Exploring Differences in Performance between Knowledge Tracing Methods & Gaming the System Behavior

Husni Almoubayyed
Stephen E. Fancsali
Carnegie Learning, Inc.
{halmoubayyed, sfancsali}
@carnegielearning.com

ABSTRACT

We report work-in-progress that aims to better understand prediction performance differences between Deep Knowledge Tracing (DKT) and Bayesian Knowledge Tracing (BKT) as well as "gaming the system" behavior by considering variation in features and design across individual pieces of instructional content. Our "non-monolithic" analysis considers hundreds of "workspaces" in Carnegie Learning's MATHia intelligent tutoring system and the extent to which two relatively simple features extracted from MATHia logs, potentially related to gaming the system behavior, are correlated with differences in DKT and BKT prediction performance. We then take a closer look at a set of six MATHia workspaces, three of which represent content in which DKT out-performs BKT and three of which represent content in which BKT out-performs DKT or there is little difference in performance between the approaches. We present some preliminary findings related to the extent to which students game the system in these workspaces, across two school years, as well as other facets of variability across these pieces of instructional content. We conclude with a road map for scaling these analyses over much larger sets of MATHia workspaces and learner data.

Keywords

gaming the system, knowledge tracing: Bayesian, knowledge tracing: Deep

1. INTRODUCTION

Intelligent Tutoring Systems (ITS) like Carnegie Learning's MATHia (previously, Cognitive Tutor; [15]) have relied on statistical knowledge tracing algorithms for decades to model students' mastery of sets of knowledge components, or skills. Recent work (e.g., [7, 8, 11]) considers questions related to conditions under which recent approaches to knowledge tracing based on deep learning predict student performance better than more "traditional" methods like Bayesian Knowledge Tracing (BKT) [1, 19]. For example, [7] measured the difference in performance of the two algorithms on over 300 MATHia workspaces. They found that there are workspaces where Deep Knowledge Tracing (DKT; [14]) outperforms BKT significantly, and others where BKT does as well or slightly outperforms DKT. Other work attempts to bridge the gap between the two algorithms, in order to provide a more interpretable algorithm that performs as well as DKT (e.g., by including a forgetfulness parameter [11]). Gervet et al. present the possibility in [8] that knowledge tracing approaches based on deep learning might "pick up" on behavior like "gaming the system" [4].

The present work describes work-in-progress investigating both performance differences between BKT and DKT as well as models of gaming the system behavior, analyzed on a workspace-by-workspace basis (what was called in [7] a "non-monolithic" approach). We begin by considering two relatively simple measures extracted from MATHia log data that may serve as *prima facie* proxies for more complex phenomena like "gaming the system" to see whether these simple proxies are correlated with the difference between DKT and BKT prediction performance over several hundred MATHia workspaces. Next, we consider a more sophisticated model of gaming the system thus far applied to six MATHia workspaces' data from two distinct school years (2018-19 and 2021-22). The six workspaces considered represent the three workspaces with the largest difference in DKT and BKT performance and the three workspaces with the smallest difference in this performance. We find roughly consistent patterns of gaming the system in each workspace across the two school years and that gaming the system may be more common in workspaces in which DKT out-performs BKT, though we take the question to be unresolved. We conclude by laying out a road map for on-going and future analyses investigating knowledge tracing prediction performance differences, gaming the system, and questions of year-over-year applications of these models to larger datasets from platforms like MATHia.

2. MATHIA

MATHia is an ITS developed by Carnegie Learning, typically used as part of a blended mathematics curriculum, primarily in middle and high schools across the United States. In such a blended curriculum, educators are typically advised to allow students to use MATHia for approximately 40% of the in-classroom instructional time. MATHia is used by hundreds of thousands of students in the United States every year. MATHia delivers instruction and practice grouped into topical "workspaces.", where each grade-level course in middle school (Grades 6-8) and high school course (Algebra I-II and Geometry) are comprised of approximately 70 to 120 workspaces. MATHia's "mastery" workspaces deliver instruction and practice on sets of fine-grained knowledge components (or skills) using a mastery learning approach [16]. Students are presented practice on skills until they reach a threshold (probability of learning equalling 0.95) for mastery of all skills within a workspace as estimated by BKT (or a pre-defined maximum number of problems, so that students who struggle to reach skill mastery are not perpetually provided additional practice). When a student is estimated to reach mastery of all skills associated with a workspace, they progress to the next workspace in the sequence of workspaces assigned to them. As students work on complex, multi-step problems in MATHia, they can request context-sensitive hints and receive just-in-time (JIT) feedback when mistakes they make are (at least roughly) aligned with known misconceptions or expected errors (e.g., that an errant numeric response appears in a problem-statement).

2.1 Knowledge Tracing Models

The present work compares the statistical prediction performance of two knowledge tracing models, BKT and DKT. For each knowledge component or skill, Bayesian Knowledge Tracing can be represented by a two-state hidden markov model (HMM) with four parameters governing transitions between the two-states, namely an "un-mastered" state and a "mastered" state, and student performance on opportunities to practice a particular skill. BKT's parameters for each skill represent students' prior knowledge, the probability that a student transitions from the un-mastered to the mastered state at a particular practice opporunity, the probability that the student may be able to guess correctly, or make an error despite mastery (or "slip"). Using student performance data, these models can be used to predict student correctness at particular practice opportunities as well as to estimate student mastery.

A newer recurrent neural network based method known as Deep Knowledge Tracing (DKT) [14] has been shown to perform better than BKT in certain cases and similarly in others [7, 8, 11]. As a part of our explorations of differences in DKT and BKT performance and the incidence of gaming the system across MATHia workspaces, we consider whether this difference in performance, at least in part, may be attributed to gaming the system behavior (addressing a question raised by Gervet et al. [8]) and relatively simple potential proxies for gaming behavior.

2.2 Gaming the System

"Gaming the System" is a widely studied phenomenon in the educational data science and learning analytics literature, referring to behavior whereby learners using educational technologies like ITS attempt to exploit features of learning technologies (e.g., hints and feedback) to make progress through instructional material without learning [3]. This literature often targets the development of data-driven "detectors" of such behavior, relying on action-level learning process data to predict instances of learner interactions in ITS that are likely examples of such behavior. Cognitive Tutor, now MATHia, has been a target platform for the development of detectors of gaming the system (e.g., [4]) as well as other behaviors and affective states (e.g., [5]). Notably, work like [2] adopted what we have called a "non-monolithic" approach to considering behaviors like gaming the system, considering characteristics of individual Cognitive Tutor "lessons" (what we call workspaces in MATHia) that may be associated with students' gaming the system behavior. Mostly recently, several implementations of a gaming the system detector were developed using data from MATHia in the 2020-21 school year [12], we use one of them, the Random Forest implementation, in this work. In addition to better understanding the extent to which students game the system across individual MATHia workspaces, we are also beginning to explore the question of the extent to which behavioral patterns like gaming the system persist across multiple school years in the same workspaces, a question related to whether such detector models "rot" over longer periods of time raised by recent work [12].

2.3 Data

In this study, we use student action-level data from both the 2018-19 and 2021-22 academic years for the 3 workspaces that were found in [7] to have the highest and lowest (non-absolute value) difference in performance between the DKT and BKT algorithms, where performance is measured in [7] on 2018-19 MATHia data, as the area under receiver operator curve (AUC ROC) of the two algorithms' predictions (probabilities) of students' correctness at opportunities to practice skills in those workspaces. [1] These 6 workspaces come from a diverse variety of curricula in terms of grade-level. Workspaces with the highest AUC(DKT) - AUC(BKT) difference implies that the DKT algorithm performs better in those workspaces, while the lowest AUC(DKT) - AUC(BKT) difference is slightly negative and implies that the BKT is slightly better in those workspaces. For brevity, we refer to the former sample as DKT+ and the latter sample as BKT+. The workspaces used in the DKT+ and BKT+ samples are shown in Tables 1 and Table 2 respectively, along with the number of students, and number of actions in the sample.[2]. The DKT-BKT performance difference is 0.31, 0.28, and 0.27 for Rewriting Radicals with Variables, Checking Solutions to Linear Equations, and Solving Literal Equations respectively; and -0.08, -0.07, and -0.03 for Solving with Addition and Subtraction (Type In), Solving Problems with Both Sales Tax and Discounts, and Solving with the Distributive Property Over Multiplication respectively. While we are working to deploy the relatively sophisticated gaming the system detector over a broader range

[1] We limit ourselves to 6 workspaces as we're currently working to "scale up" the implementation of gaming the system detectors to make predictions over hundreds of workspaces and hundreds of millions of student actions to allow for more definitive inferences about the relationship, if any, between prediction performance and gaming the system.

[2] The three BKT+ workspaces we consider have smaller sample sizes (in terms of students and transaction counts) in the 2018-19 data over which these models were trained. As Fancsali et al. [7] note a small correlation (r = 0.2) between sample size and the DKT-BKT performance difference, this may be a contributing factor to the "good" performance of BKT relative to DKT. However, there are workspaces with comparable sample sizes to the DKT+ workspaces for which BKT and DKT perform similarly. Future work will more comprehensively consider workspaces with both large sample sizes and "good" BKT performance.

of MATHia workspaces and student data, we consider two relatively simple potential proxies for gaming the system behavior, calculated over all workspaces from the 2018-19 and 2021-22 academic years: action duration and JIT feedback summary statistics.

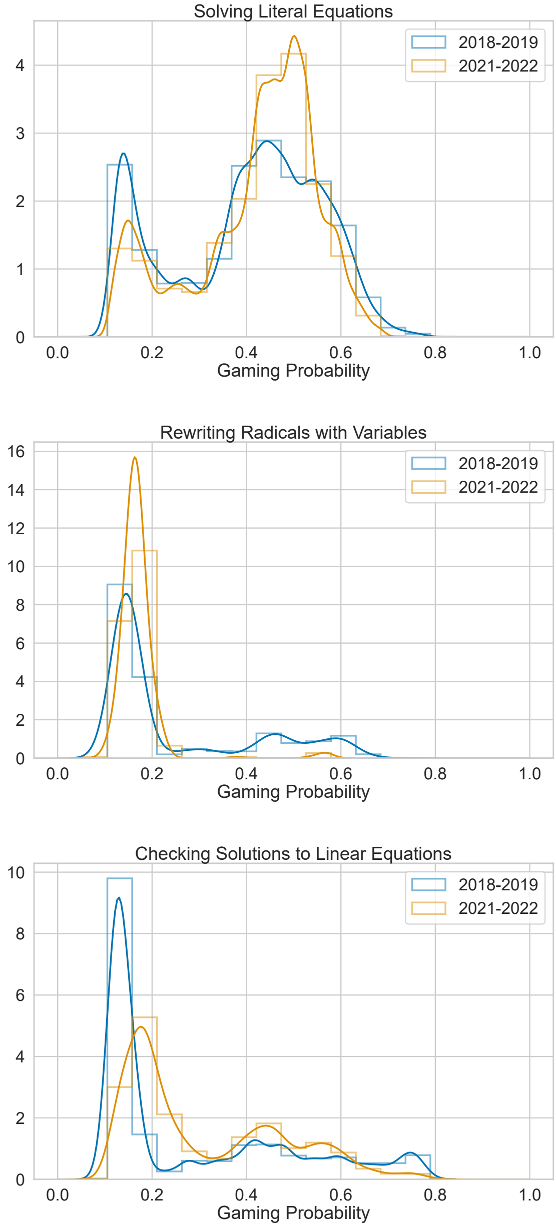
We use the recent gaming the system detector developed by [12], generating the 18 features described in [12] from MATHia transactional action-level data. We then define "clips" of data on which to run the gaming detector. We compute 9 summary statistics (mean, count, minimum, maximum, median, Q1, Q3, standard deviation, and sum) for each of the 18 features on the actions within a clip, resulting in 162 features on which we run the gaming detector. We define a clip as a single students' action-level data of up to 8 actions, and at least 2 actions, in up to 20 seconds. To avoid ambiguity, we look for series of actions satisfying exactly 8 actions in up to 20 seconds first, and identify them as clips. Once those actions have been used to generate clips, those actions are removed from the dataset, and we look for actions satisfying exactly 7 actions in up to 20 seconds, and so forth in descending order. This procedure insures that actions are not repeated across clips, and that the number of action within a clip is maximized up to 8. It is noteworthy that the 18 features are generated prior to defining clips, including certain features that targets actions on a rolling basis. For example, the `dur_sd_prev_5` feature, which is the sum of the last 5 standardized action durations, is computed on windows of actions of size 5 before potentially splitting those actions into clips.

3. GAMING THE SYSTEM PREDICTIONS

3.1 Proxy measures of gaming the system

Two *prima facie* potential correlates of gaming the system behavior include rapid student actions, and student actions that lead to JIT feedback, sometimes referred to in the literature (e.g., [2]) as "buggy" messages. Fancsali, for examples, suggests in [6] an especially high correlation between student errors that lead to JIT feedback and predictions that a student is gaming the system. While rapid student actions could also indicate attempts to game the system (e.g., by rapidly entering numbers that appear in a problem), rapid student actions may also simply be the result of the instructional design of a particular MATHia workspace. Fancsali et al. [7] noted, for example, that workspaces for which DKT vastly out-performed BKT tended to be workspaces in which students used an equation solving tool, used a numberline tool, or in which problem-steps were frequently "drop-down" menus with several choices for response. Any of these types of interface elements could lead to more rapid student actions, regardless of whether students are gaming the system.

Taking into account all workspaces in MATHia (as these proxy measures are more easily extracted from MATHia process data), we find that the median duration of student actions in each workspace is weakly negatively correlated at a statistically significant level to the DKT-BKT performance difference in both the 2018-19 and 2021-22 data (where the DKT-BKT performance difference is re-used for the 2021-22 correlation). Specifically, these correlations were r=-0.15 for 2018-19 at a p-value of 0.008 and r=-0.18 for 2021-22 at a p-value of 0.002.


Solving Literal Equations

Rewriting Radicals with Variables

Checking Solutions to Linear Equations

Figure 1: The distribution of gaming probabilities for all clips in the DKT+ sample, on a workspace-by-workspace basis, for both the 2018-19 and 2021-22 academic years. The distributions show that gaming trends remain the same over the years, even for different student samples. KS test show that the CDFs of the two distributions for each workspace differs by 0.10, 0.28, and 0.38 for Solving Literal Equations, Rewriting Radicals with Variables, and Checking SOlutions to Linear Equations respectively.

Table 1: Data volume for workspaces in the DKT+ sample, where DKT outperforms BKT the most

	Workspace Name	# of transactions	# of students
2018-19	Solving Literal Equations	443433	5911
	Checking Solutions to Linear Equations	1234234	27956
	Rewriting Radicals with Variables	86923	2750
	Workspace Name	# of transactions	# of students
2021-22	Solving Literal Equations	1134960	22191
	Checking Solutions to Linear Equations	411898	26690
	Rewriting Radicals with Variables	59326	6296

Table 2: Data volume for workspaces in the BKT+ sample, where BKT outperforms DKT the most.

	Workspace Name	# of transactions	# of students
2018-19	Solving with the Distributive Property Over Multiplication	15266	365
	Solving Problems with Both Sales Tax and Discounts	10737	529
	Solving with Addition and Subtraction (Type In)	20245	582
	Workspace Name	# of transactions	# of students
2021-22	Solving with the Distributive Property Over Multiplication	711740	20461
	Solving Problems with Both Sales Tax and Discounts	596958	23862
	Solving with Addition and Subtraction (Type In)	6830	445

We found no statistically significant between the median or average number of total JIT feedback messages that a student received in a workspace and the DKT-BKT performance difference. Our on-going work will more comprehensively consider the extent to which student actions triggering JIT feedback are correlated with detector predictions of gaming the system, in addition to whether more comprehensive predictions of gaming the system over all MATHia workspace correlate with the performance difference between these knowledge tracing methods.

The negative correlation with action durations implies that workspaces where DKT most outperforms BKT have shorter durations between student actions. Given that gaming behavior may be occurring when students are making rapid attempts, this might point towards that DKT outperforms BKT for workspaces more susceptible to gaming. However, both the possibility that rapid student actions may arise out of particular workspace design features and the lack of correlation with JITs imply otherwise, and further work is needed to confirm or reject this hypothesis.

Notably, these behavioral patterns (i.e., rapid student actions and the extent to which students trigger JIT feedback in these workspaces) stayed largely the same across years, the correlation between the median duration in workspaces from 2018-19 and 2021-22 data was 0.86 and the correlation between the number of JITs for the median student from 2018-19 and 2021-22 data was 0.87, both at p-values $< 10^{-81}$.

3.2 Persistence of gaming the system behavior across time

We generate 342,274 clips from the 2021-22 data and 183,568 clips from the 2018-19 data, each containing summary statistics of between 2 and 8 student actions. On a high level, we find that clips in the DKT+ sample are 8.4% and 3.4% more likely to be classified as gaming the system than clips in the BKT+ sample for the 2021-22 and 2018-19 data respectively.

It is unclear whether a trend will emerge when considering a larger number of workspaces, and the higher level difference is likely driven by a few workspaces, as there is large variance between gaming behavior trends in different workspaces.

Interestingly, gaming trends seem to have stayed largely the same between the 2018-19 and 2021-22 samples in these workspaces. Figures 1 and 2 show histograms and kernel density estimations (KDE) of the per-clip gaming probabilities for each workspace in both years considered. The distributions were normalized such that they are comparable across years. The KDE uses a Gaussian kernel with bandwidth selected using Scott's rule of thumb [17].

Quantitatively, we used a two-sided Kolmogorov-Smirnov (KS) tests [13] to measure the absolute maximum distance between the Cumulative Distributed Functions (CDFs) of the distributions of gaming probabilities (over all values of gaming probabilities) in the 2018-19 and 2021-22 data in each workspace. While the KS test found statistically significant differences between the distributions in each case, these differences were small, between 0.1 for Solving Literal Equations and 0.38 for Checking Solutions to Linear Equations, where values closer to 0 indicate that the two distributions were drawn from the same distribution and values closer to 1 indicate that there is no similarity in the CDFs of the two distributions. Given the large sample sizes making up the distributions, KS tests are expected to show statistically significant differences even for minor differences in the two distributions.

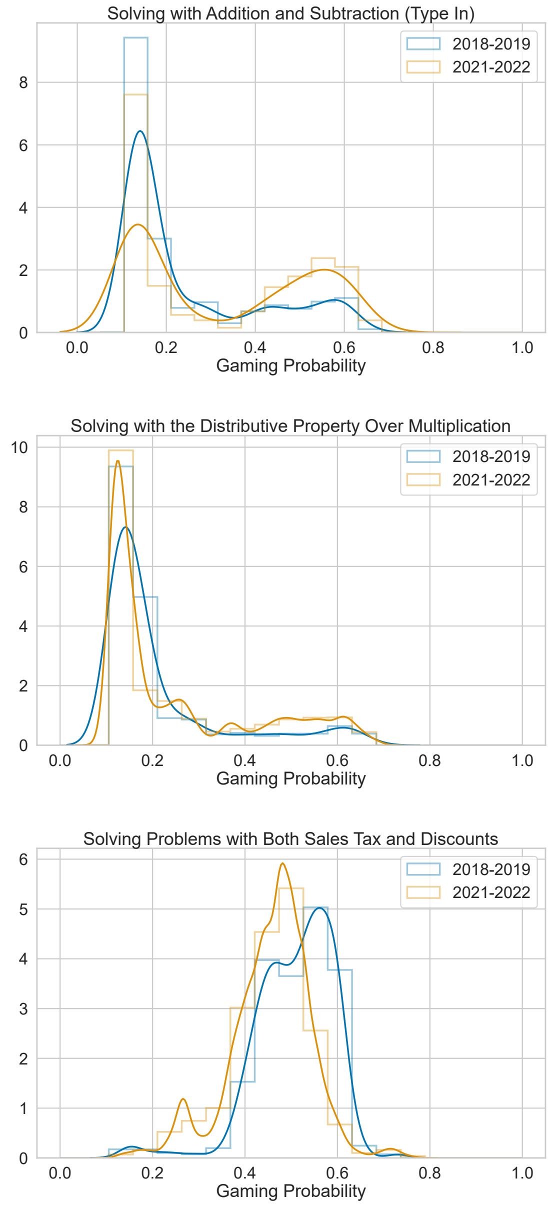


Figure 2: The distribution of gaming probabilities for all clips in the BKT+ sample, on a workspace-by-workspace basis, for both the 2018-19 and 2021-22 academic years. The distributions show that gaming trends remain the same over the years, even for different student samples. KS test show that the CDFs of the two distributions for each workspace differs by 0.23, 0.17 and 0.30 for Solving with Addition and Subtraction (Type In), Solving with the Distributive Property Over Multiplication, and Solving Problems with Both Sales Tax and Discounts respectively.

4. ROAD MAP + CONCLUSIONS

In the present report on work-in-progress, we have looked at how certain workspace characteristics vary along with the difference in DKT and BKT performance in those workspaces. We found a small, but statistically significant, negative correlation between the median duration of student actions and the difference in DKT and BKT performance over all MATHia workspaces. This is the first statistically significant correlation we have found between the DKT and BKT performance difference and a characteristic of (learner behavior in) workspaces (aside from correlation with sample size noted by [7]).

We also considered gaming the system behavior for six workspaces, selected as the three with the highest and lowest difference in DKT and BKT performance. We found that gaming behavior varies highly between workspaces, but that within each workspace, similar gaming behavior trends have persisted across years. We found that, on average, the three workspaces with higher DKT performance compared to BKT combined showed higher probability of gaming the system, for both 2018-19 and 2021-22. However, it is not yet clear whether this would generalize over all MATHia workspaces.

A more holistic (and consequently more computationally intensive) approach is planned for a future study, where we plan to develop and use a scalable and refined version of the gaming detector (i.e., constructed over more labeled examples of gaming versus non-gaming behavior), deployed on MATHia data over all workspaces for one or more school years. We also intend to better understand the extent to which DKT and BKT performance differences and gaming the system behavior persist from school year to school year when considered on a workspace-by-workspace basis, following the “non-monolithic” analysis approach for which we have advocated [7]. In the spirit of Baker’s Cognitive Tutor Lesson Variation Space [2], intended to capture “lesson” or workspace-level variation in gaming the system, we also intend to better explore the space of factors that vary workspace-by-workspace that may help explain both gaming the system as well as differences in performance among knowledge tracing methods.

5. ACKNOWLEDGMENTS

This research is funded by the National Science Foundation under the award The Learner Data Institute (Award #1934745). Opinions, findings, and results are solely those of the authors and do not reflect those of the National Science Foundation.

Figures in this paper were generated using the open-source Python libraries: matplotlib [9] and seaborn [18], and statistical tests were performed in SciPy [10].

6. REFERENCES

[1] J. R. Anderson and A. T. Corbett. Knowledge tracing: Modeling the acquisition of procedural knowledge. *User Modeling and User-Adapted Interaction*, 4:253–278, 1995.

[2] R. Baker. Differences between intelligent tutor lessons, and the choice to go off-task. In *EDM*. International Educational Data Mining Society, 2009.

[3] R. Baker, J. Walonoski, N. Heffernan, I. Roll, A. Corbett, and K. Koedinger. Why students engage in" gaming the system" behavior in interactive learning environments. *Journal of Interactive Learning Research*, 2008.

[4] R. S. Baker, A. T. Corbett, K. R. Koedinger, and A. Z. Wagner. Off-task behavior in the cognitive tutor classroom: When students "game the system". In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '04, page 383–390, New York, NY, USA, 2004. Association for Computing Machinery.

[5] R. S. Baker, S. Gowda, M. Wixon, J. Kalka, A. Wagner, A. Salvi, V. Aleven, G. Kusbit, J. Ocumpaugh, and L. Rossi. Sensor-free automated detection of affect in a cognitive tutor for algebra. In *Educational Data Mining 2012*. International Educational Data Mining Society, 2012.

[6] S. E. Fancsali. Causal discovery with models: behavior, affect, and learning in cognitive tutor algebra. In *Educational Data Mining 2014*. International Educational Data Mining Society, 2014.

[7] S. E. Fancsali, H. Li, M. Sandbothe, and S. Ritter. Targeting design-loop adaptivity. In *Proceedings of the 14th International Conference on Educational Data Mining*, volume 14. International Educational Data Mining Society, 2021.

[8] T. Gervet, K. Koedinger, J. Schneider, and T. Mitchell. When is deep learning the best approach to knowledge tracing? *Journal of Educational Data Mining*, 12(3):31–54, Oct. 2020.

[9] J. D. Hunter. Matplotlib: A 2d graphics environment. *Computing In Science & Engineering*, 9(3):90–95, 2007.

[10] E. Jones, T. Oliphant, P. Peterson, et al. SciPy: Open source scientific tools for Python, 2001.

[11] M. M. Khajah, R. V. Lindsey, and M. C. Mozer. How deep is knowledge tracing? In *Proceedings of the 9th International Conference on Educational Data Mining*, volume abs/1604.02416. International Educational Data Mining Society, 2016.

[12] N. Levin, R. S. Baker, N. Nasiar, S. E. Fancsali, and S. Hutt. Evaluating gaming detector model robustness over time. In *International Conference on Educational Data Mining 2022*. International Educational Data Mining Society, 2022.

[13] F. J. Massey. The kolmogorov-smirnov test for goodness of fit. *Journal of the American Statistical Association*, 46(253):68–78, 1951.

[14] C. Piech, J. Bassen, J. Huang, S. Ganguli, M. Sahami, L. J. Guibas, and J. Sohl-Dickstein. Deep knowledge tracing. In C. Cortes, N. Lawrence, D. Lee, M. Sugiyama, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 28. Curran Associates, Inc., 2015.

[15] S. Ritter, J. R. Anderson, K. Koedinger, and A. T. Corbett. Cognitive tutor: Applied research in mathematics education. *Psychonomic Bulletin & Review*, 14:249–255, 2007.

[16] S. Ritter, M. V. Yudelson, S. E. Fancsali, and S. R. Berman. How mastery learning works at scale. *Proceedings of the Third (2016) ACM Conference on Learning @ Scale*, 2016.

[17] D. W. Scott. On optimal and data-based histograms. *Biometrika*, 66(3):605–610, 12 1979.

[18] M. Waskom, O. Botvinnik, D. O'Kane, P. Hobson, S. Lukauskas, D. C. Gemperline, T. Augspurger, Y. Halchenko, J. B. Cole, J. Warmenhoven, J. de Ruiter, C. Pye, S. Hoyer, J. Vanderplas, S. Villalba, G. Kunter, E. Quintero, P. Bachant, M. Martin, K. Meyer, A. Miles, Y. Ram, T. Yarkoni, M. L. Williams, C. Evans, C. Fitzgerald, Brian, C. Fonnesbeck, A. Lee, and A. Qalieh. mwaskom/seaborn: v0.8.1, Sept. 2017.

[19] M. V. Yudelson, K. R. Koedinger, and G. J. Gordon. Individualized bayesian knowledge tracing models. In H. C. Lane, K. Yacef, J. Mostow, and P. Pavlik, editors, *Artificial Intelligence in Education*, pages 171–180. Springer Berlin Heidelberg, 2013.