Simulated Students, Mastery Learning, and Improved Learning Curves for Real-World Cognitive Tutors

Stephen E. Fancsali, Tristan Nixon, Annalies Vuong, and Steven Ritter

Carnegie Learning, Inc.
Frick Building, Suite 918
437 Grant Street, Pittsburgh, PA 15219

{sfancsali, tnixon, avuong, sritter}
@carnegielearning.com

Abstract. We briefly describe three approaches to simulating students to develop and improve intelligent tutoring systems. We review recent work with simulated student data based on simple probabilistic models that provides important insight into practical decisions made in the deployment of Cognitive Tutor software, focusing specifically on aspects of mastery learning in Bayesian Knowledge Tracing and learning curve analysis to improve cognitive (skill) models. We provide a new simulation approach that builds on earlier efforts to better visualize aggregate learning curves.

Keywords: Knowledge tracing, learning curves, student modeling, Cognitive Tutor, simulation, simulated students, mastery learning

1 Introduction

There are at least three general approaches to simulating students for the purposes of improving cognitive (skill) models and other features of intelligent tutoring systems (ITSs). One approach, generally connoted in discussions of "simulated" students or learners, employs aspects of cognitive theory to simulate students' learning and progression through ITS problems (e.g., via machine learning or computational agents like SimStudent [2]). Another class of simulations makes use of relatively simple probabilistic models to generate response data (i.e., Bayesian Knowledge Tracing [BKT] [1]) intended to represent a (simulated) student's evolving performance over many practice attempts. Third, there are data-driven approaches that do not easily fit into either of the first two categories.

In this work, we explicate and provide examples of each approach and briefly describe Carnegie Learning's Cognitive Tutors (CTs) [3]. We then focus on the second approach and review recent work on simulations of student learning with simple probabilistic models. These simulation studies provide novel insights into a variety of features of CTs and their practical deployment.

CTs implement mastery learning; mathematics content is adaptively presented to students based upon whether the tutor has judged that a student has mastered particular skills. Mastery is assessed according to whether the tutor judges that the probability that a student has mastered a particular skill exceeds a set threshold. We review a simulation study that provides for best and worst-case analyses (when "ground truth" characteristics of simulated learner populations are known) of tutor skill mastery judgment and efficient student practice (i.e., adaptively providing students with opportunities to practice only those skills they have not mastered). This study not only provides justification for the traditionally used 95% probability threshold, but it also illuminates how the threshold for skill mastery can function as a "tunable" parameter, demonstrating the practical import of such simulation studies.

Finally, learning curves provide a visual representation of student performance on opportunities to practice purported skills in an ITS. These representations can be used to analyze whether a domain has been appropriately atomized into skills. If opportunities correspond to practice for a single skill, we expect to see a gradual increase in the proportion of correct responses as students get more practice opportunities. If, for example, the proportion of students responding correctly to an opportunity drastically decreases after three practice opportunities, it seems unlikely that the opportunities genuinely correspond to one particular skill. Turning to the third, data-driven approach to simulating students, we provide a new method to visualize aggregate learning curves to better drive improvements in cognitive (skill) models used in CTs, This approach extends recent work that explores several problems for utilizing learning curves aggregated over many students to determine whether practice opportunities correspond to a single skill.

2 Cognitive Tutors

CTs are ITSs for mathematics curricula used by hundreds of thousands of K-12 and undergraduate students every year. Based on cognitive models that decompose problem solving into constituent knowledge components (KCs) or skills, CT implements BKT to track student skill knowledge. When the system's estimate of a student's knowledge of any particular skill exceeds a set threshold, the student is judged to have mastered that skill. Based on the CT's judgment of skill mastery, problems that emphasize different skills are adaptively presented so that the student may focus on those skills most in need of practice.

3 Three Approaches to Simulating Learners

There are at least three general simulation methods used to model student or learner performance. One simulation strategy, based on cognitive theories such as ACT-R [4], explicitly models cognitive problem-solving processes to produce rich agent-based simulated students. The SimStudent project ([2], [5]), for example, has been developed as a part of a suite of authoring tools to develop curricula for CTs, called Cognitive Tutor Authoring Tools (CTAT) [6]. SimStudent learns production rules

from problem-solving demonstrations (e.g., an author providing simple demonstrations of problem solutions or via ITS log data). These human-interpretable production rules correspond to KCs that comprise cognitive models vital to CTs. SimStudent aims to simplify development of new CT material by automating the discovery of KC models in new domains via a bottom-up search for skills that potentially explain the demonstrations.

Second, there are numerous probabilistic methods that model task performance as a function of practice, according to various task and learner-specific parameters. One may instantiate numerous such models, with varying parameters, and sample from the resulting probability distributions to obtain simulated performance data for an entire hypothetical learner population.

One common example is a Hidden Markov Model (HMM) with two latent and two observable states, that can serve as a generative BKT model, using parameters specified according to expert knowledge or inferred by a data-driven estimation procedure. Two hidden nodes in the HMM represent "known" and "unknown" student knowledge states. In practice, of course, student knowledge is latent. Simulated students are assigned to a knowledge state according to BKT's parameter for the probability of initial knowledge, $P(L_0)$, and those in the "unknown" state transition to the "known" state according to the BKT parameter for the probability of learning or transfer, P(T). Simulated, observed responses are then sampled according to BKT parameters that represent the probability of student guessing, P(G) (i.e., responding correctly when in the unknown state) and slipping, P(S) (i.e., responding incorrectly when in the known state), depending upon the state of student knowledge at each practice opportunity.

Contrary to her real-world epistemological position, simulations generally allow an investigator to access the student's knowledge state at each simulated practice opportunity. This allows for comparisons between the "ground truth" of skill mastery and any estimate derived from resulting simulated behavior. Clearly, richer cognitive agents, such as SimStudent, provide a more complete picture of the student's cognitive state at any point.

Simpler probabilistic models represent student knowledge of a skill with a single state variable, so they correspondingly scale better to larger scale simulations of whole populations. While a probabilistic model only requires a reasonable distribution over initial parameters, richer cognitive models may require training on a great deal of detailed, behavioral or demonstration data. Nevertheless, cognitive model-based simulations allow us to investigate issues like timing (i.e., response latency), sensitivity to input characteristics, and error patterns in learner responses.

There are many cases in which a relatively simple probabilistic model may be of utility, despite its impoverished nature. A simplistic representation of student knowledge provides an ideal situation to test the performance and characteristics of inference methods using data from a known generating process and parameters. One might, for example, compare the point at which simulated students acquire knowledge of a skill to the point at which the CT judges the student to have mastered the skill. The approach thus allows for students of "best" and "worst" case scenarios with respect to the relationship between how the CT models students and the actual make up

of (simulated) student populations. We can better understand the dynamics of the student sub-populations we inevitably face in practice by simulating data from diverse sub-populations, the make up of which we can specify or randomize in various ways. Furthermore, we can simulate student performance (sometimes augmenting available empirical data) both with and without mastery learning (i.e., students being removed from a population because they have mastered a skill) on learning curves constructed from aggregate data.

Previous work [7] explored a third, data-driven simulation method that "replays" empirical student performance data through CT in order to estimate the impact of a change in BKT parameters in a more substantive way. For each KC that occurred in a given problem, we sampled the next observed response on that KC from the sequence actually observed from a real student. These responses would then drive updates to CT's cognitive model, knowledge tracing, and the problem-selection mechanism. If more data were required than were observed for a given student, further observations were sampled from a BKT model initialized to the state inferred from the student's actions thus far. By repeating this process for all students in the observed data set, we could obtain estimates of the number of problems students would be expected to complete if a change to the cognitive model were implemented.

This method has the advantage of preserving characteristics of real student data rather than resorting to a theoretical model of student performance. However, it does make several assumptions about the reproducibility of that behavior under the hypothesized changes. Specifically, it assumes that the observed sequence of correct/incorrect responses would not change even given a different selection of problems, potentially emphasizing different KCs. This assumption may be justified if we believe we have complete coverage of all KCs relevant to the task in question in the cognitive model and that all KCs are truly independent of each other.

While simulation methods based on rich cognitive theory and data-driven re-play of empirical data provide many opportunities for future research, we focus in this paper on simple, probabilistic simulations in the context of the BKT framework.

4 Substantive Measures of Efficient Student Practice

Before we discuss how the BKT mastery threshold probability functions as a "tunable" parameter in an ITS like the CT, we provide "substantive" quantification of goodness of fit of cognitive/skill models for CTs beyond mere RMSE of prediction (i.e., beyond the extent to which models can predict whether students will respond correctly to particular practice opportunities) [8-11]. New error or goodness of fit measures are countenanced in terms of efficient student practice, based on the number of practice opportunities (i.e., "over-practice" or "under-practice") we might expect a student to experience in a CT. Over-practice refers to the continued presentation of new practice opportunities, despite the student's mastery or knowledge of the relevant KC.[1] Student "under-practice" instances are those in which a student has yet to

[1] One exception is an experimental study [11] that reports increased efficiency by deploying parameters estimated using a data mining method called Learning Factors Analysis (LFA).

achieve knowledge of a KC, and yet the mastery learning system has judged the student as having mastered it, ending the presentation of further learning opportunities. From estimates of expected under- and over-practice, one can calculate other meaningful measures of students gains and losses, such as time saved or wasted.

Some of this work [8, 9] uses empirical data to estimate the extent of under-practice and over-practice we might expect students to experience. Specifically, the expected numbers of practice opportunities it takes a student to reach mastery when parameters are individualized per student are compared to the expected practice when a single (population) set of parameters is used to assess all students. One individualization scheme used to study under and over-practice estimates all four BKT parameters, per student, from response data over all relevant skills (i.e., each student receives one individualized quadruple of BKT parameters for all KCs) [8]. Another approach [9] only individualizes P(T) for each student based on both per-student and per-skill components estimated from observed data [12]. Both individualization schemes provide for substantive gains (compared to using a set of population parameters to assess all students' progress to mastery) in the efficiency of practice (i.e., fewer expected under and over-practice opportunities) as well as better prediction performance judged, in the standard way, by a metric like RMSE.

5 Idealized Performance of Mastery Learning Assessment

Now we address how BKT performs with respect to efficiency of practice in idealized cases in which the composition of student (sub-) populations is known. Simulation studies can shed light on how BKT performs when mastery learning parameters used by the CT run-time system exactly match those of the generating model (i.e., the best case), and in worst cases in which student parameters either maximally differ from mastery learning parameters or vary at random for each student.

Recent work addresses these issues by adopting a probabilistic simulation regime [10]. Since we can track the point at which a simulated student acquires knowledge of a skill, we are able to compare this to the opportunity at which the mastery learning system first judges it to be acquired. Simulations were run for fourteen skills, a subset of those found by [13] to be representative of a substantial portion of skills in deployed CT curricula, across thousands of virtual students.

Even in idealized, best case scenarios (i.e., when parameters used to assess skill mastery perfectly match simulated student data-generating parameters), for most skills and a large number of students, we expect there to be one to four "lagged" practice opportunities between the point at which simulated students transition to mastery and the point at which the BKT run-time system judges mastery. That is, in general, even when a student population is modeled "perfectly," and given the traditional setting of the probability threshold for mastery at 95%, most students should be expected to see at least a few opportunities beyond the point of skill acquisition. That some "over-practice" may be inevitable provides a relevant context within which to consid-

---

Efficiency is operationalized as decreased time required to work through material in the Geometry CT without decreasing overall learning.

er empirically driven results of [8, 9]. Although a certain amount of lag may be inherent in the nature of BKT, we seek to establish a range for the "acceptable" lag, and to better appraise efficiency of practice [10].

6 Mastery Learning Threshold as a "Tunable" Parameter

In addition to lagged opportunities and over-practice, situations in which students under-practice skills are important to consider. Given the possibly inevitable lag between skill acquisition and mastery judgment, simulations [10] have also been used to explore how the mastery probability threshold might be "tuned" to influence the trade-off of over-practice and under-practice experienced by students in mastery learning systems like CTs.

Pre-mature mastery judgments can lead, for example, to students being moved along by the CT to problems that emphasize new KCs without having mastered prerequisite KCs. Other things held equal, simulations in [10] provide that pre-mature mastery judgment is more likely to occur in worst-case scenarios, when mastery-learning parameters do not match parameters for sub-populations of simulated students.

Simulations in [10] also establish that the mastery-learning threshold can function as a tuning parameter, partially governing the trade-off between the expected proportion of students pre-maturely judged to have reached skill mastery and the number of over-practice opportunities they are likely to experience. As the threshold probability is increased, the proportion of students assessed as having pre-maturely mastered skills decreases while the proportion of those that are exposed to practice opportunities after skill acquisition increases (along with the number of lagged and over-practice opportunities, i.e., those beyond a calculated acceptable lag they experience).

The results of [10] show that the traditionally used 95% threshold seems to provide for a "conservative" tutor that is more likely to present opportunities after skill acquisition rather than under-practice. Depending upon course design and practice regimes, the mastery-learning threshold might be manipulated to important, practical effect. For example, pre-mature mastery judgments might be acceptable in larger numbers when there is a mixed-practice regime that will allow students to practice KCs later in the curriculum.

7 Using Simulations to Illuminate Learning in Learning Curves

Learning curves provide a visual representation of student performance over opportunities to practice skills. For each (purported) skill, we construct a learning curve by plotting opportunities (i.e., 1st, opportunity, 2nd opportunity, and so on) on the x-axis and the proportion of students that provide correct responses at each opportunity on the y-axis. Aggregated over real-world student practice opportunity data, such

curves provide means by which to visually[2] inspect whether opportunities genuinely correspond to practice of one particular skill. If opportunities correspond to one particular skill, we expect a gradual increase in the proportion of students that respond correctly with increasing practice. Generally, for well-modeled skills (and a variety of other cognitive tasks), it is thought that such a plot should correspond roughly to a power law function (i.e., the power law of practice [14]), though this point is not without controversy [15]. Recent research [16-17] demonstrates how some aggregate learning curves can distort the picture of student learning. Aggregate learning curves may, for example, appear to show no learning, when, in fact all students are learning at different rates. Others may provide for a small rise in probability of correct response initially but then "drop," as if students were forgetting, even when individual students are consistently mastering their skills.

The learning curve of Fig. 1 illustrates aspects of both problems, with a relatively flat portion, followed by a drop, after a small increase in probability correct from its initial value. The red line, representing the size of the student population at each opportunity, illustrates that BKT is determining that students are mastering the skill relatively quickly.

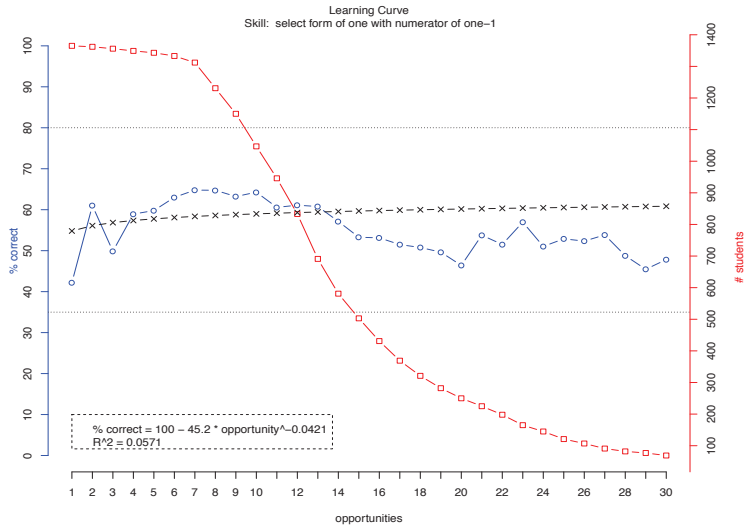

Fig. 1. Empirical Learning Curve for Skill "Select form of one with numerator of one"; the blue line represents empirical data plotted as percentage of correct responses, and the black line represents a fitted power function. The red line provides the size of the student population.

Two ways to re-visualize problematic, aggregated learning curves have been suggested [16]. One is to provide multiple learning curves (on the same plot) for individual

[2] Developers at Carnegie Learning also deploy several data-driven heuristics (that correspond to various visual features of learning curves) to analyze our large portfolio of KCs (i.e., several thousand KCs over several mathematics CT curricula) and observed student data to draw attention to those KCs that may require revision in our deployed cognitive models.

"segments" of students based upon how many opportunities students, in observed data, take to reach the mastery learning threshold for a skill. Such segmented learning curves are provided with the same x-axis and y-axis as standard learning curves (i.e., practice opportunity count on the x-axis and, e.g., percentage of student correct response on the y-axis).

The second approach suggested by [16] has the analyst plot "mastery-aligned" learning curves. In such learning curves, students are also segmented according to the number of opportunities required to reach mastery, but the end-point of the x-axis corresponds to the opportunity at which students' reach mastery (*m*) and moving left along the x-axis corresponds to the opportunity before mastery (*m-1*), the second to last opportunity before mastery (*m-2*), and so on.

Further work [17] provides a mathematical explanation, along with proof-of-concept simulation studies based on HMMs, for the dynamics of aggregate learning curves to explain how both mastery learning itself and differing student subpopulations, when aggregated, can contribute to learning curves that do not show learning (or manifest other peculiar, possible deceptive, phenomena like "negative" learning).

We illustrate an alternative to [16] by providing a method that relies on probabilistic simulation to construct aggregate learning curves that better represent learning in empirical student data. Specifically, we "pad" empirical data for student skill opportunities with simulated data to mask the effects of attrition due to mastery learning and possibly "reveal" student learning. Student opportunity data are generated with the same parameters used to track student progress and the probability of student knowledge estimated at the point at which the student crossed the mastery threshold. Such simulations provide us data after a student no longer receives practice opportunities for a particular skill because they have been judged as having achieved mastery.

For the aggregate learning curve of Fig. 1, the "padded" learning curve is Fig. 2. The fitted power-law slope parameter decreases from -0.042 to -0.363 (indicating more learning), and the goodness-of-fit of the power law function (R^2) increases from 0.0571 to 0.875. We apply the method to 166 skills identified[3] by [16] as possibly problematic in the Cognitive Tutor Algebra I (CTAI) curriculum. We find an improvement (i.e., power-fit parameter decreases from above -0.1 to below -0.1, a criterion deployed by [16]) for 98 skills (59%). While this method provides an improved visualization and understanding of fewer skills than the disaggregation procedures suggested by [16], this seems to provide evidence of the great extent to which mastery learning attrition obfuscates evidence for student learning.

Importantly, our simulation method does not eliminate the early dip in the learning curve at opportunity 3 when little attrition has yet to take place, but only masks the effects of attrition due to mastery learning. Such an approach focuses largely on a better representation or visualization of the "tail" of aggregate learning curves. This

[3] These skills were chosen because the over-whelming majority of students are judged to eventually master them (i.e., CT "thinks" the students are learning); they are not pre-mastered (i.e., $P(L_0) < 0.95$); they do not show learning in their aggregate learning curve (i.e., power-law fit parameter > -0.1); aggregate learning curves for these skills do not have multiple maxima; and we have data for at least 250 students for these skills [16].

allows us to focus on other features of the learning curve that may indicate ill-modeled KCs in a cognitive model, software bugs, and other possible problems.

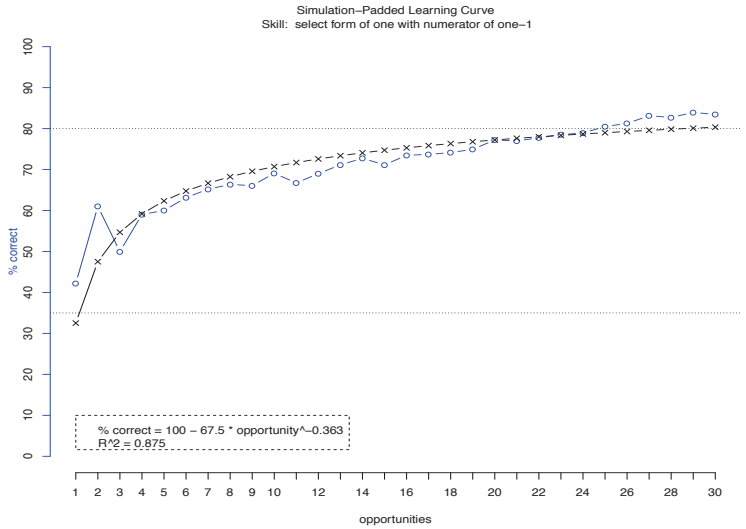

Fig. 2. Simulation-Padded Learning Curve for Skill "Select form of one with numerator of one"

8 Summary

We briefly reviewed several methods for simulating learners. We focused on ways in which simple probabilistic models, in contrast to methods that rely on rich cognitive theory, can be used to generate student performance data to help drive practical decision-making about CT deployment, focusing first on the mastery threshold probability of BKT as a tunable parameter to determine aspects of efficient practice. Then we introduced a new method for visualizing aggregate learning curves that relies on both empirical and simulated data that helps to mask the bias introduced by mastery learning attrition. Future work will further explore these methods, new simulation regimes, and their practical import.

References

1. Corbett, A.T., Anderson, J.R.: Knowledge Tracing: Modeling the Acquisition of Procedural Knowledge. User-Modeling and User-Adapted Interaction 4, 253–278 (1995)
2. Matsuda, N., Cohen, W.W., Sewall, J., Koedinger, K.R.: Applying Machine Learning to Cognitive Modeling for Cognitive Tutors. Human-Computer Interaction Institute, Carnegie Mellon University. Paper 248 (CMU-ML-06-105) (2006)

3. Ritter, S., Anderson, J.R., Koedinger, K.R., Corbett, A.T.: Cognitive Tutors: Applied Research in Mathematics Education. Psychonomic Bulletin & Review 14, 249–255 (2007)
4. Anderson, J.R.: Rules of the Mind. Erlbuam, Hillsdale, NJ (1993)
5. Matsuda, N., Cohen, W.W., Sewall, J., Lacerda, G., Koedinger, K.R.: Evaluating a Simulated Student Using Real Students Data for Training and Testing. In: Proceedings of the International Conference on User Modeling (LNAI 4511), pp. 107–116 (2007)
6. Aleven, V., McLaren, B.M., Sewall, J., Koedinger, K.R.: The Cognitive Tutor Authoring Tool (CTAT): Preliminary Evaluation of Efficiency Gains. In: Proceedings of the 8th International Conference on Intelligent Tutoring Systems, pp. 61-70 (2006)
7. Dickison, D., Ritter, S., Nixon, T., Harris, T., Towle, B., Murray, R.C., Hausmann, R.G.M.: Predicting the Effects of Skill Model Changes on Student Progress. In: Proceedings of the 10th International Conference on Intelligent Tutoring Systems (Part II), pp. 300-302 (2010)
8. Lee, J.I., Brunskill, E.: The Impact of Individualizing Student Models on Necessary Practice Opportunities. In: Proceedings of the 5th International Conference on Educational Data Mining, pp. 118–125 (2012)
9. Yudelson, M.V., Koedinger, K.R.: Estimating the Benefits of Student Model Improvements on a Substantive Scale. In: Proceedings of the 6th International Conference on Educational Data Mining (2013)
10. Fancsali, S., Nixon, T., Ritter, S.: Optimal and Worst-Case Performance of Mastery Learning Assessment with Bayesian Knowledge Tracing. In: Proceedings of the 6th International Conference on Educational Data Mining (2013)
11. Cen, H., Koedinger, K., Junker, B.: Is Over-Practice Necessary? – Improving Learning Efficiency with the Cognitive Tutor through Educational Data Mining. In: Proceedings of the 13th International Conference on Artificial Intelligence in Education, pp. 511–518 (2007)
12. Yudelson, M.V., Koedinger, K.R., Gordon, G.J.: Individualized Bayesian Knowledge Tracing Models. In: Proceedings of the 16th International Conference on Artificial Intelligence in Education (2013)
13. Ritter, S., Harris, T.K., Nixon, T., Dickison, D., Murray, R.C., Towle, B.: Reducing the Knowledge Tracing Space. In: Proceedings of the 2nd International Conference on Educational Data Mining, pp. 151-160 (2009)
14. Newell, A., Rosenbloom, P.S.: Mechanisms of Skill Acquisition and the Law of Practice. In: Anderson, J.R. (ed.) Cognitive Skills and Their Acquisition, pp. 1-55. Erlbaum, Hillsdale, NJ (1981)
15. Heathcote, A., Brown, S.: The Power Law Repealed: The Case for an Exponential Law of Practice. Psychonomic Bulletin & Review 7, 185-207 (2000)
16. Murray, R.C., Ritter, S., Nixon, T., Schwiebert, R., Hausmann, R.G.M., Towle, B., Fancsali, S., Vuong, A.: Revealing the Learning in Learning Curves. In: Proceedings of the 16th International Conference on Artificial Intelligence in Education, pp. 473-482 (2013)
17. Nixon, T., Fancsali, S., Ritter, S.: The Complex Dynamics of Aggregate Learning Curves. In: Proceedings of the 6th International Conference on Educational Data Mining (2013)