Good Fit Bad Policy: Why Fit Statistics are a Biased Measure of Knowledge Tracer Quality

Napol Rachatasumrit[1*][0000−0002−7183−8789], Daniel Weitekamp[1*][0000−0003−0079−8000], and Kenneth R. Koedinger[1][0000−0002−5850−4768]

Carnegie Mellon University, Pittsburgh, PA, 15213
{napol, weitekamp, koedinger}@cmu.edu

Abstract. Knowledge tracers are typically evaluated on the basis of the goodness-of-fit of their underlying student performance models. However, for the purposes of supporting mastery learning the true measure of a good knowledge tracer is not its goodness-of-fit, but the degree to which it optimally selects next problem items. In this context, a knowledge tracer should minimize under-practice to ensure students master learning materials and minimize over-practice to reduce wasted time. Prior work has suggested that fit-statistic-based measures of knowledge tracer quality may misrank the relative quality of knowledge tracers' item selection. In this work, we evaluate this claim by measuring over- and under-practice directly in synthetic data drawn from ground-truth learning curves. We conduct an experiment with 3 well-known student performance models: Performance Factor Analysis (PFA), BestLR, and Deep Knowledge Tracing (DKT), and find that in 43% of the synthetic datasets, the models with higher measures of overall predictive performance (e.g. AUC and MSE) were worse than a comparison model with a lower predictive performance at minimizing over-practice and under-practice. These results support the hypothesis that overall fit statistics are not a reliable measure of a knowledge tracer's ability to optimally select next items for students, and bring into question the validity of traditional methods of knowledge tracer comparison.

Keywords: Knowledge Tracing · Mastery Learning · Intelligent Tutoring System

1 Introduction

Student performance models estimate the probability that students will correctly answer the next question items given their prior correct and incorrect responses and serve both online and offline roles in education. In an online setting student performance models can be used as knowledge tracers to adaptively select next problems based on students' current abilities. In an offline setting, they can be

* These authors contributed equally to this work

used to reveal patterns in student's learning which can be used to make data-driven improvements to instructional materials [3].

Knowledge tracing is the online use of a student performance model to actively estimate students' mastery of individual knowledge components (KCs)—the pre-specified facts, skills and principles which students must understand in order to have mastered a particular domain [6]. Mastery of a KC is typically characterized as the point when a student's predicted chance of correctly answering future question items associated with the KC exceed some preset mastery threshold, typically chosen in the range 85-95% [2]. A knowledge tracer leverages its student performance model to estimate which KCs are mastered and which are not so that it can select the next practice items for students which correspond to unmastered KCs. Thus, the challenge of knowledge tracing is to actively adapt to students as they practice to optimize their use of time—giving them enough practice problems for each KC to ensure full domain mastery, but not more than this to avoid wasting time better spent practicing new material. Thus, the ideal knowledge tracer jointly minimizes over-practice, the number of prescribed practice problems given after the student has reached mastery, and under-practice, the number of practice problems which a student would still need to solve in order to achieve mastery.

Unfortunately, over- and under-practice are not directly measurable quantities. Instead, the relative quality of knowledge tracers is typically compared on the basis of the overall fit of their underlying student performance models to student data. Overall fit statistics take the form $\pi(\hat{y}, y)$ and measure the degree to which the continuous student model predictions $\hat{y}$ are a good approximation of the discrete sequence of binary correctness values $y = y_0, ..., y_n$ (correct=1, incorrect=0) collected from student transaction logs. Prior work has used a variety of fit statistics for knowledge tracer comparisons including Mean-Square Error (MSE), prediction accuracy, log-likelihood, AIC [1], BIC [7], and Area under the receiver operating characteristic curve (AUC).

In this work, we demonstrate that overall fit statistics can in fact be a biased basis for knowledge tracer comparison since there are circumstances where a model's total predictive performance can be improved without any corresponding change in the behavior of a knowledge tracer utilizing that model. A model can fit better without producing any corresponding reduction in the number of over- and under-practice problems experienced by students.

A similar concern, yet one unrelated to the claims of this work, is the debate over interpretable versus non-interpretable student performance models. The last decade of knowledge tracing research has been inclusive of a broader machine learning community which have eschewed traditional models based on Item-Response Theory (IRT) [5], hidden markov models, and logistic regression for uninterpretable yet often performant, deep-learning models. Since black-box models possess more parameters than can be practically interpreted, they are less amenable to generating defensible and actionable insights about student data. Thus proponents of deep-knowledge tracers have typically placed a greater

emphasis on the practical use of their models for online knowledge tracing over their use as tools of offline analysis.

In this work, our main objective is not to make an argument for interpretable or uninterpretable blackbox student models, but to bring into question an assumption held in common by both sides of that debate. We show through simulation that it is possible for a student performance model to fit better than a baseline model but be worse at knowledge tracing. And we contend that this raises serious doubts about whether the collective research project of trying to produce better-fitting student models is necessarily leading to knowledge tracers which are better at mastery-based item selection.

2 Over-Practice and Under-Practice

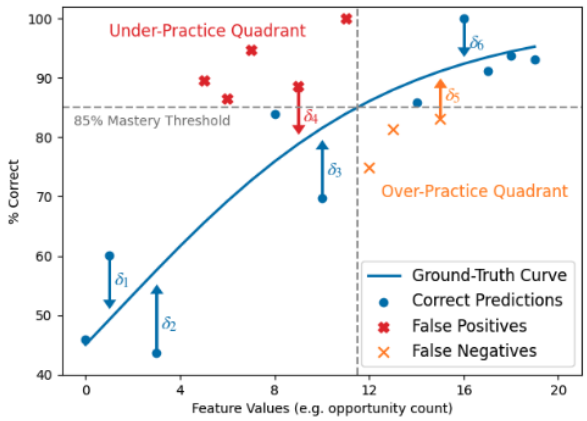

Fig. 1. Illustration of over-practice and under-practice attempts

Although counts of over- and under-practice are not directly measurable from student data, they can be defined relative to a notion of a student's ground-truth learning curve—their true probability of answering next question items correctly at each practice opportunity. Framed in non-stochastic terms, a student's ground truth curve for a given KC represents the degree to which that KC has been mastered at each learning opportunity. It captures the progression of complex cognitive factors beyond the scope of what statistical performance models typically capture. A point along the curve captures the degree to which a student has partially constructed knowledge—a notion that statistical models typically estimated solely from binary observations of correct and incorrect performance.

By reference to a ground-truth learning curve and a choice of mastery threshold, a model's instances of under-practice are those where the performance model predicts performance to be above the mastery threshold when the ground truth

is below it, and the model's instances of over practice are those where it predicts performance to be below mastery when the ground-truth is above the mastery threshold (Figure 1).

Student performance modeling can be framed as estimating students' ground-truth learning curves from the noisy sampling of performance data collected from tutoring system transactions. The logic of comparing knowledge tracers by their overall goodness-of-fit to data is motivated by the idea that an optimal recreation of the ground-truth learning curve should produce an optimal prediction of student mastery. However, this perspective conflates the logic of offline statistical modeling, in which goodness-of-fit can be used to justify hypotheses about students' learning trajectories and their relationship to learning materials, with the narrower aims of online item selection. In this context, a knowledge tracer's purpose is simply to make one critical decision: after a student completes each problem it decides whether to continue prescribing new practice problems with particular KC requirements or not. Thus, certain variations in the predictions of a student performance model simply have no bearing on the real-world quality of their knowledge tracer.

Figure 1 demonstrates how this can be the case by offering an illustration of a hypothetical set of performance model predictions relative to a ground-truth learning curve. The intersection of the ground-truth curve with the mastery threshold divides the figure into 4 quadrants. Predictions in the top-left and bottom-right quadrants are instances where the model would cause under- or over-practice. The dots and x's in Figure 1 represent the predictions $\hat{y}_A$ of a baseline model A. Consider that there is also a comparison model B with predictions $\hat{y}_B = \hat{y}_A + \delta$ perturbed by some δ which brings B closer to the ground truth than A. With these perturbations B's expected overall fit to a sample of the ground-truth curve should be better than model A's. However, only a subset of the shown perturbations would produce improvements in mastery prediction, only those perturbations which move predictions out of the over- and under-practice quadrants (e.g. like δ_4 and δ_5).

A core hypothesis of this work is that the prediction differences between different types of student performance models mostly do not correspond to differences in expected over- and under-practice like perturbations δ_4 and δ_5. Instead, we hypothesize that the majority of model improvements are like δ_1, δ_2, δ_3, and δ_6: inconsequential to levels of over- and under-practice, and generally outside the neighborhood of the ground-truth mastery threshold. One reason to expect this result is that the more data that models have about students the more similar their predictions are likely to be. We expect models to have the greatest difference in their predictions under uncertain circumstances, particularly in early practice attempts when evidence about the student's knowledge is sparse.

To test this hypothesis we utilize synthetic student data to establish ground-truth learning curves. Then we fit various student performance models on the synthetic data and utilize the ground-truth curves to measure over- and under-practice. We evaluate whether the student performance models which produce the least over- and under-practice are also the best fitting models with respect

to overall performance statistics like AUC and MSE. Finally, we graph MSE as a function of ground-truth probability to evaluate whether differences in model fit tend to be greatest within or outside the neighborhood of the mastery threshold.

3 Related Works

Prior work has argued that comparing knowledge tracers on fit statistics alone fails to estimate their relative efficacy on a substantive scale [14]. For instance, to claim that model A achieves a 5% improvement in MSE over model B fails to capture the time savings or post-test performance improvements that would be achieved by utilizing that model for adaptive item selection. Prior attempts to estimate this relationship by simulation [14] and analytically [12], have supported the conclusion that relatively small overall model improvements can yield large reductions in over- and under-practice. However, Weitekamp et al. [12] point out that in theory, it is indeed possible for a better-fitting student performance model to actually perform worse at item selection than a baseline model. The key idea is that the only predictions which matter for item selection are those in the ground-truth neighborhood of the mastery threshold—the region where a knowledge tracer makes its critical decision: to stop prescribing problems for a particular KC or not. Overall fit statistics may produce a biased sense of knowledge tracer quality because they capture the goodness of fit of a performance model on early student transactions which are unambiguously part of the unmastered region. By contrast prediction differences between models in the neighborhood of the mastery threshold are likely to be small since there is typically more supporting evidence from the student transactions preceding it.

4 Methods

We utilize 3 models for synthetic data generation and evaluation: BestLR [4], DKT [9], and PFA [8]. For each dataset, we use each model to create a simulated dataset and evaluate each generated dataset with all 3 models to create a 3x3 experiment. In all cases, we use implementations from Gervet et. al. [4].

Our synthetic data generation works by (1) fitting a generation model to the real data, (2) predicting an error rate for each transaction with a fitted model, and using the predicted value as a ground truth for an error rate in synthetic data, (3) sampling a synthetic outcome for each transaction in the synthetic data based on the corresponding error rate. In this work, we use the same 7 real-world datasets from Gervet et. al. [4], so we generated 21 synthetic datasets for our experiment using 3 generation models For each synthetic dataset, we use random cross-validation splitting by students. The data of 90% of the students are used for training and the data of the other 10% are reserved for the test set. We resample and retrain 5 times for each condition, examining the relative counts of over- and under-practice on the test set between models, and compare this to their relative AUC scores on the test set. We report the average and standard deviation for each metric across replicates.

5 Results and Discussion

Table 1 shows the average instances of over- and under-practice and Table 2 shows the average AUC for each dataset and evaluation model pair. Conventional evaluations assume that between two models the one with the higher predictive performance (e.g. higher AUC) will be the better model—the one expected to make fewer over- and under-practice errors. However, our results demonstrate that this assumption is not always true. We find that in 43% of the synthetic datasets, there are pairs of models where the higher AUC model commits more over- and under-practice errors than the lower AUC model. These results support the hypothesis that overall fit statistics are not a reliable measure of a knowledge tracer's ability to optimally select next items for students, and challenge the credibility of conventional approaches to comparing knowledge tracers.

Table 1. Average numbers of over- and under-practice for each dataset and model

Dataset	Generate	BestLR	DKT	PFA
algebra05	BestLR	**4.577 ± 0.235**	7.261 ± 0.199	10.843 ± 0.433
	DKT	13.164 ± 5.184	**8.300 ± 0.327**	32.522 ± 1.957
	PFA	9.067 ± 0.672	13.116 ± 0.835	**5.028 ± 0.563**
assistments09	BestLR	**3.355 ± 0.078**	4.488 ± 0.181	5.393 ± 0.151
	DKT	7.280 ± 0.151	**4.000 ± 0.107**	9.597 ± 0.265
	PFA	4.258 ± 0.136	5.706 ± 0.246	**3.309 ± 0.184**
assistments15	BestLR	**2.398 ± 0.045**	4.388 ± 0.323	2.961 ± 0.056
	DKT	8.096 ± 0.107	**3.963 ± 0.063**	8.233 ± 0.167
	PFA	**2.377 ± 0.118**	4.997 ± 0.186	2.425 ± 0.043
assistments17	BestLR	**2.638 ± 0.045**	3.567 ± 0.060	5.297 ± 0.085
	DKT	6.334 ± 0.226	**2.808 ± 0.027**	3.614 ± 0.098
	PFA	4.663 ± 0.280	4.738 ± 0.395	**3.495 ± 0.581**
bridge_algebra	BestLR	**3.936 ± 0.094**	5.494 ± 0.217	6.405 ± 0.132
	DKT	14.033 ± 0.368	**6.751 ± 0.165**	22.319 ± 0.712
	PFA	4.762 ± 0.300	6.539 ± 0.200	**3.759 ± 0.218**
spanish	BestLR	**2.447 ± 0.022**	4.213 ± 0.160	3.173 ± 0.083
	DKT	10.798 ± 0.222	**4.600 ± 0.194**	12.701 ± 0.345
	PFA	2.397 ± 0.041	4.324 ± 0.145	**2.109 ± 0.036**
statics	BestLR	**3.962 ± 0.205**	4.263 ± 0.185	10.559 ± 0.442
	DKT	10.379 ± 0.415	**5.095 ± 0.235**	19.067 ± 0.843
	PFA	8.333 ± 0.687	7.565 ± 0.589	**3.743 ± 0.457**

6 Conclusion and Future Works

In this work, we have utilized synthetic data generated by popular knowledge tracers to test whether models with the highest overall fit statistics necessarily produce the best predictions of student mastery. Our method allows us to answer questions of the nature: what is the quality of knowledge tracer X's item

Table 2. Average and SD of AUC for each dataset and evaluation model

Dataset	Generate	BestLR	DKT	PFA
algebra05	BestLR	**0.794 ± 0.002**	0.728 ± 0.004	0.716 ± 0.004
	DKT	**0.808 ± 0.003**	0.764 ± 0.007	0.737 ± 0.002
	PFA	0.689 ± 0.004	0.645 ± 0.004	**0.705 ± 0.002**
assistments09	BestLR	**0.712 ± 0.003**	0.636 ± 0.006	0.653 ± 0.003
	DKT	**0.736 ± 0.005**	0.696 ± 0.004	0.670 ± 0.004
	PFA	0.629 ± 0.003	0.565 ± 0.007	**0.653 ± 0.003**
assistments15	BestLR	**0.721 ±0.005**	0.702 ± 0.006	0.713 ± 0.005
	DKT	0.658 ± 0.001	**0.674 ± 0.002**	0.656 ± 0.001
	PFA	**0.659 ± 0.002**	0.630 ± 0.001	0.659 ± 0.003
assistments17	BestLR	**0.734 ± 0.004**	0.717 ± 0.005	0.654 ± 0.004
	DKT	0.702 ± 0.002	**0.728 ± 0.001**	0.617 ± 0.001
	PFA	0.636 ± 0.002	0.619 ± 0.002	**0.639 ± 0.002**
bridge_algebra	BestLR	**0.834 ± 0.031**	0.780 ± 0.033	0.780 ± 0.034
	DKT	**0.774 ± 0.003**	0.747 ± 0.008	0.705 ± 0.004
	PFA	0.699 ± 0.005	0.645 ± 0.002	**0.715 ± 0.003**
spanish	BestLR	**0.820 ± 0.003**	0.764 ± 0.001	0.811 ± 0.004
	DKT	0.808 ± 0.006	**0.813 ± 0.006**	0.788 ± 0.003
	PFA	0.813 ± 0.006	0.763 ± 0.006	**0.814 ± 0.006**
statics	BestLR	**0.799 ± 0.007**	0.785 ± 0.010	0.661 ± 0.010
	DKT	**0.804 ± 0.005**	0.801 ± 0.004	0.665 ± 0.005
	PFA	0.661 ± 0.005	0.647 ± 0.004	**0.670 ± 0.004**

selection assuming student learning behaves like model Y? Varying models X, Y, and datasets we find that in 43% of the synthetic datasets, models with higher measures of overall predictive performance (i.e. AUC) were worse than a comparison model with a lower predictive performance at minimizing over-practice and under-practice. We conclude that traditional measures of overall performance (e.g. AUC) are in fact not reliable proxies for rates of over- and under-practice. These results raise serious doubts about whether the field of knowledge tracing follows a sound logic of justification when it comes to model comparison.

As in prior works that have utilized synthetic data for analyses of student performance models [11], our method relies upon a theoretical commitment to an underlying model for generating ground-truth curves. Thus our method is not a stand-in replacement for traditional metrics of model fit which evaluate models directly on datasets. Yet, methods which draw comparisons between statistical models and synthetic ground truths have the potential to enable deeper evaluations than the simple notion of that which fits best is best. In this work, we have used statistical performance models as ground-truth generators, but more theory-driven generators such as computational models of learning [13, 10, 12] could be used in their place, to serve as more precise, predictable, and explainable generators of ground-truth learning curves and synthetic data.

Utilizing more controlled theory driven models for data generation could enable more concrete analyses of the sensitivities of different student performance models to individual student differences and domain types, and models' behav-

ior under uncertainty. For instance, while no model in our analyses stood out as decidedly better than the others, in some cases certain models performed better in terms of over- and under-practice on certain datasets. Future work may also include further analysis of the nature of the unproductive predictions that each model commits. For example, investigating the conditions when the models commit those errors and how extreme those errors are could show interesting insights that lead to a better evaluation metric for knowledge tracers.

References

1. H. Akaike. Akaike's information criterion. *International encyclopedia of statistical science*, pages 25–25, 2011.
2. M. Arlin. Time, equality, and mastery learning. *Review of Educational Research*, 54(1):65–86, 1984.
3. H. Cen, K. Koedinger, and B. Junker. Learning factors analysis–a general method for cognitive model evaluation and improvement. In *International conference on intelligent tutoring systems*, pages 164–175. Springer, 2006.
4. T. Gervet, K. Koedinger, J. Schneider, T. Mitchell, et al. When is deep learning the best approach to knowledge tracing? *Journal of Educational Data Mining*, 12(3):31–54, 2020.
5. R. J. Harvey and A. L. Hammer. Item response theory. *The Counseling Psychologist*, 27(3):353–383, 1999.
6. K. R. Koedinger, A. T. Corbett, and C. Perfetti. The knowledge-learning-instruction framework: Bridging the science-practice chasm to enhance robust student learning. *Cognitive science*, 36(5):757–798, 2012.
7. A. A. Neath and J. E. Cavanaugh. The bayesian information criterion: background, derivation, and applications. *Wiley Interdisciplinary Reviews: Computational Statistics*, 4(2):199–203, 2012.
8. P. I. Pavlik Jr, H. Cen, and K. R. Koedinger. Performance factors analysis–a new alternative to knowledge tracing. *Online Submission*, 2009.
9. C. Piech, J. Bassen, J. Huang, S. Ganguli, M. Sahami, L. J. Guibas, and J. Sohl-Dickstein. Deep knowledge tracing. *Advances in neural information processing systems*, 28, 2015.
10. N. Rachatasumrit, P. F. Carvalho, S. Li, and K. R. Koedinger. Content matters: A computational investigation into the effectiveness of retrieval practice and worked examples. In *International Conference on Artificial Intelligence in Education*, pages 54–65. Springer, 2023.
11. N. Rachatasumrit and K. R. Koedinger. Toward improving student model estimates through assistance scores in principle and in practice. *International Educational Data Mining Society*, 2021.
12. D. Weitekamp and K. Koedinger. Computational models of learning: Deepening care and carefulness in ai in education. In *International Conference on Artificial Intelligence in Education*, pages 13–25. Springer, 2023.
13. D. Weitekamp III, E. Harpstead, C. J. MacLellan, N. Rachatasumrit, and K. R. Koedinger. Toward near zero-parameter prediction using a computational model of student learning. *Ann Arbor*, 1001:48105.
14. M. Yudelson and K. Koedinger. Estimating the benefits of student model improvements on a substantive scale. In *Educational Data Mining 2013*, 2013.