Should We Give Learners Control Over Item Difficulty?

Jan Papoušek
Masaryk University
Brno, Czech Republic
jan.papousek@mail.muni.cz

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

ABSTRACT

Personalized educational systems are able to provide learners questions of specified difficulty. Since learners differ, the appropriate level of difficulty may vary and it may be impossible to find an universal setting. We implemented a version of an adaptive educational system for geography practice that allows learners to adjust difficulty of questions. We evaluated this feature using a randomized control experiment. The overall results show only a small effect of the adjustment. A more detailed analysis, however, shows that for some groups of learners the effect can be important, although not necessarily advantageous. The collected data from the experiment provide insight into how to tune question difficulty automatically.

KEYWORDS

adaptive practice, difficulty, evaluation, learning, engagement

1 INTRODUCTION

User modeling allows us to develop personalized learning environments that make learning experience tailored towards individual learners. Using learner modeling techniques [4] we can estimate probability that a learner can answer a question or solve a problem. Based on these predictions we can automatically choose items of appropriate difficulty [11].

But what is an appropriate level of difficulty? This is typically a parameter that is specified externally by developers of a learning system. The choice of this parameter has been addressed in previous research, but without clear results. The general idea that the best activity is neither too easy nor too difficult was formulated as Inverted-U Hypothesis [1]. Lomas et al. [7] found that in the context of their simple educational game easier problems lead to higher engagement, but lower learning. A similar research was done using Math Garden software [6]. The authors compared three conditions and showed that the easiest condition led to the best learning (mediated by a number of solved tasks). Our previous work [13] in the case of the adaptive practice of geography facts led to different conclusions, showing better results for both long-term engagement and learning for more difficult questions.

Moreover, it seems probable that there is not a single optimal difficulty for everyone. So a natural idea is to allow learners to manipulate the difficulty of questions. In addition to better tailored system behaviour, previous research suggests that a sense of control (or even perception of control, rather than the actual objective level of control) can increase engagement [8]. On the other hand, research on self-regulated study [3] shows that people are prone to mismanaging their own learning.

The principle of dynamic difficulty adjustment has been explored mainly in computer games (see for example [2]). In educational research the most relevant research explored self-adaptation of difficulty in math practice [5]. The authors used adaptive practice system for basic arithmetic, with possible setting of difficulty on one of three levels: 60%, 75%, 90% success rates. Their results show that preferred difficulty varies based on age and gender of students, but they did not find any impact of the availability of difficulty setting on learning, engagement, or students' self-belief. The study, however, has several limitations, e.g., setting of success rate interacted with gamification aspects of the user interface and the used sample size was small (48 students in each condition).

We present similar experiment, but for a different type of knowledge (learning geography facts) and using a large scale experiment with thousands of users and millions of answers. We allow learners to adjust the difficulty of questions and use randomized control trial to evaluate the impact of this feature. Similarly to the previous study we find patterns in learners' behaviour with respect to the setting of difficulty, but we do not find any large impact on engagement or learning, at least in the global analysis of data. Once we disaggregate the results, some interesting results emerge, particularly for a group of learners who prefer easy questions – for these learners the difficulty adjustment feature leads to lower efficiency of learning, but higher engagement. The results also provide insight for automatic setting of target difficulty. This paper is a full version of [9].

2 SYSTEM AND EXPERIMENT

We use a system for an adaptive practice of geographical facts, e.g., names and locations of countries or cities. The system is available online at outlinemaps.org. It does not work with 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 %). Learners use the system on their own or during school sessions. During the experiment we did not have any control or contact with users, specifically we did not provide any external incentives to users.

Learners can use the system with different maps and types of places (e.g., European states); these contexts differ widely in their difficulty (prior knowledge) and the number of items available to

UMAP'17 Adjunct, Bratislava, Slovakia

DOI: 10.1145/3099023.3099080

practice (from 10 to 170). Distribution of answers is highly uneven, most learners practice a few popular maps.

The system collects data about the correctness of answers and based on the collected data it estimates the current knowledge of a particular learner and personalizes the provided practice [11]. A key parameter in the adaptive algorithm is "target difficulty", which sets the average success rate of learners that the system is aiming at. The algorithm uses a learner model to construct personalized questions for each learner so that the probability of answering correctly will be close to the target difficulty [10, 11].

In previous experiment [13] we varied target difficulty between experimental groups. In the current 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 1. We call answers to these questions "ratings" (not "settings", because in a placebo condition they do not have any impact on the algorithm).

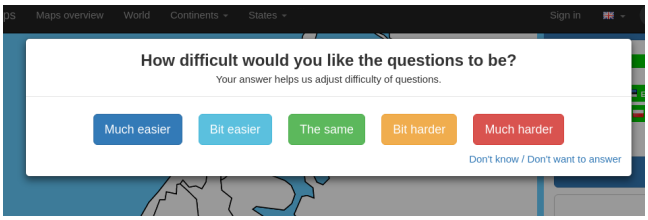

Figure 1: Dialog box shown after each practice set (10 questions) in the case of placebo and adjustment condition.

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 results of the previous experiment [13]). The experiment was performed from October 2016 to January 2017 and we have collected roughly 8 200 000 answers from 85 000 learners. To make our research reproducible we make the analyzed data set available[1].

3 ANALYSIS OF USER RATINGS

At first, we analyze ratings provided by users. Mostly, we have only one rating from a particular learner per context. Majority of learners do not provide any rating at all. Since the ratings are provided after finishing a practice set, we assume the main factor determining a learner's rating is an error rate achieved during the recent practice set, therefore we divide all ratings to buckets based on the error rate. Figure 2 shows the relation between ratings and the recent error rate (based on the last 10 questions).

[1]http://data.outlinemaps.org/2016-ab-user-difficulty-adjustment.zip

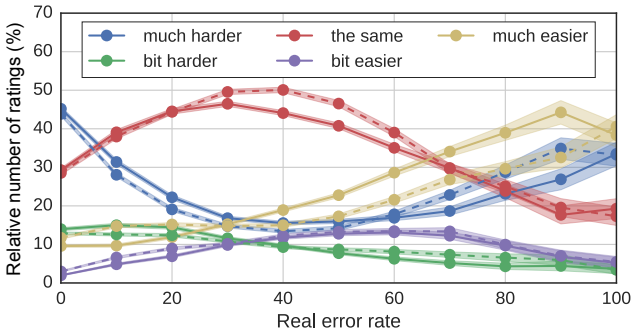

Figure 2: Learners' ratings with respect to the error rate achieved during the recent practice set. Solid lines stand for placebo condition, dashed lines stand for adjustment condition. The shaded areas shows 95% confidence intervals.

The basic relation is intuitive – successful learners want more difficult questions, unsuccessful learners want easier questions. The "appropriate" ratings have the shape of inverted-U curve with the maximum at the error rate around 35%. This result is in agreement with our previous experiments that showed that target error rate 35% is suitable [13].

For high error rates the results are intriguing. As could be expected the ratio of "bit harder" ratings is very small. Unexpectedly, highly unsuccessful learners often provide "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 presence of a systematic "irony" in responses of a subgroup of users and we hypothesize that the 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 the first sight, the lines in Figure 2 should be the same for both experimental conditions, but there is 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 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. A number of learners satisfied with it is proportionally the same for both conditions. A number of learners unsatisfied with the achieved error rate E is lower in the case of adjustment condition, because the learners where allowed to set a different difficulty. For similar reason, the adjustment condition contains also 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

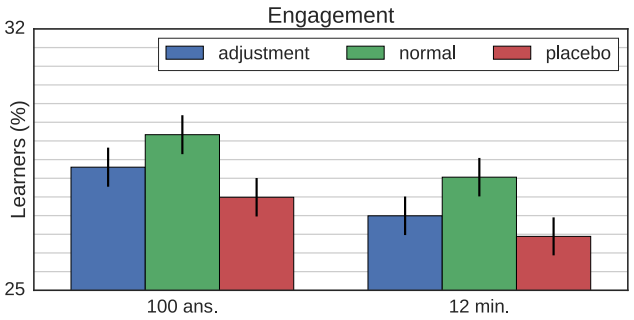

Figure 3: Comparison of engagement for the three experimental conditions. Engagement is measured as the ratio of learners who provide at least 100 questions (left) or use the system for at least 12 minutes (right). Error bars show 95% confidence intervals.

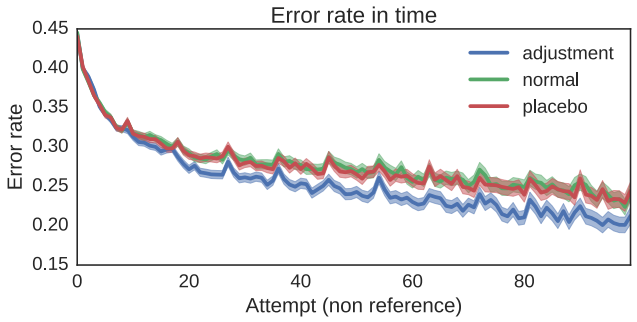

Figure 4: Average error rate by a number of attempts (reference questions are excluded). The shaded areas shows 95% confidence intervals.

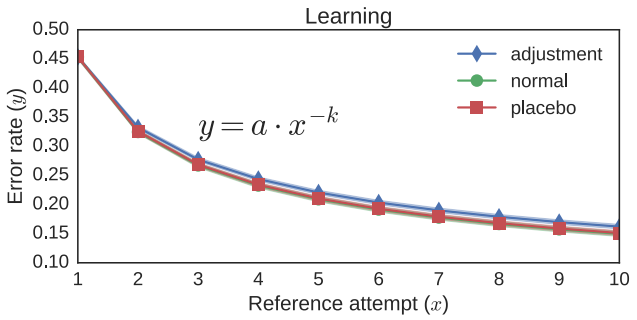

Figure 5: Learning curves for easy contexts.

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.

4 ENGAGEMENT AND LEARNING

The analysis of learners' ratings provide interesting insights, but the main point of the experiment is to find whether the dynamic adjustment leads to higher engagement and learning.

As a measure of engagement we use a survival rate – the ratio of learners who answer at least 100 questions. Since learners in the case of adjustment condition change difficulty of practice and questions with different difficulty lead to different response times, we also measure the ratio of learners practicing at least 12 minutes[2]. For discussion of the choice of metric see [12]. The results are presented in Figure 3. As expected, the engagement for the placebo condition is worse than for the control group. The dialog box asking for learners' ratings has negative impact on learners' engagement. For the adjustment condition the negative effect of the dialog is partially compensated by the benefits of the difficulty adjustment. However, the benefits of the difficulty adjustment are not sufficient to considerably overweight the disadvantage of the additional dialog box.

Allowing learners to adjust the target difficulty of their practice results in lower error rates during practice, see Figure 4. This, however, does not necessarily mean better learning, it is probably just a consequence of lower difficulty setting. To compare learning among conditions, we utilize "reference questions" – for each context separately every tenth question is selected randomly without any influence of the adaptive algorithm and we use data from these questions to construct learning curves (see [12] for more detailed discussion).

When we compare conditions in this way, the results are comparable – there are no significant differences in overall learning rates. Since many learners do not provide any ratings at all, it is not much surprising – most learners in 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 negative impact on learning. Fig. 5 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 lower difficulty even on easy contexts (e.g., by externally motivated learners from schools).

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 average difficulty setting – either the actual setting in the case adjustment condition or the hypothetical setting in the case of placebo condition. Based on this average difficulty setting we classify each answer serie into one of 4 groups as shown in Table 1. The results for these groups are not comparable to each other, because they often correspond to completely different kinds of contexts or learners, but we can compare placebo and adjustment conditions for each of these groups.

Figure 6 and Figure 7 show results for engagement and learning rate disaggregated into these four groups. Figure 6 shows that

[2]This time roughly corresponds to 100 answers on average.

Table 1: The classification of answer series.

Group	Average target error rate
G-easy	at most 25%
G-default	more than 25% and less than 45%
G-difficult	more than 45%
G-unknown	N/A (dialog is always skipped)

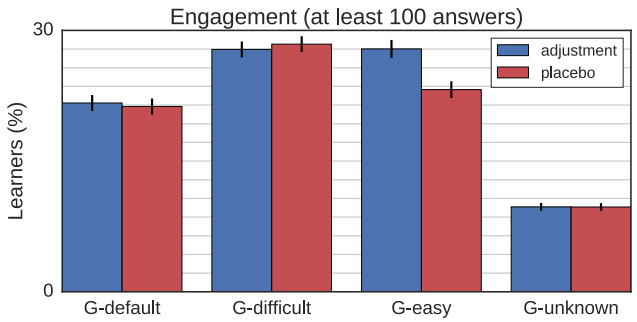

Figure 6: Analysis of survival rates with respect to groups of users based on their ratings. The survival rates are computed per context (e.g., European states). Error bars show 95% confidence intervals.

for learners who have preference for easier practice the adjustment leads to higher engagement. There is no significant difference between conditions for other groups. Figure 7 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.

So 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 "good feeling" during practice, but they do not practice knowledge they need to practice and thus their learning is slower.

5 DISCUSSION

The overall results show that in the case of the used adaptive practice system giving users control over question difficulty does not bring a straightforward advantage. The dialog box with the difficulty setting reduces engagement and the advantage brought by better tailored difficulty is not sufficient to offset this disadvantage. A more detail analysis shows significant effects on learners who prefer easy questions. For them the adjustment leads to less efficient learning, but more engagement (longer practice).

Instead of giving learner option to tune difficulty, we should probably develop methods for automatic adjustment of target difficulty. The data from the experiment provide guidance for tuning the target difficulty. They support the basic target error rate 35%. They also show that this basic target error rate could be modified based on the difficulty of the specific practice context. Another

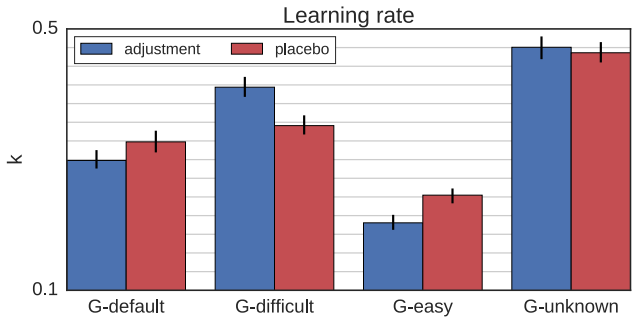

Figure 7: Analysis of learning with respect to groups of users based on their ratings. The graphs shows the learning rate k in the fitted learning function $a \cdot x^k$ (see Figure 5). Error bars show 95% confidence intervals.

factor that may be useful for automatic tuning of question difficulty is the difference between in-school and out-of-school usage of the system. Previous results [10] showed that students using the system in school prefer easier questions that out-of-school users; the data from the current experiment concord with this result.

The results also show that there is systematic "irony" in learners ratings – unsuccessful learners report that they want much harder problems. This shows limitations of dependence on collected subjective data; this issue requires more attention in research.

We have studied only one specific implementation of dynamic adjustment in a single educational system. It is possible that our results are influenced by particular choices made in the implementation or by specific features of the geography domain. However, the main results agree with previous research [5] that was done under significantly different conditions. Although the issue requires more research, the current results suggest that giving learners direct control over question difficulty is not beneficial.

REFERENCES

[1] Sami Abuhamdeh and Mihaly Csikszentmihalyi. 2012. The importance of challenge for the enjoyment of intrinsically motivated, goal-directed activities. *Personality and Social Psychology Bulletin* 38, 3 (2012), 317–330.

[2] Justin T Alexander, John Sear, and Andreas Oikonomou. 2013. An investigation of the effects of game difficulty on player enjoyment. *Entertainment Computing* 4, 1 (2013), 53–62.

[3] Robert A Bjork, John Dunlosky, and Nate Kornell. 2013. Self-regulated learning: Beliefs, techniques, and illusions. *Annual Review of Psychology* 64 (2013), 417–444.

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

[5] Brenda RJ Jansen, Abe D Hofman, Alexander Savi, Ingmar Visser, and Han LJ van der Maas. 2016. Self-adapting the success rate when practicing math. *Learning and Individual Differences* 51 (2016), 1–10.

[6] Brenda RJ Jansen, Jolien Louwerse, Marthe Straatemeier, Sanne HG Van der Ven, Sharon Klinkenberg, and Han LJ Van der Maas. 2013. The influence of experiencing success in math on math anxiety, perceived math competence, and math performance. *Learning and Individual Differences* 24 (2013), 190–197.

[7] Derek Lomas, Kishan Patel, Jodi L Forlizzi, and Kenneth R Koedinger. 2013. Optimizing challenge in an educational game using large-scale design experiments. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM, 89–98.

[8] Thomas W Malone and Mark R Lepper. 1987. Making learning fun: A taxonomy of intrinsic motivations for learning. *Aptitude, learning, and instruction* 3, 1987 (1987), 223–253.

[9] Jan Papoušek and Radek Pelánek. 2017. Evaluation of Learners' Adjustment of Question Difficulty in Adaptive Practice of Facts. In *Proc. of User Modelling, Adaptation and Personalization*. Extended abstract.

[10] Jan Papoušek and Radek Pelánek. 2015. Impact of Adaptive Educational System Behaviour on Student Motivation. In *Artificial Intelligence in Education*, Vol. 9112. 348–357.

[11] Jan Papoušek, Radek Pelánek, and Vít Stanislav. 2014. Adaptive Practice of Facts in Domains with Varied Prior Knowledge. In *Educational Data Mining*. 6–13.

[12] Jan Papoušek, Vít Stanislav, and Radek Pelánek. 2016. Evaluation of an Adaptive Practice System for Learning Geography Facts. In *Proc. of Learning Analytics & Knowledge*. ACM, 40–47.

[13] Jan Papoušek, Vít Stanislav, and Radek Pelánek. 2016. Impact of Question Difficulty on Engagement and Learning. In *Proc. of Intelligent Tutoring Systems (LNCS)*, Alessandro Micarelli, John C. Stamper, and Kitty Panourgia (Eds.), Vol. 9684. Springer.