Optimal and Worst-Case Performance of Mastery Learning Assessment with Bayesian Knowledge Tracing

Stephen E. Fancsali, Tristan Nixon, and Steven Ritter
Carnegie Learning, Inc.
437 Grant Street, Suite 918
Pittsburgh, PA 15219, USA
(888) 851.7094 {x219, x123, x122}
{sfancsali, tnixon, sritter}@carnegielearning.com

ABSTRACT

By implementing mastery learning, intelligent tutoring systems aim to present students with exactly the amount of instruction they need to master a concept. In practice, determination of mastery is imperfect. Student knowledge must be inferred from performance, and performance does not always follow knowledge. A standard method is to set a threshold for mastery, representing a level of certainty that the student has attained mastery. Tutors can make two types of errors when assessing student knowledge: (1) false positives, in which a student without knowledge is judged to have mastered a skill, and (2) false negatives, in which a student is presented with additional practice opportunities after acquiring knowledge. Viewed from this perspective, the mastery threshold can be viewed as a parameter that controls the relative frequency of false negatives and false positives. In this paper, we provide a framework for understanding the role of the mastery threshold in Bayesian Knowledge Tracing and use simulations to model the effects of setting different thresholds under different best and worst-case skill modeling assumptions.

Keywords

Cognitive Tutor, intelligent tutoring systems, knowledge tracing, student modeling, mastery learning

1. INTRODUCTION

Carnegie Learning's Cognitive Tutors (CTs) [12] and other intelligent tutoring systems (ITSs) adapt to real-time student learning to provide efficient practice. Such tutors are structured around cognitive models, based on the ACT-R theory of cognition [1-4], that represent knowledge in a particular domain by atomizing it into knowledge components (KCs). CTs for mathematics, for example, present students with problems that are associated with skills that track mathematics KCs in cognitive models. Content is tailored to student knowledge via run-time assessments that probabilistically track student knowledge/mastery of skills using a framework called Bayesian Knowledge Tracing (BKT) [8].

Even in cases in which BKT mastery learning judgments are based on parameters that perfectly match student parameters (e.g., with idealized, simulated student data), assessment of mastery or knowledge is imperfect; student performance need not perfectly track knowledge. In this context, mastery learning assessment is a kind of classification problem. Like all classifiers, an ITS is subject to two types of errors when assessing student knowledge: (1) false positives, in which a student without knowledge is judged to have mastered a skill, and (2) false negatives, in which a student is presented with additional practice opportunities after acquiring knowledge.

A false positive judgment results in pushing a student too quickly through the curriculum. Students pushed too quickly may be asked to demonstrate or use knowledge that they have not yet acquired fully. False negative judgments result in pushing a student too slowly, so the risk is that valuable instructional time is taken teaching KCs that are already mastered, rather than learning new KCs.

Depending on instructional objectives and course design, these two types of errors may not be equally important. If we present a mixed-practice curriculum in which a student will receive more practice on KCs in the future, it may be acceptable to focus on minimizing false negatives. However, if a student will receive only a single block of practice on a KC, particularly if it constitutes important pre-requisite knowledge for later material, then we will strongly prefer to minimize false positives, even at the expense of incurring additional over-practice.

Since detecting mastery typically requires a number of correct trials following the student's attainment of knowledge, a certain amount of "lag" between the point where a student acquires knowledge of a skill and the point where a tutor detects student mastery may be inevitable. We illustrate progression to mastery in Figure 1, dividing opportunities into three phases: learning before the student acquires knowledge of the skill (from opportunity 1 to K), "lag" practice opportunities immediately after knowledge acquisition (K to L), and over-practice after this lag, but before mastery judgment at opportunity M.

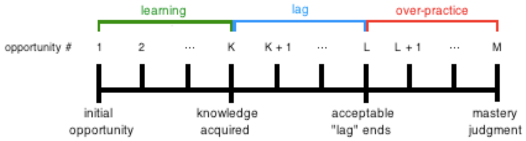

Figure 1. Progression to mastery (judgment) over M student-skill opportunities divided into three "phases"

Despite imperfect assessment of the student, adaptive tutors attempt to minimize the number of opportunities at which students practice skills they have already mastered, so they can focus student practice on skills they have yet to master. We investigate the impact of several factors on the efficiency of practice, focusing especially on the threshold used for student mastery assessments.

We provide a framework for thinking about inherent trade-offs between the two types of CT assessment errors. We quantify the notions of "lag" and over-practice and investigate their relationships with the BKT probability threshold for mastery learning, mastery learning skill parameters, and the dynamics of the student population (or sub-populations) being modeled.

Recent work also focuses on the efficiency of student practice given various methods of individualizing student parameters (e.g., [7] [9]). Increased efficiency has been quantified as the amount of time saved (e.g., by improved cognitive models), without negatively impacting learning [3]. Despite concerns to the contrary, recent work suggests that over-practice is not necessary for long-term retention [7]. Other work conceptualizes the problem of efficient practice roughly as we do, quantifying efficiency and over-practice in terms of expected counts of student opportunities [9]. This work differs in several important ways from past work, especially by focusing on *simulated* data for a variety of skills and quantifying a notion of an acceptable lag while not focusing on individualization.

2. BAYESIAN KNOWLEDGE TRACING

BKT [8] provides a method to track student knowledge acquisition, and it is the basis of the mastery learning implementation in the Cognitive Tutor. For each skill, a student can be in one of two knowledge states: "unknown" or "known." At each opportunity to practice a skill, the student generates an observable (correct or incorrect) response. Four parameters comprise the model of student behavior. The first two are called learning parameters, and the last two are performance parameters:

- $P(L_0)$: initial probability skill is known at first opportunity to practice it
- P(T): probability that student learns the skill (i.e., transition from the unknown to the known state) after an opportunity to practice the skill
- P(G): probability that student produces a correct response at an opportunity despite not knowing the skill ("guessing")
- P(S): probability that student produces an incorrect response at an opportunity despite knowing the skill ("slipping").

Corbett and Anderson [8] provide a well-known algorithm to estimate a student's knowledge state in response to each observable student action and given parameters. The Cognitive Tutor implements run-time mastery assessment using such an algorithm. Mastery of a skill is usually declared when the algorithm determines that a student has 95% probability of being in the known state for the skill. We treat this mastery threshold as a tunable parameter that controls the relative frequency of the two types of mastery assessment errors.

One way to think about false positive errors is as the proportion of students for whom at least one false positive occurred, i.e., the proportion of students for whom mastery of a skill was judged pre-maturely by the run-time algorithm. This provides an accounting of how many students the system moves on to practice new skills too quickly.

The regular occurrence of a modest "lag" (i.e., of false negatives) is both expected and vital to a tutoring system remaining "conservative" in the sense that it infrequently commits false positive errors. Some lag may be required because of uncertainty inherent in the BKT model. We can never be completely sure that correct performance results from student knowledge, rather than a guess. As we observe repeated correct performances, we become more certain that the behavior results from underlying knowledge, rather than just guessing. It is this transition from uncertainty to certainty that is the source of what we will call the "acceptable lag" after knowledge acquisition. The mastery threshold represents the point at which we consider the system to be certain enough to conclude that the student has mastered the skill.

A well-calibrated, adaptive tutoring system should not frequently prescribe large amounts of "over-practice." We quantify the notion of the acceptable lag as well as over-practice and assess the proportion of students that experience over-practice.

3. SIMULATION REGIME

We use the BKT model to generate idealized data for simulated students in a manner comparable to [6] and [11]. For example, if $P(L_0) = 0.5$, $P(T) = 0.35$, $P(G) = 0.1$ and $P(S)=0.1$, then the simulation would, for each simulated student, place the student in the known state initially with a probability of 0.5. Students in the known state would then generate correct responses with a 0.45 probability $[P(L_0)*(1-P(S))]$. Those in the unknown state would generate correct responses with probability 0.1, and have a 0.35 probability of transitioning into the known state. Percent correct on the first opportunity for all students simulated with this skill would be 0.5 $[P(L_0)*(1-P(S))+(1- P(L_0))*P(G)]$.

Since we know exactly when each virtual student transitioned into the known state, we can compare the point where this occurred to the judgment of the BKT run-time mastery algorithm, which can only observe the generated student actions. We apply this testing paradigm to scenarios where the runtime system uses the same BKT parameters as the generating model ("best-case"), and to a couple of scenarios where they are significantly different ("worst-cases").

We simulate data over skills represented by 14 unique parameter quadruples, a subset of those identified in [13] as representative of broad clusters of skills deployed in Cognitive Tutor mathematics curricula[1]. We ascertain the number of "lagged" opportunities we expect students to see, the frequency that the number of lagged opportunities can reasonably be considered over-practice (i.e., beyond the acceptable lag), and the frequency of pre-mature mastery judgment, for one best-case and two worst-case scenarios.

4. RESULTS

There are several ways of thinking about best and worst-case scenarios; we do not exhaust the space of possibilities. We begin by considering a best-case scenario.

4.1 Best Case: Homogeneous, Matching Student Population

In our first round of simulations, we simulate response data, for four mastery threshold probabilities (75%, 90%, 95%, and 98%). We assume that students are homogeneous with respect to a skill, meaning that student behavior is generated probabilistically from the same set of BKT parameters for all students.

For each skill parameter set, we simulate 10,000 students in this way, for up to thirty opportunities per student. Since we are implementing mastery learning, the actual number of opportunities generated by a simulated student depends on when and if the student reaches mastery. In this best-case scenario, the system judges mastery by the same skill parameters that are used

[1] Ritter, et al. [13] identify a total of 23 "cluster" skill quadruples inferred from empirical data collected for thousands of skills in CT curricula. We discard seven quadruples with $P(L_0) > 0.75$. Two quadruples are discarded that cover little of the empirical parameter space and seem to have implausible values for P(G) and P(S) (cf. [5]).

to generate student behaviors. That is, the system is correctly modeling student skills.

Since these are simulated students, we know when students learn each skill (by transitioning from the unknown to known state) in addition to their observed behavior, so we compare the internal knowledge state at opportunities to those at which the BKT run-time algorithm judges a student to have reached a sufficiently high probability of knowing each skill. Given the large sample of students and number of simulated opportunities, individual simulations for each mastery threshold should be comparable.

4.1.1 Efficient Practice

Figure 2 provides frequencies with which values of the median[2] number of lagged or over-practice opportunities occur over the 14 skills for four mastery thresholds. As we increase the mastery threshold, we expect the ordinary student to see more lagged opportunities (i.e., opportunities after knowledge acquisition). At the 95% mastery threshold, for example, we expect the median student to see one to four lagged opportunities on most skills. At the 98% threshold, more skills have a median lag of five or more opportunities.

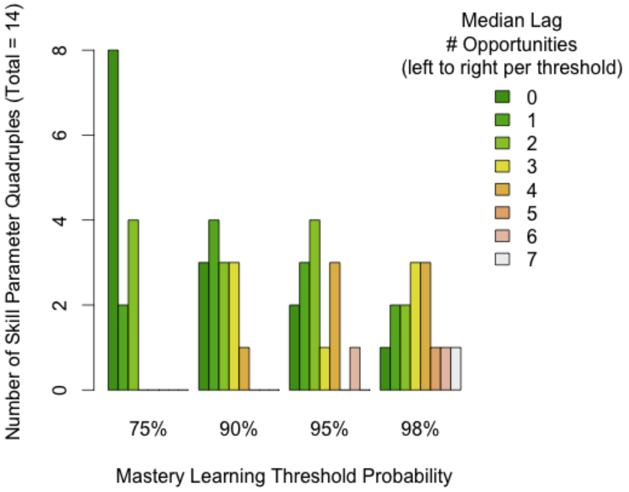

Figure 2. Frequency (# of skills) of median student lag opportunities (i.e., those beyond knowledge acquisition) [14 skills simulated for 10,000 students at each threshold; student BKT parameters match run-time mastery learning parameters.]

Figure 3 provides distributions over skills of the frequency with which student pre-mature mastery judgment (false positives) occurs for the four mastery thresholds. Coupled with Figure 2, we see a trade-off between pre-mature mastery judgments, which decrease, and lagged opportunities, which increase, as we increase the mastery threshold. We expect no more than 5% (indeed, generally less than 5%) of students to be pre-maturely judged as having acquired skill knowledge/mastery at the 95% and 98% thresholds.

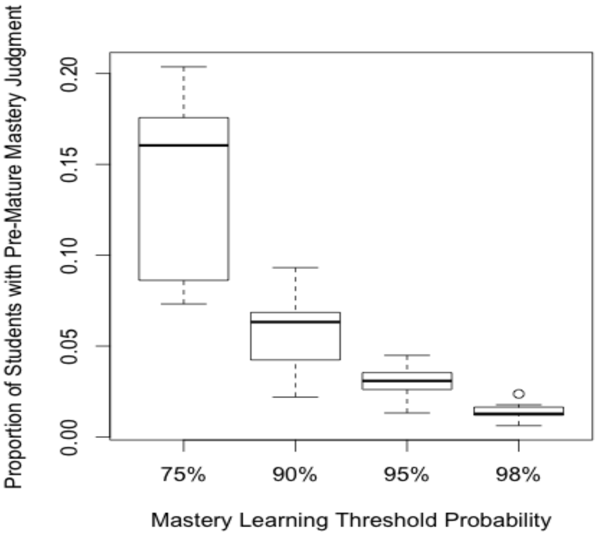

Figure 3. Distribution of proportion of simulated students (per skill) pre-maturely judged to have skill mastery [14 skills simulated for 10,000 students at each threshold; student BKT parameters match run-time mastery learning parameters]

4.1.2 Over-Practice

We seek to quantify over-practice, and the extent to which ideal students endure it, as a function of mastery thresholds and CT mastery learning parameters.

4.1.2.1 Acceptable Lag After Knowledge Acquisition

Recall that the second phase in the progression to mastery in Figure 1 begins at knowledge acquisition (opportunity K) and continues until the end of what we have called the acceptable lag at opportunity L. We define this acceptable number of lagged opportunities for each particular skill and mastery threshold so that we can quantify over-practice as opportunities after knowledge acquisition beyond an acceptable lag.

We begin by noting that properties of the BKT model entail that the inferred probability of student skill knowledge never falls to zero for non-zero $P(L_0)$, $P(T)$, and $P(G)$. Rather, each skill parameter quadruple implies a theoretical minimum probability of knowledge. We can estimate the theoretical minimum probability of knowledge per skill by simulating "runs" of consecutive incorrect student opportunities and noting the asymptotic value to which the probability of knowledge decreases.

Figure 4 shows how the BKT estimated probability of knowledge for the skill with $P(L_0) = 0.631$, $P(T) = 0.11$, $P(G) = 0.282$, and $P(S) = 0.228$, decreases over a series of consecutive incorrect responses from $P(L_0)$ to its theoretical minimum at roughly 0.161 in about six opportunities.

We determine the number of consecutive correct opportunities required to take a student from the theoretical minimum probability of knowledge, for each skill, to the mastery threshold probability. The length of this run is the acceptable amount of lagged opportunities[3]; any simulated student that encounters a lag

[2] With few exceptions, in all simulations we report, for each skill, the median, mean, and modal number of lagged opportunities are relatively close in value. The median takes on an integral value in these simulations, making it more readily interpretable.

[3] This definition presumably provides an upper bound for this value. Other definitions may be appropriate (e.g., based on analysis of empirical data), but we leave this topic for future research.

of opportunities with length greater than the acceptable number is considered to encounter over-practice for that skill. For the skill with theoretical minimum illustrated in Figure 4, the acceptable lags at the 75%, 90%, 95%, and 98% thresholds are 3, 4, 4, and 5, respectively. We determine the proportion of students who encounter over-practice and compare this to the proportion of students with pre-mature mastery judgment.

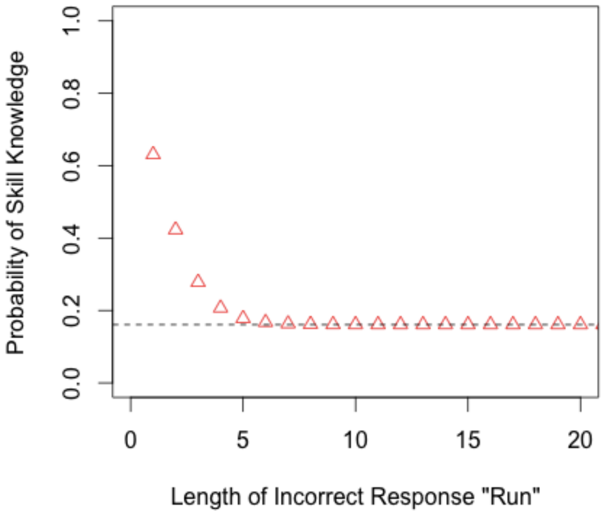

Figure 4. Illustration of theoretical minimum probability of skill knowledge over "run" of consecutive incorrect responses for a skill [$P(L_0)$ = 0.631; P(T) = 0.11; P(G) = 0.282; P(S) = 0.228]; dashed-line marks approximate asymptote at 0.161.

4.1.2.2 *Frequency and Magnitude of Over-Practice*

Figure 5 shows that by increasing the threshold for mastery we tend to increase the proportion of students to whom over-practice opportunities are provided[4] (as well as the variability of this proportion over skills). This illustrates the trade-off between pre-mature mastery judgment and over-practice (in addition to the noted trade-off between pre-mature mastery and lagged practice opportunities).

Notably, increasing the threshold does not drastically increase the *number* of over-practice opportunities the median simulated student is expected to see, only the probability that a student will get some over-practice. The median number of over-practice opportunities per student-skill interaction with over-practice is 1 for the 75%, 90%, and 95% thresholds, and 2 for the 98% threshold. While over-practice is assigned for roughly 30% of students at the traditional 95% threshold, the median student, for most skills, does not experience much over-practice. The traditional 95% mastery criterion seems to embody a conservative tradeoff: some students will receive a small amount of over-practice, but pre-mature mastery judgment is mostly avoided.

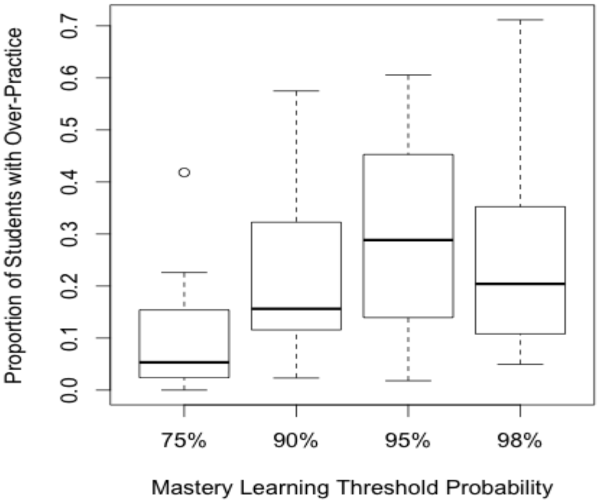

Figure 5. Distribution of proportion of simulated students (per skill) assigned over-practice opportunities (i.e., at least one opportunity beyond the acceptable lag for a particular skill) [14 skills simulated for 10,000 students at each threshold; student BKT parameters match run-time mastery learning parameters.]

4.2 Worst-Case #1: Homogenous, Non-Matching Student Population

Next, we consider one type of worst-case scenario. Simulated students are drawn from a homogenous population, but these student parameters uniformly mismatch the BKT parameters used for run-time mastery assessment. That is, the system is doing the poorest possible job of modeling the student's learning parameters. For the same 14 skills, we specify mismatched student generating parameters in the following manner where P_S stands for mismatched student parameters and P_M corresponds to mastery assessment parameters that will be used:

- $P_S(L_0) = 1 – P_M(L_0)$
- $P_S(T) = 1 – P_M(T)$
- $P_S(G) = 0.5 – P_M(G)$
- $P_S(S) = 0.5 – P_M(S)$

We generate data, in the same manner as the previous section, for 10,000 students for up to thirty opportunities.

4.2.1 *Efficient Practice*

Figure 6 provides the frequency with which particular median lagged opportunity counts occur over the 14 skills at the four mastery thresholds. We see a shift toward more skills with greater median counts of lagged opportunities, especially at the 95% and 98% thresholds. Coupled with Figure 7, we see evidence of the same trade-off between pre-mature mastery judgment and lag opportunities as in the best-case scenario, but we find (mostly) far lower proportions of students at each threshold with pre-mature mastery judgment.

Figure 7 also makes apparent that two skills have substantially greater proportions of students with pre-mature mastery. These correspond to two of the 14 skills with P(T) > 0.8. Since they have a high mastery learning $P_M(T)$ parameter and simulated students have P(T) < 0.2, the mastery learning assessment naturally counts students as acquiring knowledge pre-maturely with greater frequency.

[4] It is not clear to us why the median proportion for skills at the 98% threshold is lower than the 95% threshold. These median values are closer (with the median at 98% greater than that at 95%) under other conditions we describe later.

For this scheme of generating worst-case, mismatching student parameters (and the corresponding 14 skills' parameters), we again find evidence that BKT mastery assessment is generally conservative, erring on the side of providing students with more opportunities after knowledge acquisition, rather than prematurely judging mastery.

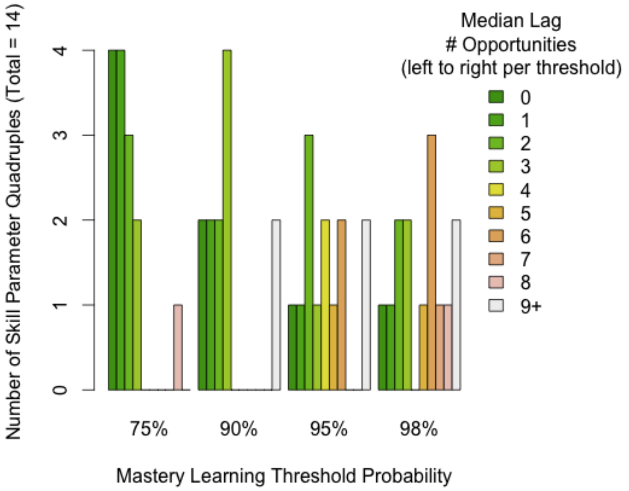

Figure 6. Frequency (count of skills) of median student opportunities beyond knowledge acquisition for four run-time mastery threshold probabilities [Student BKT parameters uniformly "mismatch" 14 run-time mastery learning parameters.]

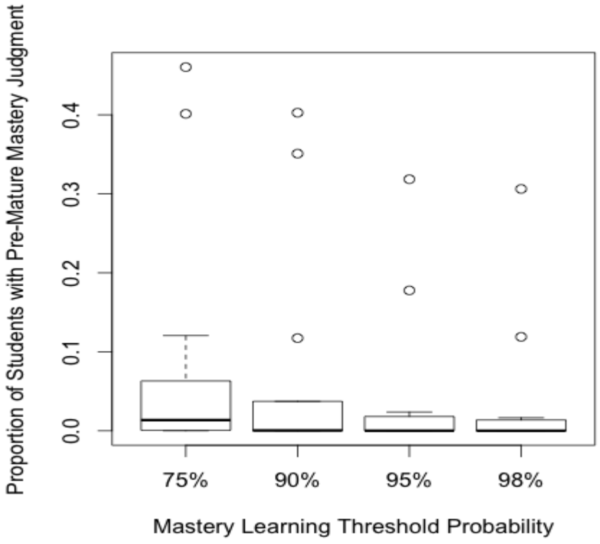

Figure 7. For "mismatched" student skill parameters, distribution of proportion of simulated students (per skill) pre-maturely judged to have skill mastery, grouped by run-time mastery thresholds

4.2.2 Over-Practice

Compared to the best case "matching" parameter scenario, Figure 8 shows that the proportions of students that experience over-practice at each mastery threshold are far greater (and increase with increasing mastery threshold). However, the amount of over-practice through which simulated students must work remains modest; the median student that experiences over-practice sees 2 over-practice opportunities per skill over all the skills at the 75% threshold, 3 at the 90% and 95% threshold, and 4 at the 98% threshold. Again, we do not witness particularly onerous over-practice in general, despite the mismatch of student parameters and mastery assessment parameters.

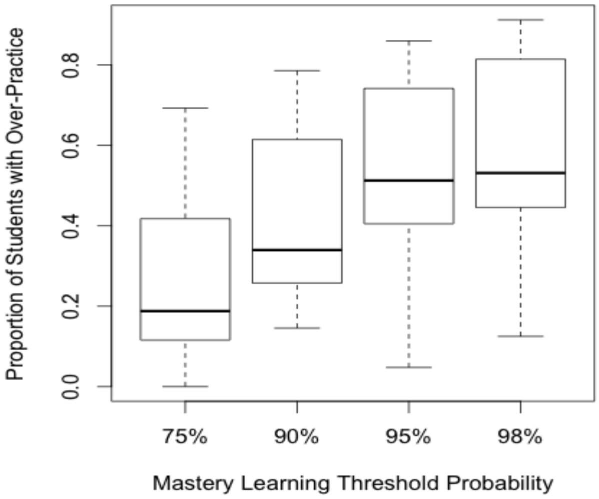

Figure 8. For "mismatched" student skill parameters, distribution of proportion of simulated students (per skill) assigned over-practice opportunities

4.3 Worst Case #2: Completely Heterogeneous Population, Random Student Parameters

We now consider simulating maximally heterogeneous populations of simulated students for the same 14 skills. Mastery learning parameters correspond to the 14 skill parameter quadruples, but for each skill and each student, BKT parameters for a generating model are randomly sampled as follows:

- $P_S(L_0)$, $P_S(T)$ ~ Uniform(0.0, 1.0)
- $P_S(G)$, $P_S(S)$ ~ Uniform(0.0, 0.5).

This corresponds to testing each CT skill parameter quadruple for robustness against a worst-case in which there are no stable sub-populations of students; data for each student are drawn from a different, random generating model. We consider how such a scenario would affect the CT's ability to assign efficient practice based on BKT mastery assessment.

4.3.1 Efficient Practice

For worst-case #2, the pattern of trade-offs in efficient practice are the same as in the previous two cases, but frequencies of median lagged opportunities (Figure 9) and proportions of students judged for mastery pre-maturely (Figure 10) fall in between the best-case scenario and the previous worst-case scenario. There is also a larger variance in the distribution of the proportion of students per skill that get pre-mature mastery judgments than in either of the previous two scenarios we have considered.

This accords with our expectations, as randomly generated parameters will sometimes be very close (and other times far removed) from each skill's mastery learning parameters. Thus, a maximally heterogeneous population is really a mixture of best-case students, worst-case students, and students somewhere in the "middle" of the two. Further, given random generation of student parameters we reasonably expect an increase in variance in pre-mature mastery judgment compared to either of the homogeneous simulated student populations we have considered. As with the mismatching student population, the two outlier skills for premature mastery judgment are those with P(T) > 0.8, but the proportion for these two skills is smaller than in the case of the mismatching population, as we expect.

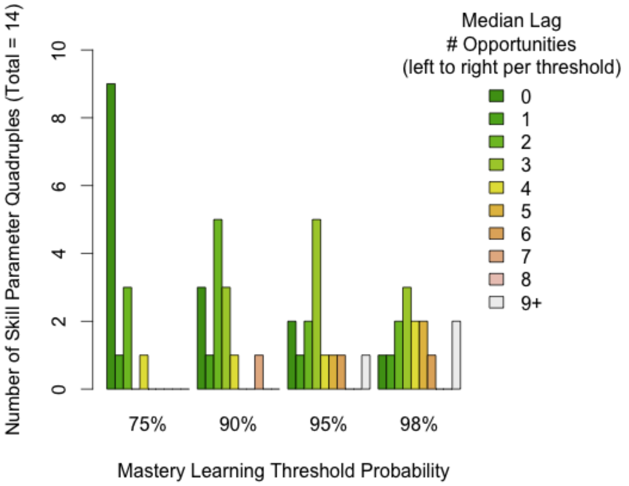

Figure 9. Frequency (count of skills) of median student over-practice opportunities for four run-time mastery threshold probabilities with random student BKT parameters & 14 "cluster" skill quadruples as mastery parameters

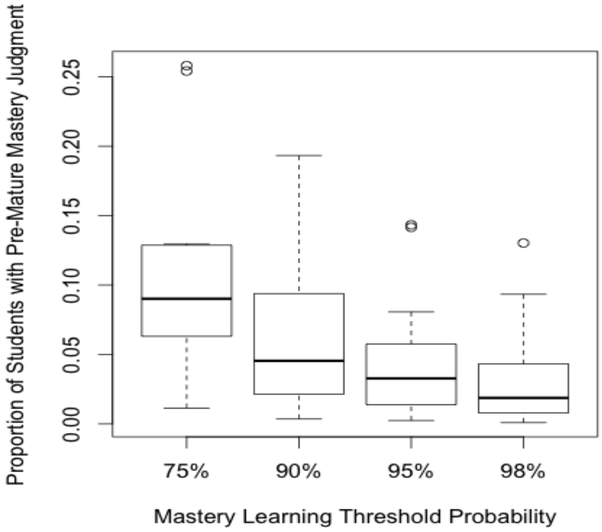

Figure 10. For random student skill parameters, distribution of proportion of simulated students (per skill) pre-maturely judged to have skill mastery

4.3.2 Over-Practice

As in the previous two cases, we see that the proportions of students experiencing over-practice per skill generally increase as we increase the mastery threshold probability (Figure 11). Further, most values of these proportions fall roughly in between the medians for the previous two cases. Median counts of over-practice opportunities over all skills at each mastery threshold are also similar to those in the other scenarios (median = 2 for 75%, 90%, and 95%; median = 3 for 98%).

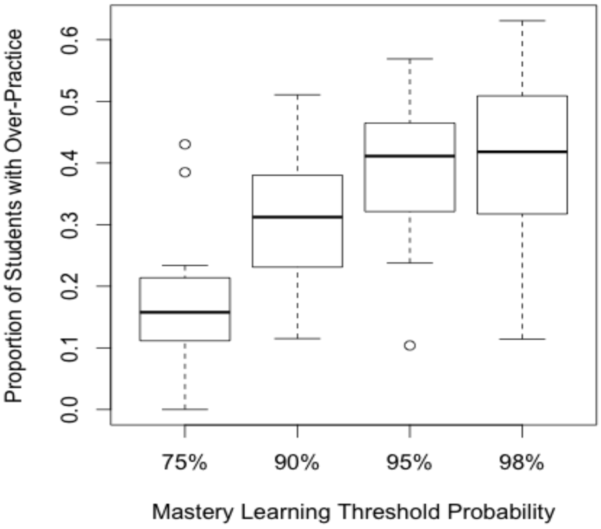

Figure 11. For random student skill parameters, distribution of proportion of simulated students (per skill) assigned over-practice opportunities

5. VISUALIZATION WITH ROC CURVES

One way to conceptualize assessing student skill mastery is as the problem of "detecting" learning from a noisy signal. With errors in mastery assessment cast in terms of false positives and false negatives, or Type I and Type II errors, a natural way to visualize mastery learning classification of student-skill opportunities and trade-offs between false positives and false negatives, as we adjust the mastery threshold, is via Receiver Operating Characteristic (ROC) curves.

Figure 12 provides scatterplots of true positive rate versus false positive rate (roughly ROC curve graphs) for each skill from our simulations with random student skill parameters, grouped by mastery threshold probability. The cluster of points at the bottom left (along with the dearth of points, two outliers aside, in the center and far right of the graph) indicate the relative "conservativeness" of BKT mastery learning for the skill parameters we consider. These points represent true positive rates that are relatively low mostly because of increased false negative errors, corresponding to opportunities that lag student knowledge acquisition (whether over-practice beyond an acceptable lag or not).

Recall that we take a greater proportion of false negatives than false positives to be a virtue of a conservative tutoring system that decreases the risk of pushing students along too quickly. Two outliers on the graph (the two right-most points of the graph) represent skills each with false positive rates near 0.5 and 1, for the 90% and 75% mastery thresholds, respectively. Those points correspond to the same two skills we identified in §4.2.1 and §4.3.1 with P(T) > 0.8. We find a similar cluster of points (and overall structure) in the same graph constructed over best-case simulations. High P(T) means we assume that students learn quickly, so these results may indicate that such an assumption for a skill is more likely to produce high false positive rates, especially for lower mastery thresholds, regardless of the makeup of the student population.

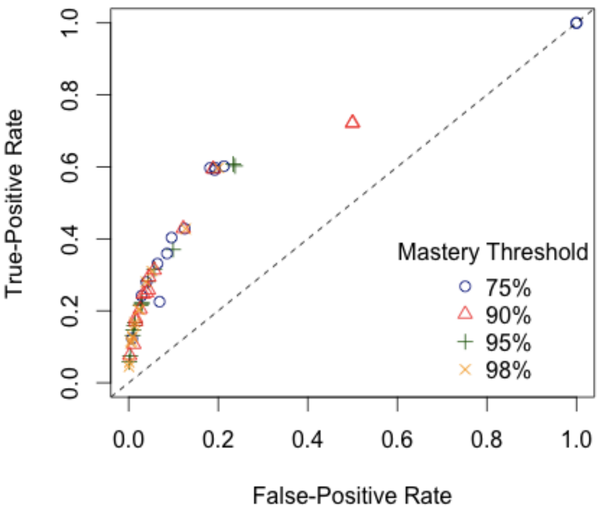

Figure 12. Scatterplots of true positive rate versus false positive rate (roughly ROC graphs) for each skill from simulations with random student skill parameters, grouped by mastery threshold probability

Since rates are calculated over student practice opportunities, false positive instances are limited to one per student per skill, because the student no longer practices a skill after mastery judgment. However, there can be many false negative errors for a single student. Simulating students without mastery learning, in general, would allow us to ascertain more balanced false positive rates for BKT mastery learning.

Analogous graphs can be constructed at the level of students. This introduces some complications, since the assessment of the classification of a student as a true positive presumably depends not just on the performance of the mastery algorithm, but also on the estimation of acceptable lag for that skill. Many students will be counted as true positives, despite the fact that the tutor is committing false negative errors within the acceptable lag at the transaction level. As such, we should expect a lower false negative rate at the student level than at the opportunity level. We leave such extensions for future research, but note that there are important similarities between the type of analysis we provide and "signal detection for learning."

6. DISCUSSION

The trade-off, as a function of mastery probability threshold, of student pre-mature mastery judgment (false positives), lagged skill opportunities, and over-practice (false negatives) is consistent across different best-case and worst-case skill modeling assumptions. The value of the mastery probability threshold and skill modeling assumptions influence the magnitude of these error rates when calculated as proportions of students pre-maturely judged to have achieved mastery or subjected to over-practice. However, we find that for the median student subjected to over-practice, regardless of the skill modeling scenario, the amount of over-practice as a count of opportunities is not, on the surface, particularly onerous. Thus, BKT and a variety of skill parameters are generally robust to committing the types of errors we have quantified, with the exception of two outlier skills we discovered with P(T) > 0.8 for which pre-mature mastery judgment occurred more frequently in the two worst-case scenarios. This suggests that without strong empirical evidence that certain skills are learned quickly, we should err on the side of setting lower values of P(T) for mastery learning.

Over all three simulation scenarios, we find that the conventional 95% mastery threshold probability leads to pre-mature mastery judgment for under 5% of students per skill for the majority of cases. Further, onerous over-practice is generally not assigned to students at these thresholds, making it less likely that student time will be "crowded out" by practice for skills they have already mastered and making it more likely that they will cover more course material over all. We emphasize that our results are limited to the skill parameter quadruples we considered that are broadly representative of those deployed in the CT [13], but broader patterns (e.g., that using P(T) > 0.8 leads to more false positives) seem to emerge and should be studied further.

Beyond this computational, pragmatic justification for the conventionally deployed mastery threshold, we provide a framework for thinking about the BKT mastery threshold as a parameter that can be tuned according to course developers' appetite for risk, in the sense of trading off false positives for false negatives, and how each type of error will affect student learning, course completion, and other instructional outcomes. There is potential that a moderate amount of over-practice might have additional value in preventing future forgetting. This framework and these (or similar) results could be used to calibrate tutors to optimize student practice for future retention. Finally, we provided a better theoretical understanding of optimal performance of BKT-based mastery learning.

This work calls for extension in several ways. Beyond considering our relatively limited best-case and worst-case scenarios, we should investigate a greater range of average-case possibilities. For example, students with diverse prior knowledge, learning rates, and other learning characteristics use real-world ITSs. How much fine-tuning of run-time mastery learning parameters, to student sub-populations or even individual students (e.g., [10], [11]), is necessary to prevent both over-practice of skills and pre-mature mastery judgment?

Future work should also address a broader, more exhaustive range of BKT parameter quadruples. Those we analyze here are important because they are representative of real-world data collected over thousands of students and skills, but we should seek a better understanding of the full parameter space and how different parameter combinations interact with factors like student sub-population composition and mastery learning thresholds, a greater number of which should also be systematically explored and tested. Finally, depending on the investigator and educator interests, the nature of particular curricula, and other concerns, exploring alternative conceptualizations of over-practice and under-practice (cf. [7]) (and their connection to our work based on practice opportunities counts) is an interesting avenue for future research.

7. ACKNOWLEDGMENTS

The authors acknowledge the helpful comments of Robert Hausmann, R. Charles Murray, Brendon Towle, and Michael Yudelson on an early draft of this paper.

8. REFERENCES

[1] Anderson, J.R. 1983. *The architecture of cognition*. Harvard UP, Cambridge, MA.

[2] Anderson, J.R. 1990. *The adaptive character of thought*. Erlbaum, Hillsdale, NJ.

[3] Anderson, J.R. 1993. *Rules of the mind*. Erlbaum, Hillsdale, NJ.

[4] Anderson, J.R., Bothell, D., Byrne, M.D., Douglass, S., Lebière, C., Qin, Y. 2004. An integrated theory of the mind. *Psychological Rev.* 111, (2004), 1036-1060.

[5] Baker, R.S.J.d., Corbett, A.T., Aleven, V. 2008. More accurate student modeling through contextual estimation of slip and guess probabilities in Bayesian Knowledge Tracing. In *Proceedings of the 9th International Conference on Intelligent Tutoring Systems* (Montreal, Canada, 2008). 406-415.

[6] Beck, J., Chang, K. 2007. Identifiability: a fundamental problem of student modeling. In *Proceedings of the 11th International Conference on User Modeling* (Corfu, Greece, 2007). 137-146.

[7] Cen, H., Koedinger, K.R., Junker, B. 2007. 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* (Los Angeles, USA, 2007). 511-518.

[8] Corbett, A.T., Anderson, J.R. 1995. Knowledge tracing: modeling the acquisition of procedural knowledge. *User Modeling & User-Adapted Interaction* 4, (1995), 253-278.

[9] Lee, J.I., Brunskill, E. 2012. The impact of individualizing student models on necessary practice opportunities. In *Proceedings of the 5th International Conference on Educational Data Mining* (Chania, Greece, 2012), 118-125.

[10] Pardos, Z.A., Heffernan, N.T. 2010. Modeling individualization in a Bayesian networks implementation of Knowledge Tracing. In *Proceedings of the 18th International Conference on User Modeling, Adaptation and Personalization* (Hawaii, USA, 2010), 255-266.

[11] Pardos, Z.A., Heffernan, N.T. 2010. Navigating the parameter space of Bayesian Knowledge Tracing models: Visualizations of the convergence of the Expectation Maximization algorithm. In *Proceedings of the 3rd International Conference on Educational Data Mining* (Pittsburgh, USA, 2010), 161-170.

[12] Ritter, S., Anderson, J.R., Koedinger, K.R., Corbett, A.T. 2007. Cognitive Tutor: applied research in mathematics education. *Psychonomic Bulletin & Rev.* 14, 2 (2007), 249-255.

[13] Ritter, S., Harris, T.K., Nixon, T., Dickison, D., Murray, R.C., Towle, B. 2009. Reducing the knowledge tracing space. In *Proceedings of the 2nd International Conference on Educational Data Mining* (Cordoba, Spain, 2009), 151-160.