An exploration of bias and fairness in Bayesian Knowledge Tracing

Alasdair Day

4th Year Project Report
Computer Science
School of Informatics
University of Edinburgh

2023

Abstract

The modelling of student performance is a task with growing engagement from the machine learning community. One of the most common methods for tackling this challenge is Bayesian Knowledge Tracing (BKT). This model aims to predict student mastery on any given topic by examining the student's answers to multiple choice questions, and while it has clear advantages it is - like most machine learning approaches - prone to bias. In this study such a model is trained using data from the Eedi challenge, and it is found that the model gives more favourable predictions to students not receiving income support when compared to those who are. The potential cause(s) for this disparity are explored and it is found that the bias present is most likely a result of the model disregarding answer speed when making predictions. To combat this flaw a new variation of the BKT model is proposed. The model is altered to take answer speed into account so that predictions on questions answered quickly are more generous, but harsher on questions answered slowly. This change results in a significant increase in fairness, thereby providing a causal link between answer speed and the bias present in the original BKT model.

Research Ethics Approval

This project was planned in accordance with the Informatics Research Ethics policy. It did not involve any aspects that required approval from the Informatics Research Ethics committee.

Declaration

I declare that this thesis was composed by myself, that the work contained herein is my own except where explicitly stated otherwise in the text, and that this work has not been submitted for any other degree or professional qualification except as specified.

(*Alasdair Day*)

Acknowledgements

There are many people without whom this project would not have been possible, and who are deserving of thanks.

To Kobi Gal, for providing me with insight, guidance, and expertise whenever I was in need of it.

To Jake Barrett, for your unwavering support, and for providing me with an invaluable wall from which I could bounce any number of bizarre ideas and theories back and forth.

To my friends and family, for your lasting patience in hearing me rant about various coding headaches and confusions. Special mention to Robert Brewer for buying me the occasional creme egg to aid in my work.

And Finally, to the staff of Pleasance Cafe, thank you for the several hundred coffees I drank while working on this project, and for occasionally playing Jazz music while I did so.

Table of Contents

Chapter 1

Introduction

The already fast-growing field of E-learning was sent into overdrive during the Covid-19 pandemic. Many existing methods of E-learning grew vastly in popularity, and several new methods cropped up to allow student learning to continue naturally from their own homes [7]. Pupils have since returned to classrooms, but the potential benefits of E-learning have not been forgotten. While these new learning opportunities are almost certainly a positive thing, they bring along with them a new set of risks and challenges. One of these risks is the potential for bias and unfairness [4].

While there has been research into the area of fairness and how it interacts with student learning and performance prediction [10] [12], the work done investigating the violations of formally defined ”fairness” within a knowledge tracing setting have been lacking, and there is limited work looking at how to address the issue.

The demographic of interest to this study is socio-economic status (measured by student eligibility for income support), this was chosen for investigation as while there are failures in the model's handling of other fields such as gender, the bias present over this field proved to be the most egregious. Unlike other demographics - such as age and gender - discrimination based on economic standing is not protected by the Equality Act 2010. Despite this, it remains a serious issue and much work has been done exploring the harm that can be caused by this sort of discrimination [6] [11]. Additionally, there have been calls for economic standing/class to be added to the protected characteristics covered by the Equality Act [13] [14]. Because of this, it is crucial that for any knowledge tracing model to be considered ”fair” it must have a consistent prediction accuracy for students with the same mastery level, irrespective of income support eligibility.

This study aims to examine and then help combat this risk within the context of a Bayesian Knowledge Tracing (BKT) model (one of the most widely used student modelling techniques [16]) by documenting, explaining, and correcting the violations of fairness that appear when the model is used without appropriate safeguards against bias. Specifically, the questions answered by this report are as follows.

- What bias is present in the predictions made by a BKT model that is trained using the Eedi data set?

- How does this bias impact the fairness of the model when considering the income support eligibility demographic?
- What aspect of the data set and/or model is causing this bias?
- How can the model predictions be improved such that the bias is reduced?

In search of answers to these questions, this study begins with the training of a standard BKT model using data from the Eedi challenge [5]. The model then makes predictions on more samples from this data set, and the accuracy of these predictions is compared across the income support eligibility demographic. From this analysis, it is seen that the model exhibits bias and unfairness over the demographic of income support eligibility. This analysis is then focused on predictions of questions answered at varying speeds, and it is hypothesised that the model's failure to consider answer speed in its calculations is the cause of the bias. Finally a solution - whereby answer speed is considered - is presented, one that has minimal impact on model accuracy yet provides a significant increase in measurable fairness. As such, this study proves a promising step in understanding and fighting the bias present in BKT models.

Chapter 2

Related Works

In this chapter the pieces of literature that proved most relevant to the study are reviewed. This includes past examinations of unfairness within BKT models, a comprehensive description of unfairness within machine learning as a whole, and an example of modifications improving the fairness of a BKT model.

2.1 Fairness and Knowledge Tracing

The core of this study is the understanding of bias and unfairness within BKT models. As such, the work that proved most helpful and influential to this study were those that also held this concept as their core.

In 2021 Barrett set out to explore the extent to which bias can manifest in student learning models and subsequently cause unfairness in the predictions these models make [2]. This was achieved by training a BKT model on the publicly available Eedi data set [5](the same data set used for this study). The predictions made by the BKT model were then compared to the widely-used N Consecutive Correct Responses (NCCR) [8] mastery predictor and it was found that the model violates fairness principles when evaluated over the fields of gender and income support eligibility. This paper provides many interesting results and was a strong initial exploration into this problem. That being said, this paper spent very little time exploring or explaining the causes of these biases, and did nothing to provide a causal link between the bias and these fields.
To build upon this, I recreated much of the same data exploration as the author in order to fully understand the nature and location of the bias found. After this, my own exploration began into the cause(s) of the bias, effectively carrying out the further research suggested in the original paper.

In 2019 Doroudi and Brunskill showed that student learning models have a tendency to be inequitable in their results, and suggested that one major reason for this is that they are typically trained on global populations. They show that this results in the models failing to take into account the variations that exist within the population (The example given in the paper is that of slow learners and fast learners being modelled in the same way) [15]. The authors reached these results after building two student learning models

(the first a Bayesian knowledge tracing model, the second an additive factor model), and then analysing their performance over data belonging to fast learners and slow learners. From this it was shown that more fast learners mastered the skills than slow learners, even though the same model prediction confidence was consistent across the learning speeds (95%).
This study effectively demonstrates an inequality in results predicted by student knowledge models, then provides and explains potential reasons for it. However, one significant flaw of the study is that it labels all its subjects as either fast or slow learners without explicitly explaining what these categories entail, how they were measured, and where distinctions were drawn between the two types. Additionally, this difference is treated as a binary class rather than a spectrum, which likely results in less accurate results, though they may be clearer. Fortunately for my own work, I am less interested in the learning speed descriptor as I aim to investigate the effect of income support eligibility, and much of the methodology used and results found were still applicable to this study. For example, while it is not an exact comparison, this study's treatment of differing answer speeds is along the lines of how the authors described modelling students of different learning speeds.

2.2 Measuring Fairness

In order to understand the problems with fairness in a BKT setting, the concept of fairness within AI in general must be understood.

In 2016 Hardt et al identified that despite the growing understanding of the potential impact of bias in AI, not much has been done to further combat it. Accordingly, they set out to define criteria for discrimination and design a framework for how to adjust learned predictors to minimize discrimination[9]. To this end, they defined two measures of fairness; those being equality of opportunity (true positive rates are consistent regardless of the protected attribute(s)), and equality of odds (both true positives and false positives must be consistent). They then define mathematically how to compute and analyse these measures, alongside demonstrating how to apply this to a learned model in order to reduce the discrimination present.
While it is doubtful that this would actually be an issue in practice, the authors only demonstrate the effectiveness of their model with one factor on one real-world data set. It is therefore possible the same success would not be achieved with factors more nuanced than race when trying to improve fairness in the ways the authors suggest. Fortunately however, this is not a flaw that impacts the usefulness of the paper to this study. Equality of opportunity and equality of odds are key concepts in this work, and many insights on their usage and meanings were drawn from this paper.

2.3 Modifying BKT for Improved Fairness

The final result of this study is a modified BKT model that provides fairer predictions. As such, it is of great help to consider previous works that also modify BKT in search of improved fairness.

In 2022 Tschiatschek et al investigated the effects of a modified BKT model that was tuned individually to the students it was making predictions on [17]. They define what constitutes an ethical AI-assisted tutoring system and then explain that a key factor for realizing this system is individualised model performance for each student. The model they propose is called Bayesian Bayesian Knowledge Tracing (B^2KT). In this model the four BKT parameters guess, slip, learn, and prior (defined and described in Section 3.1) are continually updated throughout the runtime to try and represent each student more accurately. They conclude their research by demonstrating the benefits this model can have when compared to regular BKT but concede the point that this type of research may not be enough to result in truly equitable tutoring systems.
While the modifications made by the authors are fundamentally different from the ones proposed in this work (They change the model's behaviour based on student characteristics, whereas here the model is changed based on answer speed), the motivation and end results share similarities. As such, many of the decisions made in this study were informed by the work presented here, and were this study to extend in search of further fairness, the modifications would likely be very similar to the ones described by the authors here.

Chapter 3

Methodology

In this chapter, the key pieces of background information are outlined and described. Familiarity with these subjects is required for the results presented in this study to become meaningful, so they are individually explained, and then details of how they work together are provided.

3.1 Bayesian Knowledge Tracing

Bayesian knowledge tracing (BKT) is an application of artificial intelligence that allows a student's academic performance to be modelled. Models of this kind take in a list of the students' answers (in this case answers to multiple choice maths questions), whether each answer was correct, and what subject each question was testing. This list is ordered by date and time so that the model can follow the user's path of answers.

The key difficulty in predicting mastery with any model is that it is a latent variable, meaning it cannot be directly observed from data and rather must be inferred from other variables. Because of this, a confidence value between 0 and 1 is derived describing the likelihood that the student has achieved mastery. At step n it predicts whether the student has answered the nth question correctly and then updates its predicted probability that the student has mastered the chosen skill. This is achieved using the following formulas:

$$P(C_j) = P(G)(1 - P(L_j)) + (1 - P(S))P(L_j) \tag{3.1}$$

$$P(L_j) = P(L_j - 1) + P(T)(1 - P(L_j - 1)) \tag{3.2}$$

- $P(C_j)$ is the probability that the student answers question j correctly.
- $P(L_{j]})$ is the probability that the student has mastered the skill at step j.
- $P(G)$ is the probability of a correct guess if the student does not know the skill. It is referred to as the guess parameter.
- $P(S)$ is the probability of an incorrect answer despite the student knowing the skill. It is referred to as the slip parameter
- $P(T)$ is the probability of the student learning the skill at any given step. Note that $P(T)$ is assumed to be constant over time. It is referred to as the learn parameter.

$P(G), P(S), P(T)$ and the prior probability (the likelihood that a student has mastered a given skill before answering any questions) are initialised at pre-defined values and then optimized during training using a method such as expectation maximisation (EM). After training they remain constant for future model usage.

To demonstrate the predictions made by this model, consider an example student who answers 15 questions in the following pattern: [0,1,0,1,1,1,1,1,0,1,0,0,1,0,0] (with a 1 indicating a correct answer, and a 0 indicating an incorrect answer). When this list of answers is passed into the BKT model, it produces the following result:

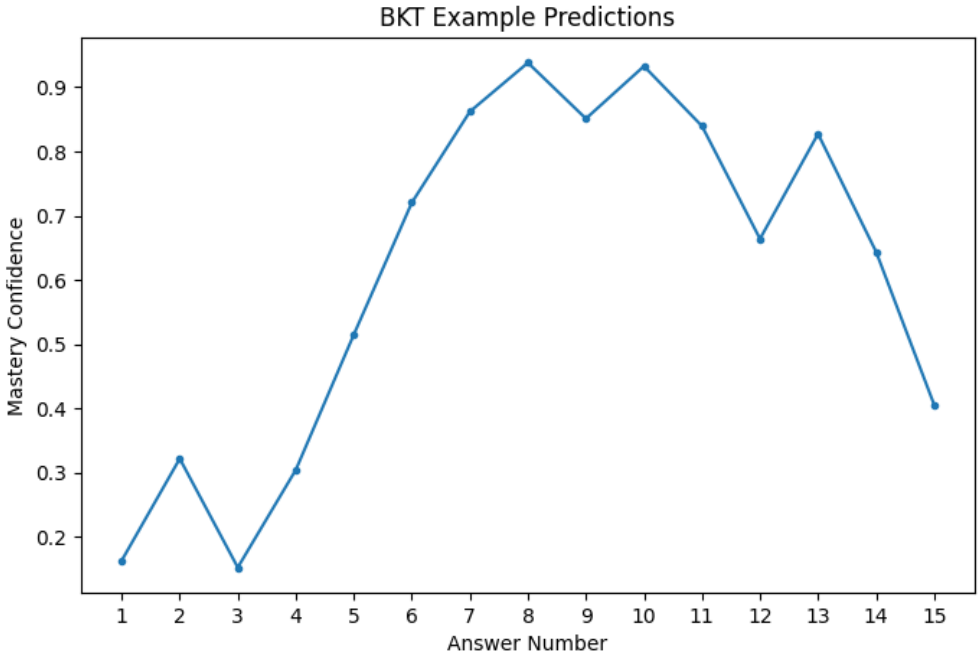

Figure 3.1: Example predictions produced by BKT model

The majority of this study utilised pyBKT (a Python implementation of BKT) to carry out modelling. This is a straightforward open-source implementation and was well-suited suited to the task at hand. However, for the final section of the research, a new implementation was built based on the formulae above and those provided in section 4.4. For more information on BKT models please see: Properties of the Bayesian Knowledge Tracing Model[18]

3.2 Eedi data set

For this project the model was trained and evaluated using the Eedi data set [5]. This data is collected via an online learning tool that contains crowd-sourced multiple-choice maths questions. This tool is primarily deployed in schools and is used by children from ages 7 – 18. Students' answers are collected along with metadata regarding the

questions, the subjects in the questions, the students, and the student's answers. More specifically, this study utilises the Eedi data for 12 to 13-year-olds.

The data is provided in the form of multiple files each pertaining to one category of information. The extensive quantity of data made it suitable for all desired analysis but also resulted in redundant data that was not useful. Accordingly, experiments were primarily performed using a master dataframe composed of the following fields:

- Order: Acts as the index for the dataframe
- QuestionID: Uniquely identifies each question in the data set
- UserId: Uniquely identifies each user in the data set
- AnswerId: Uniquely identifies each answer in the data set
- IsCorrect: A binary value, indicating whether each answer is correct or incorrect
- Name: The name of the subject that that particular question tests
- SubjectId: Uniquely identifies each subject in the data set
- DateAnswered: The date and time of the answer submission aggregated to minutes
- Gender: An integer value, indicating whether each student is male or female. (There were two other options in the data set for this field - "unspecified" and "other" - but they did not have enough entries to be useful for analysis)
- PremiumPupil: A binary value, indicating whether the student is eligible for some sort of financial aid e.g. free school meals.

3.3 Equality of Opportunity/Odds

Since fairness and bias are both core concepts for this project and will be referenced frequently, a definition is provided here. Bias – in an AI context – is when two "similar" inputs are treated dis-similarly. That is to say that if relevant factors are consistent between inputs A and B, A and B should have the same output, even if they differ on some factor x which is irrelevant to the prediction being made. Should A and B have a different output, this is identified as a bias on factor x. A level deeper than this and we get to the concept of fairness. For this report, two metrics of fairness will be considered, as defined by Hardt et al[9]: equality of opportunity and equality of odds.

Equality of opportunity is satisfied when students who have mastered a skill are generally predicted to have mastered it irrespective of any demographic variables. The other, stricter metric considered is equality of odds. This is satisfied when equality of opportunity is satisfied and additionally when students who have not mastered the skill are generally predicted not to have mastered it consistently across any demographic values.

Formally, in a knowledge tracing context, equality of opportunity requires that students who have mastered a given skill $(y = 1)$ who are a part of demographic values $D \in \{0, 1\}$

have the same predicted value for mastery in the model, $\hat{Y}$.

$$P(\hat{Y} = 1|D = 1, Y = 1) = P(\hat{Y} = 1|D = 0, Y = 1) \tag{3.3}$$

For equality of odds, the requirements are made stricter so that regardless of the "true" mastery status Y, students of different demographic groups $D \in \{0, 1\}$ have equal access to a correct prediction $\hat{Y}$.

$$P(\hat{Y} = 1|D = 1, Y = y) = P(\hat{Y} = 1|D = 0, Y = y), y \in \{0, 1\} \tag{3.4}$$

To understand the demographic bias present, an odds ratio (OR) is calculated across the demographic value in question. To compute equal opportunity in this way, only cases where the student's "true" value of mastery is positive $(y = 1)$ were examined, but for equal odds all cases regardless of "true" mastery $y \in \{0, 1\}$ are examined.

$$OR(D, y) = \frac{P(\hat{Y} = 1|D = 1, Y = y)}{P(\hat{Y} = 1|D = 0, Y = y)} \tag{3.5}$$

For demographic values $D \in \{0, 1\}$, an odds ratio above 1 shows that the model is overly assigning mastery to students of demographic group $D = 1$ irrespective of actual performance (and vice-versa a value below 1 would show favouritism to $D = 0$). Should the ratio be sufficiently far from 1 then this would be evidence that the model is biased over demographic variable D. Likewise, a model without bias would have an odds ratio of approximately 1.

3.4 N Consecutive Correct Responses (NCCR)

To evaluate the accuracy of the trained BKT model a baseline criteria for student mastery is needed. Quantifying mastery in a truly objective manner is very difficult. This report chooses to replicate the approach of Barrett[2] by recording the maximum number of consecutive correct responses (NCCR) on a per-topic basis that each student has achieved, and comparing that value to thresholds of 3, 5 and 10. The satisfaction of each threshold provides a lower bound for mastery, with weak, fair and strong confidence respectively.

If we consider the example student used in section 3.1, the highest consecutive number of correct responses they achieved was 5. Therefore the CCR data for them would look like this

Student ID	3 CCR	5 CCR	10 CCR
0	1	1	0

Table 3.1: Example Entries of NCCR

Naturally this measure is not without flaws, as NCCR will be affected by factors such as lucky guesses, forgetfulness or test-taking stress. Nonetheless, it provides a customizable strictness and is easy to implement and utilise, making it appropriate for this project. Additionally, as shown by Kelly et al [8] it provides accuracy on par or above that offered by more complex measures.

3.5 Putting it all Together

Before the model could be trained, the data had to be split into training and testing sets. To accomplish this the list of unique User Ids was randomly split such that 20% would be used for training and the remaining 80% for testing. It is necessary to split based on users rather than splitting all answers in the data set as the model requires all of any one student's answers in order to make meaningful predictions. This split provides 3,934 training users with 1,104,209 answers, and 984 test users with 274,016 answers. These testing samples were then used to analyse various aspects of the model's performance and fairness. Both the random splitting of user Ids and the parameter optimization of the model were performed with fixed seeds to maintain the necessary randomness but enable the recreation of results when necessary.

As described in section 3.1, a correctness is then predicted for every question in the training set, alongside a prediction that the student has mastered the corresponding skill at any given time. This set of predictions is grouped by User Id and Skill, then sorted by the date and time answered. This allows the final chronological prediction to be extracted for each unique user Id and skill pair.

Before an analysis of fairness can be carried out, a baseline for mastery is required. For this, the complete collection of each student's answers is examined, and the highest number of consecutive correct responses is counted for each subject. Finally, three Boolean values are added to the predictions table, indicating whether, for each student subject pair, each strictness level of NCCR mastery (3, 5, or 10) has been satisfied.

With the mastery predictions made and a baseline in place, fairness can finally be examined. Equality of odds can be computed by splitting the predictions by the income support eligibility demographic and then taking a ratio of their average values. For equality of opportunity, the same process is completed, but first the student subject pairs are filtered such that they only include the entries that satisfy that required baseline mastery (measured by the values in the three NCCR columns).

Chapter 4

Results

In this section the analysis and experiments that were carried out are described and explained. Firstly The base model performance is examined and found to be unfair. Secondly the cause of this bias is investigated and a hypothesis is drawn up. Finally, a new model is proposed that aims both to provide evidence for the hypothesis, and to improve model fairness.

4.1 Analysing fairness

To analyse the performance of the model, it must be compared to the baseline of NCCR. While this is the most suitable measure in this situation, there is still an issue; NCCR values can only ever be either 0 or 1, while the model's predictions are continuous probabilities. That being said, since the predictions will always be between 0 and 1, calculations and comparisons can still be drawn from these two measures together. Specifically, by subtracting the value in a NCCR column from the predicted mastery, we can determine to what extent the model has over or under-predicted mastery by how far the result strays from zero. For example, if the final prediction for a student skill combination is 0.9 and the value in the chosen NCCR column is 1, then

$$0.9 - 1 = -0.1 \tag{4.1}$$

showing that the model has under-predicted by a small margin, and therefore has performed well. Alternatively, if the predicted mastery is 0.9 but the value in the CCR column is 0, then

$$0.9 - 0 = 0.9 \tag{4.2}$$

showing that the model has greatly over-predicted mastery and therefore has performed very poorly. This process can be repeated for all three NCCR thresholds to provide various levels of strictness on what counts as mastery.

By applying the above calculations across the data, results were gathered to explore the model's output across the demographic of eligibility for income support.

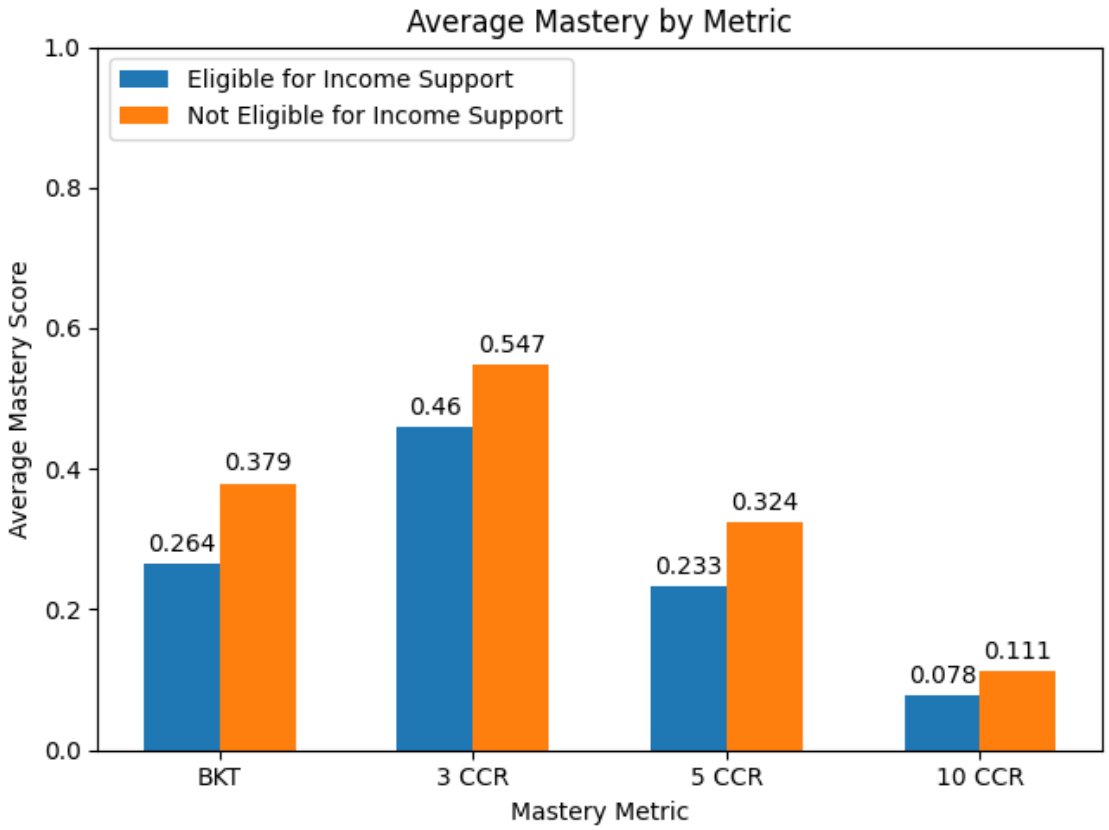

Figure 4.1: Average values for mastery when measured by different metrics (3 d.p.)

	3 CCR Difference	5 CCR Difference	10 CCR Difference
Eligible for Support	-0.196	0.031	0.186
Not Eligible for Support	-0.168	0.055	0.268

Table 4.1: Difference between average predicted and ”real” mastery at different NCCR thresholds (3 d.p.)

From table 4.1 we can conclude the model tends to over and under-predict fairly evenly when a medium strictness is chosen for the threshold (this explains why the difference is so much lower than for 3 and 10 CCR, as the over and under predictions will cancel each other out), tends to over-predict mastery when a strict threshold is used, and tends to under-predict mastery when a weak threshold is used. More interesting, however, is the effect that eligibility for income support has on predictions. When compared to the ”real” mastery value, students eligible for income support have lower mastery predictions across all CCR thresholds when compared to the same results for non-eligible students. On average, an eligible student will be predicted 30.3% lower despite only performing 15.9%, 28.1%, and 29.7% worse when judged by NCCR thresholds of 3, 5, and 10 respectively (1 d.p.). Therefore it is evident that the model does not satisfy equality of odds as the predictions are not the same across the demographic value. however, to show that equality of opportunity is violated requires investigation of only the students who achieved mastery:

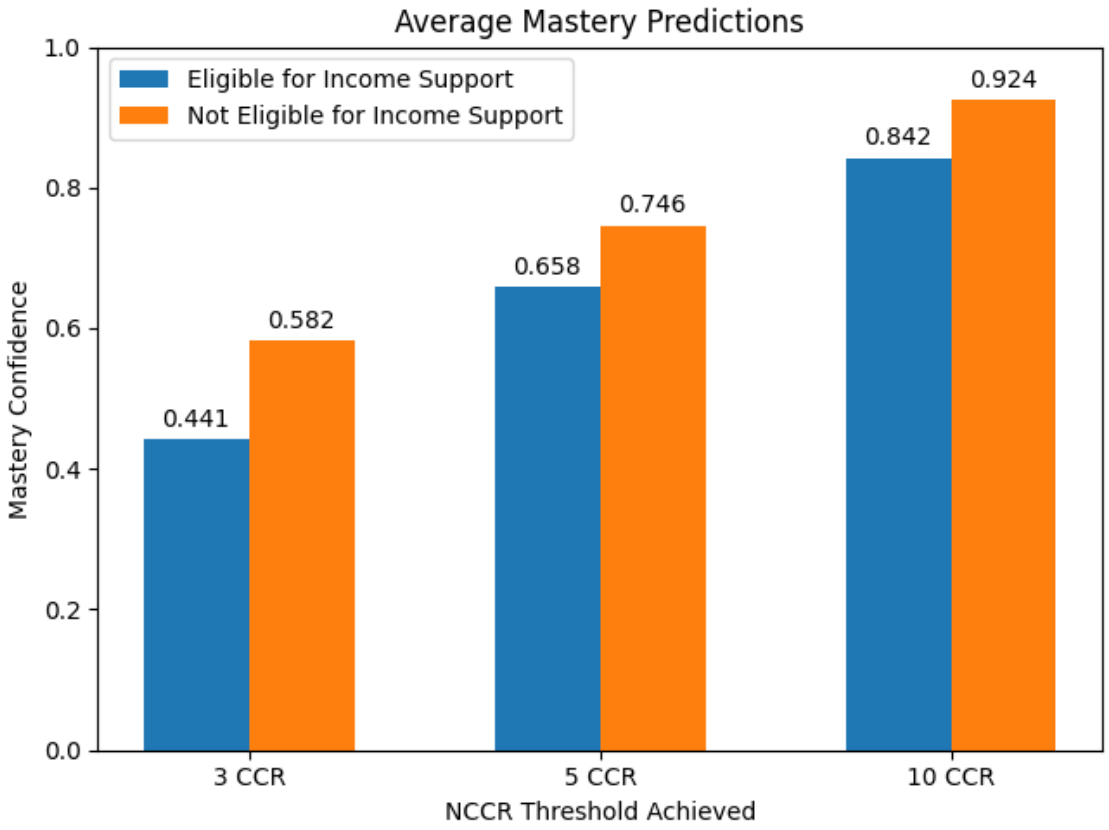

Figure 4.2: Average mastery confidences for eligible and non-eligible students who have satisfied different NCCR thresholds

From figure 4.2 it is evident that even when considering only those students who have ”mastered” the skill, favourable predictions are still given to those who are not eligible for economic support when compared to those who are. More precisely, the odds ratios of predictions for non-eligible and eligible students are 1.320, 1.134, and 1.098 for mastery thresholds of 3, 5, and 10 CCR respectively (3 d.p.). Thus this model fails to satisfy both equality of odds and equality of outcome, and is not a fair model when analysed over the income support eligibility demographic.

4.2 Students eligible for income support answer more quickly

Now that bias has been demonstrated to be present, identifying its source is key. After some data exploration answer speed became the field of interest, and thus it was hypothesised that students qualifying for income support were more prone to rushing through questions, perhaps due to a more hectic home life that results in less ideal studying conditions and/or less time available to spend on homework. These factors could negatively impact their answers causing a misrepresentation of their actual mastery. To illustrate this point, consider the example student described in section 3.1. Suppose they were assigned 15 homework questions and having performed well on the first 10

questions - achieving 5 consecutive correct answers and only making three mistakes - ran out of time and had to answer the remaining 5 questions in just a few minutes, resulting in rushed answers and only 1 correct response. Most people watching this play out would likely be happy that the student had demonstrated their mastery regardless of the final 5 questions, but the BKT model would continue lowering the predicted probability of mastery over these final questions, resulting in a lower final prediction.

To demonstrate a correlation between a pupil's premium status and answer speed, answer speed first had to be calculated using the timestamps provided for each answer. There was however the caveat that the timestamp did not include seconds and therefore all answer times were expressed as a whole number of minutes. Additionally, any answer that took over 10 minutes was viewed as the start of a new session and therefore not included in the calculations as there is no way to tell how long the student spent on it (10 minutes may seem like a short cut off, but the questions asked were mostly very short and simple).

Using this method, the average time to answer a question (t) can be calculated and it becomes possible to demonstrate the fact that higher proportions of students eligible for income support answer at fast speeds when compared to non-eligible students.

$t <$ 1 minute	$t >=$ 1 minute	$t <$ 2 minutes	$t >=$ 2 minutes
2.139	3.430	2.449	4.158

Table 4.2: ratio of the quantity of non-eligible student answers to eligible student answers at different answer speeds (3 d.p.)

As shown in table 4.2, the ratio of non-eligible students to eligible students is far greater when looking at questions that have been answered more slowly, meaning that on average, the eligible students are answering more of their questions quickly than the non-eligible students are.

4.3 Students eligible for income support perform worse when at faster speeds than non-eligible students

Of course it is clear that this difference only presents an issue of unfairness if it is the case that increased speed correlates with decreased accuracy. To examine this, the average correctness for answers at differing answer speeds is compared across the income support eligibility demographic.

$t <$ 1 minute	$t >=$ 1 minute	$t <$ 2 minutes	$t >=$ 2 minutes
1.205	1.117	1.201	1.053

Table 4.3: ratio of non-eligible student average correctness to eligible student average correctness at different answer speeds (3 d.p.)

From table 4.3 we can see that non-eligible students tend to perform better at faster speeds when compared to eligible students than they do at slower speeds.

The results provided in tables 4.2 and 4.3 strongly support the idea that students eligible for income support are likely to have lower accuracy answers as a result of insufficient time to complete questions. There are many possible ways to justify the connection between income support eligibility and more rushed answers in this setting. It seems most likely that students eligible for income support - who on average will come from lower socioeconomic backgrounds - will have less time to spend studying and fewer tools to aid with their studying such as laptops, tablets, etc. Additionally, these students will have a higher likelihood of additional responsibilities and disruptions, such as being young carers[3] which will further detract from their time available for homework.

Of course, this is all simply pointing out correlation. To demonstrate causation the potential issue (the disregarding of answer speed) must be fixed within the model so that the fairness can once again be analysed. Should this result in a fairer model then that would present evidence for a causal relationship.

4.4 An Improved Model

Having located the source of the bias, reducing its impact becomes the most pressing challenge. Working from the hypothesis that the bias manifests as a result of answer speed not being accounted for, an obvious solution appears: take account of answer speed. Fortunately, there is no need to change the model's architecture for this, as there is already a parameter in BKT fit for this purpose. The slip parameter - as described in section 3.1 - controls the probability that the student has answered a question incorrectly despite having mastered the skill. In a regular BKT model this parameter is a static value, meaning it is calculated during training time and then stays at that value whenever the model is used to make predictions. To improve upon this, the slip parameter is tweaked so that it can adapt to the speed at which each question was answered. In practice this is as simple as scaling the parameter up or down based on the time taken for the question to be answered, but the result is a slip parameter that can more accurately represent the likelihood that the student really has failed to apply a skill despite having mastered it. Thus, a new model was built that Incorporated this dynamic slip parameter.

Since pyBKT is no longer being used at this point, it becomes necessary to explain precisely how the BKT model operates at runtime. The model recursively moves through all of a student's answers for a given topic, continually updating the posterior probability that the student has mastered the skill based on each answer's correctness. If O_j (the observation of a student's answer at time j) is correct then equation 4.3 is used, and if the answer is incorrect then 4.4 is used. These formulas recurse until they hit the base case for $P(L_j)$ which is the prior parameter (a learned value indicating the probability that the student has mastered the skill before answering any questions). For clarification on the parameters of the model see section 3.1.

$$P(L_j|O_j) = 1 - \frac{(1-P(T))[1-P(L_{j-1}|O_{j-1})]P(G)}{P(G)+(1-P(S)-P(G))P(L_{j-1}|O_{j-1})} \tag{4.3}$$

$$P(L_j|O_j) = 1 - \frac{(1-P(T))[1-P(L_{j-1}|O_{j-1})](1-P(G))}{1-P(G)-(1-P(S)-P(G))P(L_{j-1}|O_{j-1})} \tag{4.4}$$

All that has changed for the model proposed here is that the slip parameter $P(S)$ is scaled up or down depending on answer speed, but this presents the question of what values to scale by. Interestingly - and unfortunately - there is very little literature on this subject. Additionally, since the answer times in Eedi do not include seconds the scaling will not evenly change, and rather will fluctuate between set values. While this is not ideal, the quality of the data set in all other areas means it is still the best available for this study overall.

After some analysis of student performance at differing speeds, the scaling values of 2.5, 1.5, and 0.5 were arrived at for questions answered in under a minute, one minute to two minutes, and two and above minutes respectively. For example, if the base slip parameter is 0.15 and the model is making predictions on a question that was answered in under a minute, then the slip parameter would be increased to 0.375, meaning the model views this answer as more likely to be the result of a careless mistake than others, and therefore calculates a more favourable prediction of mastery at that step. Alternatively, if the same model is predicting on a question answered in 4 minutes, then the slip parameter would be decreased to 0.075, meaning the model views this answer as less likely to be the result of a careless mistake than others, and therefore calculates a less favourable prediction of mastery at that step. No guarantee is provided that these values are optimal for this problem, but in the context of the question difficulty and student ages, they seem intuitively appropriate.

When choosing values for the prior, slip (before scaling), guess, and learn parameters, the optimized per-subject values calculated by pyBKT were used at first. Unfortunately, the scaling of the slip parameter combined with these optimized values caused the predictions to fluctuate significantly and therefore provided less meaningful results. As such, the average value across all subjects was calculated for each of the four parameters, and it is these values that were used for the new model. While there is certainly room for some form of per-subject optimization in this section of the model, it will likely need to be more nuanced than the current practice for BKT, as swapping to the averaged values resulted in an increase in both model accuracy and fairness.

To demonstrate the effect of this changed model, let us consider again the example student described in Section 3.1. The predictions for this student made by the default BKT model will not vary no matter how fast or slow the questions are answered. On the other hand, the modified model will provide different results depending on answer speeds. As such, we can graph the predictions made on our example student if they speed up and rush the last five questions, alongside the predictions made if they answer at a consistent speed throughout.

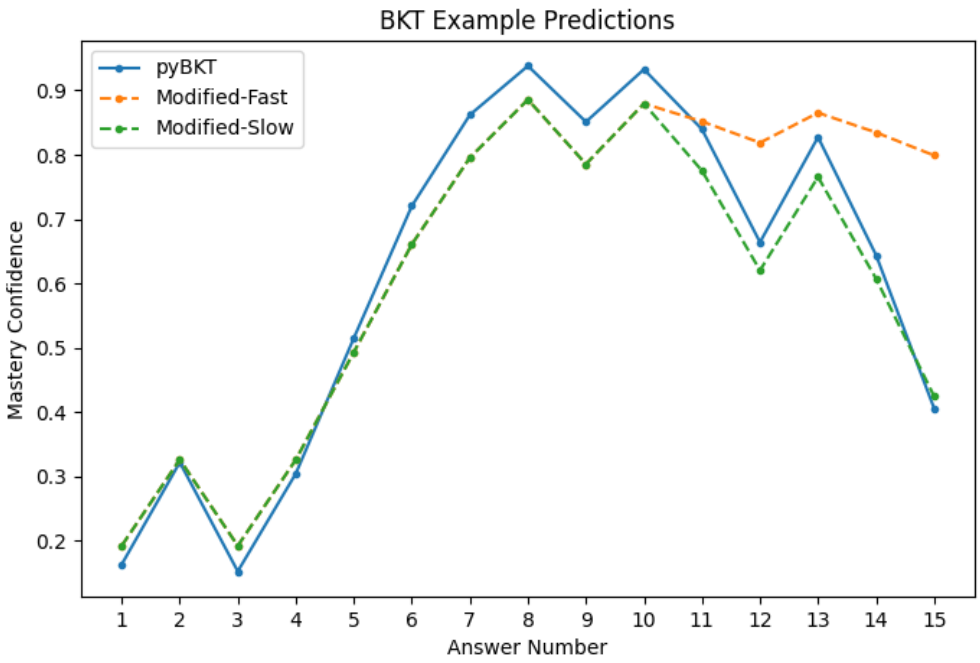

Figure 4.3: Example predictions produced by pyBKT and the modified model for a student if they maintain a steady pace throughout (slow), or if they rush the final 5 questions (fast)

As shown in figure 4.3, the modified model follows the default pyBKT model very closely for the first 10 answers. For these 10, an answer speed of 1 minute was used, meaning a scaling value of 1.5 was applied to the slip parameter. This scaling (alongside the other parameters differing slightly) explains why the predictions grow and fall slightly more slowly than the pyBKT model (see equations 4.3 and 4.4). For the final 5 answers, the Modified-Slow answers continue to be completed in 1 minute each, and therefore the performance continues to closely resemble that of pyBKT. However, for the final five answers on the Modified-Fast line, each question is answered in under a minute, representing the scenario described in section 4.2 where a student has run out of time in their homework. While the predictions still drop over the course of these five questions, the drop is far less substantial than the other two plots, resulting in a much higher (and more accurate given the hypothetical situation) final prediction of mastery.

Following the definition of this model, the same test group used for the pyBKT model was fed into it, and results on performance and fairness were gathered. Firstly, to compare the performance of the new model with pyBKT, the mean squared error(MSE) is examined.

	3 CCR	5 CCR	10 CCR
pyBKT	0.227	0.142	0.195
Modified Slip	0.176	0.147	0.242

Table 4.4: Mean Squared Error (MSE) of both models when calculated using various NCCR baselines (3 d.p.)

As demonstrated above in table 4.4, the performance of the two models is comparable; with the new model providing slightly more generous predictions overall as shown by a lower error on a weaker bound of mastery and higher error on a stricter bound of mastery. Additionally, the accuracy of the new model could almost certainly be improved upon with a more dynamically adjustable slip parameter that would be possible for data that includes seconds in its Date Answered field.

Finally then, the fairness of the new model is examined in comparison to the original model. To do this, the odds ratios of eligible and non-eligible students who have mastered skills according to different thresholds are compared across the two models.

	3 CCR	5 CCR	10 CCR
pyBKT	1.320	1.134	1.098
Modified Slip	1.179	1.063	1.057

Table 4.5: Odds ratios of both models considering student performance for those who have mastered skills according to differing NCCR baselines (3 d.p.)

It can be seen from table 4.5 that the odds ratios of the new model are lower for all thresholds of NCCR. This provides some causal evidence that the bias and unfairness present in the original pyBKT model is due - in at least some capacity - to the disregarding of answer speed. Furthermore, it shows that the odds ratio can be improved by as much as 12% - indicating a large increase in fairness - as a result of the dynamic slip parameter. While the trade-off between accuracy and fairness is a heavily studied and debated topic[1], there can be little doubt that the results presented here show the potential for improved fairness within BKT, potentially (depending on the NCCR of interest) with a minimal decrease in accuracy. This is an intuitive result, as we are feeding more information into the model, indicating that a higher overall accuracy and/or fairness should be achievable.

Chapter 5

Conclusions

In this section the study and its results are summarized, and the implications of these points are discussed. Finally, some suggestions for future work within the field are provided.

5.1 Summary

This study investigated the impact of income support on the fairness of BKT model predictions and finds that it is inherently biased when evaluated with the metrics of equal odds and equal opportunity, and therefore is not a ”fair” model. Specifically, for any two students A and B who have attained mastery, where A is eligible for income support and B is not, the model is 10-30% more likely to predict mastery for B than for A. A large part of the reason for this is that students eligible for income support tend to answer the questions more quickly (perhaps due to a more hectic and demanding home life) than those not eligible, a factor that the model fails to account for.

This finding demonstrates a significant problem with the BKT model, as students eligible for income support already have a lower average answer accuracy[3], a difference that would only be reinforced by unfair models. As such, it is clear that work must be done to incorporate more fairness considerations into BKT before it is appropriate to be deployed for widespread use in student learning tools. One such way to do this - as demonstrated in this study - is changing the model such that it actively considers answer speeds when making its predictions. This solution has shown the potential for a strong improvement in fairness while maintaining similar accuracies to the original model. Additionally, this work has been achieved with a limited data set, meaning the results gained can be seen as a lower bound on success for improvements made in this manner. This means that should similar improvements continue to be made, it is very much possible that BKT models could provide a practical, useful, and most importantly fair tool to be used within the setting of AI-assisted education.

5.2 Future Work

The next logical step in this area would be to further explore the effect that dynamic parameters can have on BKT model performance and fairness. The results provided in this report are promising but by no means comprehensive, and there is much room for both the accuracy and the fairness of this model to be improved further with more fine-tuning.

As mentioned in the report, one potential way to do this would be to train a model on data that has more precise timing information such that a direct relationship between answer speed and slip chance may be calculated, as opposed to the simplistic solution used here, made necessary by the lack of such precise time information present in the Eedi data set. Additionally, there is much room for experimentation with different optimization techniques regarding the trainable model parameters and how they interact with the slip parameter now that is no longer a static value.

Should these suggested improvements be implemented successfully, there is a real and lasting improvement to fairness within the field of AI-assisted education that can be made here. While the improvements made to BKT are the particular interest in this paper, modifications and considerations of the kind discussed here could likely have benefits and implications in many other areas of the field.

Bibliography

[1] J Sánchez-Monedero A Valdivia and J Casillas. How fair can we go in machine learning? assessing the boundaries of accuracy and fairness. volume 36, pages 1619–1643, 2021.

[2] J Barrett. A study of fairness in knowledge tracing. 2021.

[3] G. C. Bond. Social economic status and educational achievement: A review article. *Anthropology and Education Quarterly*, 12(4):227–257, 2009.

[4] B. Sonja. C. Wendy. Implicit bias and first name stereotypes: What are the implications for online instruction? volume 19, 2015.

[5] eedi.com. Neurips 2020 education challenge.

[6] A. Chouldechova. et al. A case study of algorithm-assisted decision making in child maltreatment hotline screening decision. Proceedings of FAT, 2018.

[7] B. M. Amarneh. et al. The impact of covid-19 on e-learning: Advantages and challenges. Proceedings of AICV, 2021.

[8] K. Kelly. et al. Defining mastery: Knowledge tracing versus n-consecutive correct responses. Proceedings of EDM, 2016.

[9] M. Hardt. et al. Equality of opportunity in supervised learning. Proceedings of NIPS, 2016.

[10] S. Hutt et al. Evaluating fairness and generalizability in models predicting on-time graduation from college applications. Proceedings of EDM, 2019.

[11] Y. J. Juhn. et al. Assessing socioeconomic bias in machine learning algorithms in health care: a case study of the houses index. *Journal of the American Medical Informatics Association*, 29(7):1142–1151, 2022.

[12] W. Jiang and Z. A. Pardos. Towards equity and algorithmic fairness in student grade prediction. Proceedings of AIES, 2021.

[13] C. W. Evans. M. Chance. Social class discrimination – time for a new protected characteristic for a post-covid britain? Rosenblatt, 2020.

[14] K. Malleson. Equality law and the protected characteristics. *Modern Law Review*, 81(4):598–621, 2018.

[15] E. Brunskill. S. Doroudi. Fairer but not fair enough on the equitability of knowledge tracing. Proceedings of LAK, 2019.

[16] A B Bichi S M Sani and S Ayuba. Artificial intelligence approaches in student modeling: Half decade review (2010-2015). volume 5, pages 746–754, 2016.

[17] M. Knobelsdorf S. Tschiatschek and A. Singla. Equity and fairness of bayesian knowledge tracing. Proceedings of EDM, 2022.

[18] B. Van De Sande. Properties of the bayesian knowledge tracing model. Proceedings of EDM, 2013.