Computerized Adaptive Practice of Factual Knowledge

Doctoral Thesis

Jan Papoušek

Brno, Fall 2017
Declaration

Hereby I declare that this paper is my original authorial work, which I have worked out on my own. All sources, references, and literature used or excerpted during elaboration of this work are properly cited and listed in complete reference to the due source.

Jan Papoušek

Advisor: Radek Pelánek
Acknowledgement

In 2013, I decided to change the field of my interest. Thanks to my supervisor Radek Pelánek, I appreciate this decision as successful. Working with him was an unforgettable experience which sometimes went into depressing moments, when I realized I was unable to respond to the flow of his ideas fast enough. He is a person who has greatly expanded my horizons of critical thinking. Even very simple things can introduce issues that catch you in the moments when you are least prepared. I regret, however, I have not been able to learn Radek’s optimism over the last few years. Thanks to Radek for being my supervisor.

For most of my studies, I was working with Vít Stanislav who advocates that issues can be solved easily and quickly. Although we sometimes had to extinguish a fire created in the meantime, without this approach, the Outline Maps and Practice Anatomy projects would not have arisen. With Jiří Řihák, we have undergone many stimulating debates concerning not only our research topics. Thanks to the presence of Víta and Jirka in our research group, Ph.D. study was not only professionally beneficial for me, but also incredibly entertaining. It was rewarding to work in an environment where people systematically try to find conceptual errors in their thinking. Thanks to Víta and Jirka that I could study by their side.

Last but not least, I must mention my family. Without my wife Terka and our evening discussions, I would be lost. Terka is the driving force of my life. Also, I would not stay here without my parents and their support. Thanks to them.
Abstract

The importance of online educational systems is growing. They are able to collect more and more data which opens new possibilities to apply methods known from machine learning and data analysis to improve their behavior to make them more beneficial to learners. We propose a framework for computerized adaptive practice of facts providing multiple-choice questions in domains with varied prior knowledge. The framework uses two separate models for prior and current knowledge and based on this, it tries to provide learners with question of appropriate difficulty. To construct sufficiently competitive distractors in the case of multiple-choice questions, the framework works with historical data containing past wrong answers. We implemented the proposed approach in two systems for practising geography and anatomy. These systems are widely used and serve as a platform to evaluate the framework.

New possibilities come together with new issues related to the bias in the collected data introduced by the intelligent behavior of systems and by learners interacting with the systems in an uncontrolled environment. We investigate the impact of the data collection on the evaluation of learner models. We discuss methodological issues in evaluation of an adaptive educational system as a whole, and propose specific techniques for measuring learners’ engagement and learning. To illustrate these techniques, we discuss five case studies — large scale randomized control trials performed using the adaptive learning system for geography. The results provide an interesting insight into several design aspects of an adaptive educational system. More importantly, the analysis of results highlights several general evaluation issues: the choice of what we measure (engagement versus learning, short term engagement versus long term engagement), the choice of how we measure (measuring engagement using the number of answers versus the time spent within the system), and attrition bias.
Keywords

varied prior knowledge, adaptive educational system, intelligent tutoring system, computerized adaptive practice, factual knowledge, task construction
4.1.3 Learner Models ............................. 52
4.1.4 Data ........................................... 53
4.2 Impact on the Practiced Content ................. 53
4.3 Feedback between Data Collection and Evaluation .... 55

5 Evaluation Methods .......................... 57
  5.1 Engagement ...................................... 58
    5.1.1 Survival Analysis .......................... 58
    5.1.2 Short-term vs. Long-term Engagement .......... 60
  5.2 Learning ........................................ 60
    5.2.1 Reference Questions ................. 61
    5.2.2 Learning Curves .......................... 62

6 Case Studies ................................. 65
  6.1 Adaptive vs. Random ........................... 66
    6.1.1 Data Analysis .............................. 68
    6.1.2 Attrition Bias .............................. 69
    6.1.3 Results .................................... 71
  6.2 Question Difficulty .......................... 73
    6.2.1 Data Analysis .............................. 73
    6.2.2 Results .................................... 75
  6.3 Choice of Distractors ......................... 77
    6.3.1 Results .................................... 79
  6.4 Number of Options ............................ 80
    6.4.1 Results .................................... 80
  6.5 Self-Regulated Question Difficulty ......... 80
    6.5.1 Data Analysis .............................. 81
    6.5.2 Results .................................... 84

7 Conclusion ........................................ 89
  7.1 Lessons Learned .............................. 90
    7.1.1 Monitoring of System Behavior ............... 90
    7.1.2 Choice of Metric ............................ 91
    7.1.3 Attrition Bias ............................. 92
    7.1.4 Levels of Aggregation .................... 93
    7.1.5 Design of Conditions ................... 94
  7.2 Future Work .................................... 95
    7.2.1 Comparison to Other Methods ............. 95
<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>7.2.2 Bias in Data Collection</td>
<td>95</td>
</tr>
<tr>
<td>7.2.3 Automated Parameter Tuning</td>
<td>96</td>
</tr>
<tr>
<td>A Data</td>
<td>97</td>
</tr>
<tr>
<td>A.1 Usual Traffic</td>
<td>97</td>
</tr>
<tr>
<td>A.1.1 Ethical and privacy considerations</td>
<td>98</td>
</tr>
<tr>
<td>A.1.2 Description</td>
<td>98</td>
</tr>
<tr>
<td>A.1.3 Limitations</td>
<td>100</td>
</tr>
<tr>
<td>A.2 Case studies</td>
<td>100</td>
</tr>
<tr>
<td>B Author’s Publications</td>
<td>103</td>
</tr>
<tr>
<td>B.1 Journal Papers</td>
<td>103</td>
</tr>
<tr>
<td>B.2 Conference and Workshop Proceedings</td>
<td>103</td>
</tr>
<tr>
<td>B.3 Technical Reports</td>
<td>105</td>
</tr>
</tbody>
</table>
1 Introduction

Online educational systems like Khan Academy, Coursera, or Duolingo are becoming more and more important for individual learning. The number of learners using this kind of systems is growing [81, 3] and they are expected to be a key part of education. They have the potential to provide adaptive content personalized to the needs of a particular learner. Teachers who have 20 – 30 students with various skills and knowledge in a classroom simply cannot have enough time to devote to everyone individually. The mentioned systems are expected to assist teachers and help them to save time to focus on the improvement of teaching.

Obviously, with the growing number of users, there is higher and higher pressure on educational systems to tune their behavior and assess whether the expectations are actually met in reality. Since the systems collect a huge amount of data, it should theoretically be easy. Unfortunately, there is a lot of decisions which have to be made and as the majority of these systems are online and open to everyone (not only properly labeled users coming from schools) proper evaluation of made decisions is really complicated. Learners come with different motivations, interact with the application in different environments and have different capabilities. The collected data are noisy and often biased, e.g., the departure of users from the application can depend on a lot of factors — the acquired mastery in the required skill or knowledge, or the end of the reserved time period.

At the same time, we are faced with another issue. We implement the intelligent behavior step by step making new decisions building upon previous ones. When we design a new feature, we usually work with data collected by an already somehow intelligent system we implemented in the past. Therefore, the data we work with can be easily biased and our conclusions not completely correct. Unfortunately, examining our ideas without the risk of a negative impact on learners is very difficult, sometimes even impossible. Of course, we can perform experiments using synthetic data or lab/school-scale experiments with real participants, but these methods are limited. There is often only one way to evaluate our ideas — use the only fully realistic environment we have available, deploy a new version of the system,
provide it to at least a small percentage of our users and monitor key metrics we are interested in.

In the thesis we present a framework providing computerized adaptive practice [49] focused on the acquisition of factual knowledge. Using the terminology of the knowledge-learning-instruction framework [52], we work with knowledge components with a constant application condition and a constant response. More specifically, we deal with domains where there is a high variety of learners’ prior knowledge, e.g., foreign language vocabulary, geography, or human anatomy. In comparison with large online educational applications, the system is used only by a fraction of users, but its traffic is sufficient to collect enough data to make the system intelligent, to be able to evaluate our hypotheses and to study related issues.

A key part of the thesis is the adaptive behavior of educational systems and its evaluation. Adaptation itself has been studied thoroughly in the context of computerized adaptive testing [24], where the primary goal is to estimate learners’ abilities with the most confidence. On the other hand, we focus on computerized adaptive practice where our goal is to improve learners’ skill. These areas sound similar, but in the case of practicing, an estimation is needed to provide personalized content and statistical confidence is not so important. Estimation itself is only a secondary goal.

Adaptive behavior has also been studied in the context of intelligent tutoring systems [125]. The meaning of this term varies, but it often refers to systems which are designed mainly for learning more complex cognitive skills than learning facts, e.g., mathematics [51] or physics [117]. Here, the relevant issues to study are step-by-step solution monitoring, hints, and forms of feedback. The learning of facts (or drilling the multiplication table) is less complex, so these issues are not directly relevant. For one user, we typically have data about interactions with much more tasks. Although the data about these interactions contain less information, it allows us to deal with one instance of the task solution less and focus more on the order of the presented tasks, question construction, or evaluation methods.
1. Introduction

1.1 Contribution of the Thesis

1.1.1 Systems Providing Practice of Factual Knowledge

We have developed a framework running two online adaptive educational systems providing practice of factual knowledge — geography (outlinemaps.org) and anatomy (practiceanatomy.com). Both these systems are actively used by hundreds of learners per day helping them learn a given subject. The application for geography is by 2 years older and has a much higher traffic (see Figure 1.1). It collects tens of thousands of answer per day, so it serves as the main source of educational data and the main platform for online experiments focused on several different aspects of practice (difficulty, number of distractors, et cetera). Both the data from usual traffic and data from the performed experiments are available online to make our research reproducible \(^1\). We try to close the loop, so the findings based on the collected data help to improve the systems.

![Figure 1.1: Usage per month.](image)

Main contributions:

- The systems implement a model predicting learners’ performance (see Section 3.2) and an algorithm for adaptive practice (see Section 3.3).
- The applications are widely used by elementary/high school students in the case of Outline Maps and medical students in the case of Practice Anatomy.

\(^1\) [https://data.outlinemaps.org](https://data.outlinemaps.org) and [https://data.practiceanatomy.com](https://data.practiceanatomy.com)
1. Introduction

- Outline Maps serves as a platform for online experiments and a source which can be analyzed offline, e.g., [84, 106, 126].

Author’s publications relevant to the topic:


1.1.2 Evaluation Methods

We have identified issues related to the data collection within adaptive online educational systems. The collected data are often biased, e.g., the system prefers to ask a specific type of questions or the learners’ error rate is constant regardless of the learners’ ability. Since the data are biased, we must be careful when analyzing and interpreting them. For example, designing and fitting a learner model is directly affected by the system which collects the data, so we definitely should know whether the system provides educational content adaptively or randomly. A simple comparison of two variants of the system becomes very difficult.

We propose a framework for the evaluation of online adaptive educational systems as a whole focused on learners’ engagement and learning. Unlike the commonly used evaluation methods where participants are recruited artificially or from schools and interact with the studied system in a defined way, we focus on experiments where the participants are real users of the system. Typically, we have almost no information about them and no control over their interaction with the system. This approach is already widely used in other areas, mainly in industry [41, 53]. Despite the fact that it is quite complicated in many aspects, the internet-scale experiments should suitably supplement widespread lab-scale and classroom-scale experiments [120].

Main contributions:
1. Introduction

- Description of issues related to adaptivity, data collection and evaluation.

- A framework which can be easily used to evaluate systems for computerized adaptive practice with respect to users’ learning and engagement.

Author’s publications relevant to the topic:


1.1.3 Insight into the Practice of Facts

Although the systems implemented for the purpose of the thesis are quite narrowly oriented, their design consists of a lot of attributes which are shared across many online educational systems. Using the framework for online experiments mentioned above, we ran multiple experiments examining a few aspects of practicing facts. Our findings can be more or less implemented into other educational systems to improve learners’ benefit.

Main contributions:

- We have proven that the adaptive practice is beneficial for learners. It increases both learning and engagement.

- We have confirmed the hypothesis about the relation of optimal difficulty and external motivation [1]. Learners with external motivation prefer easier content.

- We examined the impact of the practice difficulty on learning and engagement. We found there is difference between short-term and long-term engagement.
1. Introduction

- We found the optimal number of distractors for multiple-choice questions.

Author’s publications relevant to the topic:


2 Related Work

This chapter presents an overview of the research related to the thesis. In the beginning, we set up the terminology and introduce general ideas behind computerized adaptive practice. The first part of the chapter (Section 2.1) deals with several models predicting learners’ performance. These models are necessary to make an educational system adaptive. Next (Section 2.2), we present research on instructional policy and various approaches to learning not only facts. The last part (Section 2.3) is devoted to evaluation methods which are typically chosen to examine hypotheses related to computerized adaptive practice.

In the thesis, the following terminology is used. A learner is a user of a system who interacts with the learning material. The smallest unit of the material is called an item. In a lot of systems, items correspond directly to the tasks presented to learners, but it is not necessary — a system can generate questions algorithmically. By an item, we mean the smallest piece of knowledge or skill that learners want to acquire. An item can be a Latin word, the location of a country, or the ability to algorithmically multiply two numbers. Items are assigned to a hierarchy of knowledge components. Sometimes one item belongs to several knowledge components. The granularity of the knowledge components fully depends on the purpose of a particular educational system. It can be useful to create one knowledge component for each item in the system and leave it at that. On the other hand, in a lot of systems, designers build a rich hierarchy of knowledge components to be able to create better learner models, or to provide a richer user interface.

Based on recent research [9], self-testing is a good strategy that may boost learners’ performance. To get an idea what a common system for adaptive practice looks like, see a screenshot of a sample question from an application called Practice Anatomy (Figure 2.1). Using this system, people can study a narrow subarea of anatomy answering multiple-choice and open questions. An example of an item is the relation “transverse cervical artery supplies trapezius muscle”. The knowledge of this item is tested using a multiple-choice question where the distractors are generated based on the collected
2. Related Work

Figure 2.1: A sample question from practiceanatomy.com.

data. The item belongs to the knowledge component “relations” and knowledge components corresponding to terms “transverse cervical artery” and “trapezius muscle”. These components further belong to the components corresponding to organ systems and parts of the body. The interactions the system collects are answers to these questions. We are usually interested in whether the answer is correct or wrong, when it is entered into the system, the amount of time a particular learner spends answering the question, … Based on this data, a learner model builds knowledge about learners and items and this knowledge is further used by an instructional policy to provide an adaptive educational content.

In this chapter we will not cover heuristic-based systems such as [132]. Although variants of these algorithms are popular in some flashcard software to determine spacing intervals for practice, we assume parameters of this method are hand-picked [118].

2.1 Learner Models

A fundamental part of adaptive educational systems are learner models [26]. Typically, we use a model predicting a learner’s future performance based on their historical performance. We are interested in the probability of choosing the correct answer if the system shows a learner a particular task — based on binary historical data (wrong/correct
answers), we predict a real number between 0 and 1. Recent research suggests that even a small improvement in the accuracy of a learner model can have significant impact on learners’ practice [134, 98, 65]. Of course, that depends on how much an educational system relies on the model. To assess the model quality, we usually look at its accuracy. The accuracy is important, but we usually deal with a big stream of data, so the model should be easy to update. It is not realistic to assume you can “stop the world” and fit parameters of the model offline. This process usually takes quite a long time and the model must be usable online in the meantime. Ideally, the model updates its parameters locally, looking only at a piece of data (one answer). Similarly, it should be easy to add new items to the system without major interventions. Authors of educational systems often start with a proof of concept covering a small part of the curriculum and then extend their application to provide more and more content.

For the purpose of building a new system providing computerized adaptive practice, we should consider two kinds of learner models. Firstly, we need to predict a learner’s knowledge of items for which the system has no interaction records from the particular learner. In this case, the model estimates the knowledge based on the learner’s past interactions with other items, even items which are not from the same knowledge component. These models actually handle a learner’s prior knowledge which is not affected by practicing within the system. Therefore, for simplicity, we assume it is constant. Models estimating prior knowledge based on the data coming from educational systems are very similar to models widely used for testing purposes.

On the other hand, a learner may interact with the same item multiple times, so the system needs a model which handles learning (and forgetting). We say the model estimates a learner’s current knowledge. For this purpose, we can draw inspiration from learning of procedural knowledge, where a knowledge component consists of a number of items. When a learner interacts with more items from the same component in the case of procedural knowledge, they are in the same situation as a learner interacting multiple times with the same item in the case of factual knowledge. This claim is based on the assumption that one factual item corresponds to exactly one knowledge component, and vice versa.
2. Related Work

It is worth noting that predictions are not the only useful output we can get from a learner model. Sometimes it is simply the fact that we are able to model the reality that can give us insight into the data. In Duolingo, a model consisting of a mixture of individual learning curves [121], revealed a group of users consistently skipping a certain type of content. This insight can help developers modify the system to better fulfill its users’ needs.

2.1.1 Prior Knowledge

Generally, there is less attention to modeling learners’ prior knowledge than modeling their current knowledge taking learning into account within the community around an educational system [84]. Actually, this topic is much more related to testing, so in this section we describe item response theory [24] and the way of estimating its parameters online inspired by the Elo system [32, 104].

Item Response Theory

Item response theory (IRT) [5, 39, 70] originally comes from psychometrics, specifically, computerized adaptive testing. The goal is to efficiently estimate the skills of the tested people. Despite the fact that the goal of educational systems is quite different — improve learners’ skills, it seems suitable to use IRT as an estimator of the correct answer to a potential question presented by a system. There are multiple versions of IRT differing in a number of parameters describing learners and items. In all cases, the IRT model is based on the logistic function.

The basic version of the IRT model is the one-parameter logistic model, often referred to as the Rasch model [11, 133]. It consists of one parameter for each learner \((s)\) standing for their skill \((\theta_s)\) and one parameter for each item \((i)\) representing its difficulty \((b_i)\). The resulting probability of the correct answer on an item \((i)\), or question related to an item \((i)\), depends on the difference of these parameters:

\[
P_i(\text{correct}_{si} = 1) = \frac{1}{1 + e^{(\theta_s - b_i)}} \tag{2.1}
\]

If a learner’s skill is equal to the difficulty of the item, the probability of a correct answer is 50%. With increasing learner’s skill the
2. Related Work

The probability of the correct answer is higher, with increasing difficulty the probability is lower.

![Graphical representation of parameters in the three-parameter logistic model (left) and sample parameter setting (right).]

There is a natural extension of the one-parameter logistic model — a three-parameter logistic one. It introduces two extra parameters for each item — a discrimination parameter $a_i$ and a pseudo-guessing parameter $c_i$. The parameter $a_i$ represents how steeply the probability of the correct answer varies with the ability of learners and the parameter $c_i$ stands for the lower asymptote — the probability that learners with a lower skill level will answer correctly due to guessing, e.g. in the case of multiple-choice questions $c_i = \frac{1}{n}$ where $n$ stands for the number of choices. The resulting formula follows:

$$P_i(\text{correct}_s = 1) = c_i + \frac{1 - c_i}{1 + e^{-a_i(\theta_s - b_i)}} \quad (2.2)$$

To use the logistic model in a real educational system, it is necessary to estimate all its parameters. The pseudo-guessing parameter $c_i$ is often known (multiple-choice questions) and difficulty $b_i$ is sometimes provided by an expert. On the other hand, we still have to estimate at least the learner’s skill based on the collected data and the experts can be wrong, so it is useful to estimate the difficulty of items as well. The estimation of parameters related to an item should be the same regardless of whether the item is shown to a representative subset of people, or to people with high (or low) skill only [5]. It can easily
2. Related Work

happen that we observe mostly correct answers, however, the item is really difficult.

To get values of the parameters, the maximum likelihood estimation [6, 10] is usually used. This method repeatedly alternates two steps — an estimation of skills using already estimated parameters for items and an estimation of item parameters based on already estimated skills. We repeat the steps until the values of the parameters converge [5]. To get reliable estimates of item parameters, the method needs a relatively large training set containing hundreds of answers for each item. A big disadvantage is that this method looks at all the data, so it cannot be easily used online because it needs to “stop the world”. In the next section, we show an algorithm to estimate the parameters easily online. Another (minor) issue is that the absolute value of the parameters does not uniquely determine the model. It is possible to transform the parameters by adding a constant \( k \) to all skills and difficulties and get the same predictions of correct answers. This issue is usually resolved by some kind of normalization, e.g. requiring that the mean value of an ability is 0 and variance is 1.

![Information function](image)

Figure 2.3: Information function of a particular item — we get the most information when the difficulty of the presented item equals to a learner’s skill.

Similarly to all other models, the precision of a learner’s skill estimation depends on the difficulty of the presented items. From this perspective, ideal items are items where a particular learner has a 50% chance of choosing the correct answer. Presenting this kind of items is beneficial in the context of testing. On the other hand, it is question-
able whether this difficulty is also optimal with respect to learning. The differing information gain of items complicates the evaluation of systems for computerized adaptive practice, e.g., when we try to find the optimal target difficulty for our system.

Elo System

The principle of the Elo rating system [32], originally designed to rate chess players, can be described as follows. For each player \( i \) we have a rating estimate (or skill) \( \theta_i \). When a player \( i \) plays a match against a player \( j \), we get a result \( R_{ij} \in \{0, 1\} \) (0 stands for a loss, 1 for a win). The expected probability that the player \( i \) wins the match is described by the logistic function taking the difference of the estimated ratings \( \theta_i - \theta_j \):

\[
P(R_{ij} = 1) = \frac{1}{1 + e^{-(\theta_i - \theta_j)}}
\] (2.3)

After the match, we update the estimated ratings based on the result using the following rule (constant \( K \) stands for the sensitivity of the estimate to the last update):

\[
\theta_i = \theta_i + K(R_{ij} - P(R_{ij} = 1))
\] (2.4)

The introduced idea can be applied in the context of educational systems by interpreting a learner’s answer to an item as a match between the learner and the item. Although the Elo rating system was originally designed to track changing skills and in the case of the prior knowledge we assume that the skill level is constant, this way can be easily used to estimate skills \( \theta_s \) and difficulties \( b_i \) for a one-parameter logistic model.

\[
\theta_s = \theta_s + K(correct_{si} - P(correct_{si} = 1))
\] (2.5)

\[
b_i = b_i - K(correct_{si} - P(correct_{si} = 1))
\] (2.6)

---

1. The original Elo system for chess players rescales the standard logistic function:

\[
P(R_{ij} = 1) = \frac{1}{1 + 10^{-(\theta_i - \theta_j)/400}} [107].
\]}
2. Related Work

Initially, the values of skills $\theta_s$ and difficulties $b_i$ are set to 0. As we already stated, the parameter $c_i$ can often be computed directly without a complicated estimation, so it can still be easily incorporated.

To make the estimate stable and converging we need to replace the sensitivity constant $K$ by an uncertainty function ensuring that later changes have less weight. Without that we could easily lose information from past updates. This function can have several forms, but often depends directly on the number of past updates of a particular difficulty (or skill) [107, 131]. Another approach is to define uncertainty as a separate parameter for each item and each learner, so it is updated with each update of a particular difficulty (or skill) [49]. In this text, when we refer to the Elo system, we use the following uncertainty function:

$$K \sim U(n) = \frac{a}{1 + bn} \quad (2.7)$$

The variable $n$ stands for the number of past updates and $a, b$ are meta-parameters fitted to data ($a = 1, b = 0.05$ in [88, 84]).

There are already studies comparing the Elo system to other methods for estimating the difficulty of items, such as the joint maximum likelihood estimation [107], the portion of correct answers, or human judgment methods [130]. Figure 2.4 (left) shows the correlation of estimated difficulties using data from our adaptive system Outline Maps depending on the size of the collected data. This data are not optimal, but realistic. The distribution of answers over the items is uneven and the average success is far from ideal - 50%, see Figure 2.4 (right). The Elo rating system seems to be a reliable estimation method.

In an educational context, we often work with multiple knowledge components [52]. Instead of a single skill for a learner, we can measure several correlated skills. This extension has been studied in the context of adaptive experiments in psychology [28] and also for estimating a learner’s prior knowledge using data collected by Outline Maps [84].

2.1.2 Current Knowledge

As a learner practices and interacts with items even several times, we need to estimate not only their prior knowledge, but also their
2. Related Work

Figure 2.4: Correlation between difficulties estimated by the Elo rating system and the joint maximum likelihood method with respect to the size of data (left). Comparison of average success rates and estimated difficulties, the area of markers stands for the number of answers collected for a particular item (right). Dataset consists of 39 items (European countries) where the minimum number of answers per item is 139 (in the full dataset). Data comes from an adaptive system Outline maps, so there is a tendency to have roughly 75% of correct answers.

current knowledge for each knowledge component. By interacting with the items, the learner learns, and, in the case of factual knowledge, their knowledge also decreases with time because of forgetting. It is worth noting that models for the current knowledge are related to the field called mastery detection [18] where a learner interacts with one knowledge component until a system detects that the knowledge component is mastered by the learner. Then the learner is moved to another component. In these systems, one component typically consists of a big number of items, or the provided tasks are generated automatically.

Bayesian Knowledge Tracing

Probably the most popular model estimating a learner’s current knowledge is Bayesian Knowledge Tracing [21, 20, 124], a special case of the Hidden Markov Model. This model was originally developed to describe the acquisition of procedural knowledge and assumes that a learner’s knowledge is represented as a set of binary variables — one
per each skill (knowledge component). We observe binary events (answers) — a particular learner gets a problem either right or wrong. Since the skill is not continuous, but rather simplified as a binary state, the model is suitable for a specific set of educational content.

The model has the following parameters for each skill (knowledge component). Their meaning is schematically expressed by Figure 2.5.

\[ P(L_0) \] is the initial probability that the learner knows a particular skill before the skill is used;

\[ P(G) \] is the probability of guessing correctly if the learner does not know the skill;

\[ P(S) \] is the probability of making a slip if the learner knows the skill;

\[ P(T) \] is the probability of learning the skill when the skill is used and the learner knows it.

![Diagram](image.png)

**Figure 2.5:** Diagram describing Bayesian Knowledge Tracing. Circles stand for a hidden state, rectangles for observation.

If we know the probability \( P(L_n) \) that a particular learner knows a skill after \( n \) answers, the prediction whether the next answer will be correct can be described as follows (\( O_i \) stands for the observation of the \( i^{th} \) answer).

\[
P(O_{n+1} = \text{correct}) = P(G) (1 - P(L_n)) + (1 - P(S)) P(L_n) \quad (2.8)
\]
Using conditional probability we can also express how the probability that a learner knows the skill changes after \( n + 1 \) answers. It is defined by the following equations [124]:

\[
P(L_{n-1}|O_n) = \begin{cases} 
    \frac{P(L_{n-1}|O_{n-1})(1 - P(S))}{P(L_{n-1}|O_{n-1})(1 - P(S)) + (1 - P(L_{n-1}|O_{n-1})) P(G)} & \text{if } O_n = \text{correct} \\
    \frac{P(L_{n-1}|O_{n-1})P(S)}{P(L_{n-1}|O_{n-1})P(S) + (1 - P(L_{n-1}|O_{n-1}))(1 - P(G))} & \text{if } O_n = \text{wrong} 
\end{cases}
\]

(2.9)

(2.10)

(2.11)

\[
P(L_n|O_n) = P(L_{n-1}|O_n) + (1 - P(L_{n-1}|O_n))P(T)
\]

Figure 2.6: Sample scenarios showing the behavior of the Bayesian Knowledge Tracing model.

Although the model assumes the skill can only be learned and there is no parameter for the loss of the acquired knowledge, when it observes wrong answers, it updates its belief about a learner, so the predicted probability of a correct answer decreases, see Figure 2.6. The parameters are usually fit using the expectation maximization method [61], conjugate gradient search [21], or discretized brute-force search (grid search) [7]. There are also extensions covering more
2. Related Work

complex scenarios like forgetting [129, 111] or learners’ prior knowledge [97, 96, 135].

Performance Factor Analysis

One of the models working with continuous skill is the Performance Factor Analysis [102, 35] originally based on the Learning Factor Analysis [17]. Similarly to the Item response theory, the model uses a logistic function. A learner’s skill is represented as a linear combination of past successes (s) and failures (f). Given a difficulty $\beta_i$ for each knowledge component $i$, the probability of the correct answer is expressed by the following equation:

$$m = \sum_{i \in KC} -\beta_i + \gamma_i s_i + \delta_i f_i$$  \hspace{1cm} (2.12)

$$P(\text{correct} = 1) = \frac{1}{1 + e^{-m}}$$  \hspace{1cm} (2.13)

The parameter $\gamma$ stands for the change of the skill associated with a correct answer, it is usually a relatively high positive value, because the correct answer indicates a learner acquired the knowledge. The parameter $\delta$ stands for the same in the case of an incorrect answer, it is often negative and its absolute value is usually lower than the absolute value of $\gamma$. Wrong answers are indicators of an absence of knowledge. On the other hand, a learner may learn something during the interaction with an item. The difficulty parameter $\beta$ can be understood as an initial value of the skill. An advantage of this model is that it naturally works with a multidimensional skill. Since the estimation of parameters $\beta$, $\gamma$ and $\delta$ is a logistic regression problem, the estimation can be solved using maximum likelihood estimation.

A big disadvantage of the model is that it does not take the order of the answers into account. See Figure 2.7, the model cannot distinguish between the scenario where a learner firstly answers correctly and than wrongly (half correct, half wrong) and the scenario where it is conversely (half wrong, half correct). The first scenario probably indicates that a learner lost the knowledge, the second one probably means that a learner actually acquired the knowledge.
Forgetting Curves

Unfortunately, the process of learning facts does not consist of interactions with the educational content only, but also the spacing and forgetting effect. The effect of practice is highly dependent on spaces between the interactions with the same item and the acquired knowledge is not stable, it decreases over time [30]. Studies suggest that people learn better if the spacing between interactions gradually increases [76]. The forgetting curve, originally defined by Ebbinghaus [30], assumes that memory decays exponentially over time:

\[ P(\text{correct} = 1) = 2^{\Delta/h} \]

The variable \( \Delta \) denotes the lag time since the item was last presented to a particular learner, the half-life parameter \( h \) represents the memory strength in the learner’s long term memory. To get an idea how the model works, see Figure 2.8. The model was not originally developed for the purpose of educational systems, but to describe the principles of human memory.

\[ \beta = 0.85, \gamma = 0.5, \delta = -0.3 \]

Figure 2.7: Sample scenarios showing behavior of the Performance Factor Analysis model.

2. This phenomenon is called lag effect.
2. RELATED WORK

Figure 2.8: An example of the forgetting curve as a function of lag time \( \Delta \) and half-life \( h = 2 \) (left). Sample trace from Outline Maps where a learner answers questions about Gambia and illustrating forgetting curves (right).

To use forgetting curves in real systems, we need to reliably compute the parameter \( h \). Recent research [118] introduces half-life regression, where common machine learning methods can be used. If \( X \) represents the feature vector describing a learner’s previous interactions with an item and \( \Theta \) stands for the parameter vector consisting of weights where each feature corresponds to one weight, then the estimation of \( h \) can expressed by the following equation. The equation is based on the assumption that each interaction updates the half-life exponentially.

\[
    h = 2^{\Theta \cdot X} \tag{2.15}
\]

This approach allows us to use an arbitrarily large set of interesting features. Having a data set \( D = \{(o, \Delta, X)\}_{i=1}^{D} \) where each data point consists of an observation \( o \), lag time \( \Delta \) and a feature vector \( X \), the goal is to find the best model weights \( \Theta^* \) to minimize some loss function \( \mathcal{L} \):

\[
    \Theta^* = \arg \min_{\Theta} \sum_{i=1}^{D} \mathcal{L} ((p, \Delta, X)_i, \Theta) \tag{2.16}
\]

Another approach to working with forgetting curves is the multiscale context model [80]. This model assumes that each interaction
with an item creates an additional item-specific forgetting curve that decays at a different rate

2.2 Instructional Policy

Using the knowledge about a learner sitting in front of our system, we want to make them practice as efficiently as possible. Previous research [125] identified two directions of helping learners with their practice — the inner and outer loop. In the case of the inner loop we are focused on the interactions with one item. The interactions often consist of multiple steps and take a long time. To get an idea where the inner loop takes place, imagine the process of solving a mathematical equation. To solve the equation, a learner often needs to apply a series of rules, which can obviously take a lot of time. When a learner struggles with a step, the system can provide a carefully aimed hint to help them with it. Here, the key questions are: When should the system provide the hint? How many hints should be shown? How strong should the hints be? If the system provides a hint and a learner successfully solves the item, should a learner model deal with this interaction as a correct or incorrect one?

In the case of the factual knowledge, the inner loop is not so important. Of course, there are exceptions like the translation of sentences, however, in a lot of systems, tasks are simple, often multiple-choice, questions. On the other hand, the outer loop does not focus only on the solution, but on the sequencing of tasks which is really important. What task should be presented to a learner? Has the learner already mastered the concept? Should we move to another topic? Should the question be a multiple-choice, or an open one? Since the outer loop is more common in the case of factual knowledge, this chapter is focused on some existing techniques of item sequencing and aspects of multiple-choice questions which should be taken into account.

The importance of systems assisting with learners’ practice is supported by many arguments. Based on recent research [9], self-testing is a good strategy that may boost learners’ performance. On the other hand, people make a lot of mistakes during self-testing. They overestimate their knowledge and this overconfidence leads to a worse choice of tasks for practice [9, 56]. This makes the practice less efficient than
2. Related Work

it could be. Digital flashcards are more effective as a learning tool than printed word lists [44].

2.2.1 Difficulty

When a learner practices a context, one of the most observable features is their success rate or the time they need to correctly solve presented tasks. This kind of a learner’s impression is called difficulty and we typically try to change the practice difficulty using a lot of factors — e.g., by the choice of presented items, distractors, or limiting time available for the interaction. The goal of the adaptive practice is to find the optimal difficulty for a particular learner and choose proper factors modulating it, so that they are engaged and learn efficiently. With respect to the learner modeling, the optimal difficulty is a 50% error rate, as shown in Subsection 2.1.1. This difficulty is used by most systems for computerized adaptive testing to minimize the length of a test [31] — the difficulty of the selected tasks match a user’s current estimated skill. However, this does not necessarily mean that it is also optimal for the learner.

The idea that presented tasks should not be too easy or too difficult is known as the Inverted-U Hypothesis [1], see Figure 2.9 (left). Interacting with tasks which are too easy or too difficult leads to only a small benefit (e.g., learners are bored or frustrated). Another term related to the optimal difficulty is the zone of proximal development [58] representing the difference between activities a particular learner can easily do without any help and activities the learner is able to do with assistance only, see Figure 2.9 (right). This zone consists of activities that a learner can do with difficulties or a little guidance and by interacting with these activities the learner learns the most. Also, when the achieved difficulty is right, we can achieve that a learner enters the flow state [23], a state when they are concentrated and fully focused on completing the presented tasks.

There is a lot of empirical studies focused on the difficulty, but they often consider engagement only, e.g. chess players reported greatest enjoyment when the probability of their success is roughly 20% [1], children smile more after completing longer and more difficult tasks [38], or players enjoy games more if the provided difficulty is reflective of their gaming experience [2]. The engagement is important because it
2. Related Work

Figure 2.9: Inverted-U curve (left) and schema for thrzone of proximal development (right).

influences time spent within an educational system and a lack of motivation negatively affects a learner’s performance [71]. However, in the educational context we need to make the interactions not only engaging, but we would also like to maximize the effectiveness of learning. There is also research investigating the relationship between difficulty and learning (or another kind of performance). Unfortunately, individual studies sometimes contradict. One study [46] comparing three conditions based on the target difficulty (target success rate 60%, 75%, and 90%) shows that the easiest condition led to the best learning (mediated by a number of solved mathematical tasks). On the other hand, another paper [68] concludes that although lower difficulty leads to higher engagement, it has a negative impact on learning. A similar situation is in the case of methods for adaptivity focused on the difficulty of the presented tasks. One study declares [116] that comparing the difficulty increasing in time vs. the difficulty adjustment based on a learner’s mastery level, the second strategy probably performs better with respect to learning. On the other hand, in a study where an educational computer game was compared varying only the method used to adjust difficulty, the results did not show statistically significant differences in motivation and performance [86].

A system providing an educational activity can also make it possible to self-regulate the difficulty. This approach is based on the idea that there is probably no optimal difficulty for all learners. In addition to the system getting information about the needs of a particular learner, previous research suggests that a sense of control (or percep-
tion of control, rather than the actual objective level of control) can increase engagement [73]. On the other hand, a recent study shows that learners allowed to self-regulate the difficulty of their practice are prone to mismanage their own learning [45].

2.2.2 Sequencing

Learning of facts is a well studied area, especially in previous research on memory covering phenomena such as spacing and forgetting effects [100] and spaced repetition [47]. When items are practiced with more time space between individual presentations, a learner’s performance during the practice itself is reduced. However, it has a positive impact on learning in the long term [72, 55]. Evidence also suggests that manipulation between sessions can have a greater impact on long-term learning [19]. Unfortunately, the environment of this kind of experiments is often not realistic — e.g., to minimize prior knowledge, learners memorize arbitrary word lists, obscure facts, or Japanese [25, 100].

Graduated-Interval Method

One of the first methods focused on sequencing of items handling the spacing effect was the graduated-interval recall method (also known as Pimsleur method) [110]. New vocabulary is introduced and then presented at exponentially increasing intervals which are interleaved with other activities. This approach is very limited since it does not take a learner’s current knowledge into account. It completely ignores a learner’s prior knowledge and varied learning speed of different vocabulary/learners.

![Figure 2.10: Schema of Leitner system.](image-url)
2. RELATED WORK

Leitner System

An improvement of the previous method is the Leitner system [60, 113]. It is more adaptive, since the spacing intervals can not only increase, but also decrease depending on a learner’s performance, see Figure 2.10. All items intended to be learnt are divided into boxes which correspond to different practice intervals (1-day, 2-days, 4-days, ...). When a new item is introduced, it starts out in the 1-day box. When the item is practiced correctly after 1 day, the item is moved to the 2-day box. After 2 days, it is presented again and, if a learner knows it, the item is promoted to the 4-day box. If they do not know it, the item is demoted to the 1-day box. And so on. This approach makes it possible to move items quickly to long practice intervals, so it handles prior knowledge better than the previous method. The Leitner system is used in several electronic flashcard programs. It was even present in the first version of Duolingo [118].

Adaptive Character of Thought-Rational

Another approach is based on a modeling system known as the Adaptive Character of Thought-Rational [4]. For each item we have a series $t_1, \ldots, t_n$ where $t_i$ represents how long ago the $i^{th}$ practice of that item occurred. Then we can compute the memory activation $m_n$ for a particular pair of learner and item as follows\textsuperscript{3}. The decay parameter $d$ is constant.

$$m_n = \ln \left( \sum_{i=1}^{n} t_i^{-d} \right)$$

(2.17)

This model can be extended in many ways. One of the disadvantages of the model is that although it handles forgetting, it does not capture the spacing effect. In [100], authors propose making the decay parameter dependent on the previous activation.

\textsuperscript{3} To get a probability of recall, the logistic function is used again: $P(\text{recall}) = \frac{1}{1+e^{-\tau s}}$, where $\tau$ is the threshold parameter (similar to item difficulty) and $s$ stands for sensitivity.
2. Related Work

\[ d_1 = a \]  
\[ d_i = ce^{m_i-1} + a \]  
\[ m_n = \ln \left( \sum_{i=1}^{n} t_i^{-d_i} \right) \]

The equation contains two more parameters — the decay scale parameter \( c \) and the intercept of the decay function \( a \). When \( c = 0 \), the decay parameter of this extension behaves the same as the decay in the original model \( (d_i = a) \). When there is a short time interval between two presentations of an item, the activation before the presentation is high, so the resulting decay parameter is also high which makes the benefit of the second presentation lower. On the other hand, when the time interval is long, the low activation makes the resulting decay also low and the benefit from the presentation is high.

In [101], authors propose using the activation, so that activation for all unpracticed items is initialized to an advantage point \( \alpha \) \((-0.63\)). Then, when a learner wants to practice, the system recommends practicing as follows (by priority):

1. Drill the already practiced item with the lowest activation above the advantage point \( \alpha \) and below the optimality point \( \beta \) \((-0.33\)).

2. Pick an item with the lowest activation; if the activation is below \( \alpha \), then study, otherwise drill.

Mastery Learning

The order of tasks in a learner’s practice is an alternative to the classical approach using mastery detection where learners do their practice within one concept until they master it, then they are allowed to practice another one [18]. When a learner practices one topic, items are usually presented randomly. Mastery detection tries to minimize the amount of under-practice and over-practice. The system usually estimates a learner’s skill for the currently practiced knowledge component and when the estimated skill or predicted performance is above the mastery-threshold, it marks the knowledge component as
mastered. The mastery-threshold can be viewed as a parameter that controls the relative frequency of under-practice and over-practice [33]. Since educational systems tend to be conservative, over-practice is usually considered less dangerous than under-practice. There are several studies examining different values of the mastery-threshold and behavior of learner models [34, 59, 134].

2.2.3 Question Construction

When a system picks an item to be presented to a learner, it also has to decide about the form of the presentation. Sometimes it is easy because the item is actually phrased as a question. However, in some cases the items do not correspond to the questions directly, but to simple facts (vocabulary, country, anatomical structure, ...). Should the system provide the fact with its description to give a learner a chance to study it? Should it show a question — a multiple-choice or an open one? Should a learner be provided with corrective feedback after answering a question?

Studies show that practicing with multiple-choice questions improves performance on final cued-recall test, but also increases the intrusion rate of distractors [115, 74]. Results state that the positive effects are higher than the negative ones. In the case of high performance in multiple-choice testing, additional lures led to an even higher performance on subsequent tests [15]. On the other hand, in the case of low performance, testing with additional lures negatively influenced later tests. These observations should be definitely taken into account when the distractors are constructed by a system for computerized adaptive practice.

The construction of questions with an appropriate number of distractors and their appropriate competitiveness seems to be very important. We assume that distractors which are too easy make the guess probability so high that a learner is not motivated to learn anything. Studies show that competitive distractors result in higher performance on final cued-recall test [63] if learners get corrective feedback. Based on previous research, there are typically one or two distractors having at least 5% answers [122]. Arguments for three options (two distractors and the correct answer) are simple: (a) it is time efficient, because a learner spends less time on answering a question;
2. Related Work

(b) there are often not many good distractors disclosing answers to next questions [114, 127]. Although three options seem to be enough, the common practice is to show four or even five options [37].

A learner can be provided with feedback about the correctness after they answer a question. Studies show that this feedback is important, because it may serve as a motivation for learners to choose options that are expected to be correct, but with low confidence [14]. Despite this fact, a lot of research focused on the efficiency of using multiple-choice questions does not provide learner with corrective feedback in the experiments.

2.3 Evaluation Methods

Any effort to improve any kind of intelligent tutoring system would be useless if we were not able to measure some of its features. The design of these systems and the choice of the parameters for their adaptive behavior are really difficult. Together with the fact that the adaptive behavior is hard to predict, it is crucial to be sure that the implemented system meets our expectations. We want to improve a user’s learning, but due to the fact that systems for computerized adaptive practice are also meant to be used by people in their free time, we primarily have to ensure that people are willing to use them. How these things can be in contradiction to each other is studied in [68].

Since a lot of systems collect data about learners’ performance and their behavior is based on models predicting the performance, it is natural to evaluate new models using the historical data with respect to the accuracy of the model. We can also try to find whether the behavior of the system would change if it used a different model, or deploy a new feature, collect new data and then see what has changed. Unfortunately, this is really expensive because we force learners to spend their time on testing our system instead of on learning something new. The last way to examine our ideas is to rely on synthetic data. Generating synthetic data is cheap, there is no risk, but it is necessary to work with a lot of assumptions which can be quite limiting.
2. Related Work

2.3.1 Historical Data

Typical data about users’ performance contains a list of solved tasks and binary information whether a user’s solution is correct (1) or incorrect (0). A model predicting a user’s performance returns values from the interval $[0, 1]$. Based on this we can try to create a better model, fit it on a subset of data and test its predictions comparing them to the old model using the rest of data. The choice of metric for the model evaluation is crucial, we can find inspiration in many areas such as recommender systems [40], ecology [64], or weather forecasting [123]. If we are interested in relative values of predictions, *area under curve* is probably a good choice. On the other hand, if the system relies on the absolute values of the predictions, the use of this metric may have negative consequences and *root mean square error* is a better choice [105].

Results of evaluation often lead to small differences in the predictive accuracy, which leads some researchers to question the importance of model improvements and the meaningfulness of such results [8]. Perhaps we should pay more attention to the consistency of parameters [42] and focus less on the prediction accuracy. There are attempts to connect the model performance to an actual benefit for the learners, e.g., quantifying the time learners saved in achieving mastery when a better model was used [136]. The issue is similar to examining the relation between the model accuracy and the behavior of a system as a whole in the case of recommender systems [22].

The big issue of the historical data analysis is related to the process of the data collection. The data are often collected using an adaptive approach, e.g., mastery learning. In the case of the mastery learning, learners with high skill drop out earlier which influences the analysis of aggregated data — aggregated learning curves. Previous research shows illustrations of confounded learning [85], analyzes disaggregation of learning curves [82], and proposes mastery-aligned models [48] taking this bias into account. An interaction loop between the data collection and the model evaluation was previously discussed in the context of the “exploration vs. exploitation problem” (multi-armed bandits), with applications for news [62, 128] and advertisement [13, 57] selection. The potential bias in the data collection introduced by fil-
2. Related Work

Figure 2.11: Sample synthetic scenario illustrating a problem with aggregated learning curves. Learners were generated with the probability of error varying from $20 - 100\%$. After each attempt a learner’s probability of error decreases by 5%. If a learner answers correctly, they drop out.

...tering users or splitting data into a train set and a test set was also studied in the context of recommender systems [36, 40, 119].

2.3.2 New Data

Although the current research focuses mainly on the evaluation of learner models (the first layer from [95]), there is also the need of examining a system as a whole. A classic approach to examining the impact of a system on learners’ behavior is to follow the pre-test $\rightarrow$ intervention $\rightarrow$ post-test experimental design [27]. Researchers usually have full control over participants, often in a laboratory. In the test phase, the learners are assessed with respect to a measured variable and in the intervention phase, they use the system for a specified amount of time. There are already experiments of this kind performed online, e.g., in [103] authors found participants using Amazon Mechanical Turk motivating them by money. This approach leads to biased population, but it seems the experiments produce qualitatively and quantitatively
2. Related Work

Figure 2.12: A schema illustrates an interaction loop between a student model and question construction [83].

similar results to university and other online participants [87]. Although this approach is very helpful and brings a lot of interesting observations, it does not scale.

In an online environment we need to periodically deploy new features and test them only on a subset of users before we publish them completely to minimize the negative effect. Doing this kind of A/B testing [53, 54] opens new possibilities to the designers of educational systems [120]. It is possible to collect a large volume of data from realistic environment and make the cycle of hypothesis testing much faster. Using advanced techniques like methods for the multi-armed bandit problem focused on minimizing the negative effect of testing can make A/B testing even more efficient [69]. On the other hand, online testing using real traffic brings new issues which are not present in an offline environment. Real learners have their motivation to use an educational system. They enter and quit the system any time they want. They are often resistant to taking the pre/post test, so it is really hard to evaluate the impact of the system on learning. For this reason, A/B experiments in educational applications [68, 120, 78, 118] tend to focus on engagement rather than learning. Measuring the number of a learner’s interactions or time spent in the system is much easier than quantifying improvement in their performance. This holds especially in the case of an adaptive system where the collected data are highly biased. Unfortunately, engagement metrics does not have to be directly correlated with learning [68].

When online A/B experiments are performed, one should be aware of issues related to metrics behaving differently in short-term and long-

---

4. This process is also known as *canary deployment*. 
term. In the case of experiments running really long, technical issues like detecting the same users via cookies may appear [?]. Previous research describes attempts to predict long-term effects based on short-term experiments [41].

2.3.3 Synthetic Data

An alternative approach to the previously described ones is to use simulated learners to generate synthetic data. Based on real historical data, we can formulate assumptions under which the simulation is performed. This approach opens new directions for the analysis, because we know all the parameters of the simulation (the ground truth). This is already used in the field of recommender systems [40, 112].

The knowledge of the ground truth allows us to investigate issues which could not be analyzed otherwise, like detection of mastery [33], or learning curves [34]. In [33], authors analyze the impact of various values of threshold for mastery learning. Since skills of all simulated learners are known, the authors can easily compare the amount of under-practice and over-practice. It is also possible to examine the relation between performance metrics and how accurate learner models estimate actual knowledge of learners [99, 59]. Previous studies also analyze under-practice and over-practice when Bayesian knowledge tracing or a model using logistic function is used [106].
2. Related Work

Figure 2.13: An example of the simulation process from [83] taking known learners’ skills and item difficulties and examining the impact of a learner model on data collection.
3 System for Computerized Adaptive Practice

This chapter describes a framework running two online applications providing adaptive practice of factual knowledge: www.outlinemaps.org and www.practiceanatomy.com. However, most of its parts are so general that they can also be used in the context of procedural knowledge. The framework is already applied in the system for elementary mathematics www.matmat.cz (only in Czech). The text is based on [88, 108].

We assume that a developer of this kind of an educational system does not have any knowledge about its users and items representing the educational content. This especially means that the system does not rely on the input coming from experts for the given domain and the adaptivity is fully based on the collected data. Using this approach, the system tries to make the provided practice efficient and also show appropriate feedback about the acquired knowledge to a learner.

The educational content is usually built iteratively. In the beginning, the system provides only a small subset of the desired content to prove a concept, other content is added later when there is sufficient online traffic. We want to make the addition of new items as easy as possible, e.g., without the need of recomputation of a learner model.

3.1 Architecture

The provided educational content has to be structured so that a learner can navigate through it and the learner model can use the information about the relation between items to improve its accuracy. Generally, we have a directed acyclic graph where an edge \( x \rightarrow y \) means that \( x \) is a parent of \( y \). The items meant to be practiced are represented as nodes without outcoming edges (leaves). In Outline Maps (see Figure 3.1) we have four types of nodes: (a) maps (e.g., Europe or world); (b) categories (e.g., countries or cities); (c) terms (e.g., Slovakia or Prague); (d) flashcards (e.g., Czech Republic on the map of Europe). Nodes without outcoming edges are always flashcards. This structure is not dependent on the type of educational content and is implemented separately from the descriptive information about items (e.g., names).
3. System for Computerized Adaptive Practice

This allows us to manipulate with the content independently from its type.

![Diagram of structured data used in Outline Maps](image)

Figure 3.1: Part of the structured data used in Outline Maps. When a learner wants to practice European countries, the system finds all leaves reachable from the Europe node and makes an intersection with all leaves reachable from the “countries” node ($\{A, C, E\}$).

Assuming that $[x]$ returns a set of all items without outcoming edges reachable from node $x$, a learner can specify the practiced content using a filter defined as follows (similar to the conjunctive normal form without negations):

\[
\text{filter} := \bigcap_{i=0}^{k} F_i \tag{3.1}
\]
\[
F := \bigcup_{i=0}^{m} f_i \tag{3.2}
\]
\[
f := [x] \tag{3.3}
\]

An example of a valid filter is $[\text{Europe}] \cap ([\text{city}] \cup [\text{country}])$ returning all European cities and countries. It is worth noting that in the real system some of the items and edges are usually marked as disabled, e.g., we found that some of the items misbehave. Similarly,
some of the edges are available to a learner, but some of them are valid only for the purposes of the model, e.g., using the data we built an artificial concept to improve the accuracy of the model.

3.2 Learner Model

The learner model implemented in the system captures the estimation of the prior knowledge — the knowledge a particular learner has before they enter the system, and their current knowledge evolving during the practice session because of learning and forgetting. Outline Maps which serve as the main source of data to evaluate the model are usually used by learners only for one session, so the effect of forgetting is not as strong as learning.

3.2.1 Prior Knowledge

First, we focus on the estimation of prior knowledge. Our aim is to estimate the probability that a learner knows an item $i$ based on the learner’s previous answers to questions about different items and previous answers of other learners to questions about the item $i$. As a simplification (for an easier interpretation of data), we use only the first answer about each item for each learner in this step. Our key assumption is that both learners and practiced items are homogeneous; we assume that we can model learners’ overall prior knowledge in the domain by a one-dimensional parameter. This assumption is reasonable for geography and learners from one country (region), but would not hold for geography and mixed population or for a mix of facts from geography and chemistry. If homogeneity is not satisfied, we can group the learners and facts into homogeneous groups (e.g. learners by their IP address, facts by an expert or by an automatic technique [12]) and then make predictions for each subgroup independently. However, in the case of Outline Maps, most of the learners are from the Czech Republic (roughly 80%).

More specifically, we model the prior knowledge using the Rasch model, i.e. we have a learner parameter $\theta_s$ corresponding to the general knowledge of geography of learner $s$, the item parameter $b_i$ corresponding to the difficulty of the item $i$, and the probability of a correct
3. System for Computerized Adaptive Practice

first answer given by the logistic function \( P(\text{correct}|s, i) = \frac{1}{1 + e^{-(\theta_s - \theta_i)}} \).

As we mention in Subsection 2.1.1, the standard approach to the parameter estimation for the Rasch model is joint maximum likelihood estimation. This is an iterative approach that is slow for large data. It is particularly not suitable for an online application where we need to adjust estimates of parameters continuously. Therefore, we consider applying the Elo rating system in our setting, because it is much more suitable for an online application and the results obtained with simulated data suggest that it leads to similar estimates.

![Figure 3.2: Progression of difficulty estimated by Elo (solid lines) for a few selected items compared to values estimated by JMLE (dashed lines).](image)

The variant of the Elo system implemented in our system provides both fast coarse estimates of item difficulty after a few answers and stability in the long run (see Figure 3.2). It also provides nearly identical estimates for item difficulty as the joint maximum likelihood estimation (Figure 2.4, correlation 0.97). JMLE is a computationally demanding iterative procedure, the Elo system requires a single pass of the data and can easily be used online. Since the estimates of the two methods are nearly identical, we conclude that the Elo system is preferable in our context.

We examine the assumption of a single global skill by computing the skill for independent subsets of items and then checking the correlation between the obtained skills. Figure 3.3 shows the results for two
3. System for Computerized Adaptive Practice

Figure 3.3: Correlation of “subskills” (parameter $\theta_s$) computed for different sets of items. We take only the learners who have at least 10 answers in each “subconcept” into account.

such particular “subskills”. The correlation coefficient for this case and other similar pairs of subskills is around 0.6. Given that there is some intrinsic noise in the data and that the skills are estimated from a limited number of questions, this is quite a high correlation. This suggests that the assumption of a global skill is reasonable. On the other hand, we continuously add new content to Outline Maps including very specific items like districts of the Czech Republic which makes the assumption of one global skill weaker even though the majority of users are Czech.

3.2.2 Current Knowledge

We now turn to the estimation of a learner’s current knowledge, i.e. the knowledge influenced by repeatedly answering questions about
3. System for Computerized Adaptive Practice

a particular item. The input data for this estimation are an estimate of prior knowledge (provided by the above described model) and the history of previous attempts, i.e. the sequence of previous answers (correctness of answers, question types, timing information).

As we mention in Subsection 2.1.2, a disadvantage of performance factor analysis (PFA) is that it does not consider the order of answers (it uses only the summary number of correct and incorrect answers). To address this issue, we propose to combine PFA with some aspects of the Elo system (in the following text we denote this version as PFAE — PFA Elo/Extended). Having $K_{si}$ stand for the estimated knowledge of a learner $s$ about an item $i$, the initial value of $K_{si}$ is provided by the estimation of prior knowledge (difference of the learner’s prior skill and item difficulty):

$$K_{si} = \theta_s - b_i$$  \hspace{1cm} (3.4)

The probability of the correct answer to a question with $n$ options is given by the shifted logistic function:

$$P(\text{correct}|K_{si}, n) = \frac{1}{n} + (1 - \frac{1}{n}) \frac{1}{1 + e^{-K_{si}}}$$  \hspace{1cm} (3.5)

After a question with $n$ options was answered, the estimated knowledge is updated as follows:

$$K_{si} := \begin{cases} K_{si} + \gamma \cdot (1 - P(\text{correct}|K_{si}, n)), & \text{if the answer is correct,} \\ K_{si} + \delta \cdot P(\text{correct}|K_{si}, n), & \text{otherwise.} \end{cases}$$  \hspace{1cm} (3.6)

The estimation can be further improved by taking the timing information into account. If two questions about the same item are asked closely one after another, then it can be expected that the learner will answer the second one correctly because the answer is still in his short term memory. In models based on a logistic function (PFA, PFAE) we can model this effect in the following way: the skill is “locally” increased by $\frac{w}{t}$, where $t$ is the time (in seconds) between attempts and $k$ is a suitable constant (optimal $w = 80$ for our data). Figure 3.5 shows how the time effect function influences the model parameters.
3. SYSTEM FOR COMPUTERIZED ADAPTIVE PRACTICE

Figure 3.4: Sample scenarios showing behavior of the Performance Factor Analysys Elo/Extended model.

It should be possible to further improve the model by a more thorough treatment of forgetting and spacing effects, e.g., by incorporating some aspects of the ACT-R model [100] or using a more sophisticated time effect function [106].

Another piece of useful timing information is the response time. Intuitively, when a correct answer is also fast, it indicates higher knowledge of a particular item. On the other hand, in the case of wrong answers, a fast response may indicate lower knowledge. If a learner takes a lot of time to answer a question, they are probably actively trying to retrieve something from their memory. Figure 3.6 shows the relationship between time spent on answering a question about a particular item and the probability of the correct answer on the next question about the same item. This relationship suggests that the response time could be used to improve the estimation of knowledge. Indeed, even simple modification of the $\gamma$ parameter in the PFA model (by comparison of the response time and mean response time) leads to a slight improvement in predictions.
3. System for Computerized Adaptive Practice

Figure 3.5: Grid search performed to find optimal values of parameters $\gamma$ and $\delta$ for PFA ($\gamma = 1.3, \delta = -0.2$) and PFAE with time effect ($\gamma = 3.5, \delta = -0.9$). Colors correspond to RMSE.

3.3 Instructional Policy

Based on past learner’s performance we want to provide a suitable next question. In the context of our application, the construction of a question consists of several partial decisions: which item to target, which type of question to ask (“Where is X?” vs. “What is the name of this country?”), and which distractors to provide a learner to choose from. Compared to the knowledge estimation, the instructional policy is more difficult to design properly since we do not have a single, clear, easily measurable goal. The overall goal of the instructional policy is quite clear — it is the maximization of users’ learning. However, this goal can hardly be evaluated using historical data. We describe the techniques used to evaluate the system as a whole in Chapter 5 and the evaluation itself is presented in Section 6.1.

3.3.1 Criteria

The question construction process should satisfy several criteria which are partly conflicting. The criteria and their weight may depend on the particular application, the target population, and learners’ goals. We propose the following main criteria.

A question should be constructed based on the estimated difficulty of items. As Figure 2.3 shows, from the testing perspective, it is optimal to use questions with the expected probability of a correct answer
reaching 50%, because such questions provide most information about the learners’ knowledge. However, a 50% success rate is rather low and for most students it would probably decrease motivation. Thus, in our setting (adaptive practice), it is more common to aim for a higher success rate, e.g. a 75% success rate [46]. Another important issue is the repetition of questions. This aspect should be governed by research about spacing effects [25, 100]. In particular, it is not sensible to repeat questions about the same item too early. It may also be welcome to have a variability of question types. Different question types are useful mainly as a tool for tuning the difficulty of questions, but even if this is not necessary, the variability of question types may be meaningful criteria in itself, since it improves user experience if used correctly.

3.3.2 Item Selection

We start by choosing a target item, which is the correct answer to a constructed question. As a general approach we have settled on a linear scoring approach. For each relevant attribute, we consider a scoring function that expresses the desirability of a given item with respect to this attribute. These scoring functions are combined using a weighted sum; the item with the highest total score is selected as a
Figure 3.7: The desired contribution of different criteria to the selection of a target item.

target. This approach is flexible and thanks to the choice of attributes and their weights it can be adjusted for a particular application. We take the following attributes into consideration:

1. the probability that the learner knows the item,
2. time since the last question about the item,
3. the number of questions already answered by the learner about the item.

Figure 3.7 illustrates the general shapes resulting from our choice of scoring functions for these attributes. Further, we specify formulas that define these shapes using simple mathematical functions.

The first function takes into account the relation between the estimated probability of a correct answer \( P_{est} \) and the target success rate \( P_{target} \). Suppose that our goal is to ask a question where the learner has a 75% chance of a correct answer. The distance from the probability for the difficult items (nearly a 0% chance of the correct answer) is higher than for easy ones (almost 100%), so it is necessary to normalize it:

\[
S_{prob}(P_{est}, P_{target}) = \begin{cases} 
  \frac{P_{est}}{P_{target}} & \text{if } P_{target} \geq P_{est}, \\
  \frac{1-P_{est}}{1-P_{target}} & \text{if } P_{target} < P_{est}.
\end{cases} \tag{3.7}
\]

The second scoring function penalizes items according to the time elapsed since the last question about the same item — we do not
3. System for Computerized Adaptive Practice

want to repeat it when it is still in the short term memory. We use the following function, where the variable $t$ stands for time in seconds and $t_{\text{max}}$ is a threshold parameter also in seconds:

\[
S_{\text{time}}(t) = \begin{cases} 
1 & \text{if } t \geq t_{\text{max}}, \\
\log t / \log t_{\text{max}} & \text{otherwise.}
\end{cases}
\] (3.8)

Using only the above mentioned attributes, the system would ask questions for only a limited pool of items. To induce the system to ask questions about new items we introduce a third scoring function that uses the total number $n$ of questions answered by the learner for the given item:

\[
S_{\text{count}}(n) = \frac{1}{\sqrt{1 + n}}
\] (3.9)

The total score is given as a weighted sum of individual scores, with the weights being set manually based on our experiences with the prototype version of the system: $W_{\text{prob}} = 10$, $W_{\text{count}} = 5$, $W_{\text{time}} = 5$.

To strengthen the adaptivity of system behavior we propose an additional dynamic adjustment of target difficulty. With this mechanism, the target probability is modified depending on a learner’s recent performance (as a measure of recent performance we use the success rate on the last ten questions). Our system poses easier questions to less successful learners and more difficult questions to more successful ones; a specific function for a transformation of the target rate is depicted in Figure 3.8 (right). Note that by using this mechanism we also indirectly correct a potential estimation bias given by the used learner model.

3.3.3 Distractors

Even though the difficulty of the item the system asks about is already taken into account by the first scoring function, it is possible that the estimated difficulty of the selected candidate does not match the target difficulty, e.g., most European countries are easy for the users of Outline Maps (see Figure 3.8, left). Unfortunately, there is nothing that can be done in the case of too easy items. In the case of a more
3. System for Computerized Adaptive Practice

Figure 3.8: Histogram for the probability of the correct answer on open questions in the case of European countries and illustration of the scoring function for difficulty (left). Illustration of the mechanism of target difficulty adjustment based on a learner’s past answers (right).

difficult candidate, the system can use a multiple choice question to give the student a chance to guess the correct answer. For a multiple choice question, the probability of a correct answer is the combination of the probability of guessing the answer \( P_{\text{guess}} \) and knowing the item \( P_{\text{est}} \), see Equation 3.10. This is, of course, a simplification since a multiple choice question can also be answered by ruling out distractor options. But if the distractors are well chosen, this simplification is reasonable.

\[
P_{\text{success}} = P_{\text{guess}} + (1 - P_{\text{guess}}) \cdot P_{\text{est}} \quad (3.10)
\]

As the system goal is to get \( P_{\text{success}} \) close to \( P_{\text{target}} \) would like to make \( P_{\text{guess}} \) close to:

\[
G = \frac{P_{\text{target}} - P_{\text{est}}}{1 - P_{\text{est}}} \quad (3.11)
\]

For \( G \leq 0 \), we use open questions (no options), otherwise we use \( n \) closest to \( \frac{1}{G} \) as the number of options. For principal reasons the minimal possible value of \( n \) is 2, for practical reasons there is also an upper bound for \( n \) (e.g., 6), since a large number of options would be confusing.

When using multiple choice questions, we also need to choose the distractors. Unlike other systems for practice dealing with text [77, 79],
3. SYSTEM FOR COMPUTERIZED ADAPTIVE PRACTICE

Figure 3.9: Misanswers for Albania within European countries (left) and illustration of adaptiveness in number of options (right).

we work with well-structured data, so the selection of distractors is easier. Although the choice of distractors can be based on domain information, e.g., geographically close countries or countries with similar names, we choose distractors using historical data. We take items most commonly mistaken for the target item in open questions, and select from them randomly. The random choice is weighted by the frequency of mistakes with the given item — the distribution of wrong answers is typically highly skewed, see Figure 3.9 (left). For example, Albania is most often confused with Macedonia (19%), Montenegro (16%), Kosovo (16%), Moldova (9%), and Serbia (8%).

Having the set $W_i$ containing all wrong answers on questions asking for an item $i$ and $W_i^\infty = \{a \in W_i \mid |\text{distractors}(a)| = \infty\}$ consisting of all answers on open questions (in real systems there may be a finite number of distractors even in open questions), then the probability we choose distractor $d$ for a question asking for item $i \neq d$ is computed as expressed in Equation 3.12.

$$P_i(distractor = d) = \frac{|\{a \in W_i^\infty \mid answered(a) = d\}|}{|W_i^\infty|} \quad (3.12)$$

This approach relies on the fact that a system provides not only multiple-choice questions, but also open ones. Alternatively, when the data with answers on open questions are not available, the probability can be computed as described in Equation 3.13.
3. System for Computerized Adaptive Practice

\[ P_i(distractor = d) = \frac{\{a \in W^i \mid answered(a) = d\}}{\{a \in W^i \mid d \in distractors(a)\}} \] (3.13)

3.4 Usage

The described framework is deployed within two applications providing practice of factual knowledge — Outline Maps and Practice Anatomy. The applications are very similar to each other, but they differ in the target audience.

3.4.1 Outline Maps

The application Outline Maps is available at outlinemaps.org and focuses on learning the location of countries, cities, rivers, etc. It provides roughly 100 contexts (e.g., European countries) containing almost 3000 items to practice. Figure 3.10 shows sample questions shown to learners. So far, we have collected more than 40000000 answers and we collect 900000 answers per month on average, see Figure 1.1 (left). The application does not collect any personal information about learners like age or gender. It allows learners to sign up to keep their practice history, but this functionality is used only by 2% of them. The system is available in many languages (Czech, English, German, Russian, Slovak, and Spanish), but most users are from the Czech Republic (85%) and Slovakia (10%).

We do not collect any detailed information about learners, but we assume the application is widely used at Czech schools. Our assumption is based on the parallel access of multiple users from the same IP address (or similar location) at the same time. We observe this kind of behavior mainly in the morning, which corresponds to teaching hours, see Figure 3.11. We estimate that the usage directly in schools corresponds to roughly 15% of answers.

When a learner enters the system, they choose a context to practice (e.g., African rivers). These contexts differ in their difficulty (prior knowledge) and the number of items available (from 10 to 120), see Figure 3.12. Some contexts are much more popular than others, so the distribution of answers is highly uneven. Almost 50% of the data
3. System for Computerized Adaptive Practice

is collected for 4 most practiced contexts. During our analysis, we sometimes group contexts by some of their features and analyze these groups separately.

3.4.2 Practice Anatomy

Practice Anatomy is a similar application to the previous one. It provides practice of anatomical structures and is publicly available at practiceanatomy.com. It was launched two years after Outline Maps and its target audience is much smaller — mostly students of medicine. The questions are not focused only on the location of anatomical terms in pictures, but also on the relations between the terms, e.g. trapezius is supplied by transverse cervical artery. Figure 2.1 shows an example of such a question.

The application contains roughly 250 pictures taken from [43] with more than 2,500 items and almost 1,400 relation items. So far, we have collected almost 1,500,000 answers and we collect 75,000 answers per month, see Figure 1.1 (right). There are more or less 10% registered learners. The system is available with a Czech and English user interface and Czech, Latin and English terminology. Again, most learners are from the Czech Republic (84%) and Slovakia (11%).
3. System for Computerized Adaptive Practice

Figure 3.11: The distribution of answers over a day. When we detect at least 10 users accessing the system from the same IP address, we mark their answers as “in-school”.

Figure 3.12: Size vs. difficulty for 20 most practiced contexts. Difficulty is defined as an average error rate on the first question in the context. The size of circles is proportional to the number of answers for the particular context.
4 Adaptivity and Data Collection

Before we focus on the evaluation of an adaptive educational system as a whole, we provide an analysis of the impact of adaptivity on the data collection which also directly affects the evaluation of learner models. In the case of learner modeling, we mostly evaluate models based on the quality of their predictions of learners’ answers as expressed by a performance metric. The results of the evaluation often lead to small differences in predictive accuracy, which leads some researchers to question the importance of model improvements and meaningfulness of such results [8]. We explore the impact and meaning of small differences in predictive accuracy with the use of synthetic data. For our discussion and experiments in this section, we use a single performance metric — Root Mean Square Error (RMSE), which is a common choice (for rationale and overview of other possible metrics see [105]). The studied questions and the overall approach are not specific to this metric.

As we discuss in Subsection 2.3.3, simulated learners provide a good way to study methodological issues in learner modeling. When we work with real data, we can use only proxy methods (e.g., metrics like RMSE) to evaluate the quality of models. With synthetic data we know the “ground truth” so we can study the link between metrics and the true quality of models. This enables us to obtain interesting insights which may be useful for the interpretation of results over real data and for devising future experiments. We use a simple setting for our experiments based on an abstraction of our system presented in Chapter 3. We simulate an adaptive instructional policy assuming items with normally distributed difficulties, learners with normally distributed skills, and probability of the correct answer given by a logistic function of the difference between skill and difficulty. The text of this chapter is based on [83, 109]

4.1 Methodology

Figure 2.13 presents the overall setting of our experiments. The system asks learners about items, answers are dichotomous (correct/incorrect), a learner does not answer an item more than once. The system tries to
present items of suitable difficulty. In the evaluation, we study both the prediction accuracy of models and also the sets of used items. This setting is closely related to item response theory and computerized adaptive testing, specifically to simulated experiments with the Elo-type algorithm reported by Doebler et al. [28].

4.1.1 Simulated Learners and Items

We consider a set of simulated learners and simulated items. To generate learners’ answers we use a logistic function: 

\[
P(\text{correct}|\theta_s, d_i) = \frac{1}{1 + e^{-(\theta_s - d_i)}},
\]

where \(\theta_s\) is the skill of a learner \(s\) and \(d_i\) is difficulty of an item \(i\). To make the simulated scenarios more interesting, we also consider multiple knowledge components. Items are divided into disjoint knowledge components and learners have a different skill for each knowledge component. Learners’ skills and item difficulties are sampled from a normal distribution. Skills for individual knowledge components are independent from one another.

4.1.2 Item Selection Algorithm

The item selection algorithm has as the parameter of a target success rate \(t\). It repeatedly presents items to a (simulated) learner, in each step it selects the item with the best score with respect to the distance of the predicted probability of the correct answer \(p\) and the target rate \(t\) (illustrated by a gray dashed line in Figure 4.2). If there are multiple items with the same score, the algorithm randomly selects one of them.

4.1.3 Learner Models

Predictions used by the item selection algorithm are provided by a learner model. For comparison we consider several simple learner models:

- Optimal model — Predicts the exact probability that is used to generate the answer (i.e., a “cheating” model that has access to the ground truth — a learner’s skill and an item difficulty).

- Optimal with noise — Optimal model with added (Gaussian) noise to the difference \(\theta_s - d_i\) (before we apply logistic function).
4. Adaptivity and Data Collection

- Constant model — For all learners and items it provides the same prediction (i.e., with this model the item selection algorithm selects items randomly).

- Naive model — Predicts the average accuracy for each item.

- Elo model — The Elo rating system with a single skill. The used model corresponds to the version of the system as described in Subsection 2.1.1 (with a slightly modified uncertainty function).

- Elo concepts — The Elo system with multiple skills with correct mapping of items to knowledge components.

- Elo wrong concepts — The Elo system with multiple skills with a wrong mapping of items to knowledge components. The wrong mapping is the same as the correct one, but 50 (randomly chosen) items are classified incorrectly.

4.1.4 Data

We generated 5000 learners and 200 items. The items are divided into two knowledge components, each learner has two skills corresponding to the knowledge components and each item has a difficulty. Both skills and difficulties were sampled from standard normal distribution (the data collected from the geography application suggest that these parameters are approximately normally distributed). The number of items in a practice session is set to 50.

4.2 Impact on the Practiced Content

Our first set of experiments studies the differences in the behavior of the simulated system for different models. To evaluate the impact of the model, we compare sets of items selected by the item selection algorithm using different models. We make the assumption that the algorithm for item selection using the optimal model also generates optimal practice for learners. For each learner we simulate a practice of 50 items (each item is practiced at most once by each learner). To compare the set of practiced items between those generated by the optimal
4. Adaptivity and Data Collection

model and other models, we look at the size of the intersection. We assume that a bigger intersection with the set of practiced items using the optimal model indicates better practice. Since the intersection is computed per learner, we take the mean. This is, of course, only a simplified measure of item quality. It is possible that an alternative model selects a completely different set of items (i.e., the intersection with the optimal set is empty) and yet the items are very similar and their pedagogical contribution is nearly the same. However, for the current work this is not probable since we are choosing 50 items from a pool of only 200 items.

The optimal model with noise allows us to easily manipulate differences in the predictive accuracy and study its impact on the system behavior. The experiment reported on the left side of Figure 4.1 shows both the predictive accuracy (measured by RMSE) and the impact on system behavior (measured by the size of the intersection with the optimal practiced set as described above) depending on the size of noise (we use Gaussian noise with a specified standard deviation). The impact on noise on RMSE is approximately quadratic and has a slow rise — this is a direct consequence of the quadratic nature of the metric. The impact on used items is, however, approximately linear and rather steep. The most interesting part is for noise values in the interval \([0, 0.1]\). In this interval the rise in RMSE values is very small and unstable, but the impact on used items is already high.

The right side of Figure 4.1 shows the distribution of the number of answers per item when different models for item selection are used.

Figure 4.1: Size of the intersection with the optimal practiced set of items and RMSE depending on Guassian noise in optimal model (left). Distribution of answers over the items based on the given model (right).
The used models have a similar predictive accuracy (specific values depend on what data we use for their evaluation, as discussed below in Section 4.3), yet the used model can dramatically change the form of the collected data. When we use the optimal model, the collected data set covers most of the items from the item pool almost equally. In the case of worse models, the use of items is skewed (some items are used much more frequently than others). An obvious exception is the constant model for which the practice is completely random. The size of the intersection with the optimal practiced set for these models is — Constant: 12.5; Elo: 24.2; Elo, Concepts: 30.4; Elo, Concepts (wrong): 28.5; Naive: 12.0. Figure 4.2 presents a distribution of answers according to the true probability of their correctness (given by the optimal model). Again, there is a huge difference among the given models, especially between simple models and those based on Elo.

![Figure 4.2: Distribution of answers according to the true probability of the correct answer. The gray dashed line stands for the score function used by the algorithm for item selection.](image)

### 4.3 Feedback between Data Collection and Evaluation

To study the feedback between the used learner model and the collected data, we perform the following experiment: We choose one model and use it as an input for an adaptive choice of items. At the same time, we let all the other models do predictions as well and log answers together with all predictions. Figure 4.3 shows the resulting RMSE for each model in individual runs (data collected using a specific model). The figure shows several interesting results. When
4. Adaptivity and Data Collection

Figure 4.3: RMSE comparison over data collected using different learner models.

As the above presented analysis shows, different models lead to a very different choice of items and consequently to different learners’ experience. The reason for small differences in RMSE is not the similarity between models, but the characteristics of data (“good choice of suitable items”), which make predictions difficult and even a naive predictor comparatively good. Another observation concerns the comparison between the “Elo concepts” and “Elo concepts (wrong)” models. When data are collected by the “Elo concepts (wrong)” model, these two models achieve nearly the same performance, i.e., models seem to be of the same quality. But the other cases show that the “Elo concepts” model is better (and in fact it is by construction a better learner model).
5 Evaluation Methods

Current online educational systems like Khan academy, Duolingo, or edX are typically online and open to be used by anybody, anywhere. The learner population is typically very heterogeneous, often comprising students using the system compulsory within a classroom, learners using the system voluntarily as a part of their preparation for an exam, adult learners who want to refresh their knowledge, and also people who just stumbled upon the system while browsing the Internet or following a suggestion of a friend on a social network. The motivation of learners to use the system thus widely differs, the distribution of time in the system is typically highly skewed (most learners use the system for only a short time) and the departure from the system is not random. This creates attrition bias, which complicates the evaluation of learning within the system. Adaptive behavior of systems is a further obstacle in the evaluation — each learner proceeds through the system using different learning materials and questions and it is not easy to use these adaptively constructed questions for evaluation of learning gains. As discussed in Chapter 4, the behavior can even create a feedback loop between a learner model and the collection of data for evaluation.

Another aspect that makes the evaluation of online educational systems difficult is the need to optimize both learning and engagement. Unfortunately, the classical approaches to evaluation of educational interventions do not consider engagement. However, in the context of online systems, engagement is a key factor — if learners do not actually use an educational system, they cannot learn from it. Standard methods of evaluation using pre-test/post-test setting are thus not applicable to online systems — such methods could be performed, but would lack ecological validity. An advantage of online systems is that we can easily collect extensive data about learners’ actions. On the other hand, the collected data are only a noisy proxy for what we are really interested in, because learning and engagement are latent constructs that we cannot measure directly. Using large scale data, it is possible to obtain interesting and valid evaluation results, but we have to be careful about potential biases and aggregation issues.
5. Evaluation Methods

The goal of this chapter is to study methodological evaluation issues for online adaptive educational systems. Our main aim is the description of a coherent framework for the evaluation of online adaptive educational systems and the discussion of general methodological issues. To measure learning we utilize learning curves [75] which map learning during the usage of the system. In the case of adaptive systems, a straightforward application of learning curves would be misleading. To meaningfully compare different variants of a system, we use periodic “reference questions” that are constructed fully randomly. This allows us to compare different experimental conditions, but the interpretation of learning curves is still complicated by issues with aggregation and attrition bias as described in Subsection 2.3.1. We analyze the attrition bias present in our data and discuss techniques for dealing with it. To measure engagement, we utilize techniques from survival analysis, particularly basic survival curves.

5.1 Engagement

Now we describe our approach to evaluate systems for adaptive practice. First, we try to measure the engagement of learners. As the systems are typically online and open to anyone, their users vary in many aspects including their motivation for using the system. Since the engagement is a latent feature that cannot be measured directly, it is commonly studied through proxy measures, particularly “the time spent using the system”. We also use this approach, but we found that even this simple proxy measure is not straightforward to use.

5.1.1 Survival Analysis

The time spent using a system is typically very skewed and thus it is not suitable to compare conditions using averages (or even other measures of central tendency like the median). Therefore, it is useful to employ techniques from the survival analysis. Survival analysis deals with questions like “What proportion of population will survive past a given time?”, typically in the context of medical data. Once we interpret “survival” as “active usage of a system”, it is directly relevant to the evaluation of educational systems.
Figure 5.1: Aggregated survival curves using data from Outline Maps. The right graph shows the proportion of learners with respect to the number of practiced attempts. Big drops are related to the user interface of the system — the system shows a summary after each set of 10 questions. The left graph shows how much time learners spend within the system.

Figure 5.1 shows examples of survival curves, i.e., the proportion of learner population “surviving” over a specific time. In educational systems, there are two natural measures of “time” — the number of answers or the elapsed time (in seconds). The number of answers is a commonly used characteristic of learners’ engagement. It is also typically used in the analysis of learning curves and thus using it for measuring engagement makes the analysis coherent. One disadvantage of this measure is that due to its discrete nature, a survival curve can contain sharp steps due to specific features of the system. In the case of Outline Maps, a feedback message is presented after 10 questions. Since these messages interrupt learners’ practice, they create a natural point to leave the system. This leads to steps of size 10 in the survival curve (Figure 5.1 left).

If we use the elapsed time instead, the curve is smoother. In the computation of elapsed time, it is necessary to treat outliers (e.g., use a “trim-off” on time), because sometimes learners are interrupted in their use of the educational system and answer a question after an exceedingly long time. Untreated use of this time would confound the analysis. The decision whether to use the elapsed time or the number of attempts may seem like a technical detail, but as we describe in Section 6.2, it can highly influence the interpretation of experiment results.
5. Evaluation Methods

5.1.2 Short-term vs. Long-term Engagement

In our applications we observe differences between short term and long term engagement. There are techniques working quite well to increase short term engagement, but having negative impact on long term engagement. Similarly, we observe that engagement and learning may not be aligned, particularly when we look at short term engagement (which is a tempting measure, as it is the easiest to collect data for). Both these trends are illustrated in Section 6.2.

Previous research and our analysis of the data from several systems suggest that survival data from educational systems can be well fitted by standard distributions from survival analysis, particularly lognormal and Weibull distribution [29, 90]. Nevertheless, we suggest not reporting the fitted distributions, but using survival rates at specific points on the curve instead. The survival rates are easier to interpret than parameters of the fitted functions and they provide similar insights.

Specifically, we have chosen to report the following survival rates in our experiments:

- short-term survival: survival rate for 10 answers and 1 minute (typically around 85% for Outline Maps),
- long-term survival: survival rate for 100 answers and 10 minutes (typically around 30% for Outline Maps).

As an alternative measure of long term engagement, we consider the probability of returning to the system. As a “return” we consider the usage of the system after a delay of more than 10 hours (the specific duration of a delay is not important, different values lead to very similar results).

5.2 Learning

Measuring learning is even more difficult than measuring engagement, particularly for online systems where we cannot use the basic pre-test/post-test method for the evaluation of educational interventions. Moreover, the adaptive behavior of a system presents additional methodological complications. We present an evaluation methodology based on the use of reference questions and learning curves.
5. Evaluation Methods

5.2.1 Reference Questions

Our system collects data about the correctness of answers, so ideally we would like to see a decreasing error rate with respect to time. This time is often measured by the number of a learner’s interactions within the system. Unfortunately, the evaluation of learning cannot be simply based on the learners’ achieved error rate, since this error rate is by definition heavily influenced by the used experimental conditions. Consider two systems: the first one provides content with increasing difficulty, the second one presents the content randomly. In the case of the first system, it could easily happen that the achieved error rate would be constant (a learner’s skill increases with the same speed as the difficulty), or even increasing (the difficulty increases more than a learner’s skill). In the second case, the achieved error rate would probably decrease, because the difficulty does not change over time and a learner’s skill increases. It would be wrong to conclude that the second system leads to better learning.

One approach to solve the issue of adaptively constructed content would be to use model based detectors of learning, i.e., to fit a learner model (e.g., Bayesian knowledge tracing or Performance factor analysis) to data and interpret the model parameters as an evidence of learning. Such results would, however, be influenced by violations of simplifying assumptions of learner models and by feedback loops between data collections and learner models. E.g., by changing the difficulty of practice we get data providing a varied amount of information about a learner (see Figure 2.3), so the resulting estimates differ in their quality. The impact of data collection on interpretation and evaluation of learner models is described in Chapter 4.

To minimize bias during the evaluation of learning within the adaptive system we propose to collect data independently of the studied condition. In our case, we use “reference questions”. The reference questions are open questions about a randomly chosen item from a particular context (independently of the experimental condition), e.g., a question shown in Figure 3.10 (right). The questions are used periodically — every 10th question in a particular context is a reference question. The first reference question is the first question shown to a learner within the context, i.e., before the adaptive algorithm has any chance to influence the practice for the given context. Figure 5.2
5. Evaluation Methods

![Diagram illustrating reference answers collection](image)

Figure 5.2: An illustration of the mechanism by which the system collects reference answers. Light dots stand for common questions, dark dots for reference questions. When a particular user starts practicing a different context, the collection of reference answers restarts.

illustrates the idea. A similar approach based on random items has been used for evaluation previously, for example in [66, 67]. A limitation of this approach is that we are not able to compare the learning of individual learners, because their answers to reference questions are probably about different items.

5.2.2 Learning Curves

By using reference answers (answers to the reference questions described above), we construct a learning curve [75]. We put together reference answers from all the available contexts and compute an average error rate preserving their ordering within contexts, e.g., we put together all the first reference answers from all users and contexts to get the first point of the learning curve.

By default, we do not filter any data and users may quit their practice on their own, so for the first point of the learning curve we have more answers than for the second one and so on. An alternative approach is to analyze data from users who have at least a particular number of reference answers. In this case all points of the learning curve stand for the same number of reference answers. Unfortunately, by this filtering we get only a specific group of users. We can also transform the learning curve, so its last point stands for learners’ last
reference answers. In this case, we get the information about how learners learn before they quit the practice (as opposed to learning in the beginning of the practice). The last point of this learning curve is computed from the largest number of reference answers. Each of the described learning curves has some disadvantages, but in our previous research [90] all of them led to similar results.

In accordance with the previous research [75], we assume that the learning curve corresponds to the power law, i.e., the error rate can be expressed as $ax^{-k}$, where $x$ is the number of the reference answer, $a$ is the initial error rate, and $k$ is the learning rate. To get an idea of how significant the differences between conditions are, we compute a 95% confidence intervals using bootstrapping (resampling learners’ series of reference answers). The advantage of this method is that it can be used easily for both coarse and fitted learning curves.
6 Case Studies

Using the evaluation methods described in Chapter 5, we performed five experiments running on Outline Maps. These experiments are standard randomized controlled trials: we randomly assigned new users to one of several experimental conditions and then analyzed the collected data. Note that the experiments have been performed on common users of our system, they have not been externally motivated by us in any way (e.g., money) and we do not have detailed information about them. Learners present in our system before a particular experiment are provided with a standard version of the system and are not taken into account for further analysis.

The performed experiments are:

- **Adaptive vs. random**: a comparison of an adaptive vs. random choice of a stem and distractors.
- **Question difficulty**: a comparison of the adaptive algorithm with different settings of the target difficulty parameter.
- **Choice of distractors**: a comparison of different methods for the choice of distractors in multiple-choice questions (competitive vs. adaptive).
- **Number of options**: a comparison of different settings of the maximal number of options used in multiple-choice questions.
- **Difficulty adjustment**: we allowed the learner to manually adjust the target difficulty and compared this setting to the placebo and control group.

Table 6.1 provides basic statistics about the extent of experiments. To make our research reproducible, we make all analyzed data sets available\(^1\), together with a brief description and the terms of use. The overview of results for individual experiments is given in Figure 6.1, except for Choice of Distractors and Difficulty Adjustment experiments. For these experiments the overview figure does not show any difference among the experimental conditions. The columns correspond to

---

\(^1\) [http://data.outlinemaps.org](http://data.outlinemaps.org)
6. Case Studies

Table 6.1: The overview of performed experiments.

<table>
<thead>
<tr>
<th>Experiment</th>
<th>Answers (millions)</th>
<th>Time span</th>
</tr>
</thead>
<tbody>
<tr>
<td>Adaptive vs. random</td>
<td>1.4</td>
<td>Aug 2015 – Oct 2015</td>
</tr>
<tr>
<td>Question difficulty</td>
<td>3.3</td>
<td>Nov 2015 – Jan 2016</td>
</tr>
<tr>
<td>Choice of distractors</td>
<td>2.6</td>
<td>Mar 2016 – May 2016</td>
</tr>
<tr>
<td>Number of options</td>
<td>2.1</td>
<td>May 2016 – July 2016</td>
</tr>
<tr>
<td>Difficulty adjustment</td>
<td>8.2</td>
<td>Oct 2016 – Jan 2017</td>
</tr>
</tbody>
</table>

the experiments and rows to several aspects of learners’ behavior. In the first row we fit a learning curve as described in Subsection 5.2.2 and compare the value of the parameter $k$ over conditions. Secondly, we analyze short-term survival looking at the percentage of learners who have at least 10 answers and the percentage of learners in the system for at least one minute. In the last row, we focus on long-term survival. Again, we look at the number of answers (100) and the time spent in the system (10 minutes).

Note that the results are comparable only across experimental conditions within a single experiment. The results are not comparable across experiments. This is caused by different users’ behavior during different time periods (e.g., holidays), changes of user interface and system content between experiments.

6.1 Adaptive vs. Random

In the first experiment, we evaluate our approach to adaptivity. Since the question construction has 2 steps (the selection of a stem and distractors), we compare a fully adaptive practice to partially and fully random versions. This leads to four versions of the algorithm for question construction: adaptive-adaptive (A-A), adaptive-random (A-R), random-adaptive (R-A), and random-random (R-R).

The adaptive version of stem selection (A-* ) computes a score for each item taking into account its difficulty, the number of a learner’s answers and time elapsed since the last learner’s answer. The random version of an item selection (R-* ) picks the stem randomly. As for the construction of options, the adaptive version (*-A) computes a number
Case Studies

Figure 6.1: The overall summary of the performed experiments except for the Choice of Distractors and Difficulty Adjustment experiment. The first row (A) shows the fitted value of parameter $k$ of power law mentioned in Subsection 5.2.2. The second row (B) presents the percentage of learners surviving at least 10 answers (1 minute). The last row (C) shows survival for 100 answers (10 minutes). The error bars show 95% confidence intervals.

of options to make the question as close to the target difficulty as it is possible and uses the most competitive distractors. The random version (*-R) chooses the number of options and options themselves fully randomly. The target error rate was set to 20% for this experiment.

To provide better intuition behind the used experimental conditions, we discuss a specific example of question construction for a new learner choosing to practice African countries. The first construction step in A-* condition prefers Algeria (estimated error rate 25%) to Madagascar (6% — too easy) or Zimbabwe (55% — too difficult), whereas R-* condition selects countries with uniform probability. In the second step, if R-A has Zimbabwe from the first step, it reduces its difficulty by selecting only 2 options (Zimbabwe and 1 competitive distractor — Zambia), whereas A-A has Algeria from the first
step, Algeria has an appropriate difficulty, and the algorithm thus selects either an open question or a high number of options (6) with competitive distractors (Egypt, Libya, Dem. Rep. Congo, South Sudan, and Sudan). Regardless of whether the first step selected Algeria, Zimbabwe, or another country, both *-R conditions select a random number of options and random distractors (e.g., 4-options question with distractors Morocco, Tanzania, and Ghana).

6.1.1 Data Analysis

Although it was not an original intention, the experimental conditions differ in learners’ error rate (Figure 6.2). Conditions A-R and R-R have a lower overall error rate because they are more likely to use fewer options. R-* conditions exhibit a decline in the error rate throughout the use of the system (Figure 6.2, right), whereas A-* conditions by definition keep the error rate more constant. Especially the A-A error rate is distributed closely around the target error rate (Figure 6.2, left). On the other hand, R-R error rate distribution is skewed towards 0%.

![Figure 6.2: Comparison of error rates for the four conditions. Histogram of the overall error rate (left) and error rate as a function of the number of attempts (right).](image)

Error rate is influenced by the average item difficulty which varies largely among different contexts. For the 10 most practiced contexts, the error rate on the first reference question is between 30% and 80% (Figure 3.12). The relation of the average item difficulty and the error rate is different for different experimental conditions in different contexts. In R-* conditions the error rate is highly influenced by the average item difficulty of the context. A-* conditions can decrease the error rate by asking multiple-choice questions with fewer distractors.
6. Case Studies

Figure 6.3: The average usage of options by an attempt for the experiment conditions. “O” stands for open questions and darker color for more frequent usage. Reference attempts are excluded. E.g., the top left cell stands for the fact that roughly 8% of learners answering the first non-reference question see this question with 2 options.

However, when every item in a given context has prediction below the target error rate, then there is no way to increase the error rate.

Figure 6.3 shows the distribution of the number of options for the tested conditions. The conditions A-A and R-A use our algorithm for the construction of distractors, so they construct them adaptively. An interesting observation is that they lead to the dominance of questions with two options, six options, and open questions, even though it is not their direct objective. The conditions A-R and R-R construct questions randomly, so one could think that the numbers of options are distributed uniformly in the collected data. However, there is an asymmetry in the available questions (the system never asks open questions about the name of a country), and the used implementation reconciles this asymmetry in such a way that the system uses more questions with 6 options, see dark columns for 6 options and light columns for open question in the Figure 6.3.

6.1.2 Attrition Bias

As we have no control over learners, they spend different time within the system and also different time practicing different contexts. This
6. Case Studies

can be affected by the engagement of studied conditions, but also by learners’ background (e.g., students preparing for exams vs. random visitors), or by a combination of both. Previous research [82, 85] described the attrition bias as a phenomenon, where we have data from all learners in the beginning of the practice, but later we have only a subset of them. This affects the analysis of learning based on learning curves. In combination with mastery learning (system stops a learner’s current practice when it detects mastery), the attrition bias can lead to a very flat learning curve because learners with high knowledge leave and the points in the second half of the curve are based on learners with poor prior knowledge.

If the attrition was the same in all conditions, we could probably ignore it because our system does not work with mastery learning and learning itself is detectable in our analysis (the learning curves are not flat). However, we found that learners staying in the system may be of a different kind in different conditions. If this is true, looking at the learning curves for different conditions, we actually compare not only the conditions, but also different kinds of learners. This can highly influence the interpretation of results.

Since the first reference question asked by our system is constructed before a particular condition has any effect on the practice, we can look at the error rate of the first reference answer and use it as a metric for prior knowledge. We compare the prior knowledge of learners with at least \( n \) answers for different conditions. As Figure 6.4 shows, later points of the learning curve for the R-A condition are computed for learners with higher prior knowledge than the learners in the A-R condition.

To deal with the detected bias, we use a simple technique based on propensity score matching [16] to balance data from several conditions. The method works as follows:

1. Repeatedly go through conditions.

2. From a given condition: randomly select a series \( X \) of reference answers of a particular learner for a particular context.

3. From all other conditions: randomly select a series for which the first reference answer has the same correctness as the first answer of \( X \) and the series has the same length as \( X \).
Using the described technique, we obtain data for which the curves in Figure 6.4 are aligned, i.e., there is no difference in the prior knowledge of learners in different conditions. When we use such data sets to perform the same analysis based on the learning curves and compare the experimental conditions, we get results which are not significantly different from the above presented analysis of untreated data. Thus, although the attrition bias is present in our data, we assume it does not influence the interpretation of the performed experiments.

### 6.1.3 Results

The results are summarized in Figure 6.1 (the first column). With respect to engagement, there is a difference between short-term and long-term engagement. Although \( A-R \) condition is better than \( R-A \) in the case of the first practice set (10 answers), in the case of 100 answers it is the opposite. The adaptivity in the selection of the stem positively influences short-term engagement, but for long-term engagement the construction of appropriate distractors is more important. When we look at the time users spend in the system, differences are much
smaller. In all cases, A-A is a good choice when we optimize a learner’s engagement.

With respect to learning, the most important part of the adaptivity is the second step — choosing an appropriate number of options and competitive distractors (*-A are obvious winners). The result is quite surprising because the construction of distractors depends on the learner model only slightly. When we look at differences between the A-A and R-A conditions, the R-A condition is even slightly better (although the difference is not significant), i.e., it seems that with respect to learning the adaptive choice of stems could be improved.

The use of error rate as a measure of learning is a standard (and in our setting quite natural) choice. It is, however, not the only possible choice. We can also take other aspects of learner behaviour into account, e.g., response time. Our previous research shows that the time learners spend on answering questions relates to their future success, see Figure 3.6. Even though the system does not motivate learners to have low response time (in fact it does not even indicate in any way that the response time is measured), we observe an improvement of response time and systematic differences between studied conditions (Figure 6.5). With respect to this measure, we get the best results for the R-A condition.

Figure 6.5: Learning curve for response times (the reported time for each attempt is the median of corresponding times).
6.2 Question Difficulty

As we describe in Section 2.2, our system uses a target error rate and adaptively constructs questions in such a way that learners’ achieved performance is close to this target. In the second experiment, we examine the Inverted-U hypothesis looking for the optimal target difficulty for our adaptive algorithm. In the experiment, we evaluate four experimental conditions differing in the target error: 5%, 20%, 35%, 50%. In the following text we denote the conditions as E5, E20, E35, and E50.

![Figure 6.6: Top 10 mostly used contexts available for learners to practice. (A) percentage of answers in the analyzed data set, (B) number of items, (C) average number of answers in other contexts prior to the first answer in a given context, (D) average error rate per experimental condition ignoring reference answers.](image)

6.2.1 Data Analysis

Although the system tries to achieve a specific error rate, the observed error rate is not exactly the same. There are at least three causes — noisy learners’ behaviour, imperfect predictive model, and insufficient number of appropriately difficult items. The achieved error rate depends on a specific context, see Figure 6.6. Figure 6.8 illustrates that the largest differences between conditions can be observed at the beginning of the practice. These differences, however, decrease during the practice (all conditions except E05 are in most contexts really similar after 40 questions) as some learners quit their practice and others master items from a particular context. The speed of this
“convergence” differs for different contexts; it is lower for more difficult contexts with many items (e.g., Czech cities, world states).

We are also interested in the actual form of practice, especially in the order in which the system presents the items to a learner. Figure 6.7 shows a median of the first presentation order according to the difficulty of items predicted by the currently used learner model for different contexts.

![Figure 6.7: Median of the first presentation order according to the difficulty of items predicted by the currently used learner model for different contexts.](image)

We are also interested in the actual form of practice, especially in the order in which the system presents the items to a learner. Figure 6.7 shows a median of the first presentation order according to the difficulty of items predicted by the currently used learner model. E.g., Serbia (the most difficult item from European states) is typically the second item the system asks about in the case of the E50 condition; on the other hand, Russia (the easiest item from European states) is typically the first item in the case of the E05 condition. We see that in some cases conditions radically differ (e.g., for Asian states, E05 goes from the easiest item to the most difficult one, while others in the other direction), whereas in other cases the order is quite similar (Czech cities). This phenomenon probably decreases the difference in the observed metrics among the experimental conditions. We also note that E05 often asks questions with only 2 options which leads to a faster answering speed, but we assume this feature is not fundamental for the presented analysis.
6.2.2 Results

From the global viewpoint, short term engagement is better in the case of easier questions. The survival rate after 10 answers is sorted according to question difficulty, see Figure 6.1. On the other hand, the differences decrease with the number of answers. The survival rates after 100 answers are very similar in all conditions. Note that after 30 or more questions, conditions E35 and E50 no longer achieve their target error rate in a lot of contexts (see Figure 6.8). Because of these differences in behavior among contexts, we further analyze survival in contexts separately. The return rate increases with the difficulty of questions, the largest difference being between E05 and other conditions, see Table 6.2.

Table 6.2: Global comparison of conditions with respect to engagement.

<table>
<thead>
<tr>
<th>Condition</th>
<th>Return rate *</th>
</tr>
</thead>
<tbody>
<tr>
<td>E05</td>
<td>15.2%</td>
</tr>
<tr>
<td>E20</td>
<td>16.0%</td>
</tr>
<tr>
<td>E35</td>
<td>16.6%</td>
</tr>
<tr>
<td>E50</td>
<td>16.8%</td>
</tr>
</tbody>
</table>

* confidence interval ±0.75%
There are quite large differences among the contexts (see Figure 6.9), most likely caused by learners’ preferences and implementation details of the system, e.g., the system recommends 6 contexts (e.g., European states) as “quick start” options on the home page, which makes their survival rates lower than survival rates of the “self-selected” contexts (e.g., Asian states). The magnitude of differences between conditions is mostly aligned with differences in their behavior in a particular context (confront Figure 6.9 with statistics in Figures 6.6, 6.8, and 6.7).

Short-term survival (Figure 6.9 A) differs in all contexts in favor of easier conditions. Differences between conditions in the case of some contexts are lower, probably because of the attrition bias — number of other contexts practiced prior to a particular context. The extent of that effect depends on the number of items practiced in prior contexts, which varies (see Figure 6.6 C). In the case of long-term survival (Figure 6.9, B), the trend is quite the opposite, although for individual contexts the differences are typically rather small. This contrast is best seen on European states (context with most data), where we see a reliable difference between E50 and E05.

Figure 6.9: Survival analysis (A, B) and probability of return after 10 hours (C) for 10 most practiced contexts and 4 experiment conditions. Error bars represent 95% confidence intervals.

Figure 6.9 C shows the probability of return for different contexts. Differences are often not significant within a particular context, but we generally have a higher probability of return for more difficult questions (E35, E50). There is a difference between the engagement
measured by the number of answers and the engagement measured by time a learner spent within the system. Learners answer easier questions faster, so the learners’ “speed” differs for different conditions. Figure 6.1 shows how this phenomenon influences the evaluation.

Learning

When we mix data from all contexts together and analyze learning on the global level only, more difficult practice seems to lead to better learning, see Figure 6.1. Figure 6.10 shows a more detailed analysis for individual contexts. Instead of looking at the whole learning curves, we assume that the initial error rate $a$ is the same for all conditions within the same context and we compare only their learning rate (the parameter $k$ in the power law). The learning rate differs in some contexts (e.g., Czech cities vs. European states) due to differences in the number of items and other factors. Here, we are mainly interested in the comparison of our experimental conditions within individual contexts. The general trend is the same as in the case of the global learning curve with the largest differences being between E05 and other conditions. The size of differences is related to a different behavior of conditions within individual contexts — which items are practiced in which order (see Figure 6.7) and what error rate is actually achieved (see Figure 6.8).

We also performed other kinds of analysis like filtering out learners having an insufficient number of answers (to avoid attrition bias), looking only at answers after a 10-hour delay (short term vs. long term learning), or constructing a learning curve for response time (time a learner spends on answering a question). Results are very similar in all cases — E05 is clearly the worst and differences among the others are small.

6.3 Choice of Distractors

The next experiment is motivated by observing our own practice in contexts with very low prior knowledge, e.g., Chinese provinces. Although the adaptive algorithm tries to select easy questions, often using multiple-choice questions with just two options, the compet-
ittiveness of distractors can still make the practice too difficult. We hypothesize that the distractors which are too competitive can be counterproductive in this case. Previous research suggests that repeated exposure to competitive distractors may even lead to the creation of false knowledge [115]. As we describe in Subsection 3.3.3, the construction of competitive distractors is based on wrong answers in our historical data and try to choose the most competitive distractors. We designed a new algorithm taking the distribution of historical wrong answers and transform it to another one based on the predicted learner’s knowledge. If the learner’s knowledge is high, the distractors are constructed to be competitive. On the other hand, when the knowledge is almost zero there is a high probability that we choose distractors which are least competitive (e.g., France for Romania).

On one axis we compare competitive vs. adaptive distractors. On the second axis we look at the choice of the number of options, here we compare constant (4) vs. adaptive vs. random number of options. This leads to 6 conditions: adaptive-adaptive, adaptive-competitive, constant-adaptive, constant-competitive, random-adaptive, and random-competitive.

Figure 6.10: Learning rate $k$ for different contexts. Error bars stand for 95% confidence intervals computed using bootstrapping.
6.3.1 Results

In this experiment, the results do not show any significant difference among the conditions either with respect to engagement or learning. The results for competitive distractors are slightly better than for adaptive distractors, but the differences are very small and not statistically significant. This is why this experiment is not included in Figure 6.1. Our conclusion from this experiment is that using competitive distractors is a sufficient approach. It may be that our approach to construct adaptive distractors was suboptimal, but it seems unlikely that a different approach would bring significantly different results.

A likely cause of the failure of this experiment is that conditions with adaptive and competitive distractors behave very similarly — Figure 6.11 shows a summary statistic about the constructed distractors. From answers to open questions we compute the competitiveness of distractors and for each multiple-choice question we consider how competitive the used distractors are. Although in specific situations the algorithms may differ significantly, we see that the overall behavior differs only a little, so it does not come as a big surprise that we do not detect any significant differences.

Figure 6.11: The average usage of the top 10 of the most competitive distractors for adaptive conditions vs. competitive conditions.
6. Case Studies

6.4 Number of Options

The third experiment focuses on the number of options in the case of multiple-choice questions. When we launched the application we arbitrarily set the upper bound for the number of options to 6. In this experiment we varied this setting. We deployed 5 conditions: N2, N3, N4, N6, N8 setting the maximum number of options to 2, 3, 4, 6, and 8. With higher options, we hypothesized a negative impact on engagement (due to the lower speed of practice) and positive impact on learning (due to more difficult questions requiring more cognitive processing).

6.4.1 Results

The results are summarized in Figure 6.1 (the third column). Condition N4 leads to the best learning, see Figure 6.1 (right, A). We confirm our hypothesis that a higher number of options slows down the practice. On average, learners spend 5.7 seconds answering one question in the N2 condition, 6.3 seconds in the N4 condition, and 6.8 seconds in the N8 condition. This observation mainly influences the analysis of short-term engagement. The total number of learners’ interactions is higher in the case of conditions with fewer options, but the time spent in the system is almost unaffected, see Figure 6.1 (right, B and C).

6.5 Self-Regulated Question Difficulty

We hypothesize that learners may be more effective in setting the appropriate difficulty of their practice than we are. Some learners may prefer easier content, others may prefer a more difficult one. In the final experiment we let some learners modify the parameter based on their preferences. The practice within the system is presented in groups of 10 questions. After each series of 10 questions, the systems shows a summary feedback to learners. At this moment we have inserted a new dialog box with a question “How difficult would you like the questions to be?” with 5 choices: “much harder”, “harder”, “same”, “easier”, “much easier”, see Figure 6.12. We call answers to these questions “ratings” (not “settings”) because in a placebo condition they do not have any impact on the algorithm.
6. Case Studies

Figure 6.12: Dialog box shown after each practice set (10 questions) in the case of placebo and adjustment condition.

We have performed a standard randomized control trial with the following experimental conditions:

- **normal** — a control group, a standard version of the system without the dialog box,
- **placebo** — the dialog is shown but does not have any impact on the behaviour of the adaptive algorithm,
- **adjustment** — the dialog is shown and the answer changes the target difficulty setting (-20%, -10%, 0%, +10%, +20%).

In all cases the initial setting of the target error rate parameter is 35% (the value is based on the results of the previous experiment, see Section 6.2).

6.5.1 Data Analysis

First, we analyze ratings provided by learners. We mostly have only one rating from a particular learner per context. The majority of learners do not provide any rating at all, because the presented dialog box can be skipped by closing it. Since the ratings are provided after finishing a practice set, we assume the main factor determining a learner’s rating is the error rate achieved during the recent practice set, therefore we divide all ratings to buckets based on the error rate. Figure 6.13 shows the relation between ratings and the recent error rate (based on the last 10 questions).
Figure 6.13: Learners’ ratings with respect to the error rate achieved during a recent practice set. Solid lines stand for the placebo condition, dashed lines stand for the adjustment condition. The shaded areas show 95% confidence intervals.

The basic relation is intuitive — successful learners want more difficult questions, unsuccessful learners want easier questions. The “appropriate” ratings have the shape of an inverted-U curve with the maximum at the error rate around 35%. This result is in agreement with our previous experiment that showed a 35% target error rate, see Section 6.2.

For high error rates, the results are intriguing. As could be expected, the ratio of the “bit harder” ratings is very small. Unexpectedly, highly unsuccessful learners often provide the “more difficult” rating. Although the number of highly unsuccessful learners is relatively small, this trend is statistically significant and consistent for both placebo and adjustment conditions. We interpret this trend as systematic “irony” in responses of a subgroup of users and we hypothesize that this behaviour is connected to disengagement with the system. This result should serve as a caution — learners’ expressed preferences may reflect not just their true preferences with respect to the concerned question, but may also incorporate other aspects of their (affective) state.
At first sight, the lines in Figure 6.13 should be the same for both experimental conditions, but we observe a difference between the placebo (solid line) and the adjustment (dashed line) condition. Learners assigned to the adjustment condition seem to be more satisfied. To get an idea of why this is happening, consider learners that achieve an average error rate $E$ during the recent practice set. Generally, a part of learners is able to achieve this error rate $E$ using the original setting. The number of learners satisfied with it is proportionally the same for both conditions. The number of learners unsatisfied with the achieved error rate $E$ is lower in the case of the adjustment condition, because the learners were allowed to set a different difficulty. For similar reasons, the adjustment condition also contains the learners satisfied with the error rate $E$ who are not able to achieve this error rate using the original setting.

The data also show a relation between ratings and context difficulty. The percentage of “much harder” ratings increases with decreasing context difficulty (e.g., European states are easier than African cities, at least for users of the used system). This observation indicates that our algorithm for adaptive practice is not adaptive enough and there is room for improvement — the algorithm could take into account the difficulty of a particular context.
6. Case Studies

Figure 6.15: A comparison of the engagement for the three experimental conditions. Engagement is measured as the ratio of learners who provide at least 100 questions (left) or use the system for at least 12 minutes (right). Error bars show 95% confidence intervals.

Allowing learners to adjust the target difficulty of their practice results in lower error rates during practice, see Figure 6.14. This, however, does not necessarily mean better learning, it is probably just a consequence of the lower difficulty setting.

6.5.2 Results

The analysis of learners’ ratings provides interesting insights, but the main point of the experiment is to find whether the dynamic adjustment leads to higher engagement and learning. As expected, the engagement for the placebo condition is worse than for the control group, see Figure 6.15. The dialog box asking for learners’ ratings has a negative impact on learners’ engagement. For the adjustment condition, the negative effect of the dialog is partially compensated for by the benefits of the difficulty adjustment. However, the benefits of the difficulty adjustment are not sufficient to considerably outweigh the disadvantage of the additional dialog box.

When we compare conditions with respect to learning, there are no significant differences. Since many learners do not provide any ratings at all, it is not very surprising — most learners in the placebo and adjustment conditions keep the original target error rate and thus their practice is the same as for the control group. However, a
more detailed analysis shows that for specific cases there are some trends, particularly concerning easy contexts and learners who prefer easier questions — in these cases the dynamic adjustment seems to have a negative impact on learning. Figure 6.16 shows the learning curves for 30% easiest contexts (e.g., Europe states); in this case the learning is worse for the adjustment condition. It is probably caused by learners’ tendency to set a lower difficulty even in easy contexts (e.g., by externally motivated learners from schools). On the other hand, the difference is really small.

Figure 6.16: Learning curves for easy contexts.

Another more detailed analysis disaggregates the overall results with respect to learners — specifically based on their preference for easy or difficult questions. We classify each series of a learner’s answers on a particular context based on the learner’s resulting difficulty setting (real in the case of the adjustment condition, hypothetical in the case of the placebo condition) to 4 groups as shown in Table 6.3. These groups are not comparable to each other, because they often correspond to completely different kinds of contexts or learners, but using this approach we can more deeply compare placebo and adjustment conditions.

Figure 6.17 and Figure 6.18 show results for the engagement and the learning rate disaggregated into these four groups. Figure 6.17 shows that for learners who have a preference for easier practice the adjustment leads to higher engagement. There is no significant difference between conditions for other groups. Figure 6.18 shows that the
adjustment leads to better learning for those who prefer more difficult questions and to lower learning for those who prefer easier questions. The results are interesting particularly for the group of learners who prefer easy questions — for this group the adjustment hampers the speed of learning, but increases engagement.

Although the summary results do not show large differences between conditions, for specific learners the impact of the difficulty adjustment can be important. Particularly, it seems that some learners prefer easy questions which give them a “good feeling” during practice, but they do not practice the knowledge they need to practice and thus their learning is slower.
Figure 6.18: Analysis of learning with respect to groups of users based on their ratings. The graphs show the learning rate $k$ in the fitted learning function $a \cdot x^k$. Error bars show 95% confidence intervals.
7 Conclusion

The thesis deals with the application of machine learning techniques in the context of online educational systems. Using learner models which take historical data and predict current learners’ knowledge, we propose a framework providing adaptive practice of facts in domains with varied prior knowledge. The presented framework allows the system to present tasks with appropriate difficulty, which is assumed to be beneficial to learning. Learners’ knowledge is decomposed into prior and current knowledge. To model the prior knowledge, we drew inspiration from the chess player rating and applied the Elo rating system to estimate learners’ skill and difficulty of items. This method seems more suitable to be used in an online environment than other techniques like Item response theory because of its simplicity (speed of parameter estimation) and sufficient accuracy. To handle learning, we deploy a slightly modified Performance factor analysis which is again really easy to use to estimate learners’ skill online.

Using the proposed framework, we built two applications providing adaptive practice of factual knowledge covering geography and anatomy. These applications are used by hundreds of learners per day which allows us to identify several issues related to the adaptivity and data collection and its impact on the evaluation. We also use these applications as a platform for online experiments where we test our hypotheses and try to find optimal parameters of the presented framework. Since we are trying to keep a learner’s error rate roughly constant during their practice, we work with biased data and it complicates their further analysis. This phenomenon influences both the evaluation of learner models and the evaluation of the impact of adaptivity on learners. We believe the described issues (or similar ones) are not related to our approach only, but are present in a lot of other similar systems already used by a large number of learners.

To minimize the bias in the collected data, we designed a methodology to evaluate computerized adaptive practice containing data independent of the algorithm controlling the practice itself. In our case, the system collects roughly 10% of answers on open questions asking for a random stem. We call this data reference answers and use them to build learning curves describing learning in the population.
7. Conclusion

of our users. As we present, this approach still has a few issues, but we show that these issues do not outweigh the benefits. We state that learning is not the only important measure we should focus on when evaluating an online educational system. Nowadays, these systems are not used by students from schools only. Developers of these systems have to attract people without the external motivation, so we should also measure the engagement. Without learners actively using the system, the system cannot be beneficial.

Focusing on learning and engagement, we performed several online experiments. Almost every experiment revealed new issues related to the interpretation of the results. We carefully described the issues and showed techniques solving them. Step by step, we were able to prove that our algorithm for adaptive practice works well and is beneficial for learners; we found the optimal target difficulty; optimal number of distractors; and showed that the benefits of self-regulated difficulty are questionable.

7.1 Lessons Learned

Ideally, we would like experiments to lead to clear conclusions concerning the choice of the best experimental condition, so that we can use the results for a direct improvement of a particular educational system. For this purpose we typically focus on highly-aggregated data and well-defined summary metrics. Experience with our experiments, however, suggests that results are typically nuanced and that it is useful to disaggregate summary numbers and analyze the results in more detail. In the following text we present the most important lessons we learned.

7.1.1 Monitoring of System Behavior

First of all, we confirm the experience mentioned in [53]. We were not able to start most of the described experiments on the first attempt. Typically, it was necessary to run them twice because of a bug in the implementation. Since it is very easy to make very interesting conclusions from absolutely wrong data, it is definitely worth thoroughly
testing the system before the deployment and monitoring it during the experiment.

Even an implementation without bugs can easily have unintended effects during the experiment. A good example is Adaptive vs. Random. In Subsection 6.1.1, we show how our implementation of the adaptive construction of distractors leads mainly to questions with 2 options, 6 options, and open questions, although it was not our original intention. In the case of the Question Difficulty experiment, we found that the achieved error rate is far from the target error rate of the tested conditions, see Figure 6.8. Of course, the results from the experiments still hold and are important for us, but we cannot make a simple general conclusion “more difficult practice leads to better learning”. Based on this we encourage authors of experiments to analyze data from their experiments as soon as they can to be sure that the experiments prove the initial assumptions. One of the worst things which can happen is making findings from data that do not correspond to your assumptions, or throwing data from a six months running experiment after you accidentally realize there is a bug in the system.

7.1.2 Choice of Metric

Although it is tempting to summarize experiments by a single number, this could hide important nuances and may lead to misleading conclusions when the measure is not carefully selected. Our results repeatedly show how the choice of metrics matters and how we can reach a different conclusion by using different measures.

For engagement there are often differences between short term and long term engagement. What is good in a short term may not be good in a long term. Similarly, engagement and learning may not be aligned, particularly when we are focused on short term engagement (which is a tempting measure, as it is the easiest to collect data for). Both these trends are demonstrated in the Question Difficulty experiment.

Similarly, in the case of the Number of Options experiment there is a difference between engagement measured by time vs. engagement measured by the number of interactions. As questions with fewer options are easier to read, learners answer them faster, so the “speed” of learners differs in different conditions. Since we typically see smaller
differences with respect to elapsed time, we may tend to look mainly at the number of interactions, but we should not forget about learners’ effort affecting the speed of practice.

On the other hand, to make decisions effectively, it is beneficial to have one overall evaluation criterion with the possibility to make a more detailed analysis when you need it. So if you really need it, as the overall evaluation criterion we suggest using an average success rate on the last reference question of each learner. This metric naturally combines engagement and learning. Figure 7.1 shows the value of this metric for some of the performed experiments.

7.1.3 Attrition Bias

In Subsection 6.1.2 we show the detected attrition bias introduced by different attractivity of the practiced content for a similar group of learners in various conditions. Another interesting issue appeared in the Question Difficulty experiment. Our system introduces several educational categories learners can choose to practice (as our system does). For example, it allows learners to practice Asian countries, which is quite advanced for some of them. We found that the average error rate of users on the first reference answer in one condition (the easiest one) is roughly 5% higher than in other conditions. Learners
in this condition are probably more confident in their knowledge and move to more advanced topics faster than users in other conditions. Based on this, it is really difficult to analyze one global learning curve aggregating data from all contexts, because when we construct the learning curve, we assume it starts in the same point for all studied conditions.

Although we did not find any direct impact of this bias in our analysis, we assume it potentially influences the results of many published experiments.

7.1.4 Levels of Aggregation

We found that the content our system provides is not homogeneous. Some contexts (e.g., European countries) are easy for our learners, some contexts are very difficult; some contain many items to practice, others consist of only a few items (see Figure 3.12). When we aggregate all the data, we have a higher chance of having statistically significant results. On the other hand, a high-level aggregation may hide important differences or potential biases. In the worst case, it can lead to completely misleading results.

Unfortunately, in many cases we cannot use fine-grained aggregation due to the small amount of data in individual contexts. It would provide us better insight, but this kind of analysis is also susceptible to noise in data, and the statistical significance is weaker. Based on our experience, we suggest using an intermediate level of analysis by aggregating contexts into groups, e.g., groups of “easy contexts” or “contexts with more than 50 items”. This approach carries a risk of “data fishing” – analyzing different kinds of groups until some results are found. Therefore, the groups should be specified in advance or should be defined somehow naturally. If the groups are well selected, we can achieve both an interesting insight and statistical significance.

As a specific example of this approach, we discuss the analysis of the Question Difficulty experiment. The message given from the global-level analysis seems relatively simple. Looking at a number of answers, easier conditions are better with respect to short-term engagement, but long-term engagement is not affected by the choice of the target difficulty, see Figure 6.1. However, when we look at the data more closely, the results are more nuanced. As shown in Figure 7.2,
7. Conclusion

Figure 7.2: Survival curves for the easiest and the most difficult condition from Question Difficulty experiment described in Section 6.2. Curves are constructed separately for the top 25% easiest and most difficult contexts. The difficulty is defined as the average error rate on the first reference question.

the results hold for the top 25% most difficult contexts, but in the case of the top 25% easiest contexts there is a swap after 20 answers. Easier conditions are still better with respect to short-term engagement, but in the case of long-term engagement it is exactly the opposite. Results from this analysis indicate that the used algorithm for adaptive practice is not able to balance the difficulty of questions ideally.

7.1.5 Design of Conditions

When we run AB experiments and analyze the collected data, we usually fight with statistical significance. Of course, we can try to collect even more data as is already common in the case of big companies [54], but in academia we usually do not have a system with such high traffic. On average, the application (Outline Maps) is used by more than 600 learners every day, so we collect more than 45,000 answers per day. Even with this traffic we must carefully design experiments to achieve statistically significant results. Generally, more extreme conditions are better. Before running an experiment, the designed conditions often seem more extreme than they turn out to be. In the case of the Question Difficulty experiment, conditions E20, E35, and E50 have very similar behavior in many contexts (see Figure 6.8), so the biggest difference is between condition E05 and the others. When we designed this experiment, some of us thought that condition E50 is too difficult and would discourage users of the system.
The second story is the Choice of Distractors experiment. As we explain in Subsection 6.3.1, there is no big difference in the resulting behavior of the experimental conditions. The situations when the conditions differ by design are really rare. It is worth analyzing the already collected historical data and, based on this, estimate the amount of data influenced by the difference of the designed conditions. This kind of analysis can save a lot of time.

7.2 Future Work

The thesis deals mainly with one proposed method for computerized adaptive practice of factual knowledge, its evaluation, and finding optimal values of its parameters, related to the practice difficulty mainly. Working on the described issues, we are aware that there is a lot of directions for future work. Here are some of them.

7.2.1 Comparison to Other Methods

In a domain of learner modeling it is common to compare a proposed model to other, already existing models. It can be very limited and unfair, because the authors of the proposed model can be focused on its tuning more than on tuning other models. However, this kind of comparison gives a reader at least a rough idea about the benefits of the proposal. On the other hand, research papers focused on intelligent educational systems as a whole often compare the proposed methods to simple baselines only (e.g., learners using the system vs. learner reading an educational material). Our work is not an exception, we compare our approach for adaptive practice to completely random practicing. In the future, it would be worth focusing on other methods like the Leitner system (see Section 2.2.2), or mastery learning (see Section 2.2.2).

7.2.2 Bias in Data Collection

After performing five online experiments using Outline Maps, we are sure that the analysis of bias introduced by data collection should play the main role when we formulate conclusions during the evaluation process. There are already authors mentioning bias in their data when
publishing papers related to the evaluation, e.g. authors in [69] mention “seasonality” in their data related to the school lessons, or [34] describes the impact of mastery learning attrition on learning curves. However, it is not common practice. We suppose it would be very interesting to classify common kinds of bias, quantify the possibility of overlooking them using already published papers, formulate a possible impact on the results, and find ways to handle it.

7.2.3 Automated Parameter Tuning

One iteration of the online experiment can take weeks or even months. The experiments are very expensive in terms of the time spent by learners assigned to nonoptimal conditions (or even conditions with large negative effects). In practice, it would be really useful to apply methods for solving the multi-armed bandits problem as proposed in [69] to minimize this kind of price. Although we describe measures for learning short/long-term engagement (see Section 2.3, it is necessary to use a single metric only to use these methods in real systems. Since Outline Maps have relatively large traffic, it would be possible to deploy methods for solving multi-armed bandits problem to tune parameters of the adaptive practice. It would also be beneficial to analyze the impact of these methods and used metrics on the data collection, data interpretation and the resulting values of parameters.
A Data

A.1 Usual Traffic

The data set describing the usual traffic of Outline Maps is available at data.outlinemaps.org and is made available under the Open Database Licence. The system provides adaptive practice of geographical facts (e.g., names and locations countries, cities, or mountains).

The presented data set has several advantages. The data set is based on an open education system — an open source project freely available online — with available description of algorithms used [88]. Researchers can thus try the system themselves before using the data set and inspect the details of its realization. This is in contrast to many current education data sets whose origin is not completely clear or easily inspectable (e.g., data sets based on Carnegie Learning systems, which are commercial). In such cases it is hard to understand the data exactly.

The data set is also easy to interpret, with simple structures and records with clear intuitive meanings. The data set deals with geographical items that can be easily visualized. This offers possibilities for quick inspection and analysis of data. The data set contains all the important aspects of the questions asked; it does not contain any assumptions or pre-processing steps by authors (e.g., use of predefined knowledge components, as is often done in case of currently available data sets).

The content of the data set — learning of facts in a realistic setting — supplements those currently available. Most fine-grained data sets of learning processes, as available for example in DataShop [50], focus on the learning of procedural skills (e.g., math). The learning of factual knowledge has been studied thoroughly before, but mostly in laboratory experiments with small groups and artificial facts. The presented data set comes from a realistic, large-scale application used by students in schools or in preparation for exams.

1. Data for Practice Anatomy are available at data.practiceanatomy.com and their description is very similar.
2. http://opendatacommons.org/licenses/odbl/1.0/
A. Data

Although the structure of the data set is simple, the recorded student learning behaviour is complex and captures many interesting aspects of learning: widely varied prior knowledge, forgetting and short term memory effects, and the relation between response times and correctness of answers. The potential of the dataset is illustrated by previous research (in many cases the research offers just preliminary results, showing a potential for more complex modeling and deeper analysis):

1. modeling prior knowledge (considering only first answers on each item) [84];

2. modelling students’ memory (short term memory effects, forgetting) based on repeated answers utilizing time between attempts [106];

3. analysis of response times [92].

A.1.1 Ethical and privacy considerations

The educational system is used mainly by students in schools or by students preparing for exams. Nevertheless, it is an open online system, which can be used by anybody, and details about individual users are not available. Users are identified only by their anonymous identifier. Users can log into the system using their Google or Facebook account. This login is used only for identifying the user within the system and is not included in the data set. Unlogged users are tracked using web browser cookies. The system also logs the IP address from which users access the system; the IP address is included in the data set in anonymized form. We separately encode the country of origin, which can be useful for analysis, and its inclusion is not a privacy concern. The rest of the IP address is replaced by a meaningless identifier to preserve privacy.

A.1.2 Description

This is the first publicly available version of the data set. It captures learner interactions up to 21 May 2015. The basic statistics of the data set are as follows:
- 91 331 learners;
- 1 459 geographical items;
- 10 087 306 answers.

Table A.1: Usual traffic: columns of CSV file with answers.

<table>
<thead>
<tr>
<th>Column</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>id</td>
<td>answer identifier</td>
</tr>
<tr>
<td>user</td>
<td>learner’s identifier</td>
</tr>
<tr>
<td>placeAsked</td>
<td>identifier of the asked place</td>
</tr>
<tr>
<td>placeAnswered</td>
<td>identifier of the answered place, empty if the user answered “I don’t know”</td>
</tr>
<tr>
<td>type</td>
<td>type of answer: (1) find the given place on the map; (2) pick the name for the highlighted place</td>
</tr>
<tr>
<td>options</td>
<td>list of identifiers of options (the asked place included)</td>
</tr>
<tr>
<td>inserted</td>
<td>datetime (yyyy-mm-dd HH:mm:ss) when the answer was inserted to the system</td>
</tr>
<tr>
<td>responseTime</td>
<td>how much time the answer took (measured in milliseconds)</td>
</tr>
<tr>
<td>placeMap</td>
<td>identifier of the place representing a map for which the question was asked</td>
</tr>
<tr>
<td>ipCountry</td>
<td>country retrieved from the user’s IP address</td>
</tr>
<tr>
<td>ipId</td>
<td>meaningless identifier of the user’s IP address</td>
</tr>
<tr>
<td>language</td>
<td>language version of the system: (0) Czech, (1) English, (2) Spanish</td>
</tr>
</tbody>
</table>

The data set is available in the standard CSV format (commas are used as a delimiter). The core of the data set is the answer.csv file with detailed information about learners’ answers (see Table A.1). Some of the presented attributes have been implemented recently, so they may contain undefined value for some answers. A supplementary file describes the used geographical items — file place.csv describing the
items (see Table A.2) and place_type.csv with description of types of places.

Table A.2: Usual traffic: columns of CSV file with places.

<table>
<thead>
<tr>
<th>Column</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>id</td>
<td>place identifier</td>
</tr>
<tr>
<td>code</td>
<td>code of the place (ISO 3166-1 alpha-2 if it is possible)</td>
</tr>
<tr>
<td>name</td>
<td>name of the place</td>
</tr>
<tr>
<td>type</td>
<td>identifier of the place type</td>
</tr>
</tbody>
</table>

A.1.3 Limitations

The data set has very limited information about students. The system can be used by anybody and no demographic data on users is available (beyond an anonymized IP address and the associated information about user location). Moreover, logging into the system is voluntary and tracking of unlogged users is done only using web browser cookies, i.e., it may happen that the same person has multiple identifiers in the data set (this is, however, a feature of many similar systems).

The data set is not based on a randomized experiment, but on an adaptive system that uses a student model to choose questions [88]. On one hand, this is a strong point as the data correspond to real-life usage of an educational system. On the other hand, this aspect complicates the interpretation of data, e.g., care must be taken with “success rate”, since the system actively tries to achieve a predetermined success rate (by choice of suitable questions).

A.2 Case studies

We also provide data for the experiments we performed and described in Chapter 6 at the same address (data.outlinemaps.org). The data are very similar to the data described in Section A.1, but the format of the data set is slightly different. It contains two files only: answers.csv with answers and feedback.csv with learners’ feedback which is
A. Data

collected by a dialog box shown after each practice set together with the summary.

Table A.3: Case studies: columns of CSV file with answers.

<table>
<thead>
<tr>
<th>Column</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>id</td>
<td>answer identifier</td>
</tr>
<tr>
<td>user_id</td>
<td>learner’s identifier</td>
</tr>
<tr>
<td>item_asked_id</td>
<td>identifier of the asked place</td>
</tr>
<tr>
<td>term_asked_name</td>
<td>name of the asked place</td>
</tr>
<tr>
<td>item_answered_id</td>
<td>identifier of the answered place, empty if the user answered “I don’t know”</td>
</tr>
<tr>
<td>term_asked_name</td>
<td>name of the asked place</td>
</tr>
<tr>
<td>term_type</td>
<td>type of the practiced place</td>
</tr>
<tr>
<td>context_name</td>
<td>name of the practiced map (context)</td>
</tr>
<tr>
<td>direction</td>
<td>type of the answer: (t2d) find the given place on the map; (d2t) pick the name for the highlighted place</td>
</tr>
<tr>
<td>options</td>
<td>number of options (the asked place included)</td>
</tr>
<tr>
<td>time</td>
<td>datetime when the answer was inserted to the system</td>
</tr>
<tr>
<td>response_time</td>
<td>how much time the answer took (measured in milliseconds)</td>
</tr>
<tr>
<td>condition</td>
<td>identifier of the studied condition for which the asked question was constructed</td>
</tr>
<tr>
<td>ip_country</td>
<td>country retrieved from the user’s IP address</td>
</tr>
<tr>
<td>ip_id</td>
<td>meaningless identifier of the user’s IP address</td>
</tr>
</tbody>
</table>
Table A.4: Case studies: columns of CSV file with learners’ feedback.

<table>
<thead>
<tr>
<th>Column</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>id</td>
<td>answer identifier</td>
</tr>
<tr>
<td>user_id</td>
<td>learner’s identifier</td>
</tr>
<tr>
<td>inserted</td>
<td>datetime when the record was inserted to the system</td>
</tr>
<tr>
<td>value</td>
<td>learner’s feedback</td>
</tr>
</tbody>
</table>
B Author’s Publications

B.1 Journal Papers


[Author’s contribution: 25%]

I implemented the presented system, prepared the data sets and participated on the part focused on the question construction and its evaluation.


[Author’s contribution: 70%]

I prepared the data set and the final version of the paper.

B.2 Conference and Workshop Proceedings


[Author’s contribution: 10%, Best paper nomination]

This paper contains my analysis from another (workshop) paper [83]


[Author’s contribution: 45%]

I performed the experiment focused on the choice of the target difficulty, analyzed learning and presentation order of items.
B. Author’s Publications


   [Author’s contribution: 35%]

   I performed the experiment focused on the algorithm for adaptive practice and analyzed learning, attrition bias and provided global-level statistics.


   [Author’s contribution: 45%; **Best student paper**]

   I performed online experiments examining the algorithm for practice control with respect to a user’s motivation and then analyzed the collected data.


   [Author’s contribution: 35%]

   I did an analysis of the feedback loop between a model describing a user’s knowledge and an algorithm for practice control collecting the data. I also examined the impact of the model and its precision on data collection.


   [Author’s contribution: 25%]

   I analyzed the relation of response time and correctness of answers.

I implemented the predictive model and the algorithm for practice control within an application providing the practice of geography. I designed the mechanism of difficulty adjustment and selection of distractors in case of multiple-choice questions.

B.3 Technical Reports


   I performed the experiment focused on the choice of the target difficulty, analyzed learning and presentation order of items.
Bibliography


BIBLIOGRAPHY


BIBLIOGRAPHY


BIBLIOGRAPHY


[111] Yumeng Qiu, Yingmei Qi, Hanyuan Lu, Zachary A Pardos, and Neil T Heffernan. Does time matter? modeling the effect of time


