The Effect of Achievement Badges on Students' Behavior: An Empirical Study in a University-Level Computer Science Course

http://dx.doi.org/10.3991/ijet.v10i1.4221

Lasse Hakulinen, Tapio Auvinen and Ari Korhonen
Aalto University, Espoo, Finland

***Abstract*—Achievement badges are a form of gamification that are used in an attempt to increase user engagement and motivation in various systems. A badge is typically a graphical icon that appears as a reward for the user after reaching an achievement but that has no practical value. In this study, we describe and evaluate the use of achievement badges in the TRAKLA2 online learning environment where students solve interactive, automatically assessed exercises in a Data Structures and Algorithms course throughout the semester. We conducted an experiment where the students (N=281) were randomly divided into a treatment and a control group, with and without achievement badges. Students in the treatment group were awarded achievement badges, for example, for solving exercises correctly on the first attempt, doing exercises early, or solving all the exercises in a round with full points. Grading was the same for both groups, i.e. collecting badges did not affect the final grade, even though the exercise points themselves did. Students' activity in TRAKLA2 was logged in order to find out whether the achievement badges had an effect on their behavior. We also collected numerical and open-ended feedback in order to find out students' attitudes towards the badges. Our results show that achievement badges can be used to affect students' behavior. Statistically significant differences were observed in the time used per exercise, number of sessions, total time, and normalized total number of badges. Furthermore, the majority of the students reported being motivated by the badges. Based on our findings, achievement badges seem to be a promising method to motivate students and to encourage desired study practices.**

***Index Terms*—Achievement badges, E-learning, Gamification, Motivation**

I. INTRODUCTION

"A soldier will fight long and hard for a bit of colored ribbon." –Napoleon Bonaparte

Over the last years, different forms of gamification have gained a lot of attention among educators. The goal of gamification is to apply elements from games to non-game systems in order to make them more motivating and engaging, while not turning them into fully-fledged games. One popular form of gamification is the use of achievement badges. A badge typically appears to the user as a graphical icon after reaching an achievement. They do not necessarily have practical value to users, i.e. they are not worth money or open new possibilities in a game or a learning environment.

The idea of utilizing game-like elements and reward systems is not new. For example, the merit badges given by the scouts as well as military marks of rank can be seen as a way to give recognition on one's achievements. However, after the term "gameification" was introduced in 2008 [1], and later established as "gamification", the use of game elements to improve different non-game systems has been increasing rapidly. Gamification has even been noted at the *Gartner's Emerging Technologies 2013 Hype Cycle* at the top of the "Peak of inflated expectations"[1].

Gamification is a broad concept and includes methods such as achievement badges, leaderboards, points, and levels [2]. It is commonly referred to as the use of game design elements in non-game contexts [3]. However, alternative definitions have been proposed by different authors. Huotari and Hamari [1] highlight the goal of gamification rather than the methods by defining gamification as a process of enhancing a service with affordances for gameful experiences in order to support user's overall value creation. Some definitions, on the other hand, emphasize the technical aspect of gamification by restricting it to incorporating game elements into non-gaming software applications [4].

Despite the popularity of gamification, the need for more research on its effectiveness has been clearly stated in the literature. De Marcos et al. [5] claim that there is little, if any, solid empirical evidence of the effectiveness of gamification in education, and that some studies question such effectiveness pointing out potential problems that students and instructional designers face. Furthermore, Hamari et al. [6] point out in their review on gamification research that many gamification studies are descriptive in nature and more research with rigorous methodologies is needed to understand its effects.

In this study, we describe and evaluate the use of gamification in TRAKLA2, which is an online learning environment used in university-level Data Structures and Algorithms courses [7]. We added achievement badges to TRAKLA2 with the goal of motivating students to follow better study practices. In our experiment, students using TRAKLA2 were randomly divided into a treatment group with badges and a control group without them. Badges were awarded, for example, for solving exercises without

[1] http://www.gartner.com/newsroom/id/2575515

mistakes on the very first attempt, returning the exercises many days before the deadline, or completing an exercise round with full points. The badges were not tied to grading, but students could pursue them voluntarily. After the course, we analyzed students' behavior from the system logs and students' attitudes towards badges from numeric and open-ended feedback. Our research questions were:

1) What kinds of effects do achievement badges have on students' behavior?

2) What were students' attitudes towards the use of badges?

A preliminary analysis on the effect of badges on students' behavior has previously been presented in [8]. The results suggested that badges can have an effect on some aspects of students' behavior. The present study extends the quantitative analysis of students' behavior based on log data and adds the analysis of students' feedback.

This paper makes a contribution to the empirical evidence of the effect of gamification in an educational setting by reporting a controlled experiment with randomized groups. Moreover, the quantitative analysis of the effects of gamification is combined with a qualitative analysis of students' attitudes in order to further understand the relationship between the motivational factors of badges and the observed effects in behavior.

This paper is structured as follows: Related work and earlier studies are described in Section 2. Our experiment is described in Section 3, and the results of the experiment are reported in Section 4. In Section 5, we present our interpretation of the results, and finally, Section 6 concludes the paper.

II. RELATED WORK

In their review of gamification studies, Hamari et al. [6] found out that the majority of the studies reported positive effects, but the effects were greatly dependent on the users and the context. They discovered that in some cases, the same aspects of gamification were considered positive by some respondents while being disliked by others. Furthermore, in educational settings, different students have been observed to respond differently to the same gamification methods [4; 9; 10]. Therefore, understanding students' attitudes towards gamification plays a key role when evaluating its usefulness in education.

Achievement badges are one of the most commonly used gamification methods [11]. There are many definitions for achievement badges, but commonly they are seen as an additional system, which provides optional goals and challenges. Montola et al. [12] describe achievement systems as secondary reward systems with optional sub-goals that are visible to others. Hamari and Eranti [11] define them as an optional challenge provided by a meta-game that is independent of a single game session and yields possible reward(s). Morris et al. [13] identified praise (i.e. positive evaluation of performance [14]), as one reason why games are motivating. In our opinion, achievement badges can be seen as a form of praise as they often are a positive evaluation of one's performance. There can even be a fanfare or a public congratulation that further highlights the excellence of the completed achievement.

Achievement badges have been used as a gamification method in educational settings in attempts to increase learners' motivation and engagement towards the studied subject. Decker and Lawley [15] noted a significant shift in student behavior in terms of peer tutoring with the use of an achievement system. Denny [16] found out in a controlled experiment that badges had a positive effect on the quantity of students' contributions in a learning environment while the quality of the contributions remained the same. Domínguez et al. [4] concluded that students with a gamified e-learning environment got better scores in practical assignments while performing poorly on written assignments. McDaniel et al. [17] studied achievement badges in an online learning environment and found out that they had positive motivational effects, but at the same time caused frustration probably due to difficulties to get the achievements. Moreover, Anderson et al. [18] found out in a randomized experiment that even subtle changes in the way badges are represented can have an impact on their effect.

A. Risks of Gamification

Even though it seems that gamification can increase motivation and engagement, it has been criticized for focusing too much on external rewards when the actual engagement should come from students' intrinsic motivation. Nicholson [19] states that gamification might reduce internal motivation towards the activity by replacing internal motivation with external. However, he suggests that gamification can be used to improve internal motivation if the game elements can be made meaningful to users. Moreover, Lee and Hammer [20] point out that extensive gamification might teach students to study only when provided with external rewards. Some critical views on gamification even propose the use of terms as *pointsification* [21] or *exploitationware* [22] to highlight the gap between complex motivational elements in games and simplified gamification attempts.

In their meta-study about external rewards, Deci et al. [23] found evidence that external rewards might undermine intrinsic motivation. They reflect the results of their meta-study to the Cognitive Evaluation Theory (CET) [24] and state that the results provide strong support for the theory. CET states that rewards have *informational* and *controlling* aspects, and suggests that rewards perceived as controlling undermine intrinsic motivation whereas informational rewards enhance it. Mekler et al. [25] examined the effect of common gamification elements on users' intrinsic motivation in an image tagging assignment. They found no evidence that the game elements affected users' intrinsic motivation negatively. In the light of CET, they avoided the game elements from being perceived as controlling by not linking them to any external events (e.g. cash prizes). We believe that badges can be perceived as controlling or informational depending on the context, badge criteria, and individual differences.

Weaver et al. [26] have analyzed the *presenter's paradox* in a number of studies. Their findings show that perceivers make less favorable evaluations of the value of a product when mildly favorable information is added to highly favorable information. In other words, adding a cheap extra item to a valuable product may make the package appear less valuable in the eyes of the buyer. For example, a very expensive mobile phone with a one month free music service might appear as a less valuable deal for a customer than without the extra service. Adding achievement badges to learning environments might have undesirable side effects comparable to the presenter's paradox. By adding badges, we may signal that the exercises are not intrinsically motivating and hence completing the tasks is compensated with external rewards.

B. Side Effects of Automated Assessment

Distance education requires students to regulate their own learning, as there is no human tutor to give guidance. This may lead to some students to incorporate undesired study practices such as procrastinating work and returning assignments at the last minute, aiming to earn points with minimum effort, giving up after fulfilling the minimum course requirements, etc.

Edwards [27] points out that trial and error behavior is one undesirable approach observed in computer science students when using automated assessment. He claims that students will be more successful at learning if they move from *trial and error* to *reflection in action*. Karavirta et al. [28] also observed that a group of students resort to trial and error in the automatically assessed TRAKLA2 exercises. Malmi et al. [29] found that this behavior can be discouraged by limiting the number of resubmissions. However, doing so lowers the average final score, which indicates that some students benefit from unlimited attempts. We believe that it is beneficial for students to get used to checking their solutions before submitting and completing the exercises thoughtfully rather than attempting them carelessly. One of the aims of the present study is to find out if badges can be used to discourage trial and error problem solving without the drawbacks of strictly limiting the number of allowed attempts.

Another harmful habit often observed in online learning environments is procrastination i.e. starting to work near the deadline. Multiple studies on students' behavior in online learning environments have shown that there is a significant group of students who submit on the last day, and that this behavior impairs performance [30; 31; 32; 33]. Edwards et al. [31] found out in a within-subject study that this is not merely a correlation where a confounding variable, such as talent, causes both the high performance and the tendency to start working earlier. Instead, the quality of submissions from a single student are linked to how early they started to work. The authors speculate that starting early simply leaves more time to overcome problems or ask for help. Therefore, in this study, our aim is to find out if achievement badges can have a positive effect on students' time management.

III. METHODS

In this experiment, we added achievement badges to the TRAKLA2 [7] online learning environment. We used a between-subject design where students were randomly assigned to the treatment group with the badges visible, and to the control group with the badges hidden. The main view of TRAKLA2 with the badges is shown in Figure 1. Technical implementation of the badge system is described in [34].

A. Materials and Procedure

The experiment was conducted in the Data Structures and Algorithms course at Aalto University in Spring 2012. The course is mandatory for all computer science major students, as well as for minor students from some other study programmes. Computer science majors typically take the course in their first year of their Bachelor's studies, whereas minors take it in the second year. Major and minor students are combined in the analysis because we are interested in the overall effects of badges rather than the differences in the effects of badges in different contexts and student populations. The course had 56 homework exercises that were done online in TRAKLA2. The exercises were divided into 8 rounds with deadlines roughly one week apart. When registering to TRAKLA2, each student was randomly assigned to the treatment or control group with a 50/50 probability.

In a TRAKLA2 exercise, a student is shown a piece of program code (algorithm), and the task is to manipulate a data structure visualization with the mouse in order to simulate the steps that the algorithm would do. Correctness of the solution sequence is checked automatically and the student receives immediate feedback on which steps were correct. Each exercise is worth a certain amount of points. Exercises can be submitted after the deadlines as well, but points of late submissions are reduced by 50%. An example exercise is shown in Figure 2. In the example, students' task is to traverse the graph using Dijkstra's algorithm to find the shortest route.

TRAKLA2 exercises were graded on a scale from 0 (fail) to 5 so that 50% of exercise points yielded a passing grade 1 and 90% of points yielded a grade 5. Major and minor students had separate course instances. The aforementioned grade formed 20% of the final course grade for major students and 30% for minor students. The

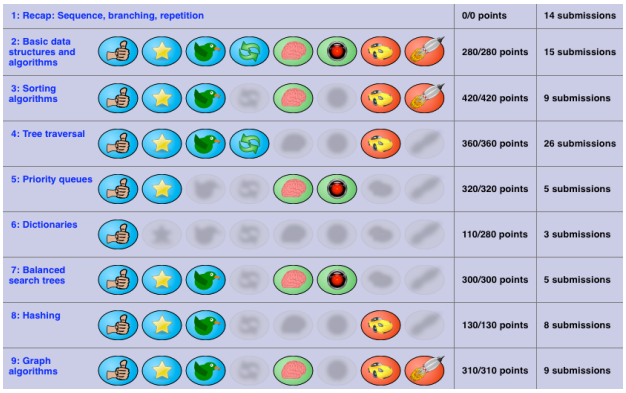


Figure 1. The main view of the TRAKLA2 system with badges.

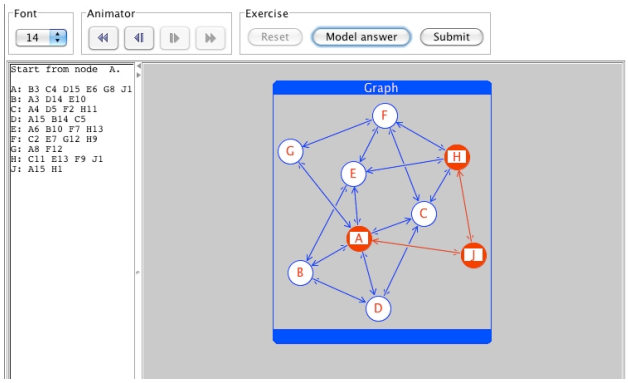


Figure 2. An example of a TRAKLA2 algorithm simulation exercise.

rest of the grade was determined by the final examination (40% weight in both courses), closed labs in majors (40% weight), and a project group work in minors (30% weight). Both courses had the same lectures as well as the same final examination.

The input data for the algorithms are randomly chosen for each student and each attempt. Thus, students can try to solve the exercises as many times as they want, and the best attempt counts. Even if an exercise is already solved correctly, it can be repeated, for example, when preparing for the final examination. The system also provides model solutions that are visualized as algorithm animations. The students may compare their own solution with the model solution after the exercise is submitted. Furthermore, students are allowed to see the model solutions at any time without even submitting. As the input data structures are initialized with random data each time, it is impossible to just copy the model answer and resubmit.

Eight different achievement badges were available on each exercise round. The badges and their criteria are shown in Table I. In the treatment group, students were able to see the badge descriptions in order to know how to earn them. They also saw the available badges as gray and blurry images, making it clear which badges are still waiting to be unlocked. After meeting the criteria for unlocking a badge, the badge icon was made visible in the student's personal TRAKLA2 main page. There were also simple statistics showing how many students in the course had earned each badge.

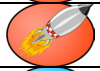
We categorized badges into three categories based on their criteria: *time management, carefulness,* and *learning*. Badges T1, T2 and T3 belong to the *time management* category, and they encourage students to complete the exercises well before the deadline. Badges T2 and T3 are competitive badges, meaning that only a fixed number of students in each round can get them. The *carefulness* category includes badges C1, C2 and C3. They encourage students to think carefully before submitting a solution and to avoid submitting incorrect answers. Finally, badges L1 and L2 form the *learning* category. They encourage students to complete the exercises with full points and to recap them, regardless of the number or time of the submissions.

There are also some connections between the badges in the same category. We wanted to make it possible to collect all the badges and therefore getting a badge with strict criteria will also result in getting the similar but easier badge. X (⇒) Y means that it is possible to get X without Y, but in most cases getting badge X results in getting also badge Y. The implications between the badges are as follows:

- Time management: T3 ⇒ T2 (⇒) T1
- Carefulness: C3 ⇒ C2, C3 ⇒ C1, C2 (⇒) C1
- Learning: L2 ⇒ L1

We did not advertise the badges in the course other than what the treatment group saw in TRAKLA2. This was to reduce contamination where the control group is aware of the badges. We did not provide any external motivation to pursue badges such as extra points. However, if a student from either group wanted to discuss the badges during the course, this was done openly and we responded by telling that there is an ongoing research about the effects of badges in this course.

TABLE I.
DESCRIPTIONS OF THE TIME MANAGEMENT (T), CAREFULNESS (C), AND LEARNING BADGES (L).

ID	Icon	Name	Description
T1		Early Bird	Complete a round with full points at least a week before the deadline.
T2		Fast & Furious	Be in the fastest 30 (majors) / 60 (minors) who complete the round with full points.
T3		Speed Machine	Be in the fastest 10 (majors) / 20 (minors) who complete the round with full points.
C1		Got it!	Get an exercise correct with the first submission (also after deadline).
C2		Brainiac	Get full points from the round and use at most 2 tries for each exercise on average.
C3		Y U No Make Mistakes?	Get full points from all the exercises in the round with the first try.
L1		Mission Accomplished	Get full points from the round.
L2		Recap paceR	Get full points from the round and do all the exercises correctly twice so that there is at least a week between the first and the last correct submission of each exercise.

After the final examination, students were asked to answer a course feedback survey. The questionnaire had Likert-scale questions as well as open-ended questions regarding badges. Students answering the survey were promised two extra final examination points in order to have a higher response rate.

B. *Analysis Methods*

To find out if students chose to pursue or ignore the badges, we analyzed the differences in the number of earned badges between treatment and control groups. Even though the control group did not see the badges, they may have met the criteria unknowingly. Furthermore, we analyzed students' behavior from the log data. Normality of the data was tested with the Shapiro-Wilk test, and non-parametric tests were used for non-normally distributed data. An alpha level of $p < 0.05$ was used for all statistical tests.

We analyzed the numerical data from the course feedback survey in order to get an overall view of students' attitudes towards the badges. Moreover, we explored the grades of the students who were the most motivated by the badges. We also categorized the open-ended feedback with thematic analysis, using a process described by Braun and Clarke [35]. We iteratively categorized the feedback comments into themes that emerged from the data until reaching a consensus.

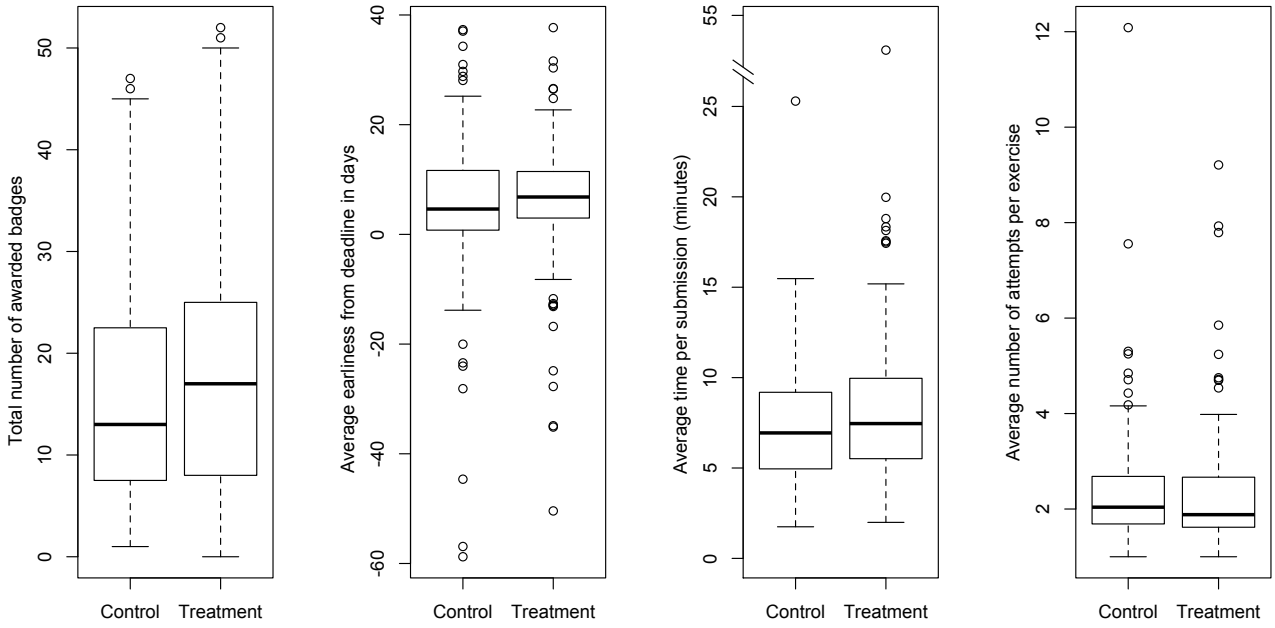
IV. RESULTS

All students who completed at least one TRAKLA2 exercise are included in the comparison between treatment and control groups (N = 281). Students who enrolled in the course but did not submit anything, as well as the lecturer, teaching assistants, etc. personnel are excluded from the results. A total of 86 students from the treatment group completed the feedback survey about the badges.

In order to study whether the students pursued the badges, we calculated the mean number of badges awarded to students in the treatment and control groups. Even though the students in the control group did not see the badges they may have earned them unknowingly. Our null hypothesis is that there are no significant differences between the number of badges earned by the treatment and control groups. If the treatment group consciously aimed to earn badges, there should be a significant increase in the number of earned badges compared to the control group. The number of awarded badges was not normally distributed, so significance was tested with a non-parametric test. We cannot assume that the effect is always positive. It is possible that there are some undesirable effects which cause the treatment group to earn fewer badges than the control group. Therefore, the two-tailed Wilcoxon rank-sum test was used.

A boxplot of the total number of badges per student is shown in Figure 3a. There are students in both groups who have earned a high or low number of badges. The mean is higher in treatment (mean treatment = 18.6, mean control = 15.9), but the difference is not statistically significant (Wilcoxon rank sum test, W=8596, $N_{treatment}$ = 142, $N_{control}$ = 139, p = 0.06, two-tailed). Figure 4 shows the mean numbers of each badge type awarded for students in treatment and control groups. It can be seen that with every badge type, the treatment group has earned more badges on average. The figure also shows the differences between the badge types as some badges have been very easy to get (e.g. C1) whereas some are very difficult (e.g. C3 and L2).

Because of the big differences between the badge types, the total number of badges does not accurately describe the effort put in pursuing the badges. Earning one of the most difficult badges may require multiple



(a) Total number of badges awarded (p=0.06)

(b) Earliness of exercise submissions (p=0.10)

(c) Average time per submission (p=0.04)

(d) Number of attempts per exercise (p=0.20)

Figure 3. Boxplots of students' behavior in treatment and control groups.

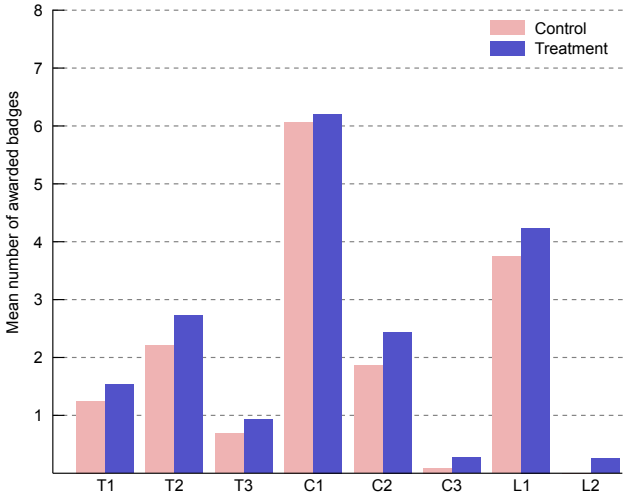


Figure 4. Mean number of awarded badges for each badge type.

times the amount of work than one of the easiest badges. Therefore, we normalized the number of awarded badges and compared the difference of the normalized values between treatment and control groups. Normalization was done by dividing the number of badges with the mean number of badges in the control group for each badge type. With the normalized data, the difference was statistically significant (mean treatment = 45.45, mean control = 8.00, Wilcoxon rank sum test, W = 7950, $p < 0.01$).

Figure 5 offers a more detailed view to the number of badges awarded on each round for the treatment and control groups. The topic of each exercise round and the mean points earned overall are also shown in the figure.

A. *Time Management*

We studied how early each student started working on the exercises by calculating the mean time from first submission of each round to the deadline (hereafter *earliness*). Boxplots for treatment and control groups are shown in Figure 3b. Differences are not statistically significant (Wilcoxon rank sum test, W=8736, p = 0.10, two-tailed) although the treatment group started slightly earlier (4.7 days in control, 6.0 days in treatment). To get an idea of students' tendency to work near the deadline, we calculated the number of submissions on the last day and after the deadline. We found that 33.2% and 39.1% of submissions were submitted within 24 hours before the deadline in treatment and control groups, respectively. Moreover, 7.6% of submissions in treatment and 10.1% in control came late (excluding recap, i.e. submissions after the deadline to exercises that had already been completed).

B. *Carefulness*

One of the goals of using badges was to reduce trial and error problem solving. We studied this by measuring the mean time spent per submission (Figure 3c) and the number of attempts per exercise (Figure 3d). The spent time was estimated by measuring the time between consecutive submissions within one session. A session is defined as a sequence of submissions with less than 2 hours between consecutive submissions. Students in the treatment group spent more time per submission on average (mean treatment = 8.59 min, mean control = 7.26 min). The difference is statistically significant (Wilcoxon rank sum test, W = 8265, p = 0.04, two-tailed). They also used less attempts (mean treatment = 2.33, mean control = 2.41) but the difference is not significant (Wilcoxon rank sum test, W = 10734, p = 0.20, two-tailed).

C. *Learning*

The point distributions of treatment and control groups are shown in Figure 6c. The distribution has slightly shifted in favor of the treatment group although the difference is not statistically significant (Pearson's χ^2

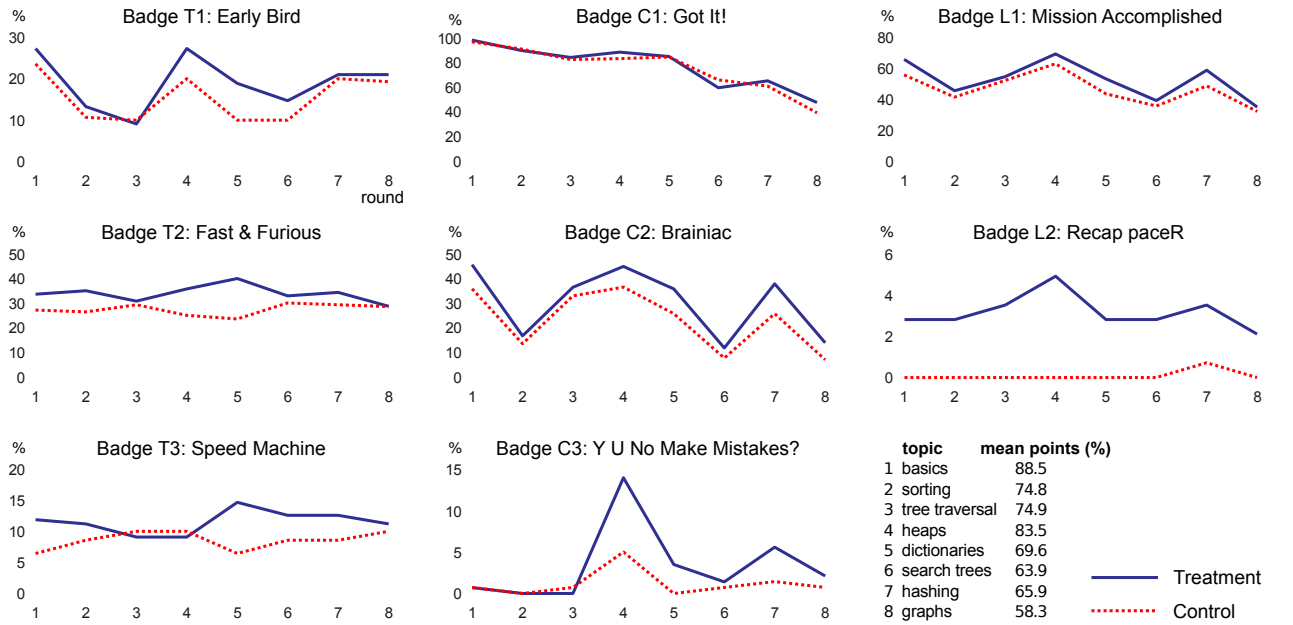
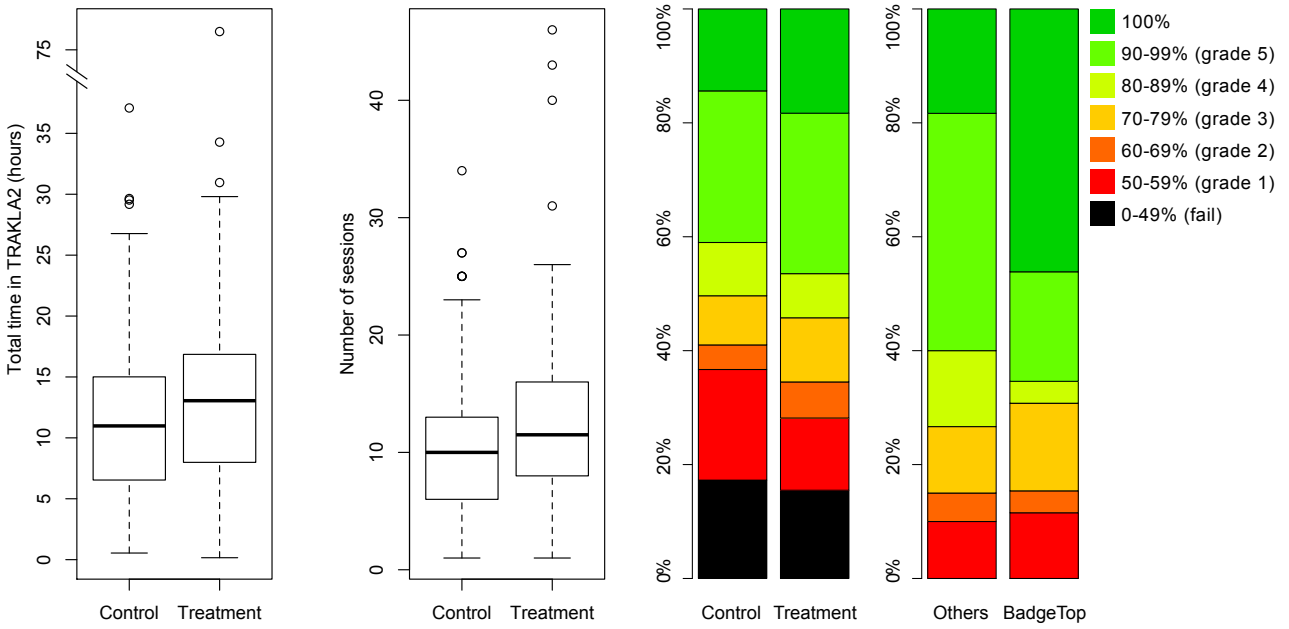


Figure 5. Percentage of students who earned each badge, round by round. In each subfigure, the X-axis represents the round number and the Y-axis represents the percentage of students that earned the corresponding badge. The mean points of each round are shown to illustrate the relative difficulty of the rounds.



(a) Total time spent in TRAKLA2 (p=0.02)

(b) Number of sessions in TRAKLA2 (p=0.01)

(c) Points on the course (p=0.66)

(d) Points of BadgeTop/Others (p=0.08)

Figure 6. Students' behavior and grades in TRAKLA2.

test, df = 6, χ^2 = 4.09, p = 0.66). Although 90% of points is enough for grade 5, 18% of students in treatment and 14% in control reached full points.

In addition to the TRAKLA2 grade, we were interested in the amount of time students spent in the learning environment. The total time spent for TRAKLA2 exercises is shown in Figure 6a (control: 11.4 hours, treatment: 13.4 hours). The difference is statistically significant (Wilcoxon rank sum test, W = 8039.5, p = 0.02, two-tailed).

The Recap paceR badge (L2) encouraged students to revisit TRAKLA2 after they had already completed the exercises. Therefore, we were also interested in whether the badges affected the number of sessions students had in the system. The number of exercise sessions over the course for each student is shown in Figure 6b. Students in the treatment group had 12.6 sessions on average while students in the control group had 10.6. The difference is statistically significant (Wilcoxon rank sum test, W = 8195, p-value = 0.01, two-tailed).

D. Numerical Feedback

Results of the numerical feedback questions are shown in Table II. The majority of the students responded that they found the badges motivating and that they had an effect on their behavior. Furthermore, the majority was satisfied with the criteria and visual appearance, and thought that badges should be used again in the following year. Moreover, only a small minority of 8% reported that badges disturbed their work.

In order to study the characteristics of the students who were the most motivated by the badges, we divided students into two groups based on their answers to the *"I found the badges motivating"* feedback question. Students who answered *"Completely agree"* were put into group *BadgeTop* (30% of population, N=26) and the rest into group *Others* (N=60).

The point distributions of *BadgeTop* and *Others* are shown in Figure 6d. The *BadgeTop* has a clearly higher proportion of students with full points. The distributions do not differ significantly (Pearson's χ^2 test, df = 5, χ^2 = 9.72, p = 0.08). However, the proportion of students with full points out of those with maximum grade is significantly higher (Pearson's χ^2 test, df = 1, χ^2 = 5.99, p = 0.01). Note that this comparison is not between randomized groups and does not imply that the difference is caused by badges.

E. Open-Ended Feedback

54 students gave open-ended feedback about the badges. The answers were categorized iteratively using thematic analysis until a consensus about the themes was reached among the authors. The themes are described in the following sections with examples representing typical answers on the theme. The feedback was in Finnish but has been translated into English. If an answer addressed multiple themes, it was split and the parts were categorized separately.

1) Increased Motivation

Some students reported that they found badges motivating. However, they did not necessarily state that their behavior was affected.

— *"I got a feeling like in a car game. It was funny how much I wanted to unlock them."*

— *"On a couple of first rounds, the badges excited me, but quite soon I got bored with them."*

2) Affected Behavior

Some students reported that badges caused a change in their behavior. The reported effects were in line with the badge criteria and no one reported negative effects.

TABLE II.
NUMERICAL FEEDBACK ANSWERS ABOUT THE ACHIEVEMENT BADGES (N=86).
(0 = COMPLETELY DISAGREE, 1 = SOMEWHAT DISAGREE, 2 = CANNOT SAY, 3 = SOMEWHAT AGREE, 4 = COMPLETELY AGREE)

Feedback item	0	1	2	3	4
I found the badges motivating.	7%	8%	10%	**44%**	30%
Badges disturbed my work.	**76%**	8%	8%	7%	1%
Trying to achieve badges had an effect on my behavior.	17%	12%	8%	**41%**	22%
Visual look of the badges was good.	0%	14%	8%	**51%**	27%
I was satisfied with the criteria for awarding badges.	0%	15%	19%	**44%**	22%
I think that badges should be used in TRAKLA2 in the next year's course as well.	1%	0%	20%	34%	**45%**

— "Because of the badges, starting from round 4 I did all the remaining rounds in one evening in order to get the Early Bird badges ! :)"
— "I often checked the answer many times before submitting to make sure it's correct."

3) Indifferent

Some students reported that they ignored the badges or that badges had no effect on their motivation or behavior.
— "They did not give me any motivation. I did [the exercises] just for the points and to learn."
— "The badges alone do not encourage much to do the exercises, because this is not primary school."

4) Concrete Reward

A few students commented on the relationship between the badges and course grading. Some wished that badges had an effect on grading and one was pleased that they did not. Some students commented that badges would have a stronger effect if there was a concrete reward, but they did not explicitly say whether they would prefer that.
— "Extra points or a concrete reward would be good instead [of just a badge]."
— "If there was some compensation in the form of points then I suppose they would motivate more to do the exercises."
— "It's good that the badges didn't affect grading."

5) Considers badges generally a good idea

Some students made positive comments about badges or gamification in general but did not necessarily report an effect on their motivation or behavior.
— "They were good and fun. Keep them!"
— "Badges suit these types of exercises very well, and they should be used in other courses as well."

6) Comments the Badge Criteria

Some students criticized that some badges were pointless or too difficult to earn but did not criticize badges or gamification per se. The same student could give negative feedback about some badges and positive about others. A few students commented that the "Y U NO MAKE MISTAKES" badge was particularly difficult to earn but did not comment whether they found it as a good or a bad thing.
— "It seemed that badges were given pretty much for the same things. There should be more variation in how you get badges."
— "I didn't understand the speed badges. Why is speed even a good thing? ... All correct at the first go badge and the recap badge were especially good!"
— "Y U NO MAKE MISTAKES?! was really difficult to get, I only got it in one round."

7) Social Aspect

A few students reported that they enjoyed comparing badges with others or studying the statistics. Another few wished that the social aspects would have a bigger role in the system.
— "Nice to see how others have done and nice to collect [badges]"
— "Results could be shown more widely and publicly?"

8) Technical Criticism

A few students had some confusion about the functionality of the system or they suspected that the system malfunctioned[2].
— "In the last round I didn't get full points so I only got the first badge - I wonder if the problem was in my points or somewhere else, that made me wonder."
— "I guess the Recap pacer doesn't work correctly? At least I didn't get a badge from a few old rounds even though I did all the exercises again before the final examination."

9) Control group

Some students from the control group gave feedback about achievement badges even though the question was addressed to the treatment group. A few of them had a clearly negative attitude towards badges or gamification while a few commented them positively.
— "I didn't notice them. I don't think achievements are suitable for 'measuring and comparing' learning because people learn in different ways. It's not good to add competitiveness."
— "I was only happy that I didn't have them, because there would have been a constant pressure to try to get them and there was enough hurry anyway."
— "Achievements are always fun, I didn't know you get to fulfill your inner perfectionist in TRAKLA2 too."
— "Achievements/trophies are always a good addition."

V. DISCUSSION

Our results show that the achievement badges had an impact on the students' behavior. With each badge type, the treatment group earned more badges than the control group (Figure 4). Furthermore, the normalized total number of earned badges was significantly higher in the treatment group, suggesting that the students in the treatment group were pursuing the badges on purpose.

Figure 5 shows the percentage of students who earned each badge type on each round. It seems that the biggest

[2] We manually inspected these cases and confirmed that the system had worked correctly.

differences between the treatment and control groups are in the badges that were hard to earn. Badge C1 (Got it!) was earned by the majority of the students from both groups, and there are no noticeable differences between them. In contrast, Badge C3 (Why Y No Make Mistakes?) and L2 (Recap paceR) were difficult to earn, and a much larger difference is observed. The reason might be that the additional challenge provided by harder badges is found motivating and therefore worth pursuing. Moreover, there may be a ceiling effect concerning the easier badges. Furthermore, some badges may be mutually exclusive. Reaching for a position in the top fastest solvers might cause those students to choose speed over carefulness. On the other hand, being cautious about making mistakes will probably lead to missing the competitive time management badges. Therefore, when designing future badge studies, it would be beneficial to select the badge criteria so that they do not cause conflicting effects.

A. Time Management

Over the years, we have noticed that a big part of the submissions to TRAKLA2 come very close to the deadlines. In order to address this issue, we introduced three *time management* badges that rewarded students for completing tasks early. Badges T2 and T3 were competitive so that only a fixed amount of fastest students were given the badge, while badge T1 (Early Bird) was awarded to everyone who completed a round at least a week before the deadline. In practice, pursuing the competitive badges also resulted in getting badge T1.

Our results suggest that the treatment group started doing the exercises slightly earlier although the difference is not statistically significant (Figure 3b). One student even reported doing the majority of the exercises in one night because of the Early Bird badges. However, going from one extreme to another is not necessarily desirable. By doing that, some students completed exercises long before the topics were introduced in the lectures. Furthermore, some students criticized that the increased competition might hinder learning. It seems that badges can be used to some extent to mitigate the deadline-oriented behavior, but the badge criteria must be more carefully chosen to avoid side effects.

B. Carefulness

In previous studies [28], we have noticed that having no resubmission limits in the exercises leads some students to resort to trial and error behavior. Three of the badges (C1, C2 and C3) rewarded students for not submitting incorrect answers and thus avoiding trial and error problem solving. Students in the treatment group spent significantly more time between submissions (Figure 3c), which suggests they checked the solutions more carefully before submitting. This behavior is also reported in the open-ended feedback.

It seems that badges can be used – at least to some extent – to reduce the trial and error phenomenon. However, a badge awarded for submitting exercises without any mistakes (C3) may not have been the best motivator against it. Once a student makes the first mistake, the badge is lost and it no longer offers an incentive to avoid resubmissions in the rest of the exercises in that round. Furthermore, the students who are capable of completing a whole round without any mistakes are not likely the ones prone to iterating in the first place. Thus, it is important to balance the achievement criteria so that they are realistically reachable by the group whose behavior we are trying to change. Optimally, the challenge provided by the badge and a student's skill level would be in a good balance, thus providing favorable conditions for *flow* [36].

C. Learning

There were two badges (L1 and L2) in the *learning* category that encouraged to complete exercise rounds with full points and to revisit the exercises afterwards. The treatment group had a significantly higher number of sessions in TRAKLA2 (Figure 6b), possibly because of pursuing the Recap paceR badge. The badge itself required considerable effort to achieve, requiring students to redo a complete round with full points. In the control group, only one student completed one round as recap. In the treatment group, 9 students completed at least one round as recap, and some of them completed several rounds. Even though the number of students who got at least one recap badge was not high in the treatment group either, it can be clearly seen that some students were pursuing the badge instead of getting it as a side effect of something they would have done anyway. Two students went as far as to complete every TRAKLA2 exercise twice with full points in order to earn all available recap badges.

In our opinion, it is beneficial to redo TRAKLA2 exercises as recap. However, the requirements for getting badge L2 were not necessarily optimal for efficient learning. It would be better to focus on the exercises that caused difficulties the first time around, rather than trying to complete whole exercise rounds. Therefore, relaxing the achievement criteria could also motivate more students to pursue the badge.

Students in the treatment group got slightly better grades from the exercises, although the difference was not statistically significant. They also spent significantly more time in total in the system (Figure 6a). Similarly, an increase in the amount of work done by students was also observed by Denny [16] in his experiment about the effects of badges. We believe that the increase in time-on-task might contribute to learning even if we could not observe statistically significant differences in the grades. Furthermore, the grades do not tell the whole truth about learning because they fail to measure important aspects such as retention and transfer of learned skills.

Interestingly, both groups had many students who earned maximum points from the exercises even though only 90% was required for the highest grade. As Montola et al. [12] point out, *completionism* can be one of the sources of motivation in gamification. Thus, collecting maximum points might be perceived as a game-like challenge in itself by some students. Naturally, students may also want to do all the exercises because of an intrinsic motivation towards the subject.

D. Feedback

We analyzed the numerical and open-ended feedback regarding badges and compared them with the findings from the log data. The themes found in the open-ended feedback support the observed effects that badges caused on students' behavior. Many students reported that they were motivated by the badges and that the badges affected their behavior, and statistically significant improvements were in fact observed in the behavior of the treatment group.

In the control group, some students made strong arguments against badges. In the treatment group, on the other hand, students' attitudes towards badges were positive or indifferent, although some implied that they belong to primary school or appeal to the "Angry Birds generation". It appears that some students were prejudiced against gamification but, interestingly, as strong negative attitudes were not expressed by students in the treatment group. Although none of the students in the treatment group reported negative effects of the badges in the open-ended feedback, 8% of them stated in the numerical feedback that badges disturbed their work.

Some students criticized the criteria for awarding badges. As suggested by Nicholson [19], if the game elements are perceived meaningful by the students, they can have a positive effect on the internal motivation. Therefore, the badge criteria have a big role in the success of the gamification. The fact that these students were not happy with all of the criteria may have reduced their motivation to collect badges. It is also possible that having even one badge criterion where the student disagrees with the beneficiality of the required behavior, can hinder the motivation to pursue other badges as well because the whole system is not seen as a worthwhile challenge.

We did not explicitly ask whether the badges should be tied to course grading. However, some students spontaneously suggested that the badges would have a stronger effect if they were. On the other hand, one student said that it was good that the badges did not affect grading. Having the badges affect the grades would change their role significantly. Instead of offering additional goals and challenges, we would explicitly reward certain behavior and implicitly punish students who do not want to pursue the badges. The fact that other students seem to like the idea that the badges would be tied to the grading while others do not, is possibly connected to the fact that people have different preferences regarding study goals, rewards, and competitions. Moreover, as Huotari and Hamari [1] point out, one of the defining aspects of gameful experience is that it is voluntary and carried out by having intrinsic motivation. In our opinion, badges are most beneficial when they offer a voluntary challenge that encourages good study practices rather than making them part of the course grading. We also believe that making them part of the course grading would increase the controlling aspect of the badges and possibly have an undermining effect on students' intrinsic motivation as suggested by the *cognitive evaluation theory* [24]. In addition, it is worth considering to make it possible for students to turn the badges off.

E. Validity Threats

Many of the students earned maximum points also in the control group. This indicates that there might be a ceiling effect when analyzing the badges' effect on learning. Moreover, exercise points can already be perceived as a gamification mechanism and therefore the effect of badges may not stand out.

We estimated the used time from the intervals between submissions. In this method, the time used for the first attempt to an exercise cannot be estimated. Furthermore, we do not know if time was spent solving the exercise or doing something unrelated. However, the values are compared to the control group which provides a baseline.

Students were rewarded two final examination points for participating in the course feedback survey. It is possible that some of the answers to the numeric questions were given without thought just to collect the points. Open-ended feedback was given by 55% of the students in the treatment group who answered the survey. In addition, seven students from the control group provided feedback even though they were not asked to do it. It is likely that certain types of people are more inclined to give open-ended feedback than others, which biases the results. Furthermore, feedback was not given by any students who dropped out of the course, which causes some bias in the studied population.

We have studied badges in the context of a single course, and the results are not necessarily directly generalizable to other contexts. Differences in the culture, course topic, arrangements, grading policy, etc. are likely to have an impact on the effects of badges. Therefore, more studies are needed to further understand the effect of achievement badges in education.

VI. CONCLUSIONS

In this experiment, we added achievement badges to the TRAKLA2 online learning environment, and studied the effects on students' behavior from the system logs. In addition, we analyzed students' attitudes towards badges from the course feedback survey. Our results show that achievement badges had a statistically significant impact on some aspects of students' behavior, and that students had generally positive attitudes towards them. Students who had the badges spent more time per exercise, suggesting that they thought about the solutions more thoroughly before submitting them. This behavior was also reported by some students in the open-ended feedback. Students in the treatment group also had a higher number of sessions and spent more time in total in the learning environment. Furthermore, the majority of the students gave positive feedback and reported being motivated by the badges.

We were able to change the behavior of some students for the better by rewarding them with badges. However, it is possible that some badges encouraged undesirable behavior as well. For example, targeting the competitive time management badges might have reduced

carefulness. Therefore, more research is needed in balancing the achievement criteria so that they maximize beneficial learning practices while minimizing harmful side effects.

Overall, the achievement badges had an impact on the students even though they merely provided voluntary extra goals and did not affect the course grading. Based on our results, badges seem like a promising way to motivate students and to encourage desirable study practices. Furthermore, badges can be a cost-efficient way to give automated feedback about study practices, without requiring additional effort from the teacher. However, the applied methods should be carefully chosen in order to fully benefit from the engaging elements and to avoid gamification from being just unnecessary eye candy.

A. *Future Work*

In this study, it was impossible to reliably measure the effects of individual badges because each badge is likely to affect multiple aspects of students' behavior. Studying achievement badges in massive open online courses (MOOCs) with thousands of students would make it possible to divide students into multiple treatment groups with different sets of badges, in order to study what kind of badges have the strongest effects.

In our implementation, students were able to see their own badges and simple statistics that showed the overall number of badges unlocked in the whole course. However, the social aspect of badges was missing. Allowing students to show their badges to others could make them more desirable and might motivate more students to pursue them. On the other hand, this could increase the competition on the course which can be demotivating for some students. Thus, more research is needed in order to understand how individual differences affect students' responses to gamification. Furthermore, because it is important to balance the badge criteria so that they are both challenging and reachable, ways to dynamically adapt the badge criteria for each student would be an interesting topic for further study.

REFERENCES

[1] K. Huotari and J. Hamari, "Defining gamification: A service marketing perspective," in *Proceeding of the 16th International Academic MindTrek Conference*, ser. MindTrek '12. New York, NY, USA: ACM, 2012, pp. 17–22.

[2] C. Muntean, "Raising engagement in e-learning through gamification," in *Proceedings of the 6th International Conference on Virtual Learning ICVL*, 2011, pp. 323–329.

[3] S. Deterding, D. Dixon, R. Khaled, and L. Nacke, "From game design elements to gamefulness: defining "gamification"," in *Proceedings of the 15th International Academic MindTrek Conference: Envisioning Future Media Environments,* ser. MindTrek '11. New York, NY, USA: ACM, 2011, pp. 9–15.

[4] A. Domínguez, J. S. de Navarrete, L. de Marcos, L. Fernández-Sanz, C. Pagés, and J.-J. Martínez-Herráiz, "Gamifying learning experiences: Practical implications and outcomes," *Computers & Education,* vol. 63, pp. 380 – 392, 2013. http://dx.doi.org/10.1016/j.compedu.2012.12.020

[5] L. de Marcos, A. Domínguez, J. S. de Navarrete, and C. Pagés, "An empirical study comparing gamification and social networking on e-learning," *Computers & Education,* vol. 75, no. 0, pp. 82 – 91, 2014. http://dx.doi.org/10.1016/j.compedu.2014.01.012

[6] J. Hamari, J. Koivisto, and H. Sarsa, "Does gamification work? a literature review of empirical studies on gamification," in *Proceedings of the 47th Hawaii International Conference on System Sciences,* Hawaii, USA, 2014.

[7] L. Malmi, V. Karavirta, A. Korhonen, J. Nikander, O. Seppälä, and P. Silvasti, "Visual algorithm simulation exercise system with automatic assessment: TRAKLA2," *Informatics in Education,* vol. 3, no. 2, pp. 267–288, 2004.

[8] L. Hakulinen, T. Auvinen, and A. Korhonen, "Empirical study on the effect of achievement badges in TRAKLA2 online learning environment," in *Learning and Teaching in Computing and Engineering (LaTiCE),* Macau, 2013, pp. 47–54.

[9] S. Abramovich, C. Schunn, and R. M. Higashi, "Are badges useful in education?: it depends upon the type of badge and expertise of learner," *Educational Technology Research and Development,* vol. 61, no. 2, pp. 217–232, 2013. http://dx.doi.org/10.1007/s11423-013-9289-2

[10] L. Haaranen, P. Ihantola, L. Hakulinen, and A. Korhonen, "How (not) to introduce badges to online exercises," in *Proceedings of the 45th ACM Technical Symposium on Computer Science Education,* ser. SIGCSE '14. New York, NY, USA: ACM, 2014, pp. 33–38.

[11] J. Hamari and V. Eranti, "Framework for designing and evaluating game achievements," in *Proceedings of DiGRA 2011 Conference: Think Design Play*, Hilversum, Netherlands, 2011.

[12] M. Montola, T. Nummenmaa, A. Lucero, M. Boberg, and H. Korhonen, "Applying game achievement systems to enhance user experience in a photo sharing service," in *Proceedings of the 13th International MindTrek Conference: Everyday Life in the Ubiquitous Era,* ser. MindTrek '09. New York, NY, USA: ACM, 2009, pp. 94–97.

[13] B. J. Morris, S. Croker, C. Zimmerman, D. Gill, and C. Romig, "Gaming science: the Gamification of scientific thinking," *Frontiers in psychology,* vol. 4, 2013.

[14] J. Henderlong and M. R. Lepper, "The effects of praise on children's intrinsic motivation: a review and synthesis." *Psychological bulletin,* vol. 128, no. 5, p. 774, 2002. http://dx.doi.org/10.1037/0033-2909.128.5.774

[15] A. Decker and E. L. Lawley, "Life's a game and the game of life: How making a game out of it can change student behavior," in *Proceeding of the 44th ACM Technical Symposium on Computer Science Education,* ser. SIGCSE '13. New York, NY, USA: ACM, 2013, pp. 233–238.

[16] P. Denny, "The effect of virtual achievements on student engagement," in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems,* ser. CHI '13. New York, NY, USA: ACM, 2013, pp. 763–772.

[17] R. McDaniel, R. Lindgren, and J. Friskics, "Using badges for shaping interactions in online learning environments," in *Professional Communication Conference (IPCC),* 2012 IEEE International. IEEE, 2012, pp. 1–4.

[18] A. Anderson, D. Huttenlocher, J. Kleinberg, and J. Leskovec, "Engaging with massive online courses," in *Proceedings of the 23rd International Conference on World Wide Web,* ser. WWW '14. Republic and Canton of Geneva, Switzerland: International World Wide Web Conferences Steering Committee, 2014, pp. 687–698.

[19] S. Nicholson, "A user-centered theoretical framework for meaningful gamification," in *Games+Learning+Society 8.0,* Madison, WI, 2012.

[20] J. Lee and J. Hammer, "Gamification in education: What, how, why bother?" *Academic Exchange Quarterly,* vol. 15, no. 2, p. 146, 2011.

[21] M. Robertson. (2010, Oct. 6). *Can't play, won't play* [Online; accessed 15-May-2014]. Available: http://www.hideandseek.net/2010/10/06/cant-play-wont-play

[22] I. Bogost. (2011, Aug. 8). *Gamification is bullshit* [Online; accessed 15-May-2014]. Available: http://www.bogost.com/blog/gamification_is_bullshit.shtml

[23] E. L. Deci, R. Koestner, and R. M. Ryan, "Extrinsic rewards and intrinsic motivation in education: Reconsidered once again," *Review of Educational Research,* vol. 71, no. 1, pp. 1–27, 2001. http://dx.doi.org/10.3102/00346543071001001

[24] E. L. Deci and R. M. Ryan, *Intrinsic Motivation and Self-Determination in Human Behavior.* Plenum, 1985. http://dx.doi.org/10.1007/978-1-4899-2271-7

[25] E. D. Mekler, F. Brühlmann, K. Opwis, and A. N. Tuch, "Do points, levels and leaderboards harm intrinsic motivation?: an empirical analysis of common gamification elements," in *Proceedings of the First International Conference on Gameful Design, Research, and Applications,* ser. Gamification '13. New York, NY, USA: ACM, 2013, pp. 66–73.

[26] K. Weaver, S. Garcia, and N. Schwarz, "The presenter's paradox," *Journal of Consumer Research,* vol. 39, no. 3, pp. 445–460, 2012. http://dx.doi.org/10.1086/664497

[27] S. Edwards, "Using software testing to move students from trial-and-error to reflection-in-action," *ACM SIGCSE Bulletin,* vol. 36, no. 1, pp. 26–30, 2004. http://dx.doi.org/10.1145/1028174.971312

[28] V. Karavirta, A. Korhonen, and L. Malmi, "On the use of resubmissions in automatic assessment systems," *Computer Science Education,* vol. 16, no. 3, pp. 229 – 240, September, 2006. http://dx.doi.org/10.1080/08993400600912426

[29] L. Malmi, V. Karavirta, A. Korhonen, and J. Nikander, "Experiences on automatically assessed algorithm simulation exercises with different resubmission policies," *Journal of Educational Resources in Computing,* vol. 5, no. 3, September, 2005. http://dx.doi.org/10.1145/1163405.1163412

[30] N. J. Falkner and K. E. Falkner, "A fast measure for identifying at-risk students in computer science," in *Proceedings of the ninth annual international conference on International computing education research.* ACM, 2012, pp. 55–62.

[31] S. H. Edwards, J. Snyder, M. A. Pérez-Quiñones, A. Allevato, D. Kim, and B. Tretola, "Comparing effective and ineffective behaviors of student programmers," in *Proceedings of the fifth international workshop on Computing education research workshop.* ACM, 2009, pp. 3–14.

[32] J. B. Fenwick Jr, C. Norris, F. E. Barry, J. Rountree, C. J. Spicer, and S. D. Cheek, "Another look at the behaviors of novice programmers," *ACM SIGCSE Bulletin,* vol. 41, no. 1, pp. 296–300, 2009. http://dx.doi.org/10.1145/1539024.1508973

[33] N. Michinov, S. Brunot, O. L. Bohec, J. Juhel, and M. Delaval, "Procrastination, participation, and performance in online learning environments," *Computers & Education,* vol. 56, no. 1, pp. 243 – 252, 2011. http://dx.doi.org/10.1016/j.compedu.2010.07.025

[34] L. Haaranen, L. Hakulinen, P. Ihantola, and A. Korhonen, "Software architectures for implementing achievement badges – practical experiences," in *Learning and Teaching in Computing and Engineering (LaTiCE),* Kuching, Malaysia, 2014, pp. 41–46.

[35] V. Braun and V. Clarke, "Using thematic analysis in psychology," *Qualitative Research in Psychology,* vol. 3, no. 2, pp. 77–101, 2006. http://dx.doi.org/10.1191/1478088706qp063oa

[36] M. Csikszentmihalyi, *Flow: The psychology of optimal experience.* New York: Harper and Row, 1990.

AUTHORS

Lasse Hakulinen is a Doctoral Candidate in the Learning + Technology research group at Aalto University. He received the Master of Science (Tech) degree in Computer Science and Engineering from Aalto University, Finland, in 2010. His research focuses on gameful approaches to enrich computer science education (e.g. gamification and alternate reality games). (e-mail: lasse.hakulinen@aalto.fi).

Tapio Auvinen is a Doctoral Candidate in the Learning + Technology research group at Aalto University. He received the Master of Science (Tech) degree in Computer Science and Engineering from Aalto University, Finland, in 2010. His research focuses on supporting self-regulating learning in online learning environments. (e-mail: tapio.auvinen@aalto.fi).

Ari Korhonen received the D. Sc. (Tech) degree in Computer Science and Engineering from the Helsinki University of Technology, Finland, in 2003. He is a Lecturer and Researcher at Aalto University and an Adjunct Professor at the University of Turku. His research focuses on developing online learning environments, learning analytics, and automatic assessment. (e-mail: ari.korhonen@aalto.fi).

This article is an extended and modified version of a paper presented at the International Conference on Learning and Teaching in Computing and Engineering (LaTiCE 2013), held 21-24 March 2013, Macau. This work was supported by the Finnish Funding Agency for Technology and Innovation (Tekes) under the grant number 40312/12. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of Tekes.

Submitted 22 October 2014. Published as resubmitted by the authors 21 February 2015.