Experimental Analysis of Mastery Learning Criteria

Radek Pelánek
Masaryk University
Brno, Czech Republic
pelanek@mail.muni.cz

Jiří Řihák
Masaryk University
Brno, Czech Republic
thran@mail.muni.cz

ABSTRACT

A common personalization approach in educational systems is mastery learning. A key step in this approach is a criterion that determines whether a learner has achieved mastery. We thoroughly analyze several mastery criteria for the basic case of a single well-specified knowledge component. For the analysis we use experiments with both simulated and real data. The results show that the choice of data sources used for mastery decision and setting of thresholds are more important than the choice of a learner modeling technique. We argue that a simple exponential moving average method is a suitable technique for mastery criterion and propose techniques for the choice of a mastery threshold.

KEYWORDS

mastery learning; learner modeling; Bayesian knowledge tracing; exponential moving average

ACM Reference format:
Radek Pelánek and Jiří Řihák. 2017. Experimental Analysis of Mastery Learning Criteria. In *Proceedings of UMAP '17, Bratislava, Slovakia, July 09-12, 2017,* 8 pages.
DOI: http://dx.doi.org/10.1145/3079628.3079667

1 INTRODUCTION

Mastery learning is an instructional strategy that requires learners to master a topic before moving to more advanced topics. A key aspect of mastery learning is a mastery criterion – a rule that determines whether a learner has achieved mastery. Mastery criteria have been studied already 40 years ago [6, 15, 24], but at that time typically only for static tests and small scale applications. Nowadays, mastery learning is used on large scale in dynamic, adaptive educational systems [11, 22].

A typical application of mastery criterion within a modern educational system is the following. A learner solves a problem or answers a question in the system. Data about learner performance are summarized by a model of learner knowledge or by some summary statistic. Mastery criterion takes this summary and produces a binary verdict: "mastered" or "not mastered". Based on this verdict, the system adapts its behavior: it either presents more problems from the same topic or moves the learner to another topic. The mastery criterion typically takes an external parameter (threshold), which specifies its strictness.

Mastery criterion gives a binary output. Most education systems use some kind of visualization (e.g., progress bars, skillometers, open learner models) that give learners a sense of progress towards the mastery goal. These visualizations are closely related to mastery criterion; in fact they can often be viewed as mastery criterion with different thresholds.

In this paper we thoroughly analyze the basic scenario for detecting mastery – we assume to have well-specified fine-grained knowledge components, i.e., sets of items related to same skill, such that these items can be treated as indistinguishable (potentially differentiated by simple parameters as difficulty or time intensity). We do not consider relation between knowledge components (e.g., prerequisites).

Simple mastery criteria are N correct in row [11, 13] or average success rate from last N attempts. More complex methods are based on models of learner knowledge and the use of mastery threshold policy. A model provides probabilistic prediction that the next answer will be correct and mastery is declared if the prediction is over a given threshold.

There exists an extensive research on learner modeling [5]. This research typically uses personalization through mastery learning as a motivation. However, evaluation of models is typically not done by evaluating the impact on mastery criterion, but instead using evaluation of predictive accuracy on historical data using metrics like RMSE or AUC [19]. Evaluation of mastery criterion is more difficult, particularly because mastery is a latent construct that cannot be directly measured.

Recent research studied impact of learner models on mastery decision. The ExpOps method [14, 23] gives an expected number of opportunities needed. This estimate is computed without using learner data, just based on assumptions of the used model, so the provided estimate may be misleading if the assumptions do not correspond to the behavior of real learners. Another proposal are effort and score metrics [10], which use historical data to estimate the effort needed to reach mastery and the performance after mastery.

Most of the research on mastery criterion was done in relation with the Bayesian knowledge tracing (BKT) model [4]; this model is also often used in practice with a standard mastery threshold 0.95. The role of this threshold was analyzed by Fancsali et al. [7, 8] by using simulated data (generated by the BKT model) to show the relation between the threshold and proportion of learners with premature mastery and over-practice. Simulated data were also used by Pardos and Yudelson [16] to study mean absolute deviation from the "true moment of learning". They focused on the analysis of a relation between predictive accuracy metrics and moment of learning detection. Baker et al. [2] also studied the moment of learning using

UMAP '17, Bratislava, Slovakia

DOI: http://dx.doi.org/10.1145/3079628.3079667

the BKT model, but they focused on "hindsight analysis" with the use of the full sequence of learner attempts. The goal was to detect at which moment learning occurred using a rich set of features (e.g., response times, hint usage). Yudelson and Koedinger [26] used several large data sets to study differences in mastery decisions done by two variants of the BKT model and showed that the impact of replacing standard model with individualized can be substantial (as measured by time spent).

Recent research [12, 23] proposed general instructional policies applicable to any predictive model: predictive similarity [23] and predictive stability [12] policies. For evaluation authors used the above described techniques: ExpOps [23] and effort and score metrics [12]. These instructional policies focus on stopping not just in the case of mastery, but also for wheel-spinning learners who are unable to master a topic [3]. These works, however, pay little attention to the choice of thresholds.

In this work we analyze different mastery criteria using experiments with both simulated and real data. We compare mastery decisions by the standard BKT model and the basic N consecutive correct criterion. We analyze the decisions of the exponential moving average method under different situations. We also explore the impact of usage of response times in mastery criterion. We explore several techniques for the analysis, including comparison with ground truth (for simulated data) and novel effort-score graphs.

Based on results of these experiments we argue that it is sufficient to use simple methods for mastery criterion; specifically, the exponential moving average method is simple and sufficiently flexible to fit many applications. Rather than focusing on the choice and parameter fitting of learner models, it is more important to focus on tuning mastery thresholds and on the choice of data that are used to make the decision (e.g., whether to use of response times).

2 TECHNIQUES FOR DETECTING MASTERY

Our aim is not to introduce new mastery criteria, but to provide insight into behavior of already proposed criteria. In this section we describe previously used methods in a single setting.

2.1 Notation

We consider only the case of learning for a single knowledge component. We assume that for each learner we have a sequence of answers to items belonging to this knowledge component. Examples of knowledge components and items are "single digit multiplication" with items like "6×7" or "indefinite articles" with items like "a/an orange".

We use the following notation: k is the order of an attempt, θ_k is a skill estimate for the k-th attempt, $P_k \in [0, 1]$ is a predicted probability of correct answer at the k-th attempt, and c_k gives the observed correctness of the k-th attempt. As a basic case we consider only correctness of answer, i.e., dichotomous $c_k \in \{0, 1\}$. It is also possible to consider "partial credit" answers (e.g., based on the usage of hints or on the response time), i.e., real valued $c_k \in [0, 1]$. A typical mastery criterion is a "mastery threshold criterion" which uses a threshold T and declares mastery when the skill estimate (or alternatively the probability of correct answer) is larger than T.

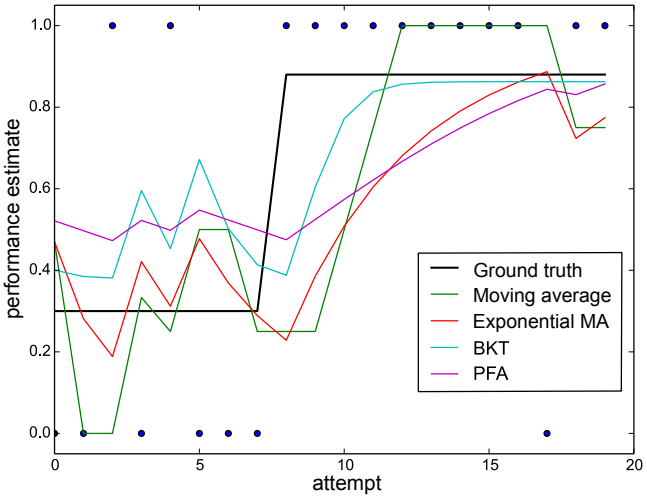

Figure 1: Illustrative comparison of performance estimation techniques for a sequence of answers of a a single simulated student (B2 from Table 1). The black line is the ground truth probability of correct answer and the dots are the simulated answers.

Figure 1 provides specific illustration based on simulated data. Since these are simulated data, we know the ground truth (the black line; mastery achieved at the 8th attempt). The colored lines show estimates by several methods. Mastery decisions would depend on particular thresholds, e.g., for a threshold $T = 0.9$ the moving average method would declare mastery at the 12th attempt.

2.2 Methods without Assumptions about Learning

Basic mastery criteria use only simple statistics about past answers without explicitly modeling the learning process.

2.2.1 Consecutive Correct. The simplest mastery criterion is "N consecutive correct answers" (NCC) (N-in-row, "streak"). With this method we simply count the number of consecutive correct answers and declare mastery once the count reaches the threshold N. As a progress bar we can simply use the current count. One of the disadvantages of this method is that any mistake (even if it is just a typo) means that the learner has to "start again from zero" and this can be demotivating. Nevertheless, this simple method is often practically used and can be successful [13].

2.2.2 Moving Average. Another simple statistics that can be used for mastery criterion is moving average. The basic average for a moving window of size n is $\theta_k = \frac{1}{n} \sum_{i=1}^{n} c_{k-i}$. In addition to n we now need a second parameter: a threshold T. Mastery is declared when $\theta_k \geq T$. One disadvantage of this approach is that it is not suitable for a progress bar. Consider a window of size $n = 6$ and a recent history of attempts $1, 1, 0, 1, 0$ ($\theta_k = 0.6$). If the learner answers correctly, the recent history becomes $1, 0, 1, 0, 1$ and the moving average remains the same ($\theta_{k+1} = 0.6$), i.e., the progress bar does not improve after the correct answers.

A natural extension, which circumvents this problem, is to use weighted average and give more weight to recent attempts, i.e., $\theta_k = \sum_{i=1}^{k} w_i \cdot c_{k-i} / \sum_{i=1}^{k} w_i$, where w_i is a decreasing function (this approach is equivalent to the "time decay" approach discussed in [18] and also closely related to [9]).

2.2.3 Exponential Moving Average. The moving average approach is often used specifically with exponential weights; this variant is called exponential moving average (EMA). This choice of weights often provides good performance [18] and it has the practical advantage of easy implementation, since it can be easily computed without the need to store and access the whole history of learners attempts.

If we choose the weights to be given by an exponential function $w_i = (1-\alpha)\alpha^{(i-1)}$, we can compute the exponential moving average θ_k after k steps as follows:

- initialization: $\theta_0 = 0$,
- update: $\theta_k = \alpha \cdot \theta_{k-1} + (1-\alpha) \cdot c_k$.

The mastery criterion remains $\theta_k \geq T$.

2.3 Methods based on Learner Models

A more sophisticated approach to detecting mastery is based on the usage of learner models. These models estimate learners knowledge and predict the probability that the next answer will be correct. These models are naturally used with the mastery threshold rule – mastery is declared once the estimate of knowledge is above a given threshold. Note that learner models can be used also with more complex instructional policies, e.g., predictive similarity [23] and predictive stability [12]. These policies deal not just with mastery, but also with wheel-spinning learners that are unable to master a topic. In this work, however, we consider only the basic mastery threshold policy.

2.3.1 Bayesian Knowledge Tracing. Bayesian knowledge tracing (BKT) [4] assumes a sudden change in knowledge. It is a hidden Markov model where skill is a binary latent variable (either learned or unlearned). The model has 4 parameters: P_i is the probability that the skill is initially learned, P_l is the probability of learning a skill in one step, P_s is the probability of incorrect answer when the skill is learned (slip), and P_g is the probability of correct answer when the skill is unlearned (guess). Note that BKT can also include forgetting; the described version corresponds to the variant of BKT that is most often used in research papers.

The estimated skill is updated using a Bayes rule based on the observed answers; the prediction of student response is then done based on the estimated skill. In the following we use θ_k and θ'_k to distinguish prior and posterior probability during the Bayesian update (θ_k is the prior probability that the skill is learned before the k-th attempt and θ'_k is the posterior probability that the skill is learned after we have taken the k-th answer into account):

$$\begin{aligned}
\theta_1 &= P_i \\
\theta'_k &= \begin{cases} \frac{\theta_k(1-P_s)}{\theta_k(1-P_s)+(1-\theta_k)P_g} & \text{if } c_k = 1 \\ \frac{\theta_k P_s}{\theta_k P_s+(1-\theta_k)(1-P_g)} & \text{if } c_k = 0 \end{cases} \\
\theta_{k+1} &= \theta'_k + (1-\theta'_k)P_l \\
P_k &= P_g \cdot \theta_k + (1-P_s) \cdot (1-\theta_k)
\end{aligned}$$

Estimation of model parameters (the tuple P_i, P_l, P_s, P_g) can be done using several techniques (the expectation-maximization algorithm, stochastic gradient descent or exhaustive search).

2.3.2 Logistic Models. Another commonly used class of learner models are models based on logistic function, e.g., Rasch model, Performance factor analysis [17], or the Elo rating system [18]. These models utilize assumption of a continuous latent skill $\theta \in (-\infty, \infty)$ and for the relation between the skill and the probability of correct answer use the logistic function $\sigma(x) = \frac{1}{1+e^{-x}}$ (the function can be easily extended to capture guessing in multiple-choice questions).

A simple technique of this type is Performance factor analysis (PFA) [17]. The skill estimate is given by a linear combination of the initial skill and past successes and failures of a student: $P_k = \sigma(\beta + \gamma \cdot s_k + \delta \cdot f_k)$, where β is the initial skill, s_k and f_k are counts of previous successes and failures of the student during the first k attempts, γ and δ are parameters that determine the change of the skill associated with a correct and incorrect answer. Parameters β, γ, δ can be easily estimated using standard logistic regression.

3 ANALYSIS AND COMPARISON OF CRITERIA

Now we compare the described mastery criteria under several circumstances and discuss general methodological issues relevant to the evaluation of mastery criteria.

3.1 Data

For our analysis we use both real and simulated data, since each of them has advantages and disadvantages. Real data directly correspond to practical applications. However, the evaluation of mastery criteria is difficult, since mastery is a latent construct and we do not have objective data for its evaluation. With simulated data we know the ground truth and thus we can perform more thorough evaluation, but the results are restricted to simplified conditions and depend on the choice of simulation parameters.

3.1.1 Simulated Data. For generating simulated data we use both the BKT model and a logistic model. We have selected parametrizations of these models in such a way as to cover a wide range of different learning situations (e.g., high/low prior knowledge, slow/fast learning, high/low guessing). Table 1 provides description of simulation scenarios used in experiments. In all cases we generate 50 answers for each learner.

The BKT model is used in its basic form of the model. It can be used in a straightforward way to generate data and the ground truth mastery is clearly defined by the model. For the logistic model we consider a simple linear growth of the skill. More specifically, for the initial skill θ_0 we assume normally distributed skill $\theta_0 \sim N(\mu, \sigma^2)$ and we consider linear learning $\theta_k = \theta_0 + k \cdot \Delta$, where Δ is either a global parameter or individualized learning parameter. In the case of individualized Δ we assume a normal distribution of its values with a restriction $\Delta \geq 0$. As a ground truth mastery for this model we consider the moment when the simulated learner has 0.95 probability of answering correctly according to the ground truth parameters.

Table 1: Specification of models used for generating simulated data. "Bn" are BKT models, "Ln" are logistic models.

	Parameters			
B1	$P_i = 0.15$	$P_l = 0.35$	$P_s = 0.18$	$P_g = 0.25$
B2	$P_i = 0.25$	$P_l = 0.08$	$P_s = 0.12$	$P_g = 0.3$
B3	$P_i = 0.1$	$P_l = 0.2$	$P_s = 0.1$	$P_g = 0.15$
B4	$P_i = 0.1$	$P_l = 0.3$	$P_s = 0.4$	$P_g = 0.05$
B5	$P_i = 0.05$	$P_l = 0.1$	$P_s = 0.06$	$P_g = 0.2$
B6	$P_i = 0.1$	$P_l = 0.05$	$P_s = 0.1$	$P_g = 0.5$
L1	$\theta_0 \sim N(-1.0, 1.0)$		$\Delta = 0.4$	
L2	$\theta_0 \sim N(-0.4, 2.0)$		$\Delta = 0.1$	
L3	$\theta_0 \sim N(-2.0, 2.0)$		$\Delta = 0.15$	
L4	$\theta_0 \sim N(0.0, 0.7)$		$\Delta \sim N(0.15, 0.1)$	
L5	$\theta_0 \sim N(-2, 1.3)$		$\Delta \sim N(0.45, 0.15)$	
L6	$\theta_0 \sim N(-0.7, 1.5)$		$\Delta \sim N(0.6, 0.3)$	

The source codes of all experiments with simulated data is available[1].

3.1.2 Real Data. We use real data from two educational systems. The first is a system for practice of Czech grammar and spelling (`umimecesky.cz`). The system implements mastery learning based on the exponential moving average method. The system visualizes progress using progress bar with highlighted thresholds (mastery levels) 0.5, 0.8, 0.95, and 0.98. The main mastery level (used for example for evaluation of homework within the system) is given by a threshold 0.95. The value of α depends on the type of exercise. For the analysis we use data from the basic grammar exercise with multiple-choice questions with two options (items of the type "a/an orange"). For this exercise the system uses $\alpha = 0.9$. The data set consist of over 40 000 answer sequences (each sequence is for a learner and particular knowledge component).

The second system is MatMat (`matmat.cz`) – an adaptive practice system for basic arithmetic with items of the type "6×7" with free-form answers. The system implements adaptive behavior even within a practice of a single knowledge component; items are chosen to be of an appropriate difficulty for a particular learner [21]. The data set was filtered to contain only learners with more than 10 answers. The used data set consist of 330 000 answers from more than 8 000 learners.

3.2 Evaluation Methods

With simulated data we have the advantage that we know the ground truth moment of learning. Clearly we want the moment when mastery is declared to be close to this ground truth, so the basic metric to optimize is mean absolute deviation between the ground truth mastery moment and detected mastery moment. This metric has been used in previous research [16]. However, in practical applications there is an asymmetry in errors in mastery decision. Typically, we are more concerned about under-practice (mastery declared prematurely) than about over-practice (lag in declared mastery). This aspect was also noted in previous work, e.g., [6] considers 'ratio of regret of type II to type I decision errors'. To take this asymmetry into account, we consider weighted mean absolute deviation (wMAD), where we put w times more weight to under-practice than to over-practice (we use $w = 5$ unless state otherwise).

Analysis of mastery criteria for real data is more difficult than for simulated data, because now we cannot analyze the decision with respect to correct mastery decisions (these are unknown). One possible approach is to compare the degree of agreement between different methods. This analysis cannot tell us which method is better, but it shows whether the decision which one to use is actually important – if mastery decisions by two methods are very similar, we do not need to ponder which one is better and we can use the simpler one for implementation in a real system. To evaluate the agreement of two methods, we use Spearman's correlation coefficient over the mastery moment for individual learners (alternative methods are also possible, e.g., using Jaccard index over sets of learners in the mastery state).

Another approach is to measure effort (how long it takes to reach mastery) and score after mastery (probability of answering correctly after mastery was declared). This type of evaluation was used in previous research [10, 11]. These metrics have to be interpreted carefully due to attrition biases in data, particularly when the used system already uses some kind of mastery learning [20].

3.3 Comparison of BKT and NCC

As a first experiment we compare the mastery threshold criterion based on the commonly used BKT model and the simplest mastery criterion N consecutive correct. We compare these methods over simulated data generated by a BKT model. Moreover, to avoid the issue of parameter fitting, we simply use the optimal ground truth BKT parameters for detecting mastery, i.e., this is the optimal case for application of the BKT model.

To make mastery decision we need to choose thresholds: N for the NCC method and T for BKT. We optimize these parameters for each simulated scenario. Since we optimize a single parameter, we use a simple grid search.

The experiments were performed as follows. We choose BKT parameters. We generate sequences of 50 answers for 10 000 simulated learners. We use this training set to fit the thresholds by optimizing the wMAD metric using the grid search. Then we generate a new set of 10 000 learners and use this test set for evaluation – computation of the metric wMAD for both methods and also correlation of their mastery decisions.

Table 2 shows results of this experiment for different scenarios from Table 1. The optimized thresholds are between 0.9 and 0.97 for BKT and between 2 and 8 for NCC. With respect to wMAD, BKT is typically better, but the difference is not large. The correlation between mastery decisions is typically very high. From the perspective of a student, these results mean that mastery is declared by both methods at the same or very similar time. Larger difference between BKT and NCC occurs only in the case with a high slip and a low guess.

The summary of this experiment is that even in the best case scenario, where data perfectly correspond to model assumptions, BKT does not bring significant improvement over the basic mastery decision criterion.

[1] `https://github.com/adaptive-learning/umap2017-mastery`

Table 2: Comparison of BKT and NCC mastery criteria over simulated data.

	Threshold		wMAD		
	NCC	BKT	NCC	BKT	Cor.
B1	2	0.92	2.56	2.42	0.88
B2	4	0.97	6.2	5.76	0.97
B3	2	0.95	2.81	2.48	0.92
B4	1	0.9	2.72	2.13	0.74
B5	4	0.97	3.77	3.62	0.99
B6	8	0.97	11.48	10.33	0.94

3.4 Role of Response Times

In the next experiment we use real data from the MatMat system and explore the relative importance of the choice of a model and the choice of input data, specifically whether to use learners response times. In the case of basic arithmetic it makes sense to include fluency (learners' speed) as a factor in the mastery decision. Does it matter whether we include response times? How much?

For the choice of a skill estimation model we consider the following two variants:

- The basic exponential moving average (EMA) method with $\alpha = 0.8$.
- A logistic learner model (denoted as M) described in detail in [21].

The basic difference between these two approaches is that the model takes into account difficulty of items, whereas the EMA approach completely ignores item information. The model thus can better deal with the adaptively collected data (learners are presented items of different difficulty).

For the choice of input data we consider also two variants:

- Only the basic correctness data, i.e., the response value is binary (0 or 1).
- Combination of correctness and response times (denoted as $+T$). The response value for wrong answers remains 0; the response value for correct answers is linearly decreasing for response times between 0 and 14 seconds; for longer times the response value is 0. The constant 14 is set as a double of the median response time, i.e., a correct answer with median response time has the response value 0.5.

We compare four models obtained as combinations along these two dimensions: EMA, EMA+T, M, M+T. We evaluate agreement between them to see which model aspects makes larger difference. To analyze mastery decision, it is necessary to choose mastery thresholds. However, the studied methods differ in the scales of their output values, e.g, EMA + T gives smaller values than EMA. It is therefore not easy to choose thresholds for a fair comparison. To avoid biasing the results by a choice of specific thresholds, we compare directly orderings of learners by different methods. For each learner we compute the final skill estimate and we evaluate agreement of methods by the Spearman correlation coefficient over these values.

Figure 2 shows correlations of the four studied methods for four knowledge components from the MatMat system. We see that

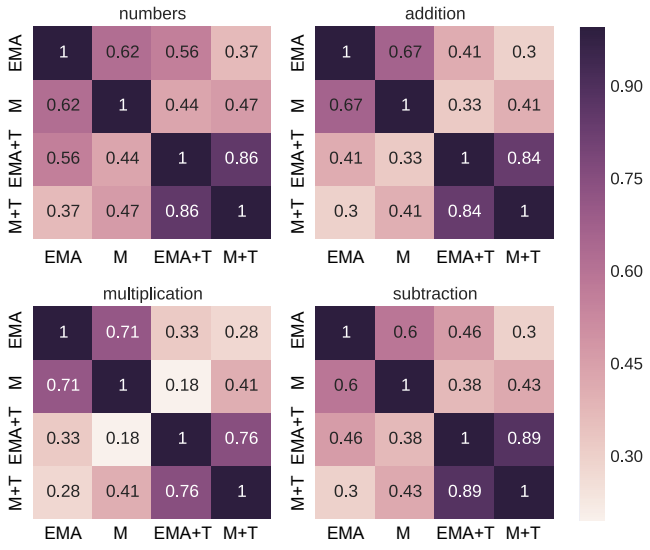

Figure 2: Spearman correlation between different learner skill estimation methods for different knowledge components (Matmat data).

the correlation between EMA and the model approach is typically higher than correlation between approaches with and without use of response time. Particularly the variants with timing information (EMA+T and M+T) are highly correlated. From this analysis we cannot say which approach is better, but we see that the impact of using response times is larger than the impact of using a learner model.

3.5 Analysis of the EMA Method

The reported results and our experience from practical application within the system for Czech grammar suggest that EMA is a reasonable method for detecting mastery. Therefore, we analyze its behavior in more detail.

EMA as a mastery criterion has two parameters: the exponential decay parameter α and the threshold T. By tuning these two parameters we can obtain different behaviors. Both parameters have values in the interval $(0, 1)$. Increase in both of these parameters leads to an increase of the length of practice, for values approaching 1 the increase is very steep.

The basic nature of this increase is apparent when we analyze the number of consecutive correct answers that guarantee passing a threshold for a given α (a sufficient, but not necessary condition): $N \geq \log_\alpha(1 - T)$. For example for a threshold $T = 0.95$ we get the following relation between α and number of attempts N:

α	0.7	0.75	0.8	0.85	0.9	0.95
N	9	11	14	19	29	59

Note that EMA can also exactly emulate the N consecutive correct criterion, e.g., when we use $\alpha = 0.5$ and $T = 1-0.5^N$, getting N consecutive correct becomes both sufficient and necessary condition for passing the threshold.

Table 3: Comparison of mastery criteria over simulated data: NCC, the EMA method with fixed $\alpha = 0.95$, and the full EMA method.

		Parameters			wMAD		
sc	N	α_{95}	α	T	NCC	EMA_{95}	EMA
B1	2	0.1	0.7	0.5	2.48	2.48	2.45
B2	4	0.5	0.75	0.75	6.45	6.23	6.07
B3	3	0.3	0.5	0.75	2.66	2.66	2.42
B4	1	0.1	0.2	0.8	2.82	3.47	2.31
B5	4	0.4	0.7	0.75	3.76	3.64	3.59
B6	7	0.7	0.75	0.92	11.04	10.45	10.41
L1	8	0.7	0.9	0.6	3.92	3.34	2.63
L2	17	0.85	0.9	0.9	9.02	8.44	7.64
L3	14	0.85	0.9	0.85	7.39	6.21	5.04
L4	15	0.85	0.8	0.98	10.28	10.7	10.3
L5	8	0.7	0.7	0.95	5.13	4.97	4.97
L6	8	0.7	0.6	0.98	6.67	7.12	6.87

We analyze EMA parameters for simulated data using the same methodology as in the experiment comparing BTK with NCC. In this case we use data generated by both BKT and logistic models, optimizing parameters and thresholds with respect to the wMAD metric. As a baseline for comparison we use the NCC method.

Table 3 shows results. We see that EMA achieves slightly better performance than NCC, for BKT scenarios the difference is typically small, for scenarios corresponding to slow learning according to the logistic model assumptions the difference can be quite pronounced. The optimal EMA parameters vary depending on the scenario – both α and T.

When we fix the threshold $T = 0.95$ and vary only the parameter α, the quality of mastery decisions (as measured by the wMAD metric) is typically better than for the NCC method, but worse than when EMA is used with full flexibility.

To explore the impact of the choice of metric, we explored different values of the weight w, which specifies the relative importance of under-practice (premature mastery) to over-practice. The key factor influencing the optimal value of α is the learning scenario, but the choice of w also has nontrivial impact. For example in the L6 scenario, the optimal value of α (for fixed threshold 0.95) varies between 0.52 and 0.73 depending on the weight w.

3.6 Effort and Score Analysis

Our results suggests that the setting of thresholds is a key aspect of mastery detection. Therefore, we need methods that could be used to choose threshold values for practical systems – the wMAD metric used in previous experiments is applicable only to simulated data for which we know the ground truth. For this purpose we explore the idea of measuring effort and score [10–12] and propose a visualization using an effort-score graph.

We measure effort and score metrics as follows: effort is the average number of attempts needed to reach mastery; score is the average number of correct answers in k attempts that follow after reaching mastery (reported experiment uses $k = 5$, the results are not sensitive to this choice). Note that there may be learners that do not reach mastery or that do not have enough attempts after mastery was reach. Treatment of these issues (e.g., whether to use value imputation as in [10]) may influence results, particularly when comparing similar methods. For the presented analysis these issues are not fundamental.

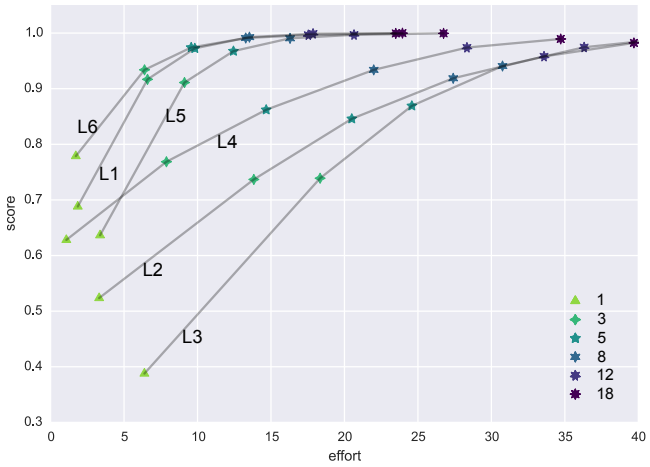

Figure 3: The effort-score graph for simulated data and the N consecutive correct method with variable N. The lines correspond to Ln scenarios from the Table 1.

To analyze the impact of the choice of threshold, we use effort-score graphs. Figure 3 shows this graph for the Ln subset of our simulated data and the basic NCC mastery criterion. The curve shows the trade-off between effort and score. By using higher mastery thresholds, the score of learners who achieved mastery improves, but at the cost of higher effort. A reasonable choice of threshold is the point at which the effort-score curve starts to level off, i.e., where additional effort does not bring improvement in performance. This is a heuristic approach, but note that it leads to similar conclusions about the choice of a threshold as experiments that utilize the ground truth (reported in Table 3).

The technique can thus be useful for setting of thresholds for real data. Figure 4 shows the effort-score graph for data from the Czech grammar and spelling system. In this case the results are provided for the EMA method with $\alpha = 0.9$ and different values of thresholds (this directly corresponds to the approach used in the actual implementation). Curves correspond to several knowledge components of varying difficulty. For easy knowledge components the score is high even for low thresholds; higher values of threshold only increase the effort, but by acceptable margin. For difficult knowledge components, the score levels off only after the threshold is over 0.95. The analysis thus suggests that the value 0.95 is a reasonable compromise.

4 DISCUSSION

We conclude with a discussion of implications of presented results. We also discuss wider context, simplifying assumptions of our experiments, and opportunities for future work.

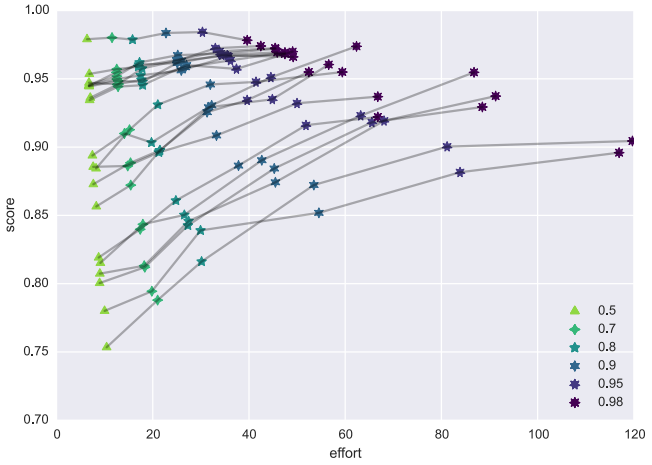

Figure 4: The effort-score graph for real data and the EMA method with $\alpha = 0.9$ and variable thresholds. The lines correspond to knowledge components of varying difficulty.

4.1 What Matters in Mastery Criteria?

Our results suggest that there is not a fundamental difference between simple mastery criteria (consecutive correct, exponential moving average) and more complex methods based on the use of learner modeling techniques. The important decisions are what data to use for mastery decision and the choice of thresholds.

The choice of mastery thresholds involves the trade-off between the risk of premature mastery and over-practice. Even small changes in thresholds can have large impact on learners practice, so setting of this parameter should get significant attention in development of systems utilizing mastery learning. The choice of thresholds depends on a particular application, because applications differ in the relative costs of premature mastery and over-practice. General research thus cannot provide universal conclusions about the choice of threshold, but it can provide more detailed guidance for techniques that can help with the choice of thresholds. As a practical tool for this choice we propose effort-score graphs, which can be easily constructed from historical data. It would be useful to further elaborate other techniques described in previous work [7, 10].

4.2 Exponential Moving Average

Our results and previous work [18] suggest that the exponential moving average method provides a reasonable approach to detecting mastery. The method has two parameters: the exponential decay parameter α and the threshold T. Together these two parameters provide enough freedom so that the method can provide good mastery decision in different situations (e.g., different speeds of learning, levels of initial knowledge, presence of guessing).

The technique is very simple to implement and use for online decisions. The technique is also directly applicable for visualization of progress to learners (typically using some kind of progress bar). It has an intuitive behavior – an increase in estimated skill after a correct answer, a decrease after a wrong answer. Such behavior may seem trivial and straightforward, but it does not necessarily hold for alternative methods. For example simple moving average often stays the same after a correct answer and some learners models may even increase skill estimate after a wrong answer (because such behavior fits training data).

4.3 Role of Learner Models

Our results suggest that learner modeling techniques are not fundamental for detecting mastery. However, that does not mean that they are not useful. Learner models are very useful for obtaining insights using offline analysis of data. One of key assumptions of our analysis is that we have well-specified knowledge components. Learner modeling techniques are useful for discovery and refinement of knowledge components and their relations. However, once this offline analysis is done, it may be better to use simpler, more robust methods for online decisions. This argument is closely related to Baker's proposal for "stupid tutoring systems, intelligent humans" [1] – using analytics tools to inform humans and then implement relatively simple, but well-selected and well-tuned methods into computer systems.

4.4 Limitations and Future Work

Our analysis uses several simplifying assumptions. Lifting these assumptions provides interesting directions for future work.

We assume well-specified, isolated knowledge components of suitable granularity. In the case of strong relations among knowledge components, the difference between learner modeling techniques and simple techniques may be larger, since learner modeling techniques may utilize information from several knowledge components for mastery decision. An interesting issue is the interaction between level of granularity of knowledge components and the choice of mastery thresholds.

We do not consider wheel-spinning learners [3] who are unable to master a knowledge component and instead of continued practice would benefit from redirection to one of prerequisite knowledge components. This issue has been addressed by policies developed in previous work [12, 23]. These policies have been evaluated for learner modeling techniques; it may be interesting to explore their combination with exponential moving average.

We do not consider forgetting. This is particularly important issue in the case of factual knowledge (e.g., foreign language vocabulary), but even in the case of mathematics previous research have shown that the mastery speed is related to future performance [25]. Instead of treating mastery as a permanent state, it would be better to treat it as a temporary state that needs reassessment. An interesting direction is an integration of mastery criteria with research on spacing effects.

Finally, in the presented analysis we ignore potential biases present in real data, particularly attrition bias [20]. This can be potentially an important issue for the analysis of effort-score graphs. It would be useful to develop techniques for detecting and overcoming such biases in the effort-score analysis.

REFERENCES

[1] Ryan S Baker. 2016. Stupid Tutoring Systems, Intelligent Humans. *International Journal of Artificial Intelligence in Education* 26, 2 (2016), 600–614.

[2] Ryan SJD Baker, Adam B Goldstein, and Neil T Heffernan. 2011. Detecting learning moment-by-moment. *International Journal of Artificial Intelligence in Education* 21, 1-2 (2011), 5–25.

[3] Joseph E Beck and Yue Gong. 2013. Wheel-spinning: Students who fail to master a skill. In *Proc. of Artificial Intelligence in Education*. Springer, 431–440.

[4] Albert T Corbett and John R Anderson. 1994. Knowledge tracing: Modeling the acquisition of procedural knowledge. *User modeling and user-adapted interaction* 4, 4 (1994), 253–278.

[5] Michel C Desmarais and Ryan SJ Baker. 2012. A review of recent advances in learner and skill modeling in intelligent learning environments. *User Modeling and User-Adapted Interaction* 22, 1-2 (2012), 9–38.

[6] John A Emrick. 1971. An Evaluation Model for Mastery Testing. *Journal of Educational Measurement* 8, 4 (1971), 321–326.

[7] Stephen E Fancsali, Tristan Nixon, and Steven Ritter. 2013. Optimal and Worst-Case Performance of Mastery Learning Assessment with Bayesian Knowledge Tracing. In *Educational Data Mining*.

[8] Stephen E Fancsali, Tristan Nixon, Annalies Vuong, and Steven Ritter. 2013. Simulated Students, Mastery Learning, and Improved Learning Curves for Real-World Cognitive Tutors.. In *AIED Workshops*.

[9] April Galyardt and Ilya Goldin. 2015. Move your lamp post: Recent data reflects learner knowledge better than older data. *Journal of Educational Data Mining* 7, 2 (2015), 83–108.

[10] José P González-Brenes and Yun Huang. 2015. Your model is predictive - but is it useful? theoretical and empirical considerations of a new paradigm for adaptive tutoring evaluation. In *Proc. of Educational Data Mining*.

[11] David Hu. 2011. How Khan academy is using machine learning to assess student mastery. (2011). http://david-hu.com/2011/11/02/how-khan-academy-is-using-machine-learning-to-assess-student-mastery.html.

[12] Tanja Käser, Severin Klingler, and Markus Gross. 2016. When to Stop?: Towards Universal Instructional Policies. In *Proc. of Learning Analytics & Knowledge*. ACM, 289–298.

[13] Kim Kelly, Yan Wang, Tamisha Thompson, and Neil Heffernan. 2015. Defining Mastery: Knowledge Tracing Versus N-Consecutive Correct Responses. In *Proc. of Educational Data Mining*.

[14] Jung In Lee and Emma Brunskill. 2012. The Impact on Individualizing Student Models on Necessary Practice Opportunities. In *Proc. of Educational Data Mining*. 118–125.

[15] George B Macready and C Mitchell Dayton. 1977. The use of probabilistic models in the assessment of mastery. *Journal of Educational and Behavioral Statistics* 2, 2 (1977), 99–120.

[16] Zachary A Pardos and Michael V Yudelson. 2013. Towards Moment of Learning Accuracy. In *AIED 2013 Workshops Proceedings Volume 4*. 3.

[17] Philip I. Pavlik, Hao Cen, and Kenneth R. Koedinger. 2009. Performance Factors Analysis-A New Alternative to Knowledge Tracing.. In *Proc. of Artificial Intelligence in Education (AIED) (Frontiers in Artificial Intelligence and Applications)*, Vol. 200. IOS Press, 531–538.

[18] Radek Pelánek. 2014. Application of Time Decay Functions and Elo System in Student Modeling. In *Proc. of Educational Data Mining*. 21–27.

[19] Radek Pelánek. 2015. Metrics for Evaluation of Student Models. *Journal of Educational Data Mining* 7, 2 (2015).

[20] Radek Pelánek, Jiří Řihák, and Jan Papoušek. 2016. Impact of Data Collection on Interpretation and Evaluation of Student Model. In *Proc. of Learning Analytics & Knowledge*. ACM, 40–47.

[21] Jirí Rihák. 2015. Use of Time Information in Models behind Adaptive System for Building Fluency in Mathematics.. In *Proc. of Educational Data Mining*.

[22] Steve Ritter, Michael Yudelson, Stephen E Fancsali, and Susan R Berman. 2016. How Mastery Learning Works at Scale. In *Proc. of ACM Conference on Learning@Scale*. ACM, 71–79.

[23] Joseph Rollinson and Emma Brunskill. 2015. From Predictive Models to Instructional Policies. In *Proc. of Educational Data Mining*.

[24] George Semb. 1974. The Effects of Mastery Criteria and Assignment Length on College-Student Test Performance. *Journal of applied behavior analysis* 7, 1 (1974), 61–69.

[25] Xiaolu Xiong, Shoujing Li, and Joseph E Beck. 2013. Will You Get It Right Next Week: Predict Delayed Performance in Enhanced ITS Mastery Cycle.. In *FLAIRS Conference*.

[26] Michael V Yudelson and Kenneth R Koedinger. 2013. Estimating the benefits of student model improvements on a substantive scale. In *EDM 2013 Workshops Proceedings*.