Accurate Predictions in Education with Discrete Variational Inference

Tom Quilter[1], **Anastasija Ilick**[2], **Karen Poon**[3], **Richard Turner**[4]
[1]University of Manchester
[2]Google DeepMind
[3]MathWorks
[4]University of Cambridge

All correspondence, `thomas.quilter@manchester.ac.uk`

Abstract

One of the largest drivers of social inequality is unequal access to personal tutoring, with wealthier individuals able to afford it, while the majority cannot. Affordable, effective AI tutors offer a scalable solution. We focus on adaptive learning, predicting whether a student will answer a question correctly, a key component of any effective tutoring system. Yet many platforms struggle to achieve high prediction accuracy, especially in data-sparse settings. To address this, we release the largest open dataset of professionally marked formal mathematics exam responses to date. We introduce a probabilistic modelling framework rooted in Item Response Theory (IRT) that achieves over 80 percent accuracy, setting a new benchmark for mathematics prediction accuracy of formal exam papers. Extending this, our collaborative filtering models incorporate topic-level skill profiles, but reveal a surprising and educationally significant finding, a single latent ability parameter alone is needed to achieve the maximum predictive accuracy. Our main contribution though is deriving and implementing a novel discrete variational inference framework, achieving our highest prediction accuracy in low-data settings and outperforming all classical IRT and matrix factorisation baselines.

Introduction

Everyone remembers an amazing teacher from their childhood. Such individuals possess a passion for their subject and the ability to convey complex ideas. However, one of the most important and sometimes overlooked skills that incredible teachers have is to be able to present the student's work at exactly the right level. This principle is echoed in Duolingo's design philosophy, where the platform aims for learners to answer around 80 percent of questions correctly. This level of difficulty is challenging enough to sustain engagement, but not so difficult as to become discouraging, occupying the so-called Goldilocks zone of learning ([Bjork 1994], [Csikszentmihalyi 1990]).

AI personal tutors hold the promise of helping to redress educational inequality, an inequality driven in large measure by unequal access to high quality personal tutoring. Although wealthier families can afford such individualised support, the majority of students cannot, exacerbating long-standing social divides. For AI tutoring platforms to succeed, their effectiveness will in part depend on whether they can deliver questions at just the right level to keep students engaged and stretched, just as the very best human tutors do.

Our aim in this paper is simply to predict whether a student will get an unseen question correct, given their previous question answers, other students' responses, and meta-data. We hereby release the largest open benchmark of marked mathematics exam paper responses to date, and create models to predict a student's question success. Moreover, educational platforms regularly have the problem of limited data, limited by the number of students for new platforms and for all platforms limited by the number of questions brand new students have answered. We produce novel methods to predict accurately in these limited data scenarios.

A natural place to start is a model incorporating the ability of the student and the difficulty of a question, a two-parameter model. We achieve remarkable accuracy with this simple ability-difficulty model. So much so that when we introduce a more complex interaction model allowing for students' individual strengths and weaknesses in certain topics of mathematics, such as algebra, data and geometry, we are unable to beat it. This suggests a potentially important result for education: that a single latent ability parameter may be the main factor driving a student's success on exam questions, and perhaps the only one needed. In our search for factors that improve prediction accuracy, we find some benefit in modelling shared topic-level strengths and weaknesses across students in the same high school class, often taught by a single teacher. We also uncover interesting interpretability results, showing that the model implicitly discovers meaningful question types such as algebra or geometry as a byproduct of its optimisation for predictive accuracy.

A major challenge for new educational platforms is the cold-start problem, having too few students on the system to make accurate predictions. It is in this low data regime that we make our most significant breakthrough. We derive a novel discrete variational inference framework, placing Gaussian priors over latent ability variables, which outperforms all our other models when student data is limited. For completeness, we also address the challenge of predicting performance for new students who have only attempted a small number of questions. We apply pool-based active learning to achieve significantly higher predictive accuracy than random selection and we also identify which questions are most informative to present, guiding personalised assess-

ment.

The main contributions of our work include:

- **When data are scarce our novel Variational Inference for discrete data yields improved accuracy** Placing uncertainty over all latent parameters delivers our most important results.
- **Open-sourcing a New Large-Scale High Quality Educational Dataset** We release the largest open dataset of professionally marked formal mathematics exam responses to date, providing a valuable benchmark.
- **Ability is all you need:** Remarkably we discover that a single parameter of student ability cannot be beaten by more complex models incorporating individual topic strengths and weaknesses.

Related Work

Item Response Theory (IRT) in Education. IRT offers a probabilistic framework for modeling the probability that a learner answers an item correctly as a function of latent traits. The simplest variant—the Rasch 1PL model—uses a single student-ability and item-difficulty parameter with a logistic link [Rasch 1960]. Classic extensions add item discrimination and guessing parameters [Birnbaum 1968]. Foundational psychometric texts firmly established IRT for large-scale assessment [Hambleton, Swaminathan, and Rogers 1991]. Early educational technology work showed its practical value: Chen and Huang [2005] used IRT to adapt e-learning content difficulty and demonstrated improved engagement. Modern computerised adaptive testing likewise selects questions in real time based on IRT estimates to maximise information gain.

Collaborative Filtering and Matrix Factorisation. Viewing students as "users" and questions as "items" yields a natural analogy to recommender systems. Matrix-factorization (MF) methods learn low-rank embeddings for students and items and, with a logistic link, are algebraically similar to Rasch. MF has repeatedly outperformed feature-engineered approaches on benchmarks such as the KDD Cup 2010; e.g. Thai-Nghe et al. [2011] reported lower prediction error than per-skill models. MF also generalises to grade prediction across courses [Polyzou and Karypis 2017]. Its flexibility allows side features or temporal components via factorization machines or tensor factorization. Notably, Vie, Kashima, and Yamada [2019] introduced *Knowledge Tracing Machines*, unifying IRT, logistic regression and performance-factor models in a factorization-machine framework that achieved state-of-the-art accuracy while remaining interpretable.

Knowledge Tracing and Sequential Models. Standard IRT and MF treat ability as static, whereas knowledge tracing (KT) models its temporal evolution. Bayesian Knowledge Tracing [Corbett and Anderson 1994] and performance-factor models are long-standing baselines. Deep KT began with recurrent networks (DKT) [Piech et al. 2015] and now includes memory networks [Zhang et al. 2017] and attention mechanisms [Ghosh and Heffernan 2020]. However, simpler latent-factor methods can rival deep KT when tuned well; Vie, Kashima, and Yamada [2019] outperformed DKT on Duolingo data. The NeurIPS 2020 Education Challenge attracted hybrid solutions mixing sequential models with meta-learning and CF components [Wang et al. 2021], underscoring a trend toward combining deep learning with probabilistic and factor models.

	Question 1	Question 2	...	Question 70
Max	5	2	...	3
Topics	Pythagoras	Fractions	...	Scattergraphs
Problem solving level	1	1	...	3
Student 1	4	1	...	2
Student 2	3	0	...	1
...	...	...	...	...
Student 49521	5	2	...	3

Figure 1: Illustration of raw data

Positioning of Our Work We introduce the first discrete variational inference framework for predicting binary student responses, achieving state-of-the-art accuracy in both full and low-data regimes. Our approach builds on interpretable IRT and collaborative filtering, while remaining effective when student data is scarce. We evaluate on the largest open dataset of marked mathematics exams to date and show that variational inference improves predictions. We also show that class effects are important and show the model is able to uncover human interpretable skill dimensions.

The Data

Our data is provided by a large UK online learning platform that provides targeted mathematics revision for students. The data contains anonymous student scores (say 1 mark out of 3) on questions from GCSE (16 year old exams) mathematics mock papers they attempted. The mock exams are sat under formal exam conditions (closed book) by the students and marked accurately by qualified teachers, which is often peer-reviewed by other teachers.

The raw dataset is illustrated in Figure 1, where the orange region represents the question metadata and the blue region represents students' raw scores.

Specifically, our models have been evaluated on data from the Edexcel GCSE 2017 Mathematics Paper 1, 2, and 3, containing 70 questions in total (24qns, 24qns, 22qns) and involving 49521 students. While most students have completed all 70 questions, some may have completed only one of the three papers, so there are fewer observed data points available for these students. The real data set may therefore be represented as a sparse matrix as illustrated in Figure 2, where the columns represent different questions and the rows represent different students, with unobserved data points spread over the data matrix.

The paper will focus on probabilistic models on predicting correctness of students' responses and hence scores are binarised: a response is marked as 1 (correct) if the student achieved more than half the available marks for that question, and 0 (incorrect) otherwise.

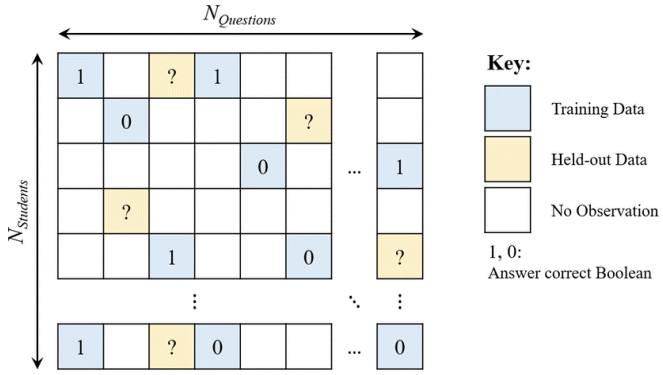

Figure 2: Illustration of preprocessed data

The Model

We consider an Item Response Theory (IRT) model in the form of a hidden latent model, where student abilities and question difficulties can be regarded as the unobserved latent that are discovered from the parameters of the model. [Magno 2009].

We look at separate formulations within the same family of models.

Ability–Difficulty Model (IRT)

The simplest model, the ability-difficulty models takes the form of a logistic function described by the following:

$$p(y_{sq} = 1|\phi) = \frac{1}{1 + e^{-(b_s + b_q)}} = \sigma(b_s + b_q) \qquad (1)$$

This is also known as the Rasch model, where the probability that a question is answered correctly by a student is now calculated through two sets of latent parameters: b_s representing the general ability of the student s and b_q representing the overall difficulty of the question q. A large positive b_s implies a high-performing student whereas a large positive b_q implies a simple question and vice versa. The parameters are optimised by stochastic gradient descent.

The Interaction Model

On top of a general ability level, students may be strong in certain areas such as algebra and weak in others such as geometry. For example, two students who receive the same mark on an exam may have achieved it through different areas of mathematics.

We therefore extend the unidimensional IRT model, which can be extended to multidimensional (MIRT) models, to explore the interactions of multiple abilities, which is of particular interest in educational contexts in uncovering different skill dimensions possessed by students and required for solving specific questions [Johannes Hartig 2022].

We now have an overall ability level for each student, along with variables that can account for their strengths and weaknesses across different areas of mathematics. Similarly, we define an overall difficulty level for each question, as well as variables indicating which mathematical topics the question involves. These variables span n dimensions; for example, three dimensions could loosely correspond to algebra, statistics, and geometry.

Formally we have:

$$\mathbf{b}_s = (b_{s0}, 1, b_{s1}, \ldots, b_{sD}), \quad \mathbf{b}_q = (1, b_{q0}, b_{q1}, \ldots, b_{qD})$$

so that the probability becomes:

$$p(y_{sq} = 1 \mid \phi) = \sigma(b_{s0} + b_{q0} + \sum_{d=1}^{D} b_{sd}\, b_{qd}) \qquad (2)$$

where b_{s0} and b_{q0} are synonymous with the overall student-ability and question-difficulty bias parameters in the Ability–Difficulty Model. b_{sd} represents the different proficiency a student may have in different areas, and b_{qd} is the question vector that contains the respective latents describing the question's involvement in each of the cognitive areas [Sandip Sinharay 2007].

The Class Interaction Model

The class interaction model is simply the interaction model with the individual skill levels of the students, b_{sd}, constrained to be the same for all students in a high school class. For example, all students in 11MA2 at say Springfield High School share the same b_{sd}.

This model allows us to examine whether a teacher of a certain class may have taught a particular topic very well or poorly, or may not have covered another topic at all.

The class interactive model is given by:

$$\mathbf{b}_s = (b_{s0}, 1, b_{s1}, \ldots, b_{sD}), \quad \mathbf{b}_q = (1, b_{q0}, b_{q1}, \ldots, b_{qD})$$

such that the probability becomes:

$$p(y_{sq} = 1 \mid \phi) = \sigma(b_{s0} + b_{q0} + \sum_{d=1}^{D} b_{cd}\, b_{qd}) \qquad (3)$$

Here, b_{s0} and b_{q0} are again synonymous with the overall student-ability and question-difficulty bias parameters from the Ability-Difficulty Model. However, each student in the same class c now shares an identical b_{cd}, where $c \in 1, 2, \ldots, n$ represents the different areas of strength and weakness for class c, and there are n unique classes.

Variational Inference

To be able to predict accurately for new platforms which have a low amount of students, we leverage variational inference treating student ability as a probability distribution. This approach allows us to capture uncertainty and individual variability in a principled way, rather than forcing a point estimate for each learner. What makes our application novel is that we derive a variational model within a discrete response setting, rather than the continuous domains found in the traditional variational inference literature. This requires special handling in both the formulation and optimisation.

Specifically instead of point estimates we place independent Gaussian variational posteriors on every latent parameters:

Algorithm 1 Discrete variational inference for binary responses for the class interaction model

Inputs, responses $y_{s,q} \in \{0,1\}$, answered item sets $\mathcal{Q}_s$, samples M, step size η, prior $p(b) = \mathcal{N}(0,1)$
Initialise, μ_s, σ_s for all students, b_q for all question items
Initialise, μ_{cd}, σ_{cd} for all class interaction terms, b_{qd} for all question interaction items
repeat
  $\mathcal{L} \leftarrow 0$
  for each student s **do**
    draw $\epsilon_s^{(1)}, \ldots, \epsilon_s^{(M)} \sim \mathcal{N}(0,1)$
    set $b_s^{(m)} \leftarrow \mu_s + \sigma_s\, \epsilon_s^{(m)}$
    draw $\epsilon_{cd}^{(1)}, \ldots, \epsilon_{cd}^{(M)} \sim \mathcal{N}(0,1)$ for all c, d
    set $b_{cd}^{(m)} \leftarrow \mu_{cd} + \sigma_{cd}\, \epsilon_{cd}^{(m)}$ for $m = 1, \ldots, M$
    $\widehat{\ell}_s \leftarrow \frac{1}{M} \sum_{m=1}^{M} \sum_{q \in \mathcal{Q}_s} \Big[y_{s,q} \log \sigma\big(b_s^{(m)} + b_q + \sum_d b_{c_s d}^{(m)}\, b_{qd}\big) + (1 - y_{s,q}) \log\big(1 - \sigma\big(b_s^{(m)} + b_q + \sum_d b_{c_s d}^{(m)}\, b_{qd}\big)\big) \Big]$
    $\mathrm{KL}_s \leftarrow \log \frac{1}{\sigma_s} + \frac{\sigma_s^2 + \mu_s^2}{2} - \frac{1}{2}$
    $\mathcal{L} \leftarrow \mathcal{L} + \widehat{\ell}_s - \mathrm{KL}_s$
  end for
  $\mathrm{KL}_{\text{class}} \leftarrow \sum_{c,d} \left(\log \frac{1}{\sigma_{cd}} + \frac{\sigma_{cd}^2 + \mu_{cd}^2}{2} - \frac{1}{2} \right)$
  $\mathcal{L} \leftarrow \mathcal{L} - \mathrm{KL}_{\text{class}}$
  Update $\{\mu_s, \sigma_s\}$, $\{\mu_{cd}, \sigma_{cd}\}$, $\{b_q\}$, and $\{b_{qd}\}$ by ascending $\nabla \mathcal{L}$ with gradient descent
until convergence
return variational parameters $\{\mu_s, \sigma_s\}$, $\{\mu_{cd}, \sigma_{cd}\}$ and item parameters $\{b_q\}$, and $\{b_{qd}\}$

$$q(b_s) = \mathcal{N}(b_s \mid \mu_s, \sigma_s^2), \quad q(b_{cd}) = \mathcal{N}(b_{cd} \mid \mu_{cd}, \sigma_{cd}^2),$$

where the parameters $\mu_s, \sigma_s^2, \mu_{cd}, \sigma_{cd}^2$ are optimised during training.

The variational objective is to maximise the evidence lower bound (ELBO), which balances the expected log-likelihood with a KL divergence regulariser:

$$\text{ELBO} = \mathbb{E}_q[\log p(\mathbf{y} \mid \mathbf{b})] - \mathrm{KL}(q(\mathbf{b}) \,\|\, p(\mathbf{b})).$$

Algorithm 1 walks through our technique for the class interaction model, as this gains our greatest prediction accuracy. We are variational over student parameters only, including the class interaction terms, but all question parameters remain as point estimates. The full novel derivation, sampling scheme, and closed-form expressions are provided in Appendix.

Results and Discussion

The first thing to notice in our results shown in Table 1 is the **strong predictive power** of the single parameter ability difficulty model.

Our model achieved a precision of 80.8% in our dataset, higher than the winning accuracy of 74.74% in the NeurIPS 2020 Education Challenge [Ghosh, Ghosh, and Lipton 2020]. This indicates both the power of the model and the quality of the dataset.

Results	1 Exam	3 Exams
Students	35,514	49,317
Questions	24	70
Ability Difficulty	80.8	80.6
Interaction (1D)	80.7	80.7
Class Interaction	82.1 (18 Dims)	81.9 (35 Dims)
InteractionVI (1D)	80.9	80.7
Class InteractionVI	82.15 (18 Dims)	80.8

Table 1: Model performance across models and dataset sizes

To evaluate whether the ability–difficulty model could be outperformed, we swept over many dimensions for the interaction model. **Surprisingly**, we did not find any setting in which the interaction model exceeded the ability–difficulty model (see Appendix). We further evaluate the interaction model using synthetic data and demonstrate that, even when only weak topic-level strengths and weaknesses are seeded into the simulated students, the model still succeeds in uncovering them to improve prediction accuracy. (see appendix)

This leads us to our most significant insight for education:

Ability alone may be the sole determinant of success prediction.

The concept that groups of students of the same overall ability may be strong in certain areas, such as algebra, and weak in others, such as geometry, may not be correct. For example, it is widely believed that students at a certain level (eg: Grade B) falls into certain groups. Some are strong in algebra and weak in geometry, and others have the opposite strengths. For example, studies such as Usiskin (1982) have shown that students often exhibit strong algebraic skills while remaining at relatively low levels of geometric reasoning [Usiskin 1982].

However, we are led to the conclusion that one parameter only, the ability of a student, is perhaps the only significant predictor of question success.

This result could have far-reaching effects on education. If general mathematical ability is the only reliable predictor of exam performance, the curricula should shift the focus from isolated subtopics to building broad reasoning and problem-solving skills. Teacher training should emphasise fostering transferable cognitive skills that underpin mathematical ability across topics rather than improving individual topics in isolation. Intervention strategies should also target general mathematical development rather than narrow skill gaps in areas like algebra.

While ability certainly dominates, there could well be other factors at play that are hard to find, although these must be much smaller. We worked hard to identify any such factors, exploring several avenues and model variations. We found one such factor that modestly improves prediction by 1.3 percent, only by introducing data on which high school class students were in. The **class interaction model**, which groups students by their school classes, out-

Dataset Size	CI (%)	CIVI (%)
Full Dataset (35,514)	82.1	82.1
50% Dataset (17,757)	81.2	81.2
25% Dataset (8,878)	79.6	79.7
15% Dataset (5,327)	78.7	79.4

Table 2: Accuracy comparison using Standard Class Interaction (CI) and the Class Interaction Variational Inference model (CIVI) across dataset sizes for one exam

performs the ability-difficulty model. This suggests that the model can, for example, detect where a class teacher is particularly strong or weak in teaching certain topics, thereby enhancing its predictive power.

Smaller Datasets Results: Discrete Variational Inference improves accuracy

While we are fortunate to have access to a large dataset, many emerging educational platforms may not share this advantage. Predicting student ability in the early stages of such platforms presents a significant challenge due to limited data availability. It is within the class interaction model that the effectiveness of our discrete variational inference approach becomes most evident, as the presence of significant interaction effects allows our discrete variational model to achieve improved predictive accuracy, in low data settings.

To simulate this scenario, we evaluated our method on progressively smaller subsets of the dataset, shown in table 2. Notably, our novel discrete variational inference begins to yield significant improvements in predictive accuracy once we reach 15 percent of the full dataset, equivalent to approximately 5,000 students. Here the 0.7 percent increase is equivalent to 840 more correct predictions, a statistically significant increase at a 1 percent significance level (see appendix).

And so here we see our new variational inference model can obtain greater prediction accuracy for smaller datasets. Note that these improvements are achieved by carefully fine-tuning the hyperparameters. Specifically, by initialising the standard deviations of student ability to values greater than 0.8, and by using a warm start for the parameters, initialised from the interaction model at its point of optimal predictive accuracy.

This suggests that, under appropriate conditions, our discrete variational inference can be a powerful tool for enhancing predictive performance in low-data regimes for educational platforms.

Interpretability: Why does the class interaction model improve accuracy?

We aim to investigate the factors contributing to the improvement in predictive accuracy observed when employing the Class Interaction (CI) model. To this end, we analyse the individual question embedding vectors b_{qd} for all 24 questions generated by the best-performing model configuration (18 dimensions) on a single exam paper. Specifically, we compute the normalised cosine similarity between these vectors to explore the structure the model has learned.

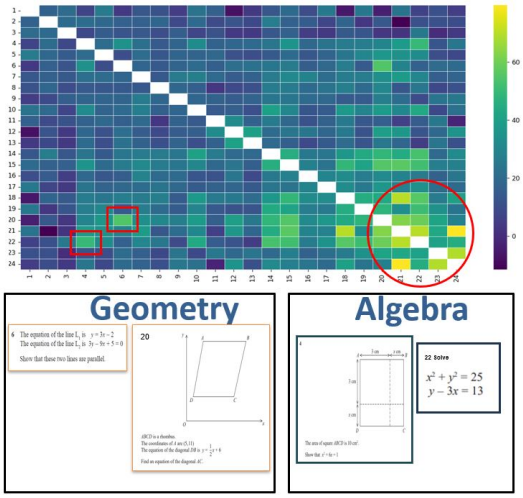

Figure 3: Normalised cosine similarity of question vectors b_{qd}. The lower left rectangle shows the model has discovered an underlying factor representing geometry, the other Algebra. The red circle shows the model has understood that only high ability classes actually get taught the hardest topics.

Our analysis reveals that the model appears to identify latent thematic structures across the test items, effectively uncovering implicit *question types*. For example, as illustrated in Figure 3, questions 4 and 20 exhibit high cosine similarity, and upon inspection, both are found to assess geometric reasoning. Similarly, questions 6 and 22 are both algebraic in nature, despite differing in difficulty. The model nonetheless assigns them closely aligned vectors, indicating it has captured an underlying algebraic theme. This pattern is repeated across several other question groupings and has been verified by fully qualified mathematics teachers.

The cosine similarity also appears to show the model has picked up on the fact that only strong ability classes are taught the hardest topics - elements of the $b_{qd}, d > 20$ vector are effectively encoding a latent query: 'Does this student belong to a high-ability class exposed to advanced topics?' and a corresponding element in the b_{cd} answering yes or no.

Active Learning

Having tackled prediction under data scarcity when a platform has few students, we now turn to the complementary problem: forecasting question success for entirely new students as soon as they join the system.

Thousands of new learners join educational platforms each day, yet most arrive with only a handful of answers on record. To cope with this data-sparse reality, we adopt pool-based active learning where the model estimates predictive uncertainty and queries the least-certain student–question pairs, updates its parameters and repeats.

Our experiments partition the data into an initial labelled set, a large unlabelled pool of new students, and a held-out test set. We first fit the Ability–Difficulty model on a limited number of initial questions. We take its uncertainty over

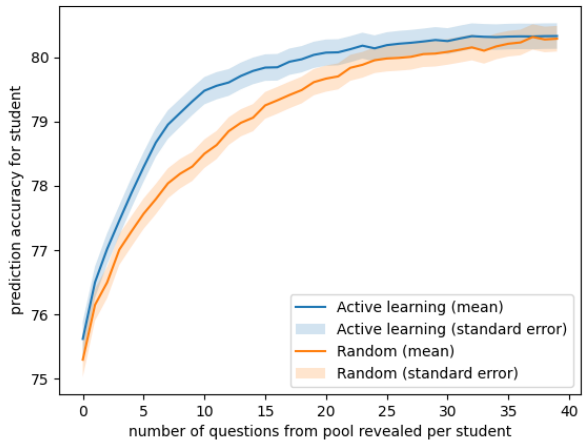

Figure 4: Prediction accuracy of student averaged across all new students in data pool plotted against the number of questions per student so far revealed to the model

each unanswered question, and iteratively choose the next data point to select based on which questions predicted probability is closest to 0.5. Retraining after every batch of newly revealed labels prevents catastrophic forgetting and lets the model refine its ability estimates on the fly.

Results on 2000 unseen students in figure 4 show that active learning surpasses random querying after only ten questions, stabilising at a one-percentage-point accuracy lift and converging with roughly 30 questions per student—far fewer than passive baselines. Gains hold across low-, medium- and high-ability groups, with the largest relative improvement in the mid-ability cohort where prediction is hardest, demonstrating faster and more economical personalisation for intelligent tutoring systems.

Limitations and Future Directions

- **Static-ability assumption.** The current models treat each student's ability as constant over the exam window, overlooking learning or fatigue effects. In the future we will look to extend the discrete VI framework to temporal knowledge tracing so ability can evolve across sessions.
- **Single-subject, UK-centric dataset.** Our benchmark consists solely of GCSE mathematics sat by UK pupils, limiting generalisability to other subjects or regions. In the future we plan to replicate the study with multilingual, multi-subject corpora to assess robustness and fairness across diverse cohorts.
- **Computational overhead of Variational Inference** Full Gaussian posteriors increase training time and memory relative to point-estimate baselines. In the future we will explore lighter variational families and amortised inference to keep the method practical for large-scale, real-time tutoring platforms.

Ethics Statement

All data were de-identified before analysis, with no personal identifiers. We will release the dataset only in a form that preserves anonymity and complies with applicable data-protection obligations.

Conclusion

We release the largest open, professionally marked GCSE mathematics dataset to date and show that a simple, interpretable ability–difficulty model already achieves strong accuracy. We introduce topic level skill variables in areas such as geometry and algebra for each student, but remarkably this fails to increase prediction power, suggesting that individual topic strengths contribute little once overall ability is accounted for. Many new platforms are unable to predict accurately due to lack of student and our main contribution is then a novel discrete variational-inference framework that delivers the biggest gains when data are scarce for new platforms and class information is available.

Overall, we release a large open-source dataset, develop a series of models that uncover a potentially pivotal piece of evidence for education, and present a novel system for dealing with data scarcity in educational settings.

References

Birnbaum, A. 1968. Some Latent Trait Models and Their Use in Test Construction. In Lord, F. M.; and Novick, M. R., eds., *Statistical Theories of Mental Test Scores*, 397–479. Reading, MA: Addison-Wesley.

Bjork, R. A. 1994. Memory and Metamemory Considerations in the Training of Human Beings. In Metcalfe, J.; and Shimamura, A. P., eds., *Metacognition: Knowing About Knowing*, 185–205. MIT Press.

Chen, C.-C.; and Huang, M.-C. 2005. Personalized Curriculum Sequencing Using Item Response Theory. In *Proceedings of the 5th IEEE International Conference on Advanced Learning Technologies (ICALT)*, 158–162.

Corbett, A. T.; and Anderson, J. R. 1994. Knowledge Tracing: Modeling the Acquisition of Procedural Knowledge. *User Modeling and User-Adapted Interaction*, 4(4): 253–278.

Csikszentmihalyi, M. 1990. *Flow: The Psychology of Optimal Experience*. Harper and Row.

Ghosh, A.; Ghosh, S.; and Lipton, Z. C. 2020. Meta-Learning Student Response Prediction. In *NeurIPS 2020 Education Challenge*. Task 4 Winner: 74.74% accuracy.

Ghosh, A.; and Heffernan, N. 2020. Context-Aware Attentive Knowledge Tracing. In *Proceedings of the 34th AAAI Conference on Artificial Intelligence*, 7342–7349.

Hambleton, R. K.; Swaminathan, H.; and Rogers, H. J. 1991. *Fundamentals of Item Response Theory*. Newbury Park, CA: Sage.

Johannes Hartig, J. H. 2022. Multidimensional IRT models for the assessment of competencies. *Studies in Educational Evaluation*, 35.

Magno, C. 2009. Demonstrating the difference between classical test theory and item response theory using derived test data. *The International Journal of Educational and Psychological Assessment*, 1(1): 1–11.

Piech, C.; Bassen, J.; Huang, J.; Ganguli, S.; Sahami, M.; Guibas, L.; and Sohl-Dickstein, J. 2015. Deep Knowledge Tracing. In *Advances in Neural Information Processing Systems*, volume 28.

Polyzou, A.; and Karypis, G. 2017. Grade Prediction with Models Specific to Students and Courses. *International Journal of Data Science and Analytics*, 3(3): 159–172.

Rasch, G. 1960. *Probabilistic Models for Some Intelligence and Attainment Tests*. Copenhagen: Danmarks Paedagogiske Institut.

Sandip Sinharay, R. A. 2007. Assessing Fit of Cognitive Diagnostic Models A Case Study. *Educational and Psychological Measurement*, 67(2).

Thai-Nghe, N.; Drumond, L.; Horv
'ath, T.; Krohn-Grimberghe, A.; and Schmidt-Thieme, L. 2011. Factorization Techniques for Predicting Student Performance. In *Proceedings of the 4th International Conference on Educational Data Mining*, 55–66.

Usiskin, Z. 1982. *Van Hiele Levels and Achievement in Secondary School Geometry*. Chicago: University of Chicago, Department of Education. Final report of the Cognitive Development and Achievement in Secondary School Geometry Project.

Vie, J.-J.; Kashima, H.; and Yamada, M. 2019. Knowledge Tracing Machines: Factorization Machines for Knowledge Tracing. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, 750–757.

Wang, Z.; Saveliev, E.; Lamb, A.; Hern
'andez-Lobato, J. M.; Turner, R. E.; and et al. 2021. NeurIPS 2020 Education Challenge: Competition Report. In *Proceedings of the NeurIPS 2021 Competition and Datasets Track*.

Zhang, J.; Shi, X.; King, I.; and Yeung, D.-Y. 2017. Dynamic Key-Value Memory Networks for Knowledge Tracing. In *Proceedings of the 26th International Joint Conference on Artificial Intelligence*, 2315–2321.

(A) Discrete Variational Inference Derivation

We retain the Bernoulli likelihood in its exact discrete form, we do not introduce a continuous surrogate, we do not bound or augment the logistic term, and we optimise the expected Bernoulli log-likelihood under a Gaussian variational posterior. This keeps the derivation faithful to binary exam responses while remaining simple and scalable.

Prior and variational family We place independent standard normal priors on student abilities,

$$b_s : \quad p(b_s) \sim \mathcal{N}(b_s; 0, 1)$$

For each observed pair (s, q) with binary response $y_{sq} \in \{0, 1\}$,

$$p\Big(\{b_s\}_{s=1}^{S} = 1, \Big| \{y_{s,q}\}_{s=1,q=1}^{S,Q}\Big) \approx \prod_{s=1}^{S} q_s(b_s),$$

$$q_s(b_s) = \mathcal{N}\big(b_s\,;\,\mu_s, \sigma_s^2\big)$$

$$\text{ELBO} = \log p\left(\{y_{s,q}\}_{s=1,q=1}^{S,Q}\Big|\{b_q\}_{q=1}^{Q}\right) - \text{KL}\left[q\left(\{b_s\}_{s=1}^{S}\right) \,\|\, p\left(\{b_s\}_{s=1}^{S}\Big|\{y_{s,q}\}_{s=1}^{S}\right)\right]$$

The evidence lower bound (ELBO) is

$$= \mathbb{E}_q\left[\log p\left(\{y_{s,q}\}_{s=1}^{S}\Big|\{b_q\}_{q=1}^{Q}, \{b_s\}_{s=1}^{S}\right)\right] - \text{KL}\left(q\left(\{b_s\}_{s=1}^{S}\right) \,\|\, p\left(\{b_s\}_{s=1}^{S}\Big|\{y_{s,q}\}_{s=1}^{S}\right)\right)$$

$$= \sum_{s=1}^{S} \mathbb{E}_{q_s(b_s)}\Bigg[\log p\left(\{y_{s,q}\}_{q=1}^{Q}\Big|\{b_q\}_{q=1}^{Q}, \{b_q\}_{s=1}^{S}\right) - \text{KL}\left(q_s(b_s) \,\|\, p_s(b_s)\right)\Bigg]$$

$$= \mathbb{E}_{q_s(b_s)}\left[\sum_{q=1}^{Q_s} \log\left(\left(\frac{e^{b_s+b_q}}{1+e^{b_s+b_q}}\right)^{y_{sq}}\left(\frac{1}{1+e^{b_s+b_q}}\right)^{1-y_{sq}}\right)\right] - \text{KL}\left(\mathcal{N}(b_s; \mu_s, \sigma_s^2) \,\|\, \mathcal{N}(b_s; 0, 1)\right)$$

We keep the Bernoulli likelihood exact and move stochasticity to the continuous latent by

$$b_s^{(m)} = \mu_s + \sigma_s\,\epsilon_m, \quad \epsilon_m \sim \mathcal{N}(0, 1)$$

A Monte Carlo estimator with M samples gives

$$\text{ELBO} \approx \frac{1}{M}\sum_{s=1}^{S}\sum_{m=1}^{M}\sum_{q=1}^{Q_s}\left[y_{sq}\left(b_s^{(m)} + b_q\right) - \log\left(1 + e^{b_s^{(m)}+b_q}\right)\right]$$

$$-\sum_{s=1}^{S} \text{KL}\left(q(b_s) \,\|\, p(b_s)\right)$$

$$\text{where} \quad b_s^{(m)} = \mu_s + \sigma_s\epsilon_s^{(m)}, \quad \epsilon_s^{(m)} \sim \mathcal{N}(0, 1)$$

$$\text{and} \quad \text{KL}\left(\mathcal{N}(x_j; \mu_1, \sigma_1^2) \,\|\, \mathcal{N}(x_j; \mu_2, \sigma_2^2)\right) = \log\frac{\sigma_2}{\sigma_1} + \frac{\sigma_1^2 + (\mu_1 - \mu_2)^2}{2\sigma_2^2} -$$

We optimise $\{\mu_s, \sigma_s\}$ and $\{b_q\}$ by gradient ascent . In practice we initialise σ_s above a small positive floor to reflect uncertainty under scarce data.

■

B Two–proportion z-test for the significance of the class interaction model with variational inference improvement over standard class interaction model for 15 percent of the dataset ($\alpha = 0.01$)

$$n_1 = n_2 = 120\,000, (number of questions) \quad x_1 = 94\,440,$$
$$x_2 = 95\,280,$$
$$\hat{p} = \frac{x_1 + x_2}{n_1 + n_2} = 0.7905,$$
$$\mathrm{SE} = \sqrt{\hat{p}(1-\hat{p})\left(\tfrac{1}{n_1} + \tfrac{1}{n_2}\right)} = 0.00166,$$
$$z = \frac{p_2 - p_1}{\mathrm{SE}} = \frac{0.794 - 0.787}{0.00166} = 4.21,$$
$$z_{0.99} = 2.576, \qquad p \approx 2.5 \times 10^{-5} < 0.01.$$

Result: $|z| > z_{0.99}$;
the 0.7 pp improvement is statistically significant at the 1 % level.

(C) Simulation Demonstrating Interaction Model Recovery

To demonstrate that our interaction model can successfully learn underlying skill dimensions when they exist, we conducted a controlled simulation with synthetic data using the same scale as our real dataset
(40,000 students and 24 questions).

We generated student and question latent variables using the following formulation:

- Student ability biases $b_s \sim \mathcal{N}(0, 1)$
- Question difficulty biases $b_q \sim \mathcal{N}(-3, 1)$
- Student skill vectors $x_s \sim \mathcal{N}(0, 1)$ in $D = 1$ dimension
- Question skill demands $x_q \sim \mathcal{N}(0, 1)$ in $D = 1$ dimension

We simulated binary responses using this model and trained both our ability–difficulty model and the 1D interaction model on the resulting dataset.
As expected, the interaction model achieved higher prediction accuracy:

- Ability–difficulty model: **95.54%**
- 1D interaction model: **96.20%**

This result confirms that our interaction model can recover underlying topic-level structure and yield improved predictions when meaningful interactions between student strengths and question demands are present.