, and impulsivity [52, 64]. Neurodiversity—which encompasses diagnoses like ADHD, autism, or dyslexia—has become a growing focus in educational inclusion, emphasizing the importance of allowing all students to leverage their unique strengths [20]. However, due to these symptoms associated with ADHD, these students may employ different learning strategies and demonstrate behavioral differences [24]. This divergence also extends to how students with ADHD interact with and use online learning platforms, which may differ from their peers without ADHD [53]. Thus, the behavioral differences that students with ADHD might exhibit are likely to be reflected in trace data, which are the digital footprints of students' interactions with learning platforms in computer-based learning environments [17].

These trace data are often used to engineer features related to learning strategies (e.g., SRL skill usage) for training predictive models. Such models include predicting students at risk of not completing courses [29], predicting student dropout [31, 38, 68], detecting affects [16, 30], and student performance prediction [2]. Therefore, if these predictive models fail to account for the unique differences in learning strategies of students with ADHD, the models may produce biased predictions for this group, which could hinder their learning experiences. For instance, if students with ADHD receive interventions—either excessive or insufficient—due to biases in the predictive model consistently over/underestimating student performance, students with ADHD may have a suboptimal learning experience that leads to decreased learning outcomes. Given that ADHD is linked to considerable educational challenges [6], investigating and addressing potential biases associated with student ADHD status within the predictive models is crucial to ensure these students do not face additional, unnecessary educational obstacles. Given this need to account for diversity in learning approaches within SRL strategies, our paper explores how the difference in learning behavior for students with and without ADHD may emerge as bias in a predictive machine learning model.

Research on algorithmic bias has primarily focused on identifying biases related to students' race [32] and gender [37], with comparatively less attention given to biases affecting students with ADHD. These biases have largely been evaluated using traditional statistical metrics, such as overall accuracy equality, statistical parity, conditional procedure accuracy equality, and conditional use accuracy equality [12, 13, 25]. Although these metrics offer a valuable quantitative assessment of biases in machine learning models, they often fail to reveal the underlying reasons and mechanisms driving these biases. Furthermore, if biases are not immediately apparent—such as those not detected by fairness metrics—they may be overlooked, despite their potential to cause the model to make biased predictions. For instance, small biases in under- and over-prediction might offset each other in aggregate metrics, masking their presence. This limitation highlights the necessity of employing multiple approaches to investigate not only the existence of potential subtle biases but also the underlying causes that may influence model outcomes. To address this need, statistical metrics can be complemented with methods that provide insights into the inner workings of machine learning models. This is where explainable artificial intelligence (XAI) methods can become invaluable. Designed to enhance the transparency and explainability of complex models [7, 28], XAI enables the interpretation and understanding of AI decision-making processes, potentially revealing sources of bias that statistical metrics alone might overlook.

Although some methods exist for understanding biases in algorithmic systems [39, 62, 70], limited research has utilized XAI methods to uncover the root causes of potential biases within predictive models. In education, for instance, XAI is commonly used to develop tools that enhance student learning by explaining to students why specific recommendations are made in computer-based learning environments [23, 36, 65]. Despite the potential of XAI to explain algorithmic bias [7], its application has largely been confined to enhancing student learning by providing explanations for personalized recommendations in educational contexts. Consequently, there remains a lack of empirical studies applying XAI to explain algorithmic bias within online learning systems or educational settings, leaving a critical gap in the practical application of these techniques.

To address this gap, our study explores whether bias exists within student performance prediction models across sensitive attributes—including ADHD status, race, and gender—by leveraging a complementary approach. Specifically, we first assess the performance of the model, which is built using features related to students' use of SRL strategies and self-reported survey data, by applying traditional fairness metrics. Recognizing that these metrics may not reveal subtle or masked biases, we further employ XAI methods, particularly SHapley Additive exPlanations (SHAP) values [43], to uncover the underlying mechanisms contributing to potential biases in the model's predictions. By combining fairness metrics with XAI techniques, we provide a comprehensive analysis that goes beyond traditional methods to identify and understand potential biases in predictive models. The results of our analysis provide actionable insights that suggest practical ways to mitigate the root causes of potential bias within the predictive model. Specifically, we discover that features designed to measure students' SRL usage may not fully capture the diverse learning strategies employed by all students, particularly those with ADHD. This finding highlights the need to refine feature engineering approaches to better reflect the varied ways students engage in SRL in computer-based learning environments. To the best of our knowledge, this is the first study to investigate and uncover the mechanisms of potential biases within predictive models concerning student ADHD status in educational contexts. Our approach not only contributes to understanding how behaviors and usage patterns associated with ADHD status may influence model predictions but also demonstrates the utility of XAI methods in identifying and explaining biases in a predictive model within educational contexts.

The paper seeks to answer the following research questions:

- **RQ1.** Does bias associated with students' ADHD status (as well as race and gender) exist in a machine learning model predicting performance grades based on student behavioral and self-reported survey data?
- **RQ2.** What causes ADHD-related biases, should they exist, in a machine learning model predicting student performance?

We anticipate the following contributions from our work:

(1) We examine whether ADHD status introduces bias to the student performance prediction model and, if so, how it does so.
(2) We present a methodological approach using SHAP values to explore whether features contribute differently to model prediction errors across sensitive attributes (i.e., race, gender, and ADHD status).

2 Related Work

2.1 Algorithmic bias in education

Algorithmic bias refers to systematic statistical differences in predictions or outcomes across groups [45]. These groups are frequently defined along the axes of legally protected identity, such as sex, race, ethnicity, or age [14, 32, 45]. These biases may stem from

errors in the machine learning model itself or reflect deeper social inequities embedded in training data [63]. If these biases are not addressed in predictive models, it could result in unequal access to resources [42], inaccurate performance evaluations [48], and misinformed interventions, which can ultimately hinder students' academic progress [40]. Therefore, recognizing the potential for bias in algorithmic systems and rigorously assessing said biases is essential for advancing equity in technology design. This type of research helps ensure that these systems uphold fairness, accuracy, and inclusivity when shaping educational outcomes. Given this necessity for evaluation, algorithmic bias has typically been measured using statistical metrics (e.g., overall accuracy equality, statistical parity, conditional procedure accuracy equality, and conditional use accuracy equality) [12, 13, 25]. These measurements offer a quantitative representation of the extent of biases, enabling researchers to compare the performance of the machine learning model in different demographics of interest. For instance, overall accuracy equality compares the overall accuracy of the model across different groups to ensure that no group is disproportionately mispredicted.

These fairness metrics are commonly used for classification models with categorical outcomes [45]. In the context of educational data, where outcomes like test scores are continuous (e.g., predicting student performance grades in this study), there is a need to adapt these metrics for regression models. Therefore, researchers have explored adapting these fairness metrics to regression settings to evaluate bias in models predicting continuous outcomes [1, 14, 41]. For example, adjusting statistical parity to consider differences in predicted means across groups or adapting accuracy measures to account for prediction errors in regression [14]. While these statistical metrics offer valuable insights into the presence of biases, they often focus on quantifying disparities without revealing the underlying reasons or mechanisms driving these biases in model predictions or decisions. Gaining such insights into how these biases manifest in the predictive models enables researchers and engineers to refine models to be fair and effective across diverse student groups, contributing to a broader understanding of bias. Without a clear understanding of these underlying factors, addressing the root causes of bias becomes particularly challenging.

In practice, numerous studies have found algorithmic bias in existing predictive models with respect to sensitive attributes such as race [18] and gender [57]. Specifically in the field of education, research has highlighted the potential for biased outcomes in predictive models for students, ranging from AI-driven technologies [70] to grade predictions [32] to year-end standardized test scores [14, 41]. These biased outcomes can have significant consequences for students, particularly those from underrepresented or marginalized groups. Moreover, such consequences not only impact individual students' educational experiences but also perpetuate systemic inequalities within the education system. These issues highlight the importance of mitigating biases in educational predictive models to prevent further entrenchment of these disparities [59, 60]. More recently, studies have expanded their focus beyond race and gender, paying attention to neurodiverse groups, such as students with ADHD [21]. Despite this progress, a significant gap remains in understanding whether predictive models exhibit bias specifically against students with ADHD, and if so, how these biases manifest. Current research still largely centers on traditional demographic categories like race and sex, leaving neurodiversity underexplored in the context of algorithmic bias. Given this need to address the gap in understanding and ensure that predictive models are equitable for all students, our work aims to investigate whether bias associated with students' ADHD status exists in machine learning models predicting performance grades based on student behavioral and self-reported survey data. By exploring potential biases and their underlying causes, we contribute to the development of fairer predictive models that more accurately reflect the diverse learning strategies of students with ADHD.

2.2 XAI for explaining algorithmic bias

Machine learning models have been employed for a variety of tasks in educational settings [19, 55]. For example, studies have used these models to predict student performance [56], identify learners at risk of not completing the course, predict dropouts in massive open online courses [68], and identify self-regulatory behaviors [5, 71] as well as affective states during learning [33]. These predictions are often used to provide adaptive content, timely scaffolds, and feedback to support learners in computer-based learning environments [66]. However, a key limitation of such recommendations from the predictive models is that students often do not understand *why* or on what basis they are receiving specific recommendations or interventions [35]. Additionally, from the perspectives of researchers and engineers, the lack of transparency of machine learning-driven recommendations limits the ability to thoroughly examine why these recommendations may be biased toward certain groups of students [8]. To address these transparency challenges, researchers have increasingly turned to XAI methods. Prior studies have explored how XAI can benefit learners [23]. The effectiveness of explanation functionality, particularly for adaptive hints, has been examined concerning user characteristics [22], as well as their impact on student success prediction models [67].

Using XAI methods to examine the decisions of models is not only beneficial to learners and teachers. It also holds promise for a wider range of applications, including checking for algorithmic biases, addressing privacy, and ensuring model transparency. Baker and Hawn [8] highlight the potential of explanation methods to make complex algorithms understandable [74], which in turn can help in identifying not only that a model is biased, but also which specific features might be contributing to it. Similarly, Khosravi et al. [36] emphasize the high interpretability of a model as a desirable feature in educational settings to mitigate any bias in the decision-making process of the model. There have been efforts to leverage XAI in domains outside of education to identify the factors causing bias in model predictions. For instance, Manresa-Yee and Ramis [44] used Local Interpretable Model-agnostic Explanations (LIME), an XAI method, to analyze the misclassification of facial images of emotions, providing insights into the important regions used by a Convolutional Neural Network (CNN) for classification. By applying LIME, Manresa-Yee and Ramis [44] were able to uncover gender-based differences in training data sets, highlighting the impact of gender bias on emotion recognition in facial expressions. Similarly, Stenwig et al. [61] used SHAP values to detect the features responsible for biased outcomes in a model predicting hospital mortality for intensive care unit patients.

While these studies demonstrate the value of XAI in identifying bias within the predictive models in image recognition and healthcare predictions, it becomes crucial to investigate how these techniques can be applied in educational settings, where the stakes and challenges are different. In education, predictive models are often built using trace data, which capture the complex and temporal aspects of learning behaviors [17]. The unique nature of trace data introduces additional complexities in model interpretation and the potential for bias, as these data reflect nuanced student learning strategies and behaviors. Given the limited empirical studies that apply XAI methods to explain bias in the context of online learning systems, there is a pressing need to explore how XAI can be leveraged to explain bias in educational contexts. In response, this study aims to fill this gap by demonstrating an approach that leverages XAI methods—specifically SHAP values—to identify and analyze potential sources of bias in predictive models used to predict student learning outcomes.

3 Research Context and Data

In this section, we discuss the online learning platform used for data collection, as well as details about the sensitive attributes of the study participants.

3.1 Self-guided online learning platform

We developed a self-guided online learning system designed for students to study four different subtopics related to introductory statistics. Students could study four different subtopics along with learning activities (i.e., Reading, Quiz, Examples, and Summary) that students were allowed to study at their own pace. Each subtopic covered different content. The first subtopic (*What is Data?*) included key statistical concepts, including descriptive and inferential statistics, the distinction between samples and populations, the concept of margin of error, and the categorization of data types (i.e., categorical or quantitative). The second subtopic (*Exploring Data with Graphs*) focused on interpreting various graphs, such as histograms, and identifying their distributions. The third subtopic (*Understanding Data with Numerical Summaries*) covered calculating central measures like mean and mode, as well as measures of dispersion, such as variance and standard deviation. The last subtopic (*Analyzing Data with Two Variables*) included contents related to response and explanatory variables, confounding variables, and associations.

Every activity was designed to serve a specific learning purpose. The reading activity, typically spanning four to six pages per subtopic, offered in-depth information on the subject matter. The quiz included approximately 10 questions per subtopic, allowing students to assess their understanding without time limits and as often as they wanted. After completing a quiz, students were informed whether their answers were correct or incorrect; however, the correct answers for any incorrect responses were not disclosed. Thus, students had to independently seek out the correct information by identifying knowledge gaps and taking proactive steps to acquire the correct answers. The examples presented students with sample questions, offering step-by-step guidance along with the correct answers to approach and solve the problems. The summaries provided a concise overview of each subtopic's key concepts, allowing students to quickly revisit and review the material for each subtopic. Prior to participating in the study, students completed a consent form approved by the Institutional Review Board (IRB protocol #21019). The study began with a demographic survey, which included questions about sensitive attributes, individual differences, and other sensitive attributes, including race (i.e., "What is your race/ethnicity? Please enter NA if you prefer not to answer"), gender (i.e., "What is your gender? Please enter NA if you prefer not to answer"), and ADHD diagnosis (i.e., "Have you ever been diagnosed with attention deficit disorder or attention deficit hyperactivity disorder (ADD or ADHD)? Please enter NA if you prefer not to answer"). Participants provided responses in open-ended text boxes and had the option not to answer any questions related to sensitive attributes if they preferred.

Additionally, we asked students to take the self-reported Online Self-Regulated Learning Questionnaire (OSLQ) survey, which was developed to measure students' self-perceived assessment of SRL skills in computer-based learning environments [9–11]. OSLQ is a 24-item scale comprising six subscales: environment structuring, goal setting, time management, help-seeking, task strategies, and self-evaluation, designed to measure students' SRL behaviors in online learning settings. Each subscale is intended to assess different aspects of students' use of SRL skills. After completing the OSLQ survey, students took a pretest (i.e., a test before learning begins) designed to measure their prior knowledge of the learning materials students would encounter in the upcoming learning session. Following the pretest, students engaged in a 60-minute, self-paced learning session. Students had the flexibility to navigate the learning activities in any order, regardless of the subtopic. While not required to complete all subtopics during the session, the learning system allowed students to revisit and repeat any activity as often as they wished. After the 60 minutes of self-paced learning, students took a posttest (i.e., a test after learning, identical in structure to the pretest but with different question variants), which assessed their understanding of the four subtopics covered during the learning session.

3.2 Data Collection and Participants

We collected behavioral data and survey responses from 277 college students, recruited through two different sampling methods. The first group consisted of 112 students from a public research university in the Midwest region of the United States. For the second group, we recruited 165 students from various U.S. colleges and universities using Prolific, an online crowdsourcing platform designed to help researchers recruit a diverse sample of participants [50]. We limited the participant eligibility to undergraduate students attending either 2-year or 4-year community colleges or universities. Upon completing the study, students in the first group received study participation credit for their enrolled courses, while Prolific participants were each compensated $15 USD. We categorized students' open-ended responses regarding sensitive attributes into groups based on their answers. For race/ethnicity, 50.2% of participants identified as White, 19.1% as Asian, 13.4% as Black, 10.8% as Latinx/Hispanic, 2.5% as Mixed, and 4.0% as Others. For gender, 57.8% of participants identified as female, 36.1% as male, and 6.1% as additional genders (grouped for anonymity). Regarding ADHD status, 78.7% of students reported that they do not have ADHD,

17.3% reported being diagnosed with ADHD, and 4.0% preferred not to answer.

4 Method

4.1 Student performance prediction

We used log data from 277 students interacting with a self-guided online learning system, capturing their real-time activities. These log data contained activities (i.e., reading, quiz, examples, and summary), activity durations, and quiz results recorded during the students' interactions with every stage of the software. To train a model predicting students' posttest grades, we extracted 12 features, including OSLQ responses, behavioral SRL measurements, quiz grades for each subtopic, and pretest grades. We had six OSLQ-related features, where each feature represented the average response for the corresponding subscale: goal setting, environment structuring, task strategy, time management, help-seeking, and self-evaluation. Each Likert-type question response ranged from 1 to 5, representing "strongly disagree" (1) to "strongly agree" (5). Thus, a high value on the OSLQ feature implies that students rated themselves as having a higher level of corresponding SRL skills.

We applied Coherence Analysis (CA), a theory-driven learning analytics approach, to engineer behavioral SRL features [58]. CA is an approach to measure students' use of SRL strategies in online learning environments, specifically metacognitive strategy, one of the common facets of SRL skills. CA measures how "coherent" students' engagement in learning activities is using the order and timing of learning activities. Coherent actions represent the active use of metacognitive strategy since students' modulation of their current action (e.g., reviewing the reading material) involves them evaluating information gained from the previous action (e.g., recognizing incorrect answers to quiz questions) [72, 73]. Using CA, we measured two types of behaviors: CA reading and CA quiz. CA reading occurs after a quiz, where students review material related to missed questions within a 5-minute window, demonstrating a reactive approach to addressing knowledge gaps. We measured this by summing the time spent on these reviews. CA quiz happens before a quiz, where students strategically engage with reading materials, examples, and summaries to prepare within a 5-minute window. This behavior reflects a strategic approach to acquiring and assessing knowledge. We measured it by calculating the total time students spent on reading activities. Pretest grades were determined by evaluating student performance on an initial pretest at the start of the study. Average quiz grades were calculated by aggregating the scores from quizzes on each subtopic. We used 5-fold cross-validation to fit a random forest regression model predicting students' posttest grades. The random forest model achieved R^2 of .481 which was calculated per testing fold and then averaged, and all features were included in training.

4.2 Fairness metrics

To measure algorithmic bias, we adapted the work by Belitz et al. [14], who modified four commonly used classification metrics (i.e., overall accuracy equality, statistical parity, conditional procedure accuracy equality, and conditional use accuracy equality) for regression tasks, as opposed to the more common classification bias metrics because our study involves continuous predictions. We provide definitions of four metrics below:

- **Overall Accuracy Equality (OAE)**: We quantified OAE using the root mean squared error (RMSE) to compare the actual values with the predicted values for the groups with specific sensitive attributes. A higher OAE value indicates a larger prediction error for the interest group(s) compared to the control group. OAE makes no distinction between negative and positive errors. That is to say, over and under predictions are weighed equally.
- **Statistical Parity (SP)**: We quantified SP using the mean of predicted values for each group. This metric examines whether groups have similar predicted outcomes, on average. SP aims to ensure the predicted mean values are consistent across each group of interest. SP helps demonstrate whether the model predicts systematically higher or lower values for different groups.
- **Conditional Procedure Accuracy Equality (CPA)**: We quantified CPA in the same way as OAE, with additional conditions. CPA is conditioned on the "true" values, separated by values where the real-world outcome is above or below a specific threshold. These errors are then compared across groups, as with OAE. CPA aims to examine a machine learning model's predictive accuracy for scores above and below the threshold. Choosing a specific value is required for the threshold, in order to condition on the continuous outcome.
- **Conditional Use Accuracy Equality (CUA)**: We quantified CUA in the same way as CPA, but conditioned on predicted values. CUA conditions on the model-predicted outcomes, rather than the real-world (i.e., ground truth) values. Thresholding is done in the same way as CPA, by choosing a specific value.

We set the threshold value for assessing CPA and CUA at the median posttest grade of 66.7%. This approach allowed us to examine whether there was any bias based on whether students perform worse (lower performers) or better (higher performers) than the median. We excluded the students who preferred not to answer questions about sensitive attributes from the analysis since we could not categorize those students into groups related to sensitive attributes. In total, 10 students did not respond to the ADHD question, 3 to the race question, and 7 to the gender question. We ran additional regression analyses to examine whether we observe statistical significance with the fairness metrics. We fit linear regression models using students' sensitive attributes, such as gender, race, and ADHD status, as predictors and used fairness metrics as response variables. However, we calculated fairness at the individual student level instead of the group-level measurements discussed in the previous subsection to run the regression models, since the regressions use student-level data. For OAE, we computed the model prediction error by calculating the difference between each student's actual posttest grades and the model's predicted posttest grades.

4.3 Correlating SHAP values with model errors

We used SHAP [43] as a post-hoc XAI method to gain fine-grained insights into how biases manifest within the model's performance with respect to sensitive variables. SHAP is based on Shapley values, which provide a way to determine the contribution of each feature to a prediction made by a model. SHAP values quantify how each feature propels the model's prediction away from the average prediction—positively or negatively—and highlight the feature's relative significance within the model.We used SHAP values to investigate the causes that lead to algorithmic bias by correlating SHAP values with model prediction errors. Each feature received an individual SHAP value for each student-level prediction. This correlational analysis could help to examine the underlying causes of bias within the model, thereby extending the application of SHAP values from merely assessing feature importance to the examination of algorithmic fairness. By analyzing the correlation between each feature's SHAP values and the model prediction errors—calculated as the discrepancy between the model's predicted values and the actual values—we can identify which features disproportionately influence the model prediction accuracy across different sensitive attributes.

We correlated SHAP values and the prediction errors in two ways: the first approach used the absolute values of both SHAP values and prediction errors, while the second approach did not account for absolute values. Analyzing absolute SHAP values allows us to measure the strength of each feature's importance regardless of its positive or negative effect on the model prediction accuracy. Conversely, analyzing the non-absolute SHAP value will allow us to gain a more fine-grained understanding of how SHAP values (either positive or negative) relate to machine learning model underpredicting (i.e., predicting lower than student's true posttest grade) or overpredicting. To examine how the causes of the model error differed for students with ADHD, we conducted the analysis for all students, students with ADHD, and students without ADHD. We calculated SHAP values for the features we used for training the random forest models. Before correlating SHAP values with the model errors, we examined the distributions by plotting box plots for each feature SHAP values and prediction errors to determine whether we need to use Pearson r or Spearman *rho* to calculate correlations. We used the Spearman *rho* for features that were not normally distributed and Pearson r for those that were.

5 Results

5.1 Bias associated with ADHD status in student performance prediction

When comparing students' actual posttest grades with those predicted by the machine learning model across sensitive attributes, we found no significant differences, indicating limited evidence of bias in the overall means. For example, regarding ADHD status, students with ADHD had an average actual posttest score of 68.6% (SD = 23.4%), while students without ADHD scored similarly at 68.2% (SD = 22.3%). The model predicted an average score of 66.8% (SD = 15.7%) for students with ADHD and 68.0% (SD = 16.4%) for students without ADHD. Although the model may have slightly under-predicted grades for students with ADHD, we did not observe a statistically significant difference between the actual and predicted grades, suggesting the model performs similarly *on average* for both students with and without ADHD—although, as explored in RQ2 below, the average may obscure important variability at a more fine-grained level. We further examined the potential presence of bias within the model across sensitive attributes using four types of fairness metrics: OAE, SP, CPA, and CUA, as detailed in Tables 1 (see Section 4.2 for interpretation of these metrics and the choice of 66.7%). While we did not observe consistently better or worse prediction accuracy across sensitive attributes, consistent variations in model performance suggest the possibility of subtle, systematic bias within the predictive model. Likewise, fairness metrics alone cannot confirm significant bias between students with and without ADHD, these metrics do provide insight into the model's behavior across sensitive attributes. As shown in Table 1, variations in OAE, SP, CPA, and CUA between ADHD groups indicate potential performance disparities that merit further investigation. Furthermore, the absence of substantial bias in the overall model does not imply that all features contribute equally to prediction errors. This observation necessitates a deeper analysis to identify which specific features might be influencing the model's performance across different groups, a question we address in RQ2.

5.2 RQ2. Causes of ADHD-related bias in student performance prediction

In this paper, when we refer to causality [49], we specifically address the causal relationships within the machine learning model used for this study, rather than making generalizable causal claims applicable outside of this model. We observed notable differences in the correlations between absolute SHAP values and model prediction performance errors across the entire sample of students, as well as within subgroups of students with and without ADHD, as illustrated in Figure 1. For the entire group of students, analyzing feature-specific correlations between absolute SHAP values and model predictions provides insight into how the importance of each feature influences the model's overall prediction accuracy. Across all students, features such as *goal setting* (r = .237, $p < .001$), *task strategy* (r = .135, p = .025), *help-seeking* (r = .212, $p < .001$), and *CA reading* (r = .219, $p < .001$) were significantly positively correlated with model prediction error (Figure 1). These positive correlations suggest that these features had a greater impact on model predictions (i.e., higher absolute SHAP values) and were associated with larger errors in model performance (i.e., higher absolute prediction errors).

We further examined these correlations based on student ADHD status to explore how bias manifests by identifying whether different features contribute more to model prediction errors for students with and without ADHD. Interestingly, when comparing correlation values between these groups, we observed notable differences in both the direction and magnitude of correlations of some of the features. For example, the CA reading feature showed weak yet statistically significant correlations for both students with ADHD (r = .354, p = .013) and those without ADHD (r = .175, p = .014), but with different magnitudes. This differing magnitude suggests that the CA reading feature had a stronger association with model errors for students with ADHD than for those without. For one

Table 1: Summary of fairness metrics across student race/ethnicity, gender, and ADHD status, including Overall Absolute Error (OAE), Statistical Parity (SP), and Conditional Performance Accuracy (CPA) and Conditional Underestimation Accuracy (CUA) for performance groups above and below the 66.7% threshold. These metrics provide insight into the model's fairness in predicting posttest grades across race/ethnicity, gender, and ADHD status.

Group	OAE	SP	CPA (≥ 66.7%)	CPA (< 66.7%)	CUA (≥ 66.7%)	CUA (< 66.7%)
Race/Ethnicity						
White	12.899	68.725	11.721	13.993	10.890	14.808
Black	13.078	55.971	9.894	14.076	11.317	13.673
Hispanic	10.292	67.183	11.274	9.677	11.021	9.506
Asian	12.181	73.417	9.265	16.757	8.569	17.150
Gender						
Female	13.244	67.638	11.370	14.772	10.621	15.479
Male	11.661	66.780	11.185	12.100	10.451	12.937
ADHD Status						
Students with ADHD	14.187	66.828	12.306	15.846	9.790	16.630
Students without ADHD	11.773	68.026	10.477	12.900	10.284	13.238

of the quiz grade features (i.e., Exploring Data with Graphs), we found a significant positive correlation for students with ADHD (r = .286, p = .049), whereas no significant correlation was observed for students without ADHD (r = -.027, p = .693). This disparity indicates that the importance of the graph interpretation quiz grade was associated with increased model error for students with ADHD, while it had minimal or no association for students without ADHD.

From **RQ1**, we did not observe a significant presence of bias in the predictive model across sensitive attributes, though smaller biases in different directions may have offset each other in the overall mean. To further examine potential underlying biases, we analyzed correlations between non-absolute SHAP values and the model prediction errors. Negative correlations between non-absolute SHAP values and model errors were found across most features, with the exception of the pretest grade feature. A negative correlation suggests that when features contribute positively (i.e., have positive SHAP values), the model tends to underestimate the true grades (leading to negative errors). Conversely, when features contribute negatively (i.e., have negative SHAP values), the model tends to overestimate the true grades (leading to positive errors). Notably, the magnitudes of these correlations varied for features with negative correlations. We observed that *CA reading* had the most negative correlation value among all other features (r = -.557, $p < .001$) followed by the *task strategy* OSLQ feature (r = -.449, p = .001). Similarly, *CA quiz* (r = -.433, p = .002) and OSLQ *environment structuring* (r = -.429, p = .002) had significant negative correlations. Given that the distributions of SHAP values for CA and OSLQ-related features are highly left-skewed (i.e., most SHAP values are positive) for students with ADHD, we can infer that these features may lead to possible underprediction in the model in some cases.

6 Discussion

We answered two research questions using data collected from 277 college students engaged in a computer-based learning environment for introductory statistics. At a high level, we found:

- **RQ1.** While fairness metrics did not reveal significant bias within our predictive model across sensitive attributes, we observed potential subtle bias indicated by variations in model performance for students with ADHD.
- **RQ2.** Features we engineered to measure students' use of SRL skills (i.e., CA reading and CA quiz) and self-reported SRL were the features that tended to increase the prediction error in the machine learning model. Further, we observed that these features (when having positive SHAP values) contributed underpredictions of the posttest grade for students with ADHD diagnosis.

6.1 RQ1. Bias associated with ADHD status in student performance prediction

We observed limited evidence of bias in the posttest grade prediction model based on traditional fairness metrics. However, potential variations in model performance across ADHD status suggest the presence of subtle biases that may not be captured by these metrics. This observation implies that ADHD status, as a sensitive attribute, warrants careful consideration in model development and validation processes to ensure fairness and accuracy in predictive models used in numerous learning settings. Given that predictive models are widely used to make or inform decisions from student placement to intervention strategies [4, 27], predictive models underestimating the performance of students with ADHD could lead to these students not receiving effective interventions or support they might otherwise benefit from based on their actual performance. For instance, if students with ADHD are consistently underpredicted by these models, they may be inappropriately placed in remedial courses or given extra interventions they do not need. Such misplacements could lead to boredom and disengagement, potentially resulting in disruptive behaviors and, ultimately, poorer educational outcomes. Therefore, it is crucial for researchers to recognize the potential for subtle biases affecting students with ADHD. We further highlight the critical need for future research to specifically address and evaluate algorithmic fairness in the context of ADHD status. This involves developing and applying rigorous methodologies to assess whether the machine learning models exhibit biases that differentially affect students with ADHD. XAI approaches can also provide insights into underlying causes, enabling the development of fairer models that better support diverse student needs.

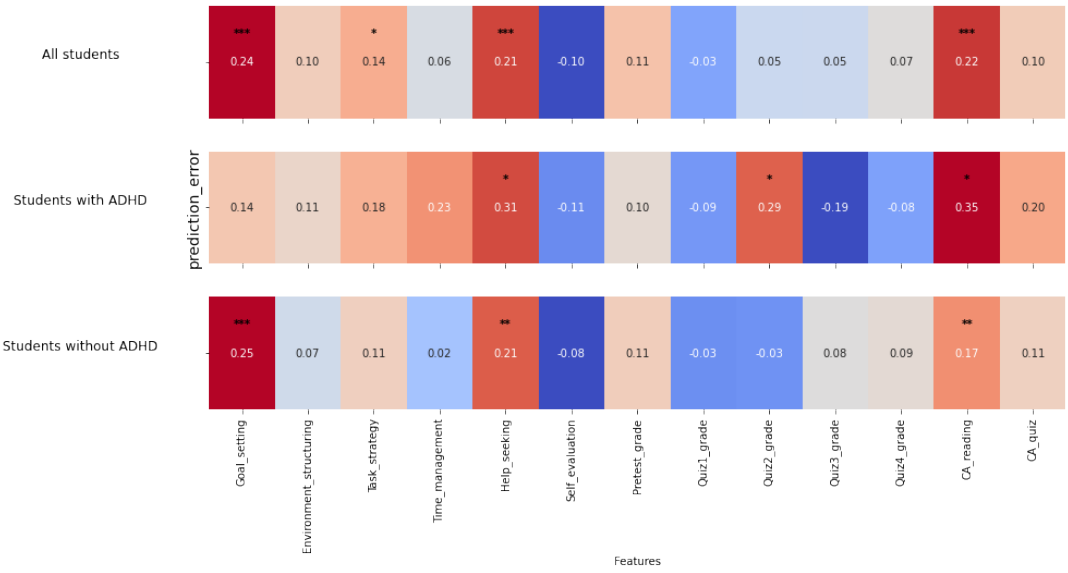

Figure 1: Heat map of correlations between absolute SHAP and prediction errors in the ML model.

6.2 RQ2. Causes of ADHD-related bias in student performance prediction

Our analysis revealed discrepancies in the correlations between absolute SHAP values and model prediction errors for all students and subgroups based on ADHD diagnosis. Specifically, CA reading—a feature engineered to measure students' use of SRL skills—had the highest correlation relates to the absolute value of the correlation between SHAP values and model prediction error; specifically, CA reading showed the highest correlation for the ADHD group compared to the all-students and non-ADHD groups. The observed higher mean CA reading and CA quiz values for students without ADHD (CA reading: 2.79 minutes, CA quiz: 9.69 minutes) compared to students with ADHD (CA reading: 1.62 minutes, CA quiz: 7.05 minutes) may reflect distinct patterns in how students with ADHD engage in learning activities. These differences could be influenced by challenges such as sustained attention or planning. While CA measures are designed to capture typical coherent learning behaviors, students with ADHD may employ alternative strategies or exhibit more variability in their engagement, leading to lower CA values overall. In sum, these findings suggest that an approach to measuring CA reading, which is assessing students' use of SRL skills, might be ineffective for students with ADHD especially, which might cause an increase in the prediction error in the predictive model.

Similarly, with the exception of goal-setting feature, other OSLQ-related features demonstrated higher correlations with prediction error for students diagnosed with ADHD compared to those without. This result further indicates that the OSLQ survey questions may be less effective at capturing SRL usage in students with ADHD, especially when these features are used to predict students' posttest grades. In sum, these observations highlight a potential mismatch between the SRL behaviors of students with ADHD and the way SRL is currently measured through behavioral data and self-reported surveys.

When we examined the correlations between non-absolute SHAP values and the model prediction errors, we further substantiated that CA features (i.e., CA reading and CA quiz) and OSLQ-related features were more strongly associated with the model underpredicting (when SHAP values are positive) and overpredicting (when SHAP values are negative). This pattern further helps us understand the potential causes of the varying performance of the predictive model across students with and without ADHD, as identified in **RQ1**. These findings from **RQ2** suggest that adjusting or optimizing how CA and OSLQ-related features are measured could be a way to mitigate bias in the posttest grade prediction model. Specifically, there is a need to refine feature engineering processes to better capture the unique SRL strategies employed by students with ADHD. Students with ADHD may adopt different SRL strategies compared to students without ADHD in computer-based learning environments. This raises important questions for future research. For instance, how do the SRL behaviors of students with ADHD differ from their peers, and how can we develop measurement instruments that more accurately reflect these differences? Investigating these questions could lead to the development of more effective tools for assessing SRL in students with ADHD, ultimately improving the accuracy and fairness of predictive models in educational settings. Moreover, our XAI approach, which focuses on identifying and addressing biases in predictive models for students with ADHD, provides a foundation that could be extended to other groups and educational contexts. For example, applying similar approaches to students from diverse backgrounds could reveal unique challenges and guide the development of fairer, more inclusive predictive models.

7 Limitations

Although we observed differences in correlations for specific features, a fundamental limitation of our study is the sample size of students with ADHD. Since ADHD affects only a fraction of the student population, obtaining a large enough sample to include a significant number of students with ADHD can be challenging. Furthermore, although we observed differences in correlations for specific features, our analysis lacks qualitative data, such as interviews, which are crucial for gaining a comprehensive understanding of how students with ADHD perceive the topic difficulties and CA measures (i.e., CA reading and CA quiz) which we engineered to assess the use of SRL skills and strategies. Such qualitative insights would provide a more nuanced understanding of how and why specific features were more associated with increasing model performance for students with ADHD compared to those without. Future work should thus consider expanding sample sizes (perhaps to thousands of students) and including more qualitative data collection to deeply understand the experiences of students with ADHD as they use machine learning-driven educational software. Moreover, our SHAP-based approach to understanding causality uncovers how features influence model predictions, which is causal with respect to the model. However, given the existence of many possible models, exploring this further could be an avenue for future research.

8 Conclusion

Using XAI to explain the underlying causes of algorithmic bias is crucial for mitigating such biases, as it can offer actionable insights that could facilitate more effective ML model refinements, such as re-engineering features. This is particularly critical in educational contexts, where comprehending the mechanisms of algorithmic unfairness related to sensitive attributes, such as ADHD, is essential to ensuring equitable outcomes for all students. Our contribution of correlating SHAP values with the machine learning prediction models allows for pinpointing the features that lead to increasing errors in the prediction model and also model under/overpredicting. Future research could explore additional XAI methods, such as counterfactual explanations [3, 46], to compare how various techniques provide unique insights into explaining the causes of algorithmic bias. Furthermore, we highlight expanding future research to investigate algorithmic biases across a broader range of sensitive attributes. Exploring whether ensuring the fairness of machine learning models across diverse groups of students can enhance learning outcomes and experiences could be another avenue for future research.

References

[1] Alekh Agarwal, Miroslav Dudik, and Zhiwei Steven Wu. 2019. Fair regression: Quantitative definitions and reduction-based algorithms. In *Proceedings of the 36th International Conference on Machine Learning*. PMLR, Long Beach, California, 120–129. https://proceedings.mlr.press/v97/agarwal19d.html

[2] Huda Al-Shehri, Amani Al-Qarni, Leena Al-Saati, Arwa Batoaq, Haifa Badukhen, Saleh Alrashed, Jamal Alhiyafi, and Sunday O. Olatunji. 2017. Student performance prediction using support vector machine and k-nearest neighbor. In *2017 IEEE 30th Canadian Conference on Electrical and Computer Engineering (CCECE)*. IEEE, IEEE, Windsor, ON, 1–4.

[3] Emanuele Albini, Jason Long, Danial Dervovic, and Daniele Magazzeni. 2022. Counterfactual Shapley additive explanations. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*. ACM, Seoul, Korea, 1054–1070.

[4] Eyman Alyahyan and Dilek Düştegör. 2020. Predicting academic success in higher education: Literature review and best practices. *International Journal of Educational Technology in Higher Education* 17, 1 (2020), 3.

[5] Eric Araka, Robert Oboko, Elizaphan Maina, and Rhoda Gitonga. 2022. Using educational data mining techniques to identify profiles in self-regulated learning: An empirical evaluation. *The International Review of Research in Open and Distributed Learning* 23, 1 (2022), 131–162.

[6] L. Eugene Arnold, Paul Hodgkins, Jennifer Kahle, Manisha Madhoo, and Geoff Kewley. 2020. Long-term outcomes of ADHD: Academic achievement and performance. *Journal of Attention Disorders* 24, 1 (2020), 73–85.

[7] Alejandro Barredo Arrieta, Natalia Díaz-Rodríguez, Javier Del Ser, Adrien Bennetot, Siham Tabik, Alberto Barbado, Salvador García, Sergio Gil-López, Daniel Molina, Richard Benjamins, et al. 2020. Explainable artificial intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. *Information Fusion* 58 (2020), 82–115.

[8] Ryan S. Baker and Aaron Hawn. 2022. Algorithmic bias in education. *International Journal of Artificial Intelligence in Education* 32, 4 (Dec. 2022), 1052–1092. https://doi.org/10.1007/s40593-021-00285-9

[9] Lucy Barnard, William Y. Lan, Yen M. To, Valerie Osland Paton, and Shu-Ling Lai. 2009. Measuring self-regulation in online and blended learning environments. *The Internet and Higher Education* 12, 1 (2009), 1–6.

[10] Lucy Barnard, Valerie Paton, and William Lan. 2008. Online self-regulatory learning behaviors as a mediator in the relationship between online course perceptions with achievement. *The International Review of Research in Open and Distributed Learning* 9, 2 (June 2008), 11. https://doi.org/10.19173/irrodl.v9i2.516

[11] Lucy Barnard-Brak, Valerie Osland Paton, and William Y Lan. 2010. Profiles in self-regulated learning in the online learning environment. *International Review of Research in Open and Distributed Learning* 11, 1 (2010), 61–80.

[12] Solon Barocas, Moritz Hardt, and Arvind Narayanan. 2023. *Fairness and Machine Learning: Limitations and Opportunities*. MIT Press, Cambridge, MA.

[13] Clara Belitz, Lan Jiang, and Nigel Bosch. 2021. Automating procedurally fair feature selection in machine learning. In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*. ACM, Virtual Event USA, 379–389. https://doi.org/10.1145/3461702.3462585

[14] Clara Belitz, HaeJin Lee, Nidhi Nasiar, Stephen E. Fancsali, Steve Ritter, Husni Almoubayyed, Ryan S. Baker, Jaclyn Ocumpaugh, and Nigel Bosch. 2024. Hierarchical dependencies in classroom settings influence algorithmic bias metrics. In *Proceedings of the 14th Learning Analytics and Knowledge Conference*. ACM, Kyoto Japan, 210–218. https://doi.org/10.1145/3636555.3636869

[15] Matthew L. Bernacki, Timothy J. Nokes-Malach, and Vincent Aleven. 2015. Examining self-efficacy during learning: Variability and relations to behavior, performance, and learning. *Metacognition and Learning* 10 (2015), 99–117.

[16] Nigel Bosch, Sidney D'Mello, Ryan Baker, Jaclyn Ocumpaugh, Valerie Shute, Matthew Ventura, Lubin Wang, and Weinan Zhao. 2015. Automatic detection of learning-centered affective states in the wild. In *Proceedings of the 20th International Conference on Intelligent User Interfaces*. ACM, Atlanta, Georgia USA, 379–388. https://doi.org/10.1145/2678025.2701397

[17] Peter Brusilovsky. 2001. Adaptive hypermedia. *User Modeling and User-Adapted Interaction* 11 (2001), 87–110.

[18] Joy Buolamwini and Timnit Gebru. 2018. Gender shades: Intersectional accuracy disparities in commercial gender classification. In *Proceedings of the 1st Conference on Fairness, Accountability and Transparency*. PMLR, New York, USA, 77–91. ISSN: 2640-3498.

[19] Lijia Chen, Pingping Chen, and Zhijian Lin. 2020. Artificial intelligence in education: A review. *Ieee Access* 8 (2020), 75264–75278.

[20] Maria Chrysochoou, Arash E. Zaghi, and Connie Mosher Syharat. 2022. Reframing neurodiversity in engineering education. *Frontiers in Education* 7 (Nov. 2022), 995865. https://doi.org/10.3389/feduc.2022.995865

[21] Lynn Clouder, Mehmet Karakus, Alessia Cinotti, María Virginia Ferreyra, Genoveva Amador Fierros, and Patricia Rojo. 2020. Neurodiversity in higher education: A narrative synthesis. *Higher Education* 80, 4 (Oct. 2020), 757–778. https://doi.org/10.1007/s10734-020-00513-6

[22] Cristina Conati, Oswald Barral, Vanessa Putnam, and Lea Rieger. 2021. Toward personalized XAI: A case study in intelligent tutoring systems. *Artificial Intelligence* 298 (2021), 103503.

[23] Cristina Conati, Kaska Porayska-Pomsta, and Manolis Mavrikis. 2018. AI in education needs interpretable machine learning: Lessons from Open Learner Modelling. *arXiv preprint arXiv:1807.00154* (2018), 21–27.

[24] David Daley and James Birchwood. 2010. ADHD and academic performance: Why does ADHD impact on academic performance and what can be done to support ADHD children in the classroom? *Child: Care, Health and Development* 36, 4 (2010), 455–464.

[25] Cynthia Dwork, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Richard Zemel. 2012. Fairness through awareness. In *Proceedings of the 3rd Innovations in Theoretical Computer Science Conference*. ACM, Cambridge Massachusetts, 214–226. https://doi.org/10.1145/2090236.2090255

[26] Anastasia Efklides. 2011. Interactions of metacognition with motivation and affect in self-regulated learning: The MASRL model. *Educational psychologist* 46, 1 (2011), 6–25.

[27] Manuela Ekowo and Iris Palmer. 2017. *Predictive analytics in higher education.* Technical Report. New America.

[28] Riccardo Guidotti, Anna Monreale, Salvatore Ruggieri, Franco Turini, Fosca Giannotti, and Dino Pedreschi. 2018. A survey of methods for explaining black box models. *ACM computing surveys (CSUR)* 51, 5 (2018), 1–42.

[29] Jiazhen He, James Bailey, Benjamin Rubinstein, and Rui Zhang. 2015. Identifying at-risk students in massive open online courses. In *Proceedings of the AAAI Conference on Artificial Intelligence.* AAAI Press, Austin, Texas, 1749–1755.

[30] Stephen Hutt, Joseph F. Grafsgaard, and Sidney K. D'Mello. 2019. Time to scale: Generalizable affect detection for tens of thousands of students across an entire school year. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems.* ACM, Glasgow Scotland Uk, 1–14. https://doi.org/10.1145/3290605.3300726

[31] Ali Shariq Imran, Fisnik Dalipi, and Zenun Kastrati. 2019. Predicting student dropout in a MOOC: An evaluation of a deep neural network model. In *Proceedings of the 2019 5th International Conference on Computing and Artificial Intelligence.* ACM, Bali Indonesia, 190–195.

[32] Weijie Jiang and Zachary A. Pardos. 2021. Towards equity and algorithmic fairness in student grade prediction. In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society (AIES '21).* Association for Computing Machinery, New York, NY, USA, 608–617. https://doi.org/10.1145/3461702.3462623

[33] Yang Jiang, Nigel Bosch, Ryan S. Baker, Luc Paquette, Jaclyn Ocumpaugh, Juliana Ma Alexandra L. Andres, Allison L. Moore, and Gautam Biswas. 2018. Expert feature-engineering vs. deep neural networks: Which is better for sensor-free affect detection?. In *Artificial Intelligence in Education: 19th International Conference, AIED 2018, London, UK, June 27–30, 2018, Proceedings, Part I 19.* Springer, London, 198–211.

[34] Sung-Hee Jin, Kowoon Im, Mina Yoo, Ido Roll, and Kyoungwon Seo. 2023. Supporting students' self-regulated learning in online learning using artificial intelligence applications. *International Journal of Educational Technology in Higher Education* 20, 1 (2023), 37.

[35] Judy Kay. 2023. Foundations for human-AI teaming for self-regulated learning with explainable AI (XAI). *Computers in Human Behavior* 147 (2023), 107848.

[36] Hassan Khosravi, Simon Buckingham Shum, Guanliang Chen, Cristina Conati, Yi-Shan Tsai, Judy Kay, Simon Knight, Roberto Martinez-Maldonado, Shazia Sadiq, and Dragan Gašević. 2022. Explainable artificial intelligence in education. *Computers and Education: Artificial Intelligence* 3 (2022), 100074.

[37] Florian Klapproth, Lucas Holzhüter, and Tanja Jungmann. 2022. Prediction of students' reading outcomes in learning progress monitoring. Evidence for the effect of a gender bias. *Journal for educational research online* 14, 1 (2022), 16–38.

[38] Marius Kloft, Felix Stiehler, Zhilin Zheng, and Niels Pinkwart. 2014. Predicting MOOC dropout over weeks using machine learning methods. In *Proceedings of the EMNLP 2014 workshop on analysis of large scale social interaction in MOOCs.* Association for Computational Linguistics, Doha, Qatar, 60–65.

[39] Binh Le, Damiano Spina, Falk Scholer, and Hui Chia. 2022. A crowdsourcing methodology to measure algorithmic bias in black-box systems: A case study with COVID-related searches. In *International Workshop on Algorithmic Bias in Search and Recommendation.* Springer, Norway, 43–55.

[40] Lin Li, Namrata Srivastava, Jia Rong, Gina Pianta, Raju Varanasi, Dragan Gašević, and Guanliang Chen. 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.* ACM, Kyoto, Japan, 608–619.

[41] Lin Li, Namrata Srivastava, Jia Rong, Gina Pianta, Raju Varanasi, Dragan Gašević, and Guanliang Chen. 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. https://doi.org/10.1145/3636555.3636920

[42] Lydia T. Liu, Sarah Dean, Esther Rolf, Max Simchowitz, and Moritz Hardt. 2018. Delayed impact of fair machine learning. In *International Conference on Machine Learning.* PMLR, Sweden, 3150–3158.

[43] Scott M. Lundberg and Su-In Lee. 2017. A unified approach to interpreting model predictions. In *Proceedings of the 31st International Conference on Neural Information Processing Systems.* ACM, Long Beach, CA, 4768–4777. https://papers.nips.cc/paper_files/paper/2017/hash/8a20a8621978632d76c43dfd28b67767-Abstract.html

[44] Cristina Manresa-Yee and Silvia Ramis. 2021. Assessing gender bias in predictive algorithms using eXplainable AI. In *Proceedings of the XXI International Conference on Human Computer Interaction.* ACM, Málaga Spain, 1–8. https://doi.org/10.1145/3471391.3471420

[45] Ninareh Mehrabi, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, and Aram Galstyan. 2021. A survey on bias and fairness in machine learning. *Comput. Surveys* 54, 6 (July 2021), 115:1–115:35. https://doi.org/10.1145/3457607

[46] Ramaravind K. Mothilal, Amit Sharma, and Chenhao Tan. 2020. Explaining machine learning classifiers through diverse counterfactual explanations. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency (FAT* '20).* Association for Computing Machinery, New York, NY, USA, 607–617. https://doi.org/10.1145/3351095.3372850

[47] Charles Magoba Muwonge, Joseph Ssenyonga, Henry Kibedi, and Ulrich Schiefele. 2020. Use of self-regulated learning strategies Among Teacher Education students: A latent profile analysis. *Social Sciences & Humanities Open* 2, 1 (2020), 100037.

[48] Zachary A Pardos, Qing Yang Wang, and Shubhendu Trivedi. 2012. The real world significance of performance prediction. In *Proceedings of the 5th International Conference on Educational Data Mining.* International Educational Data Mining Society, Greece, 192–195.

[49] Judea Pearl. 2009. *Causality* (2nd ed.). Cambridge University Press, NY.

[50] Eyal Peer, David Rothschild, Andrew Gordon, Zak Evernden, and Ekaterina Damer. 2021. Data quality of platforms and panels for online behavioral research. *Behavior Research Methods* 54, 4 (Sept. 2021), 1643–1662. https://doi.org/10.3758/s13428-021-01694-3

[51] Paul R. Pintrich. 2000. The role of goal orientation in self-regulated learning. In *Handbook of Self-Regulation.* Academic Press, San Diego, 451–502. https://doi.org/10.1016/B978-012109890-2/50043-3

[52] Guilherme Polanczyk and Luis Augusto Rohde. 2007. Epidemiology of attention-deficit/hyperactivity disorder across the lifespan. *Current Opinion in Psychiatry* 20, 4 (2007), 386–392.

[53] Abigail Reaser, Frances Prevatt, Yaacov Petscher, and Briley Proctor. 2007. The learning and study strategies of college students with ADHD. *Psychology in the Schools* 44, 6 (2007), 627–638.

[54] Michelle Richardson, Charles Abraham, and Rod Bond. 2012. Psychological correlates of university students' academic performance: a systematic review and meta-analysis. *Psychological bulletin* 138, 2 (2012), 353.

[55] Ido Roll and Ruth Wylie. 2016. Evolution and revolution in artificial intelligence in education. *International Journal of Artificial Intelligence in Education* 26 (2016), 582–599.

[56] Jennifer L. Sabourin, Lucy R. Shores, Bradford W. Mott, and James C. Lester. 2013. Understanding and predicting student self-regulated learning strategies in game-based learning environments. *International Journal of Artificial Intelligence in Education* 23 (2013), 94–114.

[57] Morgan Klaus Scheuerman, Jacob M. Paul, and Jed R. Brubaker. 2019. How computers see gender: An evaluation of gender classification in commercial facial analysis services. *Proceedings of the ACM on Human-Computer Interaction* 3, CSCW (Nov. 2019), 1–33. https://doi.org/10.1145/3359246

[58] James R. Segedy, John S. Kinnebrew, and Gautam Biswas. 2015. Using coherence analysis to characterize self-regulated learning behaviours in open-ended learning environments. *Journal of Learning Analytics* 2, 1 (2015), 13–48.

[59] Katharina Simbeck. 2024. They shall be fair, transparent, and robust: auditing learning analytics systems. *AI and Ethics* 4, 2 (2024), 555–571.

[60] Jay Sloan-Lynch and Robert Morse. 2024. Equity-forward learning analytics: Designing a dashboard to support marginalized student success. In *Proceedings of the 14th Learning Analytics and Knowledge Conference.* ACM, Kyoto Japan, 1–11. https://doi.org/10.1145/3636555.3636844

[61] Eline Stenwig, Giampiero Salvi, Pierluigi Salvo Rossi, and Nils Kristian Skjærvold. 2022. Comparative analysis of explainable machine learning prediction models for hospital mortality. *BMC Medical Research Methodology* 22, 1 (2022), 53.

[62] Isabel Straw and Honghan Wu. 2022. Investigating for bias in healthcare algorithms: A sex-stratified analysis of supervised machine learning models in liver disease prediction. *BMJ Health & Care Informatics* 29, 1 (April 2022), e100457. https://doi.org/10.1136/bmjhci-2021-100457

[63] Harini Suresh and John V. Guttag. 2021. A framework for understanding sources of harm throughout the machine learning life cycle. In *Equity and Access in Algorithms, Mechanisms, and Optimization.* ACM, New York, NY, 9. https://doi.org/10.1145/3465416.3483305 arXiv: 1901.10002.

[64] James M. Swanson, Joseph A. Sergeant, Eric Taylor, Edmund J.S. Sonuga-Barke, Peter S. Jensen, and Dennis P. Cantwell. 1998. Attention-deficit hyperactivity disorder and hyperkinetic disorder. *The Lancet* 351, 9100 (1998), 429–433.

[65] Konstantinos Tsiakas, Emilia Barakova, Javed Vassilis Khan, and Panos Markopoulos. 2020. BrainHood: Towards an explainable recommendation system for self-regulated cognitive training in children. In *Proceedings of the 13th ACM International Conference on PErvasive Technologies Related to Assistive Environments.* ACM, Corfu Greece, 1–6. https://doi.org/10.1145/3389189.3398004

[66] Duygu Umutlu and M. Emre Gursoy. 2022. Leveraging artificial intelligence techniques for effective scaffolding of personalized learning in workplaces. In *Artificial Intelligence Education in the Context of Work*, Dirk Ifenthaler and Sabine Seufert (Eds.). Springer International Publishing, Cham, 59–76. https://doi.org/10.1007/978-3-031-14489-9_4 Series Title: Advances in Analytics for Learning and Teaching.

[67] Vinitra Swamy, Bahar Radmehr, Natasa Krco, Mirko Marras, and Tanja Käser. 2022. Evaluating the explainers: Black-box explainable machine learning for student success prediction in MOOCs. In *Proceedings of the 15th International Conference on Educational Data Mining.* International Educational Data Mining

Society, Durham, England, 98–109. https://doi.org/10.5281/ZENODO.6852964

[68] Wanli Xing and Dongping Du. 2019. Dropout prediction in MOOCs: Using deep learning for personalized intervention. *Journal of Educational Computing Research* 57, 3 (2019), 547–570.

[69] Zhihong Xu, Yingying Zhao, Jeffrey Liew, Xuan Zhou, and Ashlynn Kogut. 2023. Synthesizing research evidence on self-regulated learning and academic achievement in online and blended learning environments: A scoping review. *Educational Research Review* 39 (2023), 100510.

[70] Andres Felipe Zambrano, Jiayi Zhang, and Ryan S. Baker. 2024. Investigating algorithmic bias on bayesian knowledge tracing and carelessness detectors. In *Proceedings of the 14th Learning Analytics and Knowledge Conference (LAK '24)*. Association for Computing Machinery, New York, NY, USA, 349–359. https://doi.org/10.1145/3636555.3636890

[71] Jiayi Zhang, Juliana Ma Alexandra L Andres, Stephen Hutt, Ryan S. Baker, Jaclyn Ocumpaugh, Nidhi Nasiar, Caitlin Mills, Jamiella Brooks, Sheela Sethuaman, Tyron Young, et al. 2022. Using machine learning to detect SMART model cognitive operations in mathematical problem-solving process. *Journal of Educational Data Mining* 14, 3 (2022), 76–108.

[72] Yingbin Zhang, Luc Paquette, Ryan S. Baker, Jaclyn Ocumpaugh, Nigel Bosch, Anabil Munshi, and Gautam Biswas. 2020. The relationship between confusion and metacognitive strategies in Betty's Brain. In *Proceedings of the 10th International Conference on Learning Analytics & Knowledge (LAK '20)*. Association for Computing Machinery, New York, NY, USA, 276–284. https://doi.org/10.1145/3375462.3375518

[73] Yingbin Zhang, Luc Paquette, Nigel Bosch, Jaclyn Ocumpaugh, Gautam Biswas, Stephen Hutt, and Ryan S. Baker. 2022. The evolution of metacognitive strategy use in an open-ended learning environment: Do prior domain knowledge and motivation play a role? *Contemporary Educational Psychology* 69 (2022), 102064.

[74] Tongyu Zhou, Haoyu Sheng, and Iris Howley. 2020. Assessing post-hoc explainability of the BKT algorithm. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*. ACM, New York NY USA, 407–413. https://doi.org/10.1145/3375627.3375856

[75] Barry J. Zimmerman and Adam R. Moylan. 2009. Self-regulation: Where metacognition and motivation intersect. In *Handbook of Metacognition in Education*. Routledge/Taylor & Francis Group, New York, NY, US, 299–315.