Analysis and Design of Mastery Learning Criteria

Radek Pelánek and Jiří Řihák

Masaryk University Brno, Czech Republic

ARTICLE HISTORY
Compiled May 14, 2018

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 already 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 the 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 discuss techniques for the choice of a mastery threshold. We also propose an extension of the exponential moving average method that takes into account practical aspects like time intensity of items and we report on a practical application of this mastery criterion in a widely used educational system.

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

1. Introduction

Mastery learning is a concept originally proposed by Benjamin Bloom as an educational philosophy based on the believe that nearly all students can master a studied subject when given enough time and support (Bloom, 1968). This philosophy is a key motivation for the development of individualized educational systems.

Nowadays, the notion "mastery learning" is used within widely different contexts. Particularly, there is an important difference in applications concerning testing and practice, which have significantly different costs associated with false positives and false negatives in the mastery decision. In the context of high-stakes testing, for example in medical education, it is important to avoid false positives, since premature declaration of mastery can cause harm. In the context of practice, particularly in online educational tools like Khan Academy, which are used by learners on voluntary basis, it is important to avoid false negatives – if learners are forced to do unnecessary practice, they may lose interest and stop using the system. In this work we focus on the context of online educational systems.

In this context, 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 already achieved mastery.

CONTACT R. Pelánek. Email: pelanek@fi.muni.cz

A typical application of a mastery criterion within a modern educational system is the following. A learner solves a problem or answers a question in the system. Data about the learner's performance are summarized by a model of learner knowledge or by some summary statistic. Educational systems rely on a mastery criterion to take this summary and produce 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) that specifies its strictness.

The use of mastery criteria is interwoven with the way the educational content is structured into knowledge components (also called concepts or skills; we use the terminology of the Knowledge-Learning-Instruction framework (Koedinger, Corbett, & Perfetti, 2012)). For mastery criteria to provide meaningful verdicts, the knowledge components need to be well-specified.

Educational systems that use mastery criteria typically use the following setting. The learner picks a knowledge component for practice; this choice is often guided by the system. The system then presents the learner items belonging to the knowledge component, either in fixed order or randomly chosen. The system monitors performance of the learner and after each attempt it makes the mastery decision (stop/continue). An alternative approach to adaptation is to leave the length of practice fixed (or determined by the learner) and focus on the adaptive choice of items (Pelánek, Papoušek, Řihák, Stanislav, & Nižnan, 2017).

A more general approach to adaptive learning is to address a general "What next?" question. After each answer, the system considers all possible items (not just from the current knowledge component) and decides whether it is suitable to continue with the practice of the current knowledge component (and if so, what item to use) or whether it is more beneficial to switch to practice of another knowledge component. This decision should take into account factors like learner priorities, forgetting (Mozer & Lindsey, 2016), and pedagogical aspects like "blocked vs. interleaved practice" (Rau, Aleven, & Rummel, 2010).

Such fully general approach is currently beyond realistic reach, since the scope of possible intervention is extremely large (Koedinger, Booth, & Klahr, 2013). Even the much simpler case of mastery criteria – deciding when to stop practice – is far from properly explored and understood. We believe that it is necessary to start by developing solid approaches based on mastery criteria. Once we have detailed understanding of this simpler case, we may be able to tackle the more general "What next?" question.

In this work we focus on thorough analysis of the basic use case of mastery criteria: based on a sequence of answers to a single knowledge component we aim to decide whether the learner has already mastered the knowledge component. In our discussion we use examples and data on knowledge components with "simple items", i.e., items where answering an item takes a short time (from 1 second to 1 minute) and consists of a single step. This covers a wide variety of knowledge components typically used in educational systems, e.g., spelling, basic grammar rules (both first and second language learning), or basic computation problems in math (multiplication, fractions, decimals, simple word problems). Moreover, most discussed techniques can be extended to multistep items (e.g., more complex math problems).

The main contributions of this work are the following:

- systematic discussion of mastery criteria, providing overview of related work and discussion of the problem specification;
- overview of basic techniques for mastery criteria and discussion of their proper-

ties;
- experimental analysis of basic techniques using both real and simulated data;
- results that show that from the perspective of mastery criteria it is more important to focus on the choice of input data and mastery thresholds, rather than on learner modeling techniques;
- discussion of advantages of the exponential moving average method, extension of this method for practical application, and report on its usage in a practical system.

This work is an extended version of an UMAP paper (Pelánek & Řihák, 2017); the paper is thoroughly updated, specifically it provides more detailed discussion of related work (Section 2), general discussion of the addressed problem (Section 3), and a description of a design of a specific mastery criterion employed in a widely used application (Section 6).

2. Related Work

Mastery criteria have been studied widely in the context of testing. The first studies were conducted more than 40 years ago (Emrick, 1971; Macready & Dayton, 1977; Semb, 1974), but at that time typically only for static tests and small scale applications. Later, mastery criteria were studied in the context of computerized adaptive testing (Lewis & Sheehan, 1990). Today, mastery criteria in testing are important part of education, particularly in areas like medical education (Yudkowsky, Park, Lineberry, Knox, & Ritter, 2015).

The focus of this work is, however, on application of mastery criteria within educational systems used primarily for practice (not testing). Mastery learning is today used in large scale educational systems (Hu, 2011; Ritter, Yudelson, Fancsali, & Berman, 2016).

In the context of educational systems for practice, there exists an extensive research on learner modeling, for recent overviews see Desmarais and Baker (2012); Pelánek (2017). This research typically uses personalization through mastery learning as a motivation. However, evaluation of models is typically not done by evaluating the impact of mastery criteria, but instead using evaluation of predictive accuracy on historical data using metrics like RMSE or AUC (Pelánek, 2015). Evaluation of mastery criteria is more difficult, because mastery is a latent construct, which cannot be directly measured.

Some recent research, however, studied impact of learner models on mastery decisions. The ExpOps method (Lee & Brunskill, 2012; Rollinson & Brunskill, 2015) 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 (González-Brenes & Huang, 2015), which use historical data to estimate the effort needed to reach mastery and the performance after mastery.

Most of the research on mastery criteria was done in relation with the Bayesian knowledge tracing (BKT) model (Corbett & Anderson, 1994); this model is also often used in practice with a standard mastery threshold 0.95. Fancsali, Nixon, and Ritter (2013); Fancsali, Nixon, Vuong, and Ritter (2013) analyzed the role of this threshold by using simulated data (generated by the BKT model) to show the relation between the threshold and the proportion of learners with premature judgment of mastery and with over-overpractice. Simulated data were also used by Pardos and Yudelson (2013) to study the 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, Goldstein, and Heffernan (2011) also studied the moment of learning using 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 (2013) 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 proposed general instructional policies that can be combined with wide range of learner models: predictive similarity (Rollinson & Brunskill, 2015) and predictive stability (Käser, Klingler, & Gross, 2016) policies. For evaluation authors used the above described techniques: ExpOps (Rollinson & Brunskill, 2015) and effort and score metrics (Käser et al., 2016). 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 (Beck & Gong, 2013). These works, however, pay little attention to the choice of thresholds.

3. Mastery Criteria: Problem Specification and Requirements

Our aim is to investigate different mastery criteria, to compare their behavior, and to explore methods for and selecting a suitable criterion for a specific practical application. Before starting this exploration, it is useful to explicitly formulate the problem that we are trying to solve. We analyze mastery criteria in the context of automated educational practice systems, i.e., a learner repeatedly answers question about a particular knowledge component (possibly getting explanations or hints during solutions). A mastery criterion provides a stopping criterion that decides when the learner should stop practice and move to another topic.

For illustration of the basic problem consider an illustrative example in Table 1. The example shows a sequence of learner's answers on questions about addition of fractions. Given this sequence of answers, should the system declare mastery and let the learner practice more advanced topics (e.g., equations with fractions)? Or should the learner rather be given more examples on the current topic?

3.1. Goals

As the example in Table 1 illustrates, the mastery decision problem does not have a clear-cut correct solution. The problem involves an inherent trade-off between the certainty of mastery decision and the efficiency of practice. We try to achieve two antagonistic goals:

- *Minimize false positives*: We want to minimize the cases when the system declares mastery, but the learner does not have proper understanding of the topic. Specifically, the criterion should be resistant to "gaming" (e.g., repeated random guessing).
- *Fast recognition of mastery*: Those who really know the topic should be able to

Table 1. Illustration of the mastery criteria use case. The learner answered a sequence of questions for the knowledge component "addition of fraction with the same denominator". Given this history, should the learner continue the practice or move to another topic?

item	answer	time	correct?
$\frac{1}{3}+\frac{4}{3}$	$\frac{5}{3}$	8s	correct
$\frac{4}{6}+\frac{1}{6}$	$\frac{6}{5}$	6s	incorrect
$\frac{1}{5}+\frac{2}{5}$	$\frac{3}{5}$	9s	correct
$\frac{3}{7}+\frac{5}{7}$	$\frac{8}{7}$	5s	correct
$\frac{4}{5}+\frac{3}{5}$	$\frac{7}{5}$	7s	correct

reach mastery quickly.

The relative importance of these two goals depends on the particular application. Minimizing false positives is more important in formal educational setting than in systems that are used by learners voluntarily in their free time (e.g., adult learners learning a new language). These systems may prefer fast recognition of mastery at the cost of more false positives.

3.2. Input Data

What data do we use for making the mastery decision? This is a key question – as we will show in our analysis, the choice of input data can have significant impact on mastery decisions.

The basic type of data that is necessary for mastery criterion is the *data on learner's performance*, specifically the correctness of answers. The correctness of answers is used basically in all applications of mastery learning. It is typically treated as binary variable (correct/incorrect). In some cases it is possible to further analyze wrong answers or usage of hints and use "partical credit" evaluation of answers – some answers are less wrong than others (Pelánek & Řihák, 2016; Wang & Heffernan, 2013). Another source of performance data are response times (Pelánek & Jarušek, 2015).

Another type of data that may be used for mastery criterion is *contextual data*, i.e., taking into account the context in which the learner's performance is happening. This may include the time between individual answers, so that we can take forgetting or fatigue into account. We may also take into account relations between knowledge components and factor into mastery criterion for a given component performance on related components (particularly for prerequisite components).

In this work we focus on the very basic setting – considering only basic performance data. Before designing and evaluating more complex criteria, it is necessary to properly understand this case. As our analysis shows, even the basic case is not easy.

3.3. Visualizing Progress Towards Mastery

Systems that use mastery criteria typically employ some kind of visualization to show learners their progress towards mastery ("progress bar", "skillometer"). The goal of this visualization is to support the "flow" experience (Csikszentmihalyi, 1991) by providing a clear goal and feedback on progress towards this goal. The visualization of progress also increases transparency of the mastery decision by providing some insight into how the mastery decision is made; this transparency can help to build trust of users.

The visualization should satisfy several requirements:

- The visualization should enable learners to estimate how long they need to practice to achieve mastery.
- Although learners do not need to understand the exact rule behind the progress bar, the basic behavior should be intuitive: it should increase after a correct answer and decrease or stagnate after a mistake.
- The progress bar should have "motivating behavior", e.g., large decrease after a typo is demotivating.

3.4. Practical Considerations

While designing mastery criteria for practical applications, it is also necessary to take into account the pragmatic aspects of the development of educational systems. A realistic educational application can contain hundreds of knowledge components, which undergo continuous updates. To apply mastery criteria in such realistic settings, the mastery criteria need to have small "deployment and maintenance costs":

- The mastery criterion should be as "universal" as possible – it should work with different types of exercises and different types of knowledge components with minimal number of tunable parameters.
- The mastery criterion should be robust – small change of parameters should not drastically change the behavior of the educational system.
- The mastery criterion should have efficient implementation so that it can scale to a large number of users and their answers.

4. Basic Techniques for Mastery Detection

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".

4.1. Notation

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 a 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 provides a specific illustration based on simulated data. Since these are simulated data, we know the ground truth (the black line; mastery achieved at the 8th

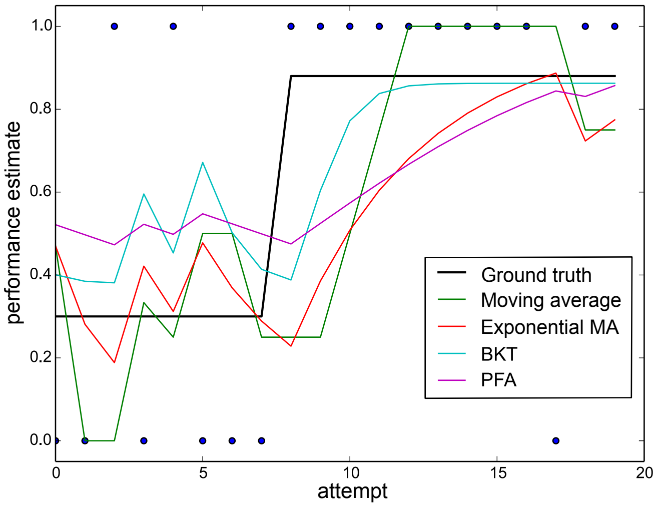

Figure 1. Illustrative comparison of performance estimation techniques for a sequence of answers of a a single simulated learner (B2 from Table 2). The black line is the ground truth probability of correct answer and the dots are the simulated answers.

attempt). The colored lines show estimates by several methods, which are described below. 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.

4.2. Methods without Assumptions about Learning

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

4.2.1. Consecutive Correct

The simplest mastery criterion is "N consecutive correct answers" (NCC) (also called N-in-row or "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 (Kelly, Wang, Thompson, & Heffernan, 2015).

4.2.2. 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 by Pelánek (2014) and also closely related to method proposed by Galyardt and Goldin (2015).

4.2.3. 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 (Pelánek, 2014) 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. We can compute the exponential moving average θ_k after k steps using a recursive rule utilizing a decay parameter $\alpha \in (0, 1)$:

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

This approach is equivalent to the weighted moving average with weights given by an exponential function $w_i = (1 - \alpha)\alpha^{(i-1)}$. The mastery criterion remains $\theta_k \geq T$.

4.3. 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 (Rollinson & Brunskill, 2015) and predictive stability (Käser et al., 2016). 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.

4.3.1. Bayesian Knowledge Tracing

Bayesian knowledge tracing (BKT) (Corbett & Anderson, 1994) assumes a sudden change in knowledge. It is a two state hidden Markov model where the state corresponds to the mastery status (either mastered or unmastered). The model has 4 parameters: P_i is the probability that the skill is initially mastered, P_l is the probability of learning a skill in one step, P_s is the probability of an incorrect answer when the skill is mastered (slip), and P_g is the probability of a correct answer when the skill is unmastered (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 learner 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 mastered before the k-th attempt and θ'_k is the posterior probability that the skill is mastered after we have taken the k-th answer into account:

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

Estimation of model parameters (the tuple P_i, P_l, P_s, P_g) can be done using techniques like discretized brute-force search, the expectation-maximization algorithm, or gradient descent (Yudelson, Koedinger, & Gordon, 2013).

4.3.2. Logistic Models

Another commonly used class of learner models are models based on the logistic function $\sigma(x) = \frac{1}{1+e^{-x}}$. This class of models includes for example the Rasch model (Rasch, 1960), Performance factor analysis (Pavlik, Cen, & Koedinger, 2009), or the Elo rating system (Pelánek, 2014). These models utilize assumption of a continuous latent skill $\theta \in (-\infty, \infty)$ and for the relation between the skill and the probability of a correct answer use the logistic function – in the simplest form the probability of a correct answer is given by $\sigma(\theta)$; this can be easily extended to incorporate for example difficulty of items or guessing in multiple-choice questions.

A simple technique of this type is Performance factor analysis (PFA) (Pavlik et al., 2009). The skill estimate is given by a linear combination of the initial skill and past successes and failures of a learner: $\theta_k = \beta + \gamma \cdot s_k + \delta \cdot f_k$, where β is the initial skill, s_k and f_k are counts of previous successes and failures of the learner during the first k attempts, γ and δ are parameters that determine the change of the skill associated with a correct and incorrect answer. The predicted probability of a correct answer for the $(k+1)$-th attempt is given by $\sigma(\theta_k)$. Parameters β, γ, δ can be easily estimated using the standard logistic regression.

5. 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.

5.1. Data and Evaluation Methods

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 have access to the ground truth, which is established by the data generating process under our control. We can thus perform more thorough evaluation, but the results are restricted to simplified conditions and depend on the choice of simulation parameters.

5.1.1. 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 2 provides description of simulation scenarios used in experiments. In all cases we generate 50 answers for each learner. The source code of all experiments with simulated data is available[1]. The implementation contains the parameters described in Table 2 in a configuration file. The below reported experiments can be easily executed for other parameter values by just changing the parameter values in the configuration file.

The BKT model is used in its basic form. 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.

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 (Pardos & Yudelson, 2013). 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., Emrick (1971) 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 stated otherwise).

5.1.2. Real Data

We use real data from two educational systems. The first is a system for practice of Czech grammar and spelling (`umimecesky.cz`). For the analysis we use data from the academic year 2016/2017. At this time the system implemented mastery learning based on the basic exponential moving average method. The system visualized 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) was given by the threshold 0.95. The value of α depended on the type of exercise. For the analysis we use data from basic grammar exercises with multiple-choice questions with two options (items of the type “a/an orange”). For this exercise the system used $\alpha = 0.9$. The data set consist of over 40 000 answer sequences (each sequence is for a learner and particular knowledge component, the median length of a sequence is 60).

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;

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

Table 2. 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	$\theta_0 \sim N(-0.7, 1.5)$		$\Delta \sim N(0.6, 0.3)$	

items are chosen to be of an appropriate difficulty for a particular learner (Rihák, 2015). 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.

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 as they 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 of 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 (González-Brenes & Huang, 2015; Hu, 2011). These metrics have to be interpreted carefully due to attrition biases in data (Pelánek, Řihák, & Papoušek, 2016). Specifically, when the system used for data collection employs some kind of mastery learning, the estimated score and effort measures are influenced by missing data due to attrition after mastery.

5.2. Importance of Learner Models

In recent years researchers proposed many techniques for modeling learners knowledge, for overviews see Desmarais and Baker (2012); Pelánek (2017). These models serve several different purposes, one of the common goals for developing these models is to use them as a basis for mastery criteria. We explore the following question: Do learner modeling techniques bring a fundamental advantage to mastery criteria? To study this question we compare mastery criteria based on learner modeling with a simple criteria based on simple statistics of learner performance.

Table 3. 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

5.2.1. 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 complications with the interpretation of results due to issues with 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 3 shows results of this experiment for different scenarios from Table 2. The optimized thresholds are between 1 and 8 for NCC and between 0.9 and 0.97 for BKT. 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 learner, 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.

5.2.2. 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 by Rihák (2015).

The basic difference between these two approaches is that the logistic model M 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 the 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 their input data and consequently also in the scales of their output values, e.g, EMA+T gives smaller values than EMA, since for a moderately fast correct answer the value of the answer is 1 for EMA and around 0.5 for EMA+T. 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 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 the correlation between EMA and the model approach is typically higher than the correlation between approaches with and without use of response time. Particularly the variants with timing information (EMA+T and M+T) are highly correlated. In other words, changing a modeling approach (from EMA to M) leads to a small change of learners' skill estimates, whereas changing input data (from not using to using response times) leads to nontrivial change of skill estimates. From this analysis we cannot say which approach is the right one for mastery decisions, but we see that the decision whether to use response times is more consequential than the decision whether to use a learner model.

5.3. 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 $T = 0.95$

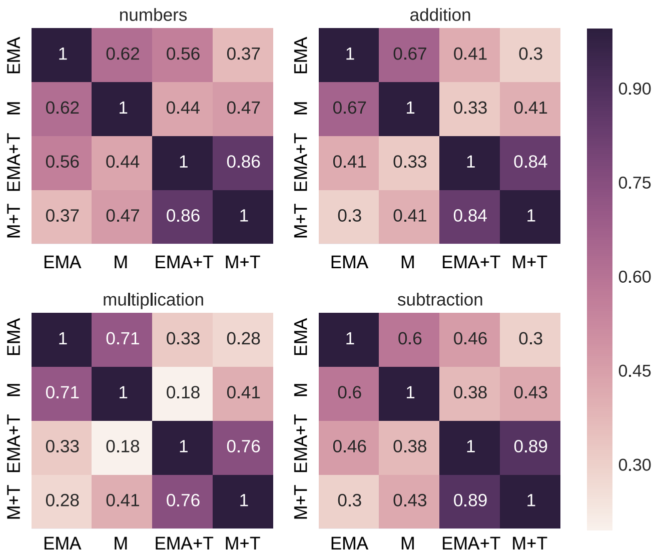

Figure 2. Spearman correlation between different learner skill estimation methods for different knowledge components (Matmat data).

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.

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 4 shows results. We see that EMA achieves slightly better performance than NCC. For BKT scenarios the difference is typically small, whereas 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 the fixed threshold 0.95) varies between 0.52 and 0.73 depending on the weight w. With higher values of α the EMA method takes into account longer history of learner's attempts and thus is more cautious in its mastery decision. The required degree of cautiousness is determined by the value w. With lower

Table 4. Comparison of mastery criteria over simulated data: NCC, the EMA method with fixed $T = 0.95$, 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

values of $w = 1$, the optimal value of α decreases – as we decrease the relative penalty for under-practice, it becomes sufficient to make the mastery decision based on recent attempts.

5.4. Choice of Thresholds: 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 (González-Brenes & Huang, 2015; Hu, 2011; Käser et al., 2016) and propose a visualization using an effort-score graph. Recently, (Käser, Klingler, Schwing, & Gross, 2017) have analyzed effort and score metrics using graphs, but using a slightly different presentation mode.

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. For the reported experiment we use a fixed value $k = 5$; 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 reached. Treatment of these issues may influence results, particularly when comparing similar methods. However, for the basic analysis presented here these issues are not fundamental.

To analyze the impact of the choice of a threshold, we propose *effort-score graphs*. An effort-score graph shows effort and score metrics for a given mastery criterion with different choices of a threshold. Figure 3 (left) shows this graph for the *Ln* subset of our simulated data and the basic NCC mastery criterion. Each curve corresponds to one simulated scenario and shows how the choice of a threshold influences the trade-off between effort and score. By using higher mastery thresholds, the score of learners who achieve mastery improves, but at the cost of higher effort. A reasonable

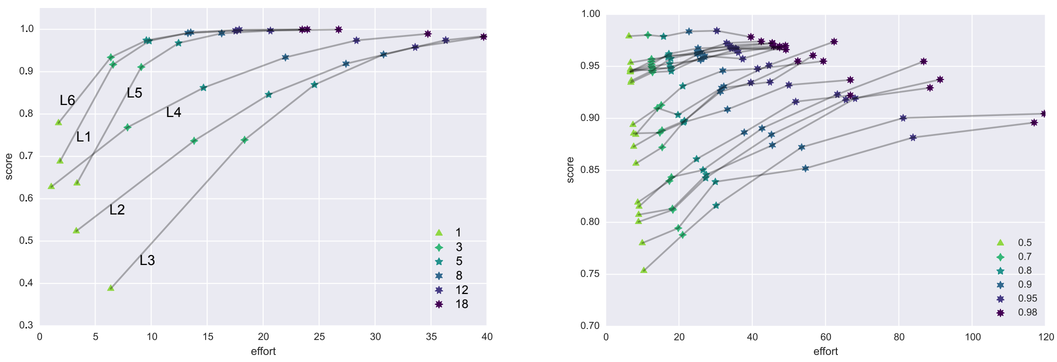

Figure 3. Left: The effort-score graph for simulated data and the N consecutive correct method with variable N. The curves correspond to Ln scenarios from the Table 2. Right: The effort-score graph for real data and the EMA method with $\alpha = 0.9$ and variable thresholds. The curves correspond to knowledge components of varying difficulty.

choice of a 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 4).

The technique can thus be useful for setting of thresholds for real data. Figure 3 (right) 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.

6. Design of Mastery Criteria for Practical Application

Our analysis of basic mastery criteria suggests that exponential moving average is a suitable basic criterion. In this section we describe an extension of this approach, which takes into account other practical aspects mentioned in the Section 3.

The described approach is used in the "Umíme" educational system (`umimeto.org`) – a Czech educational system for practicing mathematics, Czech grammar, English vocabulary, and other domains. The system contains many different forms of exercises over hundreds of diverse knowledge components. The main goal of the extension is to take into account the time intensity of problems and the chance of guessing a correct answer, so that the mastery criterion works well for widely different exercises as math word problems (typical response time larger than 20 seconds, small chance of guessing the answer) and the choice of correct spelling from two variants (typical response time around 2 seconds, high chance of guessing).

To make the system sustainable in realistic setting, we need to minimize the number of ad-hoc parameters, e.g., we want to avoid setting specific parameter values and thresholds for each exercise type or knowledge component. An alternative would be to have separate parameters for each exercise and fit these parameters from data. However, since we do not have labeled data with correct mastery decisions, parameters cannot be fitted in a simple way. Moreover, to use the data fitting approach, we need sufficiently large data. Such approach would thus require addressing the “cold start problem”, which would further complicate the development of the system.

6.1. Basic Approach

The basic approach is the same as in the exponential moving average. Based on the learner’s performance we compute the skill estimate $\theta \in [0, 1]$. The skill estimate is computed as follows:

- initialization: $\theta := 0$,
- update after each answer: $\theta := (1 - w)\theta + wC$,
  where C is the correctness of the answer ($C \in \{0, 1\}$) and the parameter $w \in [0, 1]$ gives the “weight” of the last attempt.

Mastery criterion is the classic threshold based comparison: $\theta > T$. As a value of the threshold parameter we use $T = 0.95$. This value is based on the experiments reported in previous section. Other reported parameters of the approach are tuned with respect to this choice. There are other choices of T and other parameters that would lead to very similar behavior of the system.

6.2. Dynamic Weight

If w is a fixed constant the approach corresponds exactly to exponential moving average. This works well if we set a suitable value of w for each exercise type and knowledge component. However, that would mean a large number of parameters, which is something we want to avoid.

Instead, we compute w dynamically for each answer depending on just few global parameters and easily available characteristics of the currently solved item. The weight is given as a product of “time intensity” I of the item and “persuasiveness” P of the answer: $w = I \cdot P$.

Time intensity expresses how long it typically takes to solve the item. The update of the skill should be larger when solving item requires more time. Persuasiveness quantifies how much information about the learner’s knowledge the answer carries, e.g., the update of the skill should be large, when there is small chance of guessing; if the answer can be guessed (as in multiple choice questions), the skill update should be small.

In the following we propose specific parameter values. These values were set with the following goals for mastery learning: Mastery should be achievable in around 3 minutes for those who understand the topic and are able to answer all questions correctly; at the same time the chance of reaching mastery by luck (guessing) should be minimized.

6.2.1. Time Intensity

Time intensity of an item is based on the median response time for this item. For simplicity and robustness, this parameter is discretized and computed offline (time intensity for each item is stored in database and updated periodically based on offline logs). Specific values are proposed in Table 5.

Alternatively we can express the intensity I as a function of response time t. We can derive this from the requested “target time” *Total* – how long it should take to reach mastery when all answers are correct and $P = 1$. In this case we have just a single

Table 5. Time intensity of an item as determined by the median response time of the item.

response time (s)	value of I
0–5	0.12
5–10	0.17
10–20	0.25
20–40	0.35
> 40	0.5

Table 6. Possible classification of responses and their persuasiveness.

class	description	value of P
correct, fast	response time < median time for item	1
correct, slow	response time > median time for item	0.75
correct, MCQ	cases with significant guess factor	0.75
near miss	nearly correct answer	0.2
don't know/care	"empty" answer or extremely low time	1
reasonable attempt	other wrong answers	0.75

parameter concerning time intensity and the time intensity is given by the function $I(t) = 1 - e^{\ln(1-T)t/\mathit{Total}}$.

6.2.2. Persuasiveness

Persuasiveness of the answer is based on the classification of answers into several classes. For a correct answer, we want to take into account how probable is that the answer was obtain by guessing. Particularly, we want to distinguish answers to multiple choice questions and open questions. For open questions it also makes sense to take into account response time – fast correct answers are more persuasive than slow correct answers.

For wrong answers, the persuasiveness specifies "how persuasive is the answer in showing that the learner does not understand the studied topic". The persuasiveness in this case determines the degree of decrease in estimated skill. It is thus reasonable to try to distinguish at least the following cases:

- "near miss" – the provided answer is wrong, but it is very close to the correct answer (e.g., small edit distance in case of spelling, ±1 mistake in math),
- "reasonable attempt" – the provided answer is wrong, but it is still a reasonable attempt; the learner showed at least some effort to answer the question,
- "don't know/care" – the learner did not provide any answer or the provided answer clearly shows that the learner did not think about the question (e.g., string answer to numerical question or response time under 0.5 second).

Table 6 provides a proposal for specific parameter values. To reduce the number of ad-hoc values, the proposal uses the same value 0.75 for several cases, nevertheless there is no fundamental reason why these values should be the same.

6.3. Progress Bar

As described in Section 3, typically we want to display the progress towards mastery using a progress bar. With the proposed approach we can directly use the current value of the skill θ, which lies in the $[0, 1]$ interval and thus can be straightforwardly visualized as a progress bar.

This visualization satisfies most requirements mentioned in Section 3. One disadvantage of the direct use of θ is that due to the exponential weighting used in the computation of θ it leads to "big jumps at the beginning, small increases when close to the end". It is more natural for the progress bar to have more "linear behavior", which makes it easier to set expectations ("How many correct answers do I need to reach mastery?").

Such linearization of the progress bar can be done by displaying "the ratio of the number of correct answers needed from the current state to the number of correct answers needed in the initial state". For a current θ we compute the number N_θ, such that N_θ correct consecutive answers would lead to mastery (assuming a typical weight w). A sequence of consecutive correct answers leads under the exponential moving average method to increase in skill which can be expressed as a sum of a geometric series. Based on this sum, we can calculate $N_\theta = \log_{1-w} \frac{1-T}{1-\theta}$. The ratio N_θ / N_0 finally leads to the following expression:

$$\theta' := 1 - \log_{1-T} \frac{1-T}{1-\theta}$$

As long as $\theta < T$ the value of θ' belongs to the $[0, 1)$ interval and can be easily visualized as a progress bar. If $\theta \geq T$, mastery should be declared (θ' should be set to 1, i.e., the goal is achieved).

6.4. Analysis

Fig. 4 provides analysis of the decisions of the proposed mastery criterion from the application in the Umíme systems. The graph shows data for 5 diverse exercises: Czech commas (determining correct location of commas in Czech sentences), Czech spelling and grammar (fill-in-the-blank exercises with 2 options), English vocabulary (multiple-choice questions with 4 options), Math calculation (writing the result of a math example) and Math matching (given 6 or 8 pairs of cards, match together pairs with the same value). Each type of exercise covers a wide range of knowledge components of different difficulty.

The used exercises differ quite widely in their typical time intensity, which is reflected by the number of attempts needed to reach mastery (shown in the left graph in Fig. 4). However, the distribution of time required to reach mastery (shown in the right graph in Fig. 4) is very similar for all exercise types.

This type of descriptive analysis of collected data cannot tell us whether the mastery decisions are "correct". It shows that the proposed approach to mastery criteria is robust and works in stable fashion across wide range of exercises and domains used in online educational systems.

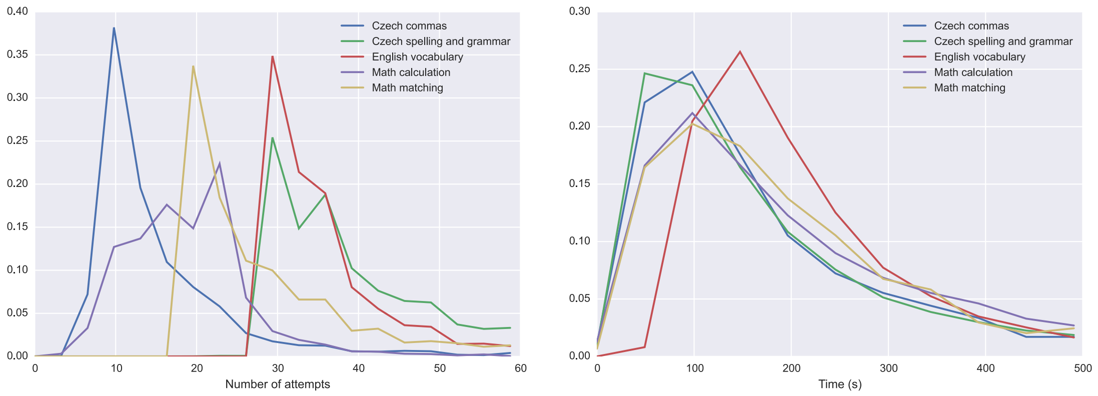

Figure 4. Data from the Umíme system: the distribution of the number of attempts to reach mastery (left) and the distribution of the time to reach mastery (right).

7. 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.

7.1. 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 the 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 thresholds, 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 constructed from historical data. It would be useful to further elaborate other techniques described in previous work (Fancsali, Nixon, & Ritter, 2013; González-Brenes & Huang, 2015).

7.2. Exponential Moving Average

Our results and previous work (Pelánek, 2014) 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. 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 the available training data.

The basic exponential moving average method utilizes only the binary correctness of answers. It can be quite easily extended to incorporate other types of data about learner's performance. We present one specific proposal, which takes into account response times and classification of answers. An important practical advantage of the approach is that it can be used for wide variety of exercises and knowledge components with just a small number of tunable parameters.

7.3. 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" (Baker, 2016) – using analytics tools to inform humans and then implement relatively simple, but well-selected and well-tuned methods into computer systems.

7.4. 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 practice, however, knowledge components are not isolated – they are interrelated, for example by prerequisite relations. Moreover, individual items may be related to several knowledge components. In these cases 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 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 (Xiong, Li, & Beck, 2013). 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.

Both forgetting and relations between knowledge components can be incorporated into the proposal presented in Section 6 by modifying the initial value of skill or the persuasiveness parameter. Such extension deserves further attention and detailed analysis.

We also do not consider wheel-spinning learners (Beck & Gong, 2013) – learners who are unable to master a knowledge component and instead of continued practice would benefit from redirection to one (or more) prerequisite knowledge components. This issue has been addressed by policies developed in previous work (Käser et al., 2016; Rollinson & Brunskill, 2015). These policies have been evaluated for learner modeling techniques; it may be interesting to explore their combination with the exponential moving average method.

Finally, in the presented analysis we ignore potential biases present in real data, particularly attrition bias (Pelánek et al., 2016). This can be potentially an important issue in the evaluation of mastery criteria, particularly in the analysis of effort-score graphs. It would be useful to develop techniques for detecting and overcoming such biases in the effort-score analysis.

Acknowledgements

We thank Petr Jarušek for assistance with the used data and for discussions about the design and implementation of the mastery criteria described in Section 6. We also thank anonymous reviewers for their valuable comments and suggestions.

References

Baker, R. S. (2016). Stupid tutoring systems, intelligent humans. *International Journal of Artificial Intelligence in Education*, *26*(2), 600–614.

Baker, R. S., Goldstein, A. B., & Heffernan, N. T. (2011). Detecting learning moment-by-moment. *International Journal of Artificial Intelligence in Education*, *21*(1-2), 5–25.

Beck, J. E., & Gong, Y. (2013). Wheel-spinning: Students who fail to master a skill. In *Proc. of Artificial intelligence in education* (pp. 431–440).

Bloom, B. S. (1968). Learning for mastery. *Evaluation comment*, *1*(2), 1-12.

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

Csikszentmihalyi, M. (1991). *Flow: The psychology of optimal experience.* Harper Perennial.

Desmarais, M. C., & Baker, R. S. (2012). A review of recent advances in learner and skill modeling in intelligent learning environments. *User Modeling and User-Adapted Interaction*, *22*(1-2), 9–38.

Emrick, J. A. (1971). An evaluation model for mastery testing. *Journal of Educational Measurement*, *8*(4), 321–326.

Fancsali, S. E., Nixon, T., & Ritter, S. (2013). Optimal and worst-case performance of mastery learning assessment with bayesian knowledge tracing. In *Educational data mining.*

Fancsali, S. E., Nixon, T., Vuong, A., & Ritter, S. (2013). Simulated students, mastery learning, and improved learning curves for real-world cognitive tutors. In *AIED workshops.*

Galyardt, A., & Goldin, I. (2015). Move your lamp post: Recent data reflects learner knowledge better than older data. *Journal of Educational Data Mining*, *7*(2), 83–108.

González-Brenes, J. P., & Huang, Y. (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.*

Hu, D. (2011). *How khan academy is using machine learning to assess student mastery.* (http://david-hu.com/2011/11/02/how-khan-academy-is-using-machine-learning-to-assess-student-mastery.html)

Käser, T., Klingler, S., & Gross, M. (2016). When to stop?: Towards universal instructional policies. In *Proc. of learning analytics & knowledge* (pp. 289–298). ACM.

Käser, T., Klingler, S., Schwing, A. G., & Gross, M. (2017). Dynamic bayesian networks for student modeling. *IEEE Transactions on Learning Technologies*, *10*(4), 450–462.

Kelly, K., Wang, Y., Thompson, T., & Heffernan, N. (2015). Defining mastery: Knowledge tracing versus n-consecutive correct responses. In *Proc. of Educational data mining.*

Koedinger, K. R., Booth, J. L., & Klahr, D. (2013). Instructional complexity and the science to constrain it. *Science*, *342*(6161), 935–937.

Koedinger, K. R., Corbett, A. T., & Perfetti, C. (2012). The knowledge-learning-instruction framework: Bridging the science-practice chasm to enhance robust student learning. *Cognitive science*, *36*(5), 757–798.

Lee, J. I., & Brunskill, E. (2012). The impact on individualizing student models on necessary practice opportunities. In *Proc. of educational data mining* (p. 118-125).

Lewis, C., & Sheehan, K. (1990). Using bayesian decision theory to design a computerized mastery test. *ETS Research Report Series*, *1990*(2).

Macready, G. B., & Dayton, C. M. (1977). The use of probabilistic models in the assessment of mastery. *Journal of Educational and Behavioral Statistics*, *2*(2), 99–120.

Mozer, M. C., & Lindsey, R. V. (2016). Big data in cognitive science. In M. N. Jones (Ed.), (chap. Predicting and improving memory retention: Psychological theory matters in the big data era). Taylor & Francis.

Pardos, Z. A., & Yudelson, M. V. (2013). Towards moment of learning accuracy. In *AIED 2013 workshops proceedings volume 4* (p. 3).

Pavlik, P. I., Cen, H., & Koedinger, K. R. (2009). Performance factors analysis-a new alternative to knowledge tracing. In *Proc. of Artificial intelligence in education* (Vol. 200, pp. 531–538). IOS Press.

Pelánek, R. (2014). Application of time decay functions and elo system in student modeling. In *Proc. of Educational data mining* (p. 21-27).

Pelánek, R. (2015). Metrics for evaluation of student models. *Journal of Educational Data Mining*, *7*(2).

Pelánek, R. (2017). Bayesian knowledge tracing, logistic models, and beyond: an overview of learner modeling techniques. *User Modeling and User-Adapted Interaction*, *27*(3), 313–350.

Pelánek, R., & Jarušek, P. (2015). Student modeling based on problem solving times. *International Journal of Artificial Intelligence in Education*, *25*(4), 493–519.

Pelánek, R., Papoušek, J., Řihák, J., Stanislav, V., & Nižnan, J. (2017). Elo-based learner modeling for the adaptive practice of facts. *User Modeling and User-Adapted Interaction*, *27*(1), 89–118.

Pelánek, R., & Řihák, J. (2017). Experimental analysis of mastery learning criteria. In *Proc. of User modelling, adaptation and personalization* (pp. 156–163). ACM.

Pelánek, R., & Řihák, J. (2016). Properties and applications of wrong answers in online educational systems. In *Proc. of Educational data mining* (p. 466-471).

Pelánek, R., Řihák, J., & Papoušek, J. (2016). Impact of data collection on interpretation and evaluation of student model. In *Proc. of Learning analytics & knowledge* (pp. 40–47). ACM.

Rasch, G. (1960). *Studies in mathematical psychology: I. probabilistic models for some intelligence and attainment tests.* Nielsen & Lydiche.

Rau, M. A., Aleven, V., & Rummel, N. (2010). Blocked versus interleaved practice with multiple representations in an intelligent tutoring system for fractions. In *International conference on intelligent tutoring systems* (pp. 413–422).

Rihák, J. (2015). Use of time information in models behind adaptive system for building fluency in mathematics. In *Proc. of Educational data mining.*

Ritter, S., Yudelson, M., Fancsali, S. E., & Berman, S. R. (2016). How mastery learning works at scale. In *Proc. of acm conference on learning@scale* (pp. 71–79).

Rollinson, J., & Brunskill, E. (2015). From predictive models to instructional policies. In *Proc. of educational data mining.*

Semb, G. (1974). The effects of mastery criteria and assignment length on college-student test performance. *Journal of applied behavior analysis*, *7*(1), 61–69.

Wang, Y., & Heffernan, N. (2013). Extending knowledge tracing to allow partial credit: Using continuous versus binary nodes. In *Proc. of Artificial intelligence in education* (pp.

181–188).

Xiong, X., Li, S., & Beck, J. E. (2013). Will you get it right next week: Predict delayed performance in enhanced its mastery cycle. In *Proc. of FLAIRS conference.*

Yudelson, M. V., & Koedinger, K. R. (2013). Estimating the benefits of student model improvements on a substantive scale. In *EDM 2013 workshops proceedings.*

Yudelson, M. V., Koedinger, K. R., & Gordon, G. J. (2013). Individualized bayesian knowledge tracing models. In *International conference on artificial intelligence in education* (pp. 171–180).

Yudkowsky, R., Park, Y. S., Lineberry, M., Knox, A., & Ritter, E. M. (2015). Setting mastery learning standards. *Academic Medicine*, *90*(11), 1495–1500.