Impact of Adaptive Educational System Behaviour on Student Motivation

Jan Papoušek and Radek Pelánek

Faculty of Informatics, Masaryk University Brno
{jan.papousek,xpelanek}@mail.muni.cz

Abstract. In this work we try to connect research on student modeling and student motivation, particularly on the relation between task difficulty and engagement. We perform experiments within widely used adaptive practice system for geography learning. The results document the impact of the choice of a question construction algorithm and target difficulty on student perception of question suitability and on their willingness to use the system. We also propose and evaluate a mechanism for a dynamic difficulty adjustment.

1 Introduction

The goal of adaptive educational systems is to make learning more effective and engaging by tailoring the behaviour of the system to a particular student. The adaptive behaviour is based on student models which estimate the knowledge of students (and potentially other characteristics like their affective state). While a lot of research has focused on development and evaluation of student models, relatively little attention has been devoted to the way the outputs of models are actually used in educational systems. The typical use of student models is for mastery learning, e.g. research studies [3,9,6] have evaluated the impact of used models and their thresholds on over-practice and under-practice.

The use of student models only for judging mastery is, however, only one possible way of making a system adaptive to behaviour of its users. Adaptive educational systems have the potential to make learning more engaging by keeping students in the concentrated flow state [4]. One of the conditions for the flow state is the balance between skills and difficulty of presented problems. The Inverted-U Hypothesis predicts that maximum engagement occurs with moderate challenge [10]. There is extensive research on this topic (e.g. [1]); the research is, however, based mainly on laboratory studies, the results of research are to a certain degree contradictory (see e.g. the discussion in [10]), and it is not clear how to apply the hypothesis in the development of a practical educational application.

In this work we connect the use of student models with the research on optimal level of challenge. We study the impact of adaptive behaviour of an educational system on student motivation in a widely used educational system for learning geographical facts.

In the previous works the specification of the adaptive behaviour was based mainly on intuition of system developers and was not evaluated [8,11] or was evaluated using only comparison to a control group without any tuning of the difficulty [2]. The most similar research is by Lomas et al. [10] who evaluated the Inverted-U Hypothesis by testing many variants of an educational game (numberline estimation). They failed to find the U-shaped relation between difficulty and motivation. For their study the relation was monotone (simpler problems were more engaging). Explaining the result they state that maybe they "never made the game easy enough" [10]. Our experiments are similar, the main difference is that we use a more realistic educational application. Another similar research was done using Math Garden software [7]. The authors compared three conditions (target success rate 60%, 75%, 90%) and showed that the easiest condition led to the best learning (mediated by a number of solved problems).

For our work we use a widely used application [11] for learning geography. We have performed randomized online experiments (multivariate testing) to evaluate the impact of the adaptive behaviour on student motivation. The appropriate difficulty of questions is evaluated using proxy measure of student motivation (number of questions answered) and student self-reports (perception of question difficulty). The results show that the adaptive behaviour is advantageous and that the suitable portion of correct answers per user (success rate) is around 65% (with students who used the system in school preferring easier questions). We also propose a dynamic difficulty adjustment of the target success rate and we show that this mechanism improves the adaptive system behaviour and makes it more robust to misalignment of the parameter setting.

2 Question Construction

We start by describing a question construction module for adaptive practice of facts (e.g. vocabulary, geography, human anatomy). Different variants of this module are used for the below reported experiments. The process of question construction has two phases. In the first phase we select a target item, which the question is concerned with, and in the second phase we construct the question itself.

2.1 Selecting a Target Item

The selection of a target item needs to balance several criteria. The main focus of the current work is on appropriate difficulty – according to the flow theory (Inverted-U hypothesis), questions should be adequately hard to ask, since with easy questions students can get bored and with difficult questions students may be frustrated. Another criterion is that questions concerning the same item should not repeat in a close succession [5]. Finally, no item should be left out while practicing, i.e. even students with high knowledge should be asked at least once about each item (our experience suggests this is an intuitive expectation of students). We combine these criteria using the mechanism of scoring functions.

Each item is evaluated by a scoring function according to each criterion and the item with the highest weighted sum is used as a candidate to ask about.

The difficulty aspect is taken into account with the use of a student model. The actual system used for experiments [11] uses a combination of Elo rating system [13] and Performance factor analysis [12]. For the purposes of question construction the details of the used student model are not important – we use it as a black box which provides for each item estimated probability P_{est} that a particular student will answer correctly. The first scoring function depends on the distance between the estimated probability for the given item and the target success rate P_{target}. Assume that our goal is to ask a question in case of which the student has 75% chance of answering correctly. The distance from the probability for the difficult countries (nearly 0% chance of the correct answer) is higher than for easy ones (almost 100%), so it is necessary to normalize it. We use the following scoring function:

$$S_{prob}(P_{est}, P_{target}) = \begin{cases} \frac{P_{est}}{P_{target}} & \text{if } P_{target} \geq P_{est} \\ \frac{1-P_{est}}{1-P_{target}} & \text{if } P_{target} < P_{est} \end{cases}$$

The second scoring function penalizes items based on the time elapsed since the last question, because we do not want to repeat items in a short time interval when they are still in short term memory. We use the function $S_{time}(t) = -1/t$, where t is time in seconds. Using just 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 the third scoring function that uses the total number n of questions for the given item answered by the student: $S_{count}(n) = 1/\sqrt{1+n}$. The total score is given as a weighted sum of individual scores, the weights are currently set manually, reflecting experiences with the system: $W_{prob} = 1$, $W_{count} = 1$, $W_{time} = 12$.

2.2 Dynamic Adjustment of Target Difficulty

One of the key parameters whose role we experimentally evaluate is the target success rate P_{target}. In the context of computerized adaptive testing the optimal success rate is 50% – such choice leads to the most informative answers and to the best estimate of a student's skill. We are, however, primarily interested in student practice. In this context student motivation is crucial and 50% success rate does not seem very encouraging. The default success rate of our system is 75% (similarly to other applications, e.g. [8]), below we report experiments with different values of the target success rate.

To strengthen the adaptivity of system behaviour we propose an additional dynamic adjustment of target difficulty. With this mechanism the target probability is modified depending on student'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 students and more difficult questions to more successful ones; a specific function for transformation of the target rate is depicted in Fig. 1. Note that by using this mechanism we also indirectly correct a potential estimation bias given by the used student model.

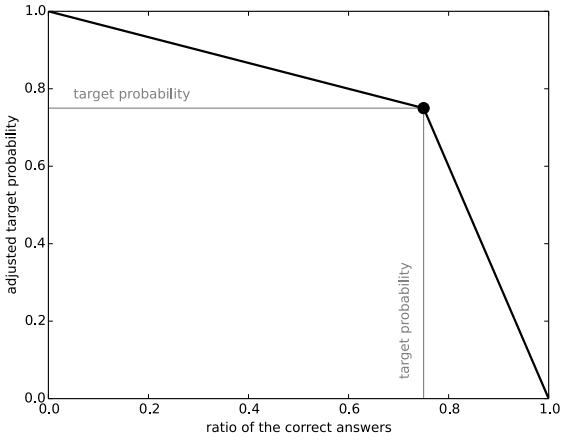

Fig. 1. Adjustment of the target probability of the correct answer.

2.3 Choice of Options

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 predicted difficulty of the selected candidate does not match the target difficulty. Unfortunately, there is nothing to do in case of too easy items. In case of a more difficult candidate the system can use a multiple choice question to give the student a chance to guess the correct answer. With the probability of guessing P_{guess} the probability of answering correctly is $P_{guess} + (1 - P_{guess}) \cdot P_{est}$, where P_{est} is the estimated probability of the correct answer on an open question asking about the given item.

Our goal is to make the probability of the correct answer close to P_{target}. This can be achieved by making P_{guess} close to:

$$G = \frac{P_{target} - P_{est}}{1 - P_{est}}$$

For $G \leq 0$ we use an open question without options, otherwise we use a multiple choice question with n options, where n is the closest integer to $\frac{1}{G}$. For obvious reasons the minimal possible value of n can be 2, for practical reasons there is also an upper bound for n (more than 6 options would look cluttered). To ensure that the options are not easy to disregard, the algorithm takes into account what other items are most commonly mistaken with the given question candidate in open questions. For example, in case of our application, Cameroon is most often confused with Niger (38%), Nigeria (27%), Central African Republic (10%), the Republic of the Congo (9%), Gabon (6%), the Ivory Coast (5%), Uganda (3%), and Guinea (2%). This ratio determines the probability that a given item appears among options in the constructed multiple choice question.

3 Experimental Setting

For experiments we use an adaptive educational system `slepemapy.cz` – an application for learning geography [11]. Students[1] can choose a specific map (e.g. Africa, the United States) and a type of places (e.g. countries, regions, cities, rivers). The system offers adaptively selected questions to students who answer them using an interactive map. After a series of 10 questions the system provides feedback on student's progress. Students can also access a visualization of their knowledge using an open learner model. The application is currently used by hundreds of students per day, majority of students is from the Czech Republic and Slovakia since the interface was originally only in Czech. English and Spanish are currently also available.

3.1 Available Data

So far we have collected almost 6 million answers. For each answer we log all details about the question (target item, options), the student ID, the chosen answer, and also the timing information. We have no personal information about students, we only log their IP address. Part of the data is made public[2].

The system is available to anybody, free of charge. We have no control over the number of answered questions, the time when students practice, or whether they ever return to the system after one session of practice. Thus we assume that the data about students' usage of the system is a reasonable proxy for their motivation to learn using the system. There is one important exception – the system is also used in some schools directly during the class time, in this case the usage of the system may not be related to student motivation. Therefore, for most of the reported experiments, we did not consider these students (an exception is an analysis in Section 4.2). To detect the 'in-school usage' we currently use only a coarse method based on IP address (a group of at least 5 students who started using our system from the same IP address). The 'in-school' usage represents about 20% of the data.

The student model currently used in the application has been calibrated using data containing mainly countries [11]. There are few areas (e.g. Czech cities) where the quality of prediction is worse than the quality of predictions for countries. Since the algorithm for question construction uses predictions as its input, we suppose its behaviour is worse for this kind of areas. Therefore we filtered the data to contain information only about users who were answering solely questions concerning countries.

To perform experiments we used multivariate testing where we randomly divided students into several groups and assigned to each of these groups a different version of the algorithm for the question construction. For analysis we filtered out students using our system before the experiments started. We have

[1] Note that the system is publicly available and can be used by anybody, for terminological consistency we use the word 'student' to denote any user.

[2] `https://github.com/adaptive-learning/data-public`

also removed students having less than 10 answers, since in this case there are not enough opportunities for differences among the tested versions to emerge.

3.2 Metrics

For our experiments we need to measure effects of different variants of the adaptive algorithm on students. We have chosen two different ways to do so. Firstly, we collect students' subjective evaluation of the provided practice. Secondly, we look at the students' behaviour.

In the first case we ask students to evaluate the difficulty of questions. After 30, 70, 120, and 200 answers the system shows the dialog "What is the difficulty of asked questions?", students choose one of the following options: "Too Easy", "Appropriate", "Too Difficult". We look at the ratio of students choosing the given option for the first time and call this metric an *explicit feedback*. In the second case we measure the total number of answered questions. The distribution of number of answers across students is highly skewed, therefore we use the median as a summary statics. For testing statistical significance between distributions we use the t-test over logarithm of number of answers (logarithm transformation is used to reduce the skew).

The two metrics are related. When we divide the students according to their first evaluation, the median number of answers is 91 for group "Too Easy", 110 for group "Appropriate", and 97 for group "Too Difficult". The difference between the group "Appropriate" and both other groups is statistically significant. The observed median is much higher than the median in the following experiments because we include only users with at least one evaluation record.

4 Experiments

Performed experiments corespond to the main aspects of the question construction algorithm. Each experiment was run only for a certain time within the system, so for each of them we report the size of data set used in evaluation.

4.1 Impact of Question Construction Algorithm

The key question of our first experiment is: "Is the proposed algorithm better than a random construction of questions?". As we already mentioned, the mechanics behind the system for constructing questions for a student consists of two main parts. Firstly, the algorithm selects the target item (i.e. which country to ask about). Secondly, it chooses the number of options and options themselves. In the first experiment we evaluated the role of both of these parts. For each part we considered the proposed adaptive mechanism and a random choice. For this experiment we collected more than 30,000 answers.

Table 1. shows how the given versions of the algorithm differ according to the median of the number of answers per student in the given group. The results show that adaptivity brings improvement, and that it is necessary to make both parts

Table 1. Algorithm variants for the question construction used in the first experiment, for each variant we report the median of the the number of answers per student in the given group.

Target item	Options	Answers
adaptive	adaptive	33.0
adaptive	random	20.0
random	adaptive	20.0
random	random	19.5

of the algorithm adaptive. The difference between completely adaptive version and other tested versions is in two cases statistically significant ($p < 0.01$). In the case of comparison with the the completely random algorithm the results are on the edge of statistical significance ($p = 0.06$) due to relatively small number of students (26) in the corresponding multivariate group. Unfortunately during this experiment we were not collecting the explicit feedback from students, so only the implicit one is available.

4.2 Impact of Difficulty

In the second experiment we study the question: "Does the difficulty of the questions matter?". The Inverted-U Hypothesis suggests that really easy and really hard questions should have negative impact on students' motivation. In this experiment we deployed several variants of the adaptive algorithm which differ only in the target probability of the correct answer. For this experiment we do not consider the mechanism for adjustment of the target probability (to simplify interpretation of the results).

The results of this experiment and their interpretation are not straightforward. With regard to the total number of answers we have not discovered any statistically significant trend. The biggest issue is probably the relation between the target probability parameter and the real success rate of students. The success rate is only partially influenced by the target rate, other factors include for example students choice of maps. Although the target probability is from the interval $[50\%, 95\%]$, the average real success rate varies only from 65% to 90%. On several maps (e.g. countries in Europe, for which we have most data) there are not sufficiently difficult items to achieve 50% success rate (for most students).

On the other hand, the relation between achieved success rate and perceived difficulty of questions shows a clear U-shaped pattern[3] (Fig. 2.). The curve does not have a sharp peak, but there is a clear dynamics between the classes. With the increasing difficulty the growth of the number of "Too Easy" votes is compensated by the drop of "Too Difficult" votes. The peak of the "Appropriate"

[3] This analysis is less dependent on the calibration of student model, thus we take into account all answers, not only countries.

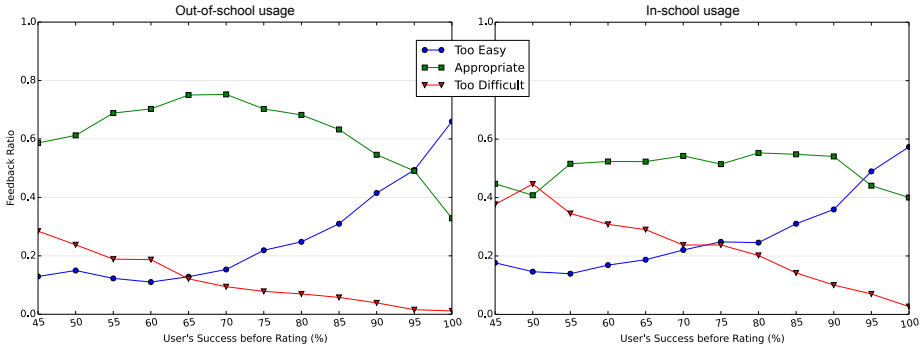

Fig. 2. Explicit feedback given by students according to their previous real success rate. The used data set consists of more than 1,700,000 answers and 12,000 feedback records.

answers as well as the equal votes for "Too Easy" and "Too Difficult" occur at the success rate 65%. This experiment thus suggests that the value 65% may be a suitable target rate for this kind of application.

Previous research [1] suggests that the optimal difficulty may differ depending on the type of motivation (internal, external), particularly that in school-related activities students prefer lower levels of challenge. To examine this hypothesis we compared results for out-of-school usage of the system with in-school usage. Fig. 2. shows that there is really a substantial difference. The in-school group prefers easier questions (the optimal difficulty is around 75%) and they are also generally less satisfied with the practice in the system. Note that we currently use only very simplified detector of in-school/out-of-school usage, therefore it is probable that the real difference is even higher.

Finally, Fig. 3. shows the relation between algorithm's target probability and obtained explicit feedback. As we already mentioned the target probability parameter influences the real student's success rate only partially, thus the relation is less pronounced than in the previous graph. The maximum satisfaction with the practice is reached when the target probability is 60%-65%, which corresponds to the 70% from the perspective of the real success rate.

4.3 Impact of Difficulty Adjustment

The last studied question concerns the difficulty adjustment and is: "Does the difficulty adjustment mechanism have an impact on the student's behaviour?" Although the average success rate is not affected by the adjustment (about 75% for both variants), student's experience is different. Median of the number of answers is 28 for enabled adjustment and 21 for disabled, the difference between the two variants is significant ($p = 0.02$).

As Fig. 3. shows, the main reason for the better effect is that the adjustment increases the robustness of the algorithm with respect to the target probability of

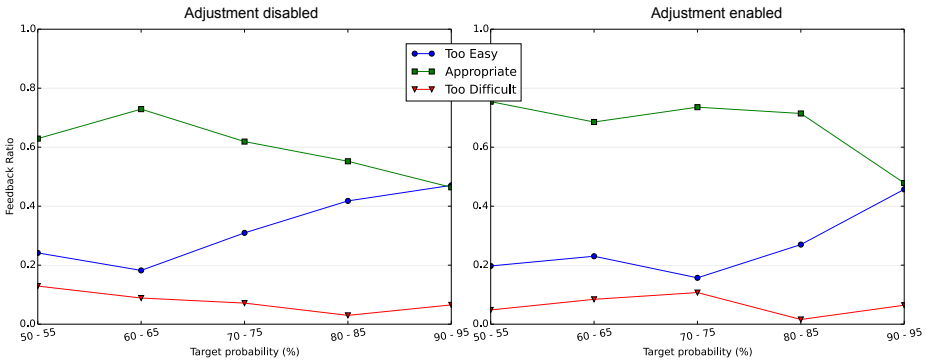

Fig. 3. Explicit feedback given by students according to the target probability used in the algorithm, depending on whether the dynamic adjustment of difficulty is disabled or enabled. The used data set for this experiment consists of 176,000 answers and 1,450 feedback records.

the correct answer. For this reason the choice of the target probability becomes less important. Regarding the explicit feedback, the ratio of "Appropriate" votes is 5% higher when the adjustment is turned on (68% vs. 63%).

5 Conclusions

We performed experiments with a widely used adaptive practice system to evaluate the impact of system behaviour on student motivation and perception of question difficulty. The results show that the adaptive algorithm for construction of questions which is based on a student modeling [11] has a positive impact on student willingness to use the system.

Based on student self-reported perception of difficulty of questions it seems that a good target success rate is around 65%. There is a difference between in-school and out-of-school usage of the system. Students using the system in schools prefer easier questions, which is in accordance with previous literature [1]. Nevertheless in the artificial intelligence in education community this aspect is worth attention, since it is usually not studied or taken into account.

For the actual behaviour of students (number of answered questions), however, we did not get significant trend with respect to the target rate (as predicted by Inverted-U Hypothesis). It seems that even for our relatively simple application for learning geography there are enough interacting factors (e.g. success rate, student skill, choice of maps) to obfuscate the relation between difficulty and motivation. Moreover, the used proxy for motivation – the number of answered questions – may be insufficient.

We also proposed a dynamic mechanism for adjustment of question difficulty. The results of experiments show that this mechanism is effective (improves students' willingness to use the system) and that it makes the behaviour of the

system robust with respect to the choice of the target success rate. This mechanism can thus both simplify the development and improve the behaviour of new adaptive practice systems.

References

1. Sami Abuhamdeh and Mihaly Csikszentmihalyi. The importance of challenge for the enjoyment of intrinsically motivated, goal-directed activities. *Personality and Social Psychology Bulletin*, 38(3):317–330, 2012.
2. Michal Barla, Mária Bieliková, Anna Bou Ezzeddinne, Tomáš Kramár, Marián Šimko, and Oto Vozár. On the impact of adaptive test question selection for learning efficiency. *Computers & Education*, 55(2):846–857, 2010.
3. Hao Cen, Kenneth R Koedinger, and Brian Junker. Is over practice necessary?-improving learning efficiency with the cognitive tutor through educational data mining. *Frontiers in Artificial Intelligence and Applications*, 158:511–518, 2007.
4. Mihaly Csikszentmihalyi. *Flow: The psychology of optimal experience*. Harper Perennial, 1991.
5. Peter F Delaney, Peter PJL Verkoeijen, and Arie Spirgel. Spacing and testing effects: A deeply critical, lengthy, and at times discursive review of the literature. *Psychology of learning and motivation*, 53:63–147, 2010.
6. Stephen E Fancsali, Tristan Nixon, Annalies Vuong, and Steven Ritter. Simulated students, mastery learning, and improved learning curves for real-world cognitive tutors. In *AIED Workshops*, pages 11–20. Citeseer, 2013.
7. Brenda RJ Jansen, Jolien Louwerse, Marthe Straatemeier, Sanne HG Van der Ven, Sharon Klinkenberg, and Han LJ Van der Maas. The influence of experiencing success in math on math anxiety, perceived math competence, and math performance. *Learning and Individual Differences*, 24:190–197, 2013.
8. S Klinkenberg, M Straatemeier, and HLJ Van der Maas. Computer adaptive practice of maths ability using a new item response model for on the fly ability and difficulty estimation. *Computers & Education*, 57(2):1813–1824, 2011.
9. Jung In Lee and Emma Brunskill. The impact on individualizing student models on necessary practice opportunities. In *Educational Data Mining (EDM)*, pages 118–125, 2012.
10. Derek Lomas, Kishan Patel, Jodi L Forlizzi, and Kenneth R Koedinger. Optimizing challenge in an educational game using large-scale design experiments. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 89–98. ACM, 2013.
11. Jan Papoušek, Radek Pelánek, and Vít Stanislav. Adaptive practice of facts in domains with varied prior knowledge. In *Educational Data Mining (EDM)*, pages 6–13, 2014.
12. Philip I. Pavlik, Hao Cen, and Kenneth R. Koedinger. Performance factors analysis-a new alternative to knowledge tracing. In *Proceedings of Artificial Intelligence in Education (AIED)*, volume 200 of *Frontiers in Artificial Intelligence and Applications*, pages 531–538. IOS Press, 2009.
13. Radek Pelánek. Time decay functions and elo system in student modeling. In *Educational Data Mining (EDM)*, pages 21–27, 2014.