

Data-driven goal setting: Searching optimal badges in the decision forest

Julian Langenhagen

Goethe University Frankfurt, Theodor-W.-Adorno-Platz 4, Frankfurt am Main, 60323, Hessen, Germany



ARTICLE INFO

Keywords:

Goal setting
Gamification
Badges
Learning analytics
Educational data mining
Decision trees

ABSTRACT

Goal setting is vital in learning sciences, but the scientific evaluation of optimal learning goals is underexplored. This study proposes a novel methodological approach to determine optimal learning goals. The data in this study comes from a gamified learning app implemented in an undergraduate accounting course at a large German university. With a combination of decision trees and regression analyses, the goals connected to the badges implemented in the app are evaluated. The results show that the initial badge set already motivated learning strategies that led to better grades on the exam. However, the results indicate that the levels of the goals could be improved, and additional badges could be implemented. In addition to new goal levels, new goal types are also discussed. The findings show that learning goals initially determined by the instructors need to be evaluated to offer an optimal motivational effect. The new methodological approach used in this study can be easily transferred to other learning data sets to provide further insights.

1. Introduction

How many steps should you walk a day? If 10,000 is the first number that comes to your mind, it is probably because many fitness devices and media sources recommend this number as a default and give this figure the appearance of a proven truth. However, this threshold was not determined with scientific evidence but originated in a Japanese marketing campaign from the 1960s [1]. Recent studies have shown that this figure is rather too high. According to the latest research, the optimal number of steps lies between 7000 and 8000 steps [2–4].

Comparable objectives can also be found in learning sciences. A well-known example with the identical number is the “10,000-hour rule” which was formulated by Malcolm Gladwell in his best-selling book “Outliers”. The rule states that it takes 10,000 hours of deliberate practice to master a skill in a given field [5, pp. 35ff.]. Although Gladwell quotes a scientific study to support his claim [6], one of the study’s authors commented in a later publication that more factors need to be considered and there is “no evidence for a magical number” [7, p. 2]. So there seems to be a demand for concrete target numbers on the one hand, but they seem not to be generated with research so far on the other hand. Moreover, even corresponding research that validates them ex-post is scarce [8]. This claim is also valid for the fields of learning analytics and educational data mining [9]. However, such numeric thresholds can be very interesting for students, especially those who have to

study for several courses simultaneously and aim for an optimal distribution of their learning effort. How much additional study time does a student need to achieve a better grade? This information can also be important because some students only want to pass an exam while others want to achieve the highest grade possible. In simplified terms, there is no point in answering 100 more questions in a learning tool if experience has shown that 1000 more are needed to make a difference. In this area of research, it is not only important to determine the threshold of the goals but also the ways the goals are visualized to the users.

Gamification – “the use of game design elements in non-game contexts” [10, p. 10] – is often used in education as a motivational tool, and badges are one game element used to visualize progress and goals [11, p. 91]. Together with points and leaderboards, badges are among the most used elements in game-based learning [12]. Yet, little is known about their optimal design in the educational domain. Optimal design in this context means that a badge can only be earned by following a learning strategy that leads to improved learning outcomes. The inspiration for the design of badges in game-based learning often comes from entertainment games [13,14]. However, the goals of such games are usually different from those of learning tools used in school or higher education [15]. This difference also applies to commercial learning apps such as Duolingo. In Duolingo, for example, it is possible to earn a badge called “Weekend Warrior” if you use the app on a Saturday or Sunday and a badge called “Photogenic” if you upload a profile

E-mail address: langenhagen@econ.uni-frankfurt.de

<https://doi.org/10.1016/j.teler.2023.100072>

Received 21 February 2023; Received in revised form 30 May 2023; Accepted 24 June 2023

2772-5030/© 2023 The Author(s). Published by Elsevier B.V. This is an open access article under the CC BY-NC-ND license

(<http://creativecommons.org/licenses/by-nc-nd/4.0/>)

picture.¹ These two badges are presumably intended to maximize usage time or identification with the app and are not necessarily aimed at an optimal learning strategy. Since designers of game-based learning tools may use such apps as a blueprint, they run the risk of copying badges that motivate non-optimal learning strategies. This study should help educators that use badges in their learning tools to evaluate whether their chosen design motivates learning strategies that lead to better learning outcomes.

The data set used in this study consists of usage data from a gamified learning app that has been in use for four semesters in a management accounting course at a large German university. Among other game elements, the quiz app uses badges with corresponding goals to motivate the students to learn. The required achievements to earn the badges include specific amounts of consecutive usage days or answered questions. The related goals were developed based on learning theories and practical experiences from past lectures but without any empirical validation. The purpose of this study is to answer the question of whether students who aligned their learning strategy to the given badge goals performed better on the exam. Based on the results, the underlying metrics but also the levels of the badges will be discussed. Following the definition of John Gerring, this case study is “an intensive study of a single unit for the purpose of understanding a larger class of units” [16, p. 341]. In other words, in this study, the badges that were implemented in the before-mentioned app are analyzed with a procedure that can be easily transferred to other comparable settings. However, the numeric results of this study cannot be generalized to other settings without thorough consideration, as the badges considered in this study are case-specific. For example, the badge awarded for the amount of answered questions in the data set has three levels. The highest level is earned when a student answers 1000 questions. Unless a large proportion of the top students answered that many questions in the app, the conclusion might be that the level needs to be lowered. If the level is too high, earning that badge could even be a predictor for a bad performance in the exam, i.e., students who answer too many questions might overestimate the app as a learning tool and neglect other learning materials. Past studies have shown that learning performance does not necessarily increase with learning time [17]. In addition, the analysis might reveal optimal learning strategies that have not yet been incentivized by any of the badges.

Recently conducted meta-studies consistently report an overall positive significant small to medium effect size of game-based learning on learning outcomes [18–20]. As badges belong to the most used game elements [12], it can be assumed, that badges have the power to motivate students to better learning outcomes. The motivational effect of badges was documented in multiple studies [e.g., 21–24] and motivation is considered as an essential factor for academic success [25].

In the initial example regarding the number of steps, it is relatively straightforward that the activity itself – walking – is beneficial to health and that it is merely a question of the number of steps. However, in learning sciences, it is not so obvious which actions lead to an increase in knowledge or help in passing an exam. One central question in this area is whether a deep or a surface learning strategy is better suited for preparing for an exam [26]. A deep learning approach means critically examining the topics in question and linking them to already known topics. A surface learning approach is more like uncritical memorization of facts without understanding the underlying principles with the goal of being able to recall the content in an exam situation [27]. According to numerous studies, students with a deep learning approach have a significantly higher academic performance than students with a surface learning approach [e.g., 28,29]. The initially designed badges of the app under consideration in this study aimed to motivate learning strategies that lead to better exam results. However, since answering short quiz questions repeatedly tends to be considered more as surface learning,

it needs to be examined whether using the app in line with the badge goals can be connected to better grades on the exam. Therefore, the first research question is:

RQ 1 Do the initial badges in the gamified learning app incentivize learning strategies that lead to good exam results?

As previously stated, prior research agrees that badges have a motivational effect. However, it is still unclear which factors contribute most to the success or failure of game-based education [20]. In addition, the meta-studies cited above shed light only on the overall effects of combinations of different game elements. A more in-depth analysis of different badge designs (or other single game elements) does not exist yet. Easley and Ghosh [30] analyzed optimal badge design with a game-theoretic approach from a general but not an educational point of view. Empirical studies that investigate optimal badge design in a learning context are rare. Facey-Shaw et al. [31] provide an overview of different badge designs and their functions in various educational applications. Still, they do not link the different designs to the achieved effects on the learning outcomes. However, as badges (at least in the setting examined in this paper) represent goals set by the instructor, the goal-setting literature could add relevant insights to this research gap.

In goal-setting research, goals can be classified as distal (long-term, end-) goals and proximal (short-term, sub-) goals [32]. The idea behind distal goals is to break down larger, abstract goals into smaller, more tangible goals. Latham and Seijts [33] suggest that none of the two goal types is superior, but a combination of both is preferable. In the setting of this paper, a good exam performance can be considered a distal goal, and the goals connected to the badges can be considered proximal goals (given that they lead to a better exam performance as discussed in the first research question). One aspect that needs further consideration is the level of the goals. After achieving a specific goal, people tend to decrease their effort and pause or relax [34]. Therefore, the goal level should not be too low. However, if a goal is too high, it fosters unethical behavior and increases the willingness to take risks [34]. While these consequences are not such a problem in a learning app than in other areas, goals should still be considered important reference points for the students [35,36]. They probably think that there is a reason behind the chosen goal height and therefore adjust their learning behavior to achieve (but not overachieve) the set goals. A focus on the goal can have a positive effect on performance but also trigger inattention blindness, i.e., losing the focus on learning activities that are not measured with a goal [37]. Although all these research results indicate that the level of a goal is critical, studies that examine how to determine the optimal height are scarce [8]. Therefore, this study will contribute to this research gap with the second research question:

RQ 2 Are the levels of the initial badges optimally set, i.e., does it make a significant difference if a user has reached a specific badge level or not?

In the section on the first research question, it was already discussed that there can be many different learning strategies and that there is, for example, intensive research comparing deep and surface learning approaches. The initially integrated badges in the app under consideration in this study were designed with the intention to motivate learning strategies that lead to better exam results. For example, one of the badges motivates a high number of answered questions, which can be a proxy for intensive use of the app. In past studies, intensive use of learning materials was regularly among the most important features of successful exam preparation [38]. The second research question aims to investigate whether the level of the goals of the already implemented badges is optimal. The aim of the third research question is to investigate whether there are other learning strategies that distinguish good students from the others that are not connected to badges so far. Additional badges could be implemented in a future app update to motivate these strategies. In most recent studies, learning strategies are extracted

¹ <https://duolingo.fandom.com/wiki/Achievements>.

Table 1
Data sources and sample sizes.

Data Source	SS19	WS19	SS20	WS20
App	559	595	447	546
Exam	575	648	616	644
App+Exam	230	243	190	190
Teaching	face-to-face		online	

from trace data from learning management systems with labels like “intensive active users” [39, p. 2018] or “task-focused users” [40, p. 79]. However, those two studies do not examine whether the user types significantly differ regarding the learning outcomes. Gašević et al. [41] as well extract user types from trace data and confirm that a deep learning approach leads to better learning outcomes than a surface learning approach – a result that prior studies found based only on self-reported survey answers and not real learning (trace) data [e.g., 42,43]. Multiple studies that analyze trace data suggest that those students that do not focus on a single one but experiment with different learning strategies tend to have a higher course performance [e.g., 44–46]. Therefore, the learning strategies that the initial badges motivate might only be a subset of the optimal learning strategies in the app. However, prior studies have also shown that too many goals can negatively affect performance [36]. Therefore, it is important to implement the right badges and not as many as possible. Therefore, the third research question contributes to this research gap:

RQ 3 Should additional badges be implemented to incentivize alternative optimal learning strategies?

As stated before, although goal levels are critical for the motivational effects, research on the optimal thresholds is scarce [8]. This study uses an innovative methodological approach to determine optimal goals in the gamified app. In the first step, the performance measures that impact the current (and possible future) badge goals are analyzed using a decision tree algorithm. The goal of a decision tree is to find thresholds in the independent variables (here: performance measures for badge goals) that can be used to divide the sample into groups that are as homogeneous as possible with respect to the dependent variable (here: exam performance). Therefore, if a student has passed such a threshold, the probability of a better grade should significantly increase. In the last step, those newly found thresholds are transformed into new badge candidates, which are then evaluated in a traditional linear regression as a robustness check. Although decision trees are among the most widely used methods in learning analytics [47,48], they have not yet been used to determine optimal learning goals. Decision trees are also frequently used in the natural sciences, e.g., to generate diet recommendations for patients with diabetes [49]. However, no study exists where decision trees were used to determine optimal goal thresholds. Therefore, this is a further contribution of this study.

2. Research setting

2.1. Data sources and methodology

The data for this study comes from the exam of a management accounting course at a large German university and the gamified learning app (called BaccUp) developed for this lecture (see Table 1). The lecture belongs to the undergraduate program and should be attended in the third semester. It is the first contact for the students with management accounting, as they only had one lecture on financial accounting in the first two semesters. The course is attended by approximately 600 students per semester and consists of lectures and exercise sessions with all students and smaller tutorials with fewer students in each session. The learning materials (in addition to the app) consist of a script, a collection of exercises with solutions, as well as a recommended textbook and exercise book. Students are graded solely on the final exam at the

end of the semester. Attendance of the previously mentioned sessions is voluntary and not recorded. Since the summer semester 2019, students have the possibility to use the before mentioned learning app to prepare for the exam. There is already a published study on the use of the app, which also contains further details on the lecture [50]. The app is a quiz app enriched with game elements. For example, students receive points for answering questions and have to unlock further chapters. The questions are grouped into the same nine chapters that are used to structure the lecture and the script. If a question is not answered correctly, a special algorithm, the so-called Skill Level Indicator (SLI), repeats the same questions until they are answered correctly. If a student only gives wrong answers, the same set of questions will be displayed until the correct answers are given. Users can answer the questions either in chapter mode, where they can decide which chapter the questions come from, or in random mode, where random questions are selected. In the third mode, the Weekly Challenge, 25 random questions are selected per week, corresponding to the lecture’s latest progress. Here, users also have the opportunity to compare themselves with other students in a leaderboard. Furthermore, students can earn various badges, described in more detail in the following section. Due to data protection rules, the two data sources can only be matched with each other for those students who voluntarily provided their student ID in the app to make the connection of usage data and exam results possible. As Table 1 shows, not all students provided their ID, which results in a final sample size of 853 students.

To analyze the badges implemented initially, the badges and the corresponding learning metrics are discussed below in the first step. In the second step, to answer the first research question, I will analyze whether students who followed the learning strategy motivated by the badges performed better in the exam by comparing the average grades and conducting a multiple regression with the badges as independent variables. Even though the exam style remained identical over the semesters, it cannot be guaranteed that the difficulty level will remain the same between semesters. This circumstance is considered in the evaluation and assignment of grades, i.e., a specific number of points can result in a different grade in different semesters. Therefore, the analysis in this study uses grades rather than points as the dependent variable to control for the effect of varying difficulty. Grades in this setting range from 1.0 (very good) to 5.0 (failed).

To answer the second research question, a decision tree is built to illustrate the influence of certain limits of the previously mentioned learning metrics on the average grades. In general, decision tree algorithms use a recursive approach to stepwise split a given data set into groups. This breakdown is done according to a splitting rule that maximizes the homogeneity of the dependent variable in each of the resulting groups [51]. At each tree node, a threshold to divide the node is determined that maximizes the homogeneity within the resulting subgroups. In the present study, this means that the difference between the group that achieved a badge and the group that did not is maximized. In terms of performance on the exam, this means that there is a statistically significant better grade to be expected if a student has earned the corresponding target badge. There are several alternatives regarding the splitting rule and the exact process, with CART, ID3, and C4.5 being among the most commonly used variants [52]. The three algorithms differ in a few details, while none outperforms its competitors in all settings [53,54]. For example, ID3 can only handle categorical features, while C4.5 is prone to noise, and CART requires more time to calculate the tree compared to ID3. C4.5 and CART can handle missing values, while ID3 tends to be the fastest algorithm among the three.

A main advantage of the decision tree method, in general, is that its results are easy to interpret and explain. This benefit is particularly useful if the results and the corresponding analysis framework are to be used in practice. For this, the graphical representation method is very suitable, especially compared to a traditional regression output. Moreover, decision trees have no problem with multicollinearity, so there is no need for a correlation analysis of the used features [55]. The disadvan-







No.	Achievement				Badge
<i>single-level</i>					
1	Entered matriculation number				
2	Answered each question correctly once				
3	Unlocked chapter (1-9)				
<i>multi-level</i>		<i>bronze</i>	<i>silver</i>	<i>gold</i>	
4	X consecutive usage days	3	5	15	
5	X total answered questions	50	500	1000	
6	X perfect weekly challenges	1	5	10	

Fig. 1. Initial badge design.

tages of the method are the lower predictive accuracy (e.g., compared to random forests) and the fact that they are less robust than other approaches [56, p. 340]. Decision trees are widely used in learning analytics and educational data mining [57,58], as well as in other disciplines [59].

All analyses for this study were conducted in R (4.2.0). To create the decision tree, the packages rpart (4.1.16) and rpart.plot (3.1.1) – which are based on the CART algorithm – were used. The resulting flow chart of a decision tree is automatically generated based on the input data and can be used to evaluate and adjust the limits of the multi-level badges to increase the motivational effect if necessary. For this study, the default parameters of rpart (4.1.16) were used. In other cases, where the resulting decision tree might be too complex, the parameters can be adjusted to prune the tree [56, p. 331]. To answer the third research question, additional learning metrics are discussed and added to the decision tree to evaluate whether badges based on these metrics should be added to increase the motivational structure. Finally, the results from these decision trees are used to create an optimized badge set which is then evaluated with a linear regression.

2.2. Initial badge design

The initial badge set included badges in six categories (see Fig. 1). The badges can be divided into single-level badges (one learning metric results in one single badge) and multi-level badges (one learning metric results in multiple badges divided into different stages). The first badge is awarded to students entering their matriculation number to share their usage data for research purposes. Since this is not a learning strategy or a measure of learning success, this badge will not be discussed further in this study. The remaining five badge categories and the corresponding learning metrics are discussed in the following sections.

2.2.1. Unique right questions

The second badge in Fig. 1 is awarded if a student answers every question in the database correctly once. The influence of this badge is evaluated with the learning metric “Unique Right Questions”. This metric measures how many different questions a user has answered correctly. Prior research has shown that the more unique sets of multiple-choice questions a student has completed, the better the exam performance [60]. As explained in Section 2.1, the SLI is designed to ensure that a correctly answered question is only displayed again after a certain period of time. However, each question can be answered more than once. The metric “Unique Right Questions” therefore provides valuable additional information. The further a user has progressed in the question database, the higher this value is. Suppose a student uses the app regularly but only answers questions from the first chapter. In that case, this may result in a high number of total answers but probably does not reflect a successful learning strategy in terms of exam performance. A user with the same number of total answered questions, but more unique

right questions, has likely learned more and is therefore more likely to do well on the exam.

2.2.2. Maximal chapter

As described in Section 2.1, according to the rules of the app, a user must work through one chapter at a time. Each user starts with the first chapter and must complete one chapter to unlock the next. The corresponding badge is awarded to the student when a chapter is unlocked. The learning metric “Maximal Chapter” measures the highest chapter a user has unlocked. As discussed in the previous section, a learning strategy that covers higher chapters is presumably more successful regarding exam preparation than intensive usage only in lower chapters. Prior research suggests that especially those questions related to more challenging concepts are important for a good exam performance [61]. Therefore, a student with the same number of answered questions but a higher maximal chapter likely prepared better for the exam than a student with the same (or a higher) number of questions that only covered the topics of the first chapters.

2.2.3. Highest streak

According to prior research, the sequence of learning days can play a decisive role in determining learning success [62]. Continuous learning tends to be better for retaining knowledge than concentrating the same amount of learning time over fewer days. Therefore, the badge in the fourth category in Fig. 1 was developed to motivate a continuous learning strategy. Learning with the app on several consecutive days can help to create a habit and make learning easier on the days that follow. The learning metric “Highest Streak” measures the highest number of consecutive usage days of a user in a given semester. This is the first multi-level badge in Fig. 1. The three stages are 3, 5, and 15 consecutive days.

2.2.4. Total answers

The number of total answers offers a basic measure of the quantitative intensity of app use. If a student has used the app extensively, this is inevitably indicated by a high number of questions answered. Such a conclusion is not as straightforward in comparable studies with data from learning management systems [63,64]. Here, documents can also be read, but the actual reading time is not necessarily reflected in the log data since it is not recorded whether the student actually looks at the screen. In most cases, data collection is based only on clicks in the learning interface. In the app, there is no information to be passively consumed besides feedback messages after a question. Therefore, answering questions is considered equivalent to using the app and if a student uses the app more, it presumably means that he or she learns more. On the one hand, it is proof of the fact that the student learns directly in the app. On the other hand, more intensive app use can also be a proxy for the fact that the student generally learns more and, for example, also engages more with the other learning materials. Nevertheless, this metric captures no qualitative assessment of the learning strategy but

Table 2
Average grades for unique right questions badge.

Badge	Grade Average	n
No	2.83	810
Yes	2.02	43

Table 3
Average grades for maximal chapter badge.

Badge	Grade Average	n
0	3.37	138
1	3.34	87
2	3.21	88
3	3.22	79
4	2.74	54
5	2.54	43
6	2.61	27
7	2.92	25
8	2.56	45
9	2.13	267

is a purely quantitative measurement. For example, if a student used a trial and error strategy by simply clicking through the different answer options, this would result in a high value of total answers but probably not a good exam performance. Therefore, it is vital to also collect qualitative aspects to capture a complete picture of a student's learning strategy (see Section 3.3.3). Still, as more learning time is considered to lead to better performance, this badge was designed to motivate the partial aspect of a higher degree of usage time [65,66]. This badge is the second multi-level badge in Fig. 1 and is awarded for the stages of 50, 500, and 1000 total answers.

2.2.5. Weekly challenges

The last badge in Fig. 1 should motivate to participate in the Weekly Challenge. Competitive students are, in most cases, more committed to achieving the highest possible exam grade [67,68]. The influence of this badge is analyzed with the learning metric "Weekly Challenges", measuring the number of Weekly Challenges a student participated in. Intensive use of this mode is considered a good proxy for a student to be highly competitive, as the Weekly Challenge is entirely voluntary, and students cannot earn any rewards for the exam by participating. This multi-level badge is awarded for the stages 1, 5, and 10 Weekly Challenges.

3. Results

3.1. Evaluating initial badges (RQ 1)

The following analyses examine whether the badges initially implemented in the app motivate successful learning strategies. For this purpose, I examine whether the students who earned the badges achieved a better grade on average than the students who did not. Before further processing, two outliers were removed: a user with over 8000 *totAnswers* (next highest value: 3,187) and a user with a *highestStreak* of 80 (next highest value: 27). Tables 2 and 3 show the average grades of the students who received the different badges based on the learning metrics "Unique Right Questions" and "Maximal Chapter".

Table 2 shows that the average grade of the students who answered all 551 questions correctly is better than that of the comparison group. However, it can also be seen that only 43 of 853 users (5%) achieved this badge, indicating a goal that might be too high. A similar picture emerges for the learning metric "Maximal Chapter" in Table 3. The fact that although 267 out of 853 (31%) users unlocked the ninth chapter, only 5% answered each question correctly once could indicate that unlocking chapters had a higher motivating factor than answering all questions correctly. In addition, the ninth chapter was probably only

unlocked at the end of the semester in most cases, so that time pressure in the main learning phase shortly before the exam could also have caused that not every single question was answered correctly.

The general trend shows that the more chapters have been unlocked, the better the average grade of the students. Although the trend is not ascending in every chapter step but has a small bump at chapters 3, 6, and 7, the result suggests that students who unlock more chapters have a better average grade.

The following analyses examine the multi-level badges in the initial badge set (see Table 4). Students are divided into four groups: Those who did not achieve even the smallest stage and those who achieved at least bronze, silver, or gold, respectively. Then, the average grades in the final exam are compared group by group. Again, the trend for all three badges is that the higher a student moves up the badge levels, the better the average grade on the exam. However, the effect size differs from badge to badge. For the "Highest Streak" badge, the grade improvement from 3.00 (no badge) ranges from 0.54 to 0.59, depending on which level was achieved. Nevertheless, it should be noted that the gold stage was only achieved by 5 students. This indicates that the highest stage may have been set too high. The average grade difference between the students who did not achieve any badge in the category "Total Answers" and those who achieved gold status is exceptionally high. Students who have not achieved any of the three levels obtained an average grade of 3.38 while students with a gold badge had an average grade of 2.21. Students with a bronze or silver badge obtained an average grade of 2.94 or 2.36, respectively. The badge "Weekly Challenges" also shows that students at higher levels tend to have better average grades. However, students with a gold badge have a slightly lower average grade (2.07) than students with a silver badge (1.93). The analysis of this badge also shows (although not quite as strongly as "Highest Streak") that the higher levels were only achieved by comparatively few students suggesting that the thresholds were set too high.

In summary, however, all initially created badges seem to motivate learning strategies that lead to better grades. For a deeper analysis, I also performed a regression analysis with the grade as the dependent variable and dummy variables for all the badges as independent variables (see Table 5). A total of six regression models were estimated. The first five each contain the different badge categories separately (see Table 1), and the sixth model contains all badges implemented so far simultaneously. In the case of the badges for the individual chapters, chapters 1, 4 and 8 are only significant at the level of $p < 0.05$ (1 and 4) or even $p < 0.1$ (8). The badge for *uniqueRightQuestions* in Model 2 is also significant at $p < 0.05$. For the badges on the Weekly Challenges, the first two levels are significant at $p < 0.01$, but the gold level is not significant. An identical picture is shown for the badges for total answers. In the case of the badges for the highest streaks, only the first level is significant ($p < 0.01$). In the sixth model, all badges are included simultaneously. This has the consequence that the significance of some badges changes. In the overall model, only the badge for chapter 4 ($p < 0.05$), the badge for chapter 8 ($p < 0.1$), the bronze ($p < 0.01$) and the silver ($p < 0.05$) badge for the Weekly Challenges, and the bronze badge for the total answers ($p < 0.1$) are significant. It should be noted here that all significant badges have a negative sign, which is a desirable relationship when the grade is the dependent variable. In other words, there is a positive relation with better grades in the exam for these badges. The result that out of 19 badges, only 5 show (and most of them only weak) significant correlations with better grades shows a need for optimization. This will be examined in more detail in the following sections.

3.2. Evaluating stages of initial multi-Level badges (RQ 2)

As indicated in the previous analyses, the designs of the initial badges should be further investigated to answer the second research question. Since the thresholds of the multi-level badges were initially set without empirical validation due to a lack of data, I analyze in the next step whether other stages are more appropriate to provide an optimal moti-

Table 4
Average grades for multi-level badges.

Badge	Highest Streak		Total Answers		Weekly Challenges	
	Grade Average	n	Grade Average	n	Grade Average	n
none	3.00	540	3.38	227	3.13	432
bronze	2.43	200	2.94	284	2.55	335
silver	2.41	108	2.36	134	1.93	56
gold	2.46	5	2.21	208	2.11	30

Table 5
Regression analysis with initial badges.

	<i>Dependent variable:</i>					
	grade					
	(1)	(2)	(3)	(4)	(5)	(6)
Constant	3.324*** (0.092)	2.808*** (0.052)	3.129*** (0.070)	3.381*** (0.095)	3.001*** (0.064)	3.403*** (0.094)
badgeChap1	- 0.345** (0.163)					0.257 (0.326)
badgeChap2	0.043 (0.187)					0.133 (0.188)
badgeChap3	0.028 (0.212)					- 0.053 (0.218)
badgeChap4	- 0.497** (0.251)					- 0.596** (0.280)
badgeChap5	- 0.444 (0.304)					- 0.442 (0.305)
badgeChap6	0.580 (0.355)					0.514 (0.359)
badgeChap7	0.017 (0.340)					- 0.104 (0.342)
badgeChap8	- 0.548* (0.293)					- 0.525* (0.294)
badgeChap9	- 0.161 (0.217)					- 0.225 (0.223)
uniqueQuestionsBadge		- 0.560** (0.275)				0.240 (0.289)
weeklyChallengesBronze			- 0.579*** (0.106)			- 0.321*** (0.120)
weeklyChallengesSilver			- 0.618*** (0.211)			- 0.428** (0.211)
weeklyChallengesGold			0.181 (0.330)			0.313 (0.324)
totAnswersBronze				- 0.438*** (0.128)		- 0.564* (0.326)
totAnswersSilver				- 0.588*** (0.150)		0.091 (0.258)
totAnswersGold				- 0.148 (0.159)		0.275 (0.209)
highestStreakBronze					- 0.573*** (0.123)	0.091 (0.150)
highestStreakSilver					- 0.022 (0.177)	0.246 (0.174)
highestStreakGold					0.054 (0.678)	0.164 (0.656)
Observations	853	853	853	853	853	853
R ²	0.113	0.005	0.064	0.094	0.035	0.137
Adjusted R ²	0.104	0.004	0.061	0.091	0.031	0.118
Residual Std. Error	1.425 (df = 843)	1.502 (df = 851)	1.458 (df = 849)	1.435 (df = 849)	1.481 (df = 849)	1.414 (df = 833)
F Statistic	11.932*** (df = 9; 843)	4.146** (df = 1; 851)	19.423*** (df = 3; 849)	29.436*** (df = 3; 849)	10.136*** (df = 3; 849)	6.976*** (df = 19; 833)

Note: *p<0.1; **p<0.05; ***p<0.01

vational effect. For this purpose, a decision tree will be generated with the different aspects of app usage data as independent variables and the exam grades as the dependent variable. The idea of such a decision tree is to determine the boundaries of the learning metrics that are best suited to divide the group into the most homogeneous subgroups. In other words, at what level of the metrics is the difference between the group that exceeds the level and the group that does not exceed the level the greatest. For example, if all students who answered more than 3000 questions have the best grade possible in the exam, it would seem unreasonable to include another stage at 4000 answered questions. Since the discussed learning metrics are at least partially related (e.g., a higher maximum chapter necessarily results in higher unique right questions), all learning metrics are examined simultaneously. Moreover, all badges, not only those created initially as multi-level, are considered in the calculations. This will allow to identify whether there are suitable thresholds for single-level badges as well, and whether they should therefore be converted into multi-level badges. The result of the corresponding decision tree is shown in Fig. 2.

The decision tree can be interpreted as follows. “Maximum Chapter” is the most important learning metric for group classification in the initial badge set. The average grade for all students is 2.8. For students who have only worked on chapters 1 to 3, the grade average is 3.3, and for students who have worked on at least chapter 4, the grade average is 2.4. The result can be interpreted in such a way that, for example, the badge on chapter 4 should be particularly highlighted, as it seems to mark an

important threshold. In the next step, the learning metric “Unique Right Questions” is used for further subdivision. If students have answered at least 479 unique questions correctly, the grade average improves to 2.0; if not, it declines to 2.6. The badge for “Unique Right Questions” has so far only been awarded if all questions, i.e., 551, have been answered correctly. The result suggests that a change should be considered, and the badge should be changed from a single-level to a multi-level badge with one level at around 479. In the next step, the group with less than 479 unique right questions is subdivided according to the metric “Weekly Challenges”. If a student participated in less than 2 Weekly Challenges, the average grade worsens from 2.6 to 2.8. If the number of Weekly Challenges is at least 2, the average grade increases to 2.2. Since the smallest level of this multi-level badge was 1, the result not necessarily indicates a need for action at this point. In the last step, the group is subdivided based on the “Total Answers”. However, the subdivision goes in the opposite direction at this point. The average score improves to 2.0 if less (!) than 1130 questions were answered. If more than 1130 questions were answered, the average grade deteriorates to 3.0. Therefore, it does not seem to be a good strategy to simply answer as many questions as possible (under certain conditions). However, it should be emphasized that this advice includes all the subdivisions made previously, so the limit of 1130 answers should not be considered in isolation. In addition, it should be emphasized that “Highest Streak” was also included in the analysis but is apparently not suitable as a criterion for subdividing the groups.

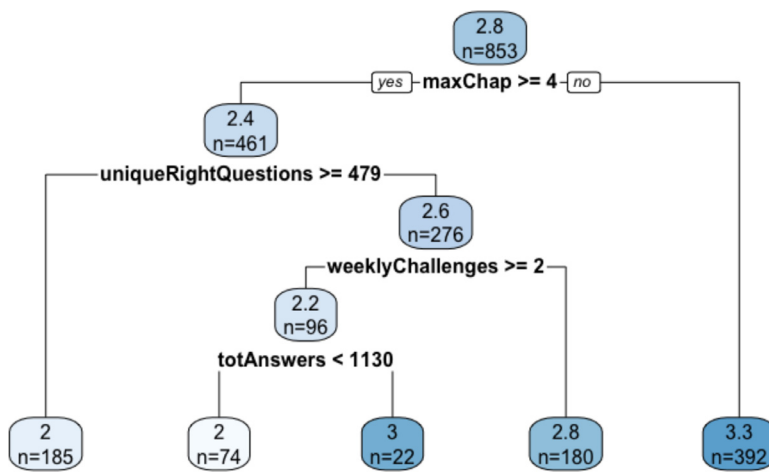


Fig. 2. Decision tree with initial badges.

3.3. Evaluating additional badges (RQ 3)

The previous analysis has shown that while the initially implemented metrics are generally related to good exam results, the set thresholds of the multi-level badges do not seem to be optimally chosen. In the following, I will examine additional not yet implemented metrics for their suitability as a basis for badges to be introduced in the future. With the help of the decision tree method, thresholds for possible multi-level badges will be determined. First, I will discuss the new metrics in groups and then analyze suggestions for additional badges based on the results. The analysis always follows the same pattern. New metrics are added in groups to the decision tree from Fig. 2 and tested to see whether they are better suited for determining targets than the metrics used previously. If so, the decision tree is adjusted, and the newly found goals are discussed. Finally, a proposal for a new badge set is determined from the candidates identified in this way.

3.3.1. Active days

In the badges available so far, the number of usage days has only been relevant for the *highestStreak* parameter. While only consecutive usage days were recorded here, the next step is to discuss whether a certain number of usage days should also be motivated with a badge. For this analysis, I use the learning metric *activeDays*, which states on how many days in the semester a student has answered at least one question.² Most of the initially implemented badges can theoretically be earned in one single day. A more realistic assumption is probably that it needs multiple days to earn all badges. However, those days could be, e.g., directly at the beginning of the semester or shortly before the exam indicating a cramming behavior [69]. If a student follows a cramming strategy, this will result in a lower amount of *activeDays*. According to prior research, spaced repetition is considered a better learning strategy for preparing for an exam than cramming [70,71]. If a student follows this strategy, this will result in a higher amount of *activeDays* [72]. In principle, it can be assumed that, on average, more learning days lead to better performance in the exam. However, since there may also be an upper limit above which the marginal benefit decreases, a multi-level badge could be a suitable solution.

The decision tree in Fig. 3 shows that the new measure replaces *totAnswers* in the initial decision tree (Fig. 2). Interestingly, *activeDays* also has a negative signed threshold. This indicates that learning on too many days under certain conditions seems to decrease the performance in the exam.

² An alternative measure where an active day was considered a day with at least five answered questions (*activeDaysFiltered*) did not lead to different results.

3.3.2. Temporal aspects of answers

The total number of answers was already considered in the initial badge set. In the following, additional aspects of the answers will be analyzed. The badges of the initial set included *maxChap* and *uniqueRightQuestions*, which necessarily require correct answers to be given. The metric *totAnswers* as such simply counts the number of answers given, regardless of any other aspects. However, past studies have shown, for example, that the time of day a student learns can influence learning success [73,74].

Therefore, I broke down the total answers according to temporal aspects for the following analyses. I first divided the day into its 24 hours and recorded how many answers a user submitted in each of the 24 hours. For example, the metric *answers06* counts answers submitted between 06:00 and 06:59. With these metrics, new thresholds are added to the initial decision tree (see Fig. 4). Answers between 6 and 7 o'clock, as well as between 20 and 21 o'clock seem to be positively related to good exam results. However, too many answers between 21 and 22 o'clock seem to be negatively related. In a follow-up analysis, I clustered the previously generated metrics into four temporal ranges of the day: responses between 00:00 and 05:59, between 06:00 and 11:59, between 12:00 and 17:59, and between 18:00 and 23:59. However, these condensed versions of the previously generated metrics do not result in new thresholds in the decision tree.

Not only the time of day when the app is used could play a role, but also the day of the week. Therefore, I also record how many answers were submitted on one of the seven weekdays. Based on this, I calculate other key figures, such as the relative distribution, i.e., how many percent of a user's total responses were submitted on Mondays, for example (*mondayShare*).

Fig. 5 shows the result of this analysis. Accordingly, it seems to make a difference in certain constellations how much of the learning time with the app falls on a Monday. If this proportion is greater than 36%, this is related to poorer grades in the exam. For example, this result could be interpreted as such that students who primarily use the app during the lecture (which took place every Monday) study too little on the other days for the exam. To examine this result more closely, I built two dummy variables based on the previously determined metrics. The variable *lectureLearner* is 1 if the percentage of learning with the app on Mondays and Tuesdays (when the lecture and exercise session took place) is higher than the sum of the remaining days. The variable *weekendLearner* is 1 if the percentage of learning with the app on Saturdays and Sundays is higher than the sum of the remaining days. In addition, I also determined the variable *learningDays*, which records on which days of the week students answered at least one question. Thus, if a student learns exclusively on Mondays, this metric is 1, and if a student answered at least one question on each of the seven days of the week, it is

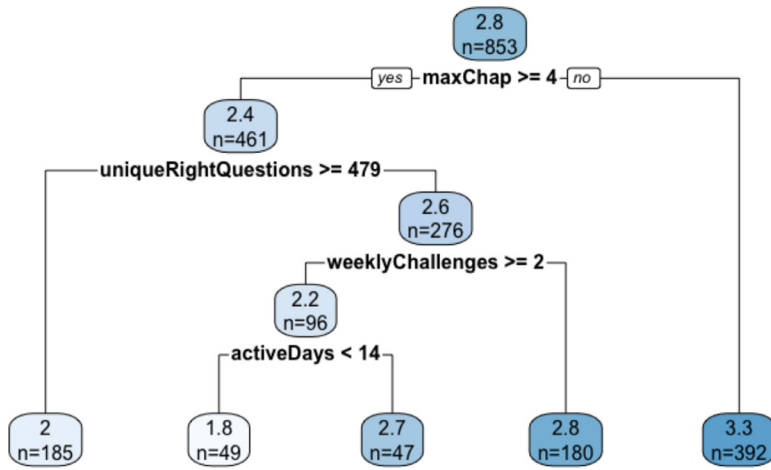


Fig. 3. Decision tree with active days.

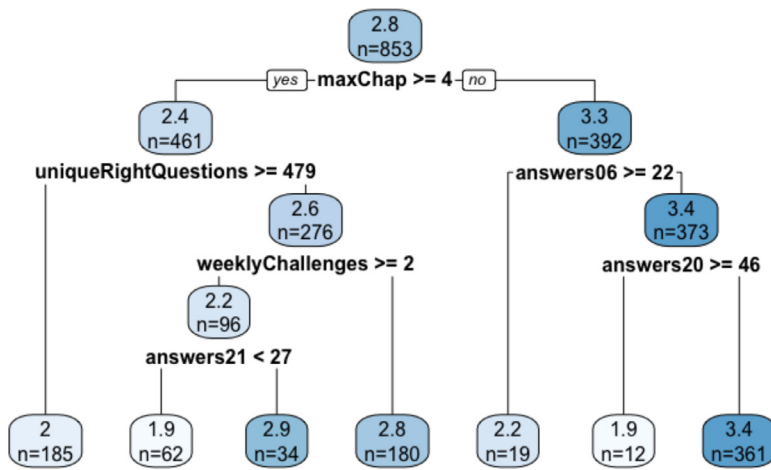


Fig. 4. Decision tree with answers per hour.

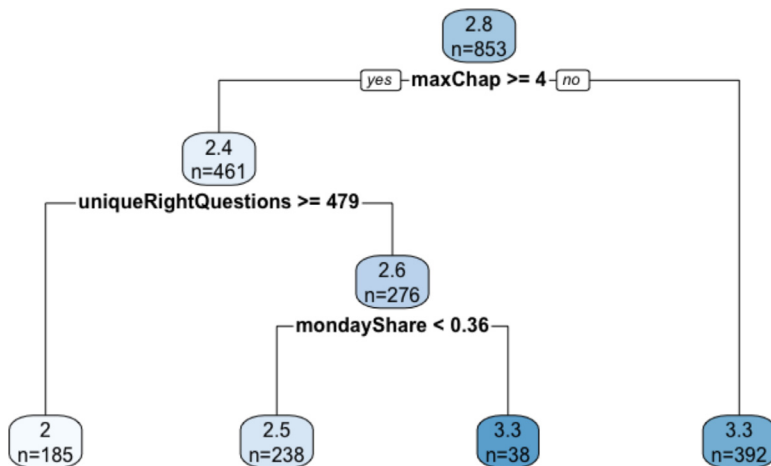


Fig. 5. Decision tree with answers per weekday.

7. However, including these new metrics does not change the decision tree.

3.3.3. Qualitative aspects of answers

It is not only the temporal factor that could be examined in more detail but also the qualitative factor of the answers. Therefore, I also record how many answers a user has given per chapter and how many

correct and incorrect answers a user has given. I additionally calculate the rate of correct answers in relation to the total number of submitted answers. Moreover, I also determine a variant of this indicator where only the first answers of a user to each of the 551 questions are included. The ulterior motive of this variant is to measure the extent to which a student has already engaged with the other learning materials of the course before using the app. Suppose a student has already learned with

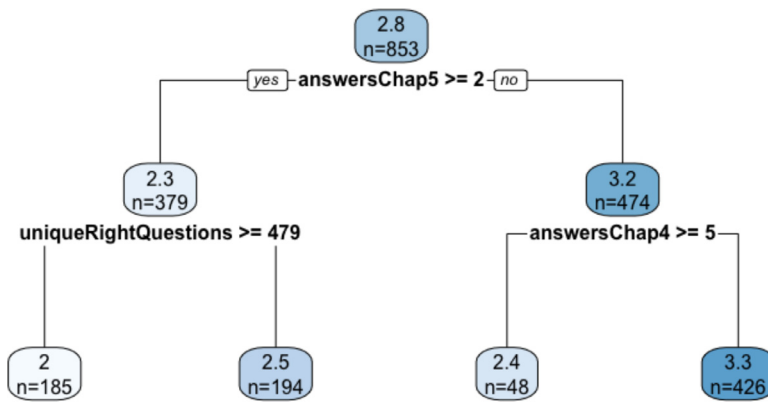


Fig. 6. Decision tree with answers per chapter.

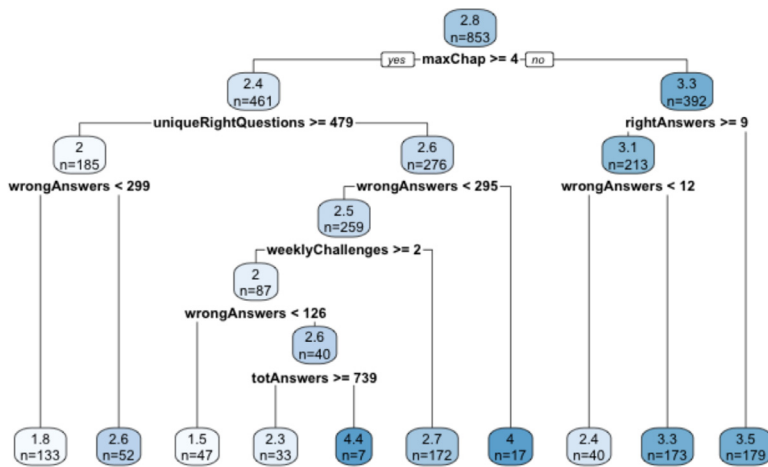


Fig. 7. Decision tree with absolute right and wrong answers.

other learning materials before using the app or has worked well in lectures and exercises. In that case, a higher success rate can be expected for his or her first answers than for students who primarily learn with the app, i.e., who try to learn the content primarily or at least partially based on the feedback that is displayed for incorrect answers.

The first tree, where I only included the answers per chapter (see Fig. 6) indicates that the number of answers in chapters 4 and 5 seem to be important thresholds. This partly confirms the regression results in Table 5, where the badge for the fourth chapter had a significant coefficient ($p < 0.05$).

The decision tree that includes the absolute number of right and wrong answers offers multiple options for additional badges (see Fig. 7). The tree suggests that several thresholds of wrong answers indicate bad results in the exam. Moreover, additional thresholds for total answers ($>= 739$) as well as right answers ($>= 9$) could be discussed as the basis for new badges. Including the share of right answers also generates a new version of the decision tree (Fig. 8) and indicates that a share of right answers of 80% seems to be a significant threshold. The decision trees where only the first answers of each user are included offer additional insights.

Fig. 9 shows that for the first right answers, three thresholds (128, 361, and 437) could be used for additional badges. Moreover, two thresholds for wrong first answers (37 and 64), as well as a new threshold for unique right questions (9), seem to be important to determine the average exam result. Additionally, Fig. 10 shows two possible thresholds for the share of right first answers (69% and 79%) that could be considered in the design of new badges.

3.3.4. Sessions

Finally, the sessions are also examined in more detail. A session in this study is defined as follows: A session always starts with a given an-

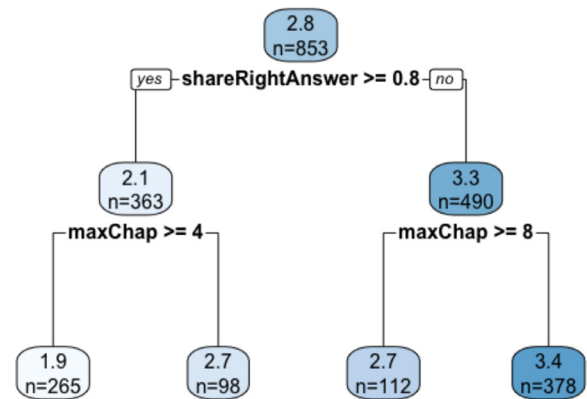


Fig. 8. Decision tree with share of right answers.

swer. There is no minimum number of answers, but if no additional answer follows for 30 minutes, the session is over [41,75]. Sessions were determined using the R package TraMineR (2.2–3) [76]. In addition to the number of sessions and the total usage time (in seconds), I also recorded the average time (in seconds) and the average number of answers per session, as well as the corresponding standard deviation to measure the consistency of the learning habits [77].

Fig. 11 shows that different aspects of the sessions seem to be important for the exam results. It seems to be negative if the time per session has a high standard deviation, i.e., if the lengths of the individual sessions strongly differ, indicating an inconsistent learning strategy. A high number of answers per session (indicating longer sessions) seems to be

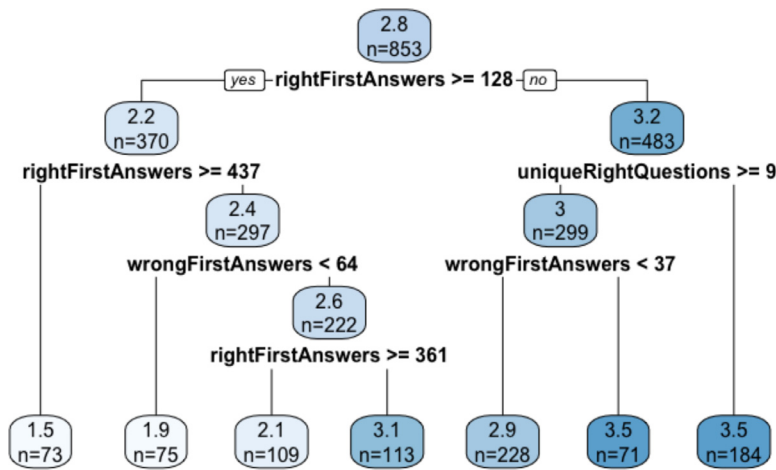


Fig. 9. Decision tree with absolute first right and wrong answers.

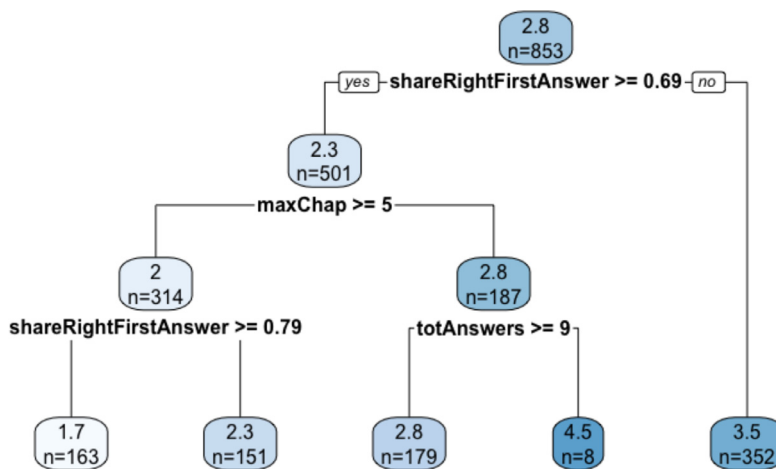


Fig. 10. Decision tree with share of first right answers.

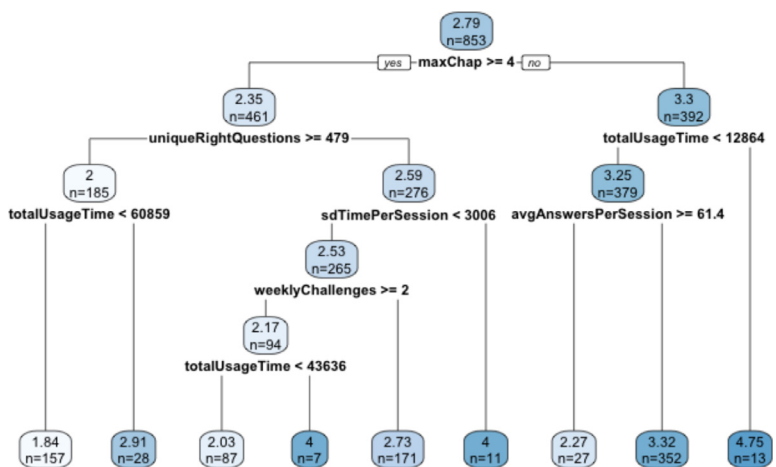


Fig. 11. Decision tree with session data.

positive. Moreover, a total usage time that is relatively high seems to have a negative effect on the exam result.

3.3.5. New badge candidates

Building on the results from the decision trees in the previous section, I will now discuss the implementation of new badges. Table 6 shows all thresholds that were included in at least one decision tree.

I sorted the thresholds into four categories: effort, negative, dynamic, and irreversible goals. In the case of an effort goal, the user must exceed

a specific threshold. Once the user has done that, he or she has definitely reached the goal and nothing can change it in the course of time. For example, if the user has answered at least 739 questions, he or she has reached the last goal in the first column of Table 6. All badges implemented so far are based on effort goals. So if newly designed badges were based on the other three categories, not only new badges but even new badge types would be implemented. Negative goals are about not exceeding certain limits. Based on the previous analyses, it seems that a particular total usage time should not be exceeded. However, such a

Table 6
Overview of new goal candidates.

Effort Goals	Negative Goals	Dynamic Goals	Irreversible Goals
weeklyChallenges >= 2	totAnswers < 1130	avgAnswersPerSession >= 61.4	rightFirstAnswers >= 128
uniqueRightQuestions >= 9	answers21 < 27	shareRightAnswer >= 0.8	rightFirstAnswers >= 361
uniqueRightQuestions >= 479		mondayShare < 0.36	rightFirstAnswers >= 437
answers06 >= 22			shareRightFirstAnswer >= 0.69
answers20 >= 46			shareRightFirstAnswer >= 0.79
answersChap5 >= 2	wrongAnswers < 12		
answersChap4 >= 5	wrongAnswers < 295		
rightAnswers >= 9	wrongAnswers < 126		
totAnswers >= 9	wrongAnswers < 299		
totAnswers >= 739	wrongFirstAnswers < 37		
	wrongFirstAnswers < 64		
	activeDays < 14		
	totalUsageTime < 12864		
	totalUsageTime < 43636		
	totalUsageTime < 60859		
	sdTimePerSession < 3006		

goal would probably be implemented in the app as a warning message for example, rather than as a badge. Badges based on dynamic goals – once earned – can also be lost again. If, for example, the limit of 80% for the share of right answers is exceeded, this can be reversed again with a larger number of wrong answers. With such badges, the motivational effect would therefore not lie in the one-time achievement of a goal but in maintaining an achieved status. The last category includes irreversible goals. Here, the focus lies on the number of correct answers to questions a user has seen for the first time. The goals are irreversible because once a user has seen all the questions and has not answered all of them correctly on the first attempt, he or she may no longer be able to reach certain target levels. Three of the goals found in the decision trees simultaneously belong to two of the goal categories. The goal of not giving (relatively) too many answers on Mondays is both a dynamic and a negative goal. The two goals for the share of right first answers are simultaneously dynamic but also irreversible since the opportunities to answer questions for the first time are used up at some point.

3.3.6. Creation and test of new badges

Based on the goals listed in Table 6, in the following section, I discuss the creation of new badges and then test how the regression from Table 5 would change with the addition of the new badges. In Section 3.2, the thresholds of the existing badges were analyzed. Here, potential new thresholds were identified for Weekly Challenges (>=2), unique right questions (>= 479), and total answers (< 1,130) (see Fig. 2). A negative goal for total answers does not seem to make much sense, as this would probably be rather difficult to explain to the users. However, a new threshold could be implemented for the Weekly Challenges and the unique right questions. For the Weekly Challenges, the result of at least two challenges can be used. For the unique right questions, rounding up to 480 seems to make sense to make the threshold less arbitrary and more comprehensible for the user. With these two updates, there would be still only badges with effort goals in place. In the following, badges with goals of the other categories will also be implemented.

First of all, the irreversible goals are discussed. The metric *rightFirstAnswers* offers interesting possibilities since three different thresholds were identified here. This would fit an implementation of a new multi-level badge with the levels bronze, silver, and gold, which are already contained in the initial badge set. For the reasons mentioned above regarding the new badge for the unique right questions, the levels 128, 361, and 437 are rounded as follows: 130, 360, and 440. For the dynamic goals, the metric *shareRightAnswer* is implemented in the test set with its calculated value of 80%. Finally, one of the negative goals should also be tested. As discussed before, in practice this will probably not be a badge but a warning message. For this warning message, the

metric *totalUsageTime* is probably the most suitable. With a comprehensible explanation, a warning about too many hours of usage seems more appropriate than a warning about too many (wrong) answers, for example. To reach as many users as possible with the warning, the lower bound of the three thresholds available (12,864 vs. 43,636 vs. 60,859) is implemented and rounded to 13,000, as discussed before. The effects of the new badges can be seen in Table 7.

Model 1 contains all old badges at the same time, as in Table 5, in order to be able to compare the following models with the initial badge set. Model 2 contains all old badges as well as all new badges mentioned before. With the introduction of the new badges, the old badge for chapter 1 becomes slightly significant ($p < 0.1$). However, the badge for chapter 4 loses its significance, as well as the one for chapter 8 and the silver badge for the Weekly Challenges. The significance of the bronze badge of the Weekly Challenges decreases from $p < 0.01$ to $p < 0.05$, and that of the bronze badge of the total answers increases from $p < 0.1$ to $p < 0.05$. The newly introduced badge for the Weekly Challenges is not significant as well as the one for the number of total answers. However, the newly introduced badge types are all significant, partly with $p < 0.1$ (first answers silver and gold) and partly with $p < 0.01$ (first answers bronze and share of right answers). As expected, the warning for the total usage time has a positive sign in contrast to the other badges, i.e., it is related to a deterioration of the exam grade. However, it is also highly significant ($p < 0.01$). Since the new badges for Weekly Challenges and unique right questions were actually planned as replacements, the old badge for unique right questions and the bronze badge for Weekly Challenges (since 1 is closest to the new limit of 2) are removed in Model 3. However, this does not change the fact that the new badges are not significant. The significance of the newly introduced badge types does also not change. Model 4 contains all badges that were significant in Model 3. As a result, the significance of *firstAnswersBadgeSilver* and *firstAnswersBadgeGold* increases to $p < 0.05$ and *totAnswersBronze* loses its significance. Finally, only all new badges are included in Model 5. Here, the new badge regarding the Weekly Challenge is slightly significant ($p < 0.1$), and *firstAnswersBadgeSilver* loses its significance. However, Models 2 to 5 show that the newly determined badges (or warnings) represent potential for meaningful additions to a new badge set.

4. Discussion

In summary, it can be stated that the initially implemented badges did motivate learning strategies that lead to good exam results. However, the exact thresholds could be optimized and further learning strategies exist that could also be integrated as badges. Comparable results from other studies do not exist because, first, there is only little research on the connection of gamified learning apps and learning outcomes and, second, the methodological approach is new.

Table 7
Regression analysis with new badges.

	<i>Dependent variable:</i>				
	grade				
	(1)	(2)	(3)	(4)	(5)
Constant	3.403*** (0.094)	3.474*** (0.091)	3.454*** (0.091)	3.462*** (0.091)	3.383*** (0.069)
badgeChap1	0.257 (0.326)	0.532* (0.314)	0.514 (0.315)		
badgeChap2	0.133 (0.188)	0.213 (0.181)	0.188 (0.181)		
badgeChap3	- 0.053 (0.218)	0.018 (0.222)	0.037 (0.222)		
badgeChap4	- 0.596** (0.280)	- 0.314 (0.291)	- 0.293 (0.291)		
badgeChap5	- 0.442 (0.305)	- 0.437 (0.293)	- 0.477 (0.293)		
badgeChap6	0.514 (0.359)	0.557 (0.351)	0.552 (0.352)		
badgeChap7	- 0.104 (0.342)	- 0.140 (0.330)	- 0.099 (0.330)		
badgeChap8	- 0.525* (0.294)	0.169 (0.360)	0.150 (0.361)		
badgeChap9	- 0.225 (0.223)	- 0.012 (0.232)	0.025 (0.228)		
uniqueQuestionsBadge	0.240 (0.289)	0.359 (0.279)			
weeklyChallengesBronze	- 0.321*** (0.120)	- 0.271** (0.129)			
weeklyChallengesSilver	- 0.428** (0.211)	- 0.350 (0.216)	- 0.343 (0.217)		
weeklyChallengesGold	0.313 (0.324)	0.368 (0.309)	0.350 (0.310)		
totAnswersBronze	- 0.564* (0.326)	- 0.682** (0.311)	- 0.754** (0.310)	- 0.207 (0.129)	
totAnswersSilver	0.091 (0.258)	- 0.089 (0.262)	- 0.108 (0.262)		
totAnswersGold	0.275 (0.209)	0.278 (0.216)	0.277 (0.213)		
highestStreakBronze	0.091 (0.150)	0.031 (0.144)	0.034 (0.144)		
highestStreakSilver	0.246 (0.174)	0.189 (0.167)	0.177 (0.168)		
highestStreakGold	0.164 (0.656)	0.021 (0.628)	0.138 (0.627)		
newChallengeBadge		- 0.028 (0.153)	- 0.172 (0.137)		- 0.229* (0.117)
newQuestionBadge		- 0.303 (0.304)	- 0.275 (0.305)		- 0.013 (0.198)
firstAnswersBadgeBronze		- 0.754*** (0.223)	- 0.741*** (0.223)	- 0.806*** (0.179)	- 0.812*** (0.177)
firstAnswersBadgeSilver		- 0.411* (0.242)	- 0.423* (0.242)	- 0.359** (0.166)	- 0.319 (0.201)
firstAnswersBadgeGold		- 0.421* (0.225)	- 0.382* (0.224)	- 0.444** (0.206)	- 0.429** (0.211)
shareRightAnswerBadge		- 0.713*** (0.113)	- 0.715*** (0.113)	- 0.743*** (0.110)	- 0.755*** (0.108)
totalUsageTimeWarning		0.695*** (0.233)	0.692*** (0.234)	0.601*** (0.176)	0.578*** (0.177)
Observations	853	853	853	853	853
R ²	0.137	0.221	0.215	0.195	0.197
Adjusted R ²	0.118	0.197	0.193	0.190	0.190
Residual Std. Error	1.414 (df = 833)	1.349 (df = 826)	1.352 (df = 828)	1.355 (df = 846)	1.355 (df = 845)
F Statistic	6.976*** (df = 19; 833)	9.036*** (df = 26; 826)	9.476*** (df = 24; 828)	34.210*** (df = 6; 846)	29.523*** (df = 7; 845)

Note: *p<0.1; **p<0.05; ***p<0.01

The results of this study are subject to the following limitations. An assumption of this analysis, based on past studies, is that students were motivated by the badges to use the appropriate learning strategies. However, since there is no control group without badges, it cannot be empirically verified that it was the badges that motivated the users to learn in certain ways. Furthermore, badges are only one game element in the app under consideration here. Since the metrics used to capture the badge goals are at least partially related to the other game elements, the effect of the badges cannot be considered in isolation. In the case of the Weekly Challenge badge, for example, it is unclear whether the badge motivated the users to participate in a Weekly Challenge or whether it was the Weekly Challenge itself. In addition, the design and display of the badges can vary significantly from app to app. The last column in Fig. 1 shows what the badges in BaccUp look like. Here it becomes apparent that the badges in BaccUp have a minimalistic look, while they can be much more colorful and pictographic in other apps. In future studies the design of the badge pictures could be altered to evaluate the corresponding effect [36]. However, in contrast to other apps, the badges in BaccUp are prominently positioned. While in other apps, users can only access their badge collection via a separate menu item, a user of BaccUp always sees them on his or her start screen. This could have a positive effect, as past studies have shown that regular monitoring of goals can foster their completion [78]. Nonetheless, without a control group, this effect cannot be examined.

Another simplifying assumption of the analyses in this study is that users learned exclusively with the app. Other learning activities are excluded from the analyses because no data are available on them, as the data only show how much students learned using the app. Learning with a quiz app can tend to be understood as surface learning. In the exam,

however, a more in-depth understanding is requested since no multiple-choice questions are included here. Therefore, if the students actually only learned with the app, one would not expect a positive influence on the exam result. Nonetheless, since there is no data on learning with the other learning materials available, no valid conclusions can be drawn about the student’s overall learning strategy.

As explained before, goals can influence the focus. In the current setting, this was only discussed using the single course the app was developed for. Badges in one course, however, can also have the effect that less is learned for other courses. Yet, I cannot examine this effect due to lack of data. According to the Octalysis framework, badges are among the game elements that motivate extrinsically [11]. If this is also true for the present setting, past studies suggest that this negatively affects intrinsic motivation [79,80]. For example, if students enjoy working with the script but cannot earn badges for this activity, motivation for this activity could decrease, and motivation for learning activities in the app could increase. This could have a negative impact on performance in the exam.

Fig. 12 shows how the procedure of this case study could be transferred to comparable research projects. While the specific learning metrics and badges might be unique to the app analyzed in this study, the general optimization process can be easily used in other settings. The starting point is the creation of an initial badge set. Here, the learning metrics and corresponding thresholds are selected, and then the respective usage data is collected for one term. The target variable could be the exam grade as in the present study, or it could be the grades of assignments that have to be handed in or the grades of oral exams. However, it is important that the learning strategies that the badges are intended to motivate have an impact on the corresponding target variable. After

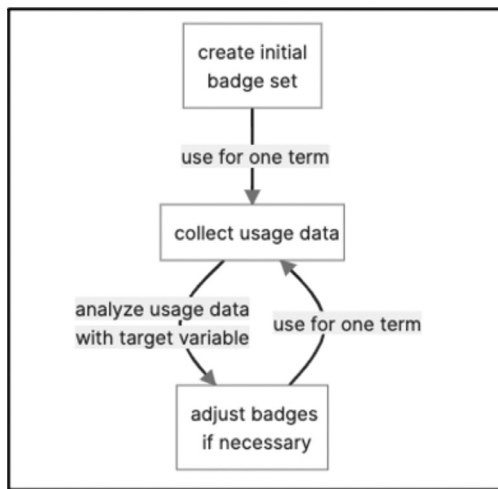


Fig. 12. Iterative Optimization Process.

one term the dependent and independent variables are analyzed with decision trees as explained in this study. Then, the effect of the motivated learning strategies and the defined thresholds are compared with the results. If necessary, the badges can be adjusted and the new badge set can then be retested for one term and the corresponding results subsequently analyzed.

The results of this study also inspire possible next steps in the development of the given app. The badges in BaccUp are currently only visible to the respective user. Since particularly public goals have a motivating effect, the developers could think about increasing the visibility of the badges in the app to the other users [81]. Furthermore, the goals in this study are evaluated as dummy variables, i.e., either they are met or not. According to goal-setting theory, however, it can be optimal to set very high (stretch) goals with the expectation that they are never met but still increase the performance [82]. For some goals that were motivated by badges, it would therefore maybe be better to be displayed in a progress bar or something similar. For example, Huynh et al. have found, that in Duolingo the “Winning Streak increases the motivation of advanced users when the attractiveness of Badge decreases” [83, p. 1]. Here the winning streak is simply a counter that displays the current winning streak and can therefore increase infinitely. Such a display is not integrated in BaccUp so far. Finally, the positive effect of goal setting is more substantial when goals are set publicly [84]. In a possible new feature, students could set goals for themselves (limited to the beginning of the semester) and have everyone observe how they follow the goals. This would also have a further competitive aspect to the motivational effect of the badges [85].

5. Conclusion

The evaluation of the initial badge set implemented in a gamified learning app showed that the badges do motivate learning strategies that lead to higher exam performances on average. However, a deeper analysis revealed that the threshold of the implemented multi-level badges could be improved. Moreover, additional learning strategies were identified that are not motivated by any badges so far. As these new learning strategies were connected to four different goal types, there is even potential for new (game) elements in the app that motivate the new strategies. This is an essential contribution to the gamification literature, where mostly points, badges, and leaderboards are used. These results were obtained using a novel methodological combination of decision trees and linear regressions. While the data set itself is unique, and the exact numeric results will only help for updating the badge set used in the app under consideration, the procedure as well as the new found goal types can be used in other studies. One important takeaway

is that not only effort goals should be implemented, but also the other goal types, as different goal styles might also trigger students with different personality traits. Moreover, the newly found method could also be tested with other data sets outside of the field of learning sciences to evaluate if optimal goals can also be found here.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

The authors do not have permission to share data.

References

- [1] C. Tudor-Locke, D.R. Bassett, How many steps/day are enough? *Sports Med.* 34 (1) (2004) 1–8.
- [2] P.F. Saint-Maurice, R.P. Troiano, D.R. Bassett, B.I. Graubard, S.A. Carlson, E.J. Shirota, J.E. Fulton, C.E. Matthews, Association of daily step count and step intensity with mortality among US adults, *JAMA* 323 (12) (2020) 1151–1160.
- [3] A.E. Paluch, K.P. Gabriel, J.E. Fulton, C.E. Lewis, P.J. Schreiner, B. Sternfeld, S. Sidney, J. Siddique, K.M. Whitaker, M.R. Carnethon, Steps per day and all-cause mortality in middle-aged adults in the coronary artery risk development in young adults study, *JAMA Network Open* 4 (9) (2021) e2124516.
- [4] I.M. Lee, E.J. Shirota, M. Kamada, D.R. Bassett, C.E. Matthews, J.E. Buring, Association of step volume and intensity with all-cause mortality in older women, *JAMA Intern. Med.* 179 (8) (2019) 1105–1112.
- [5] M. Gladwell, *Outliers: the story of success*, Little, Brown, 2008.
- [6] K.A. Ericsson, R.T. Krampe, C. Tesch-Römer, The role of deliberate practice in the acquisition of expert performance, *Psychol. Rev.* 100 (3) (1993) 363.
- [7] K.A. Ericsson, K.W. Harwell, Deliberate practice and proposed limits on the effects of practice on the acquisition of expert performance: why the original definition matters and recommendations for future research, *Front. Psychol.* 10 (2019) 2396.
- [8] S. Jain, Self-control and optimal goals: a theoretical analysis, *Market. Sci.* 28 (6) (2009) 1027–1045.
- [9] H. Aldowah, H. Al-Samarraie, W.M. Fauzy, Educational data mining and learning analytics for 21st century higher education: a review and synthesis, *Telemat. Informat.* 37 (2019) 13–49.
- [10] S. Deterding, D. Dixon, R. Khaled, L. Nacke, From game design elements to gamefulness: Defining gamification, in: *Proceedings of the 15th International Academic MindTrek Conference: Envisioning Future Media Environments*, ACM, 2011, pp. 9–15.
- [11] Y.-K. Chou, *Actionable gamification: beyond points, badges, and leaderboards*, Octalysis Media, Fremont, CA, 2015. ISBN 978-1-5117-4404-1.
- [12] A. Antonaci, R. Klemke, M. Specht, The effects of gamification in online learning environments: A systematic literature review, in: *Informatics*, vol. 6, Multidisciplinary Digital Publishing Institute, 2019, p. 32.
- [13] L. Blair, What video games can teach us about badges and pathways, *Digit. Badges Educ.: Trends Iss. Cases* (2016) 62–70.
- [14] R. McDaniel, *What We Can Learn about Digital Badges from Video Games*, in: *Foundation of Digital Badges and Micro-Credentials*, Springer, 2016, pp. 325–342.
- [15] C.S. Loh, Y. Sheng, D. Ifenthaler, *Serious Games Analytics: Theoretical Framework*, in: *Serious Games Analytics*, Springer, 2015, pp. 3–29.
- [16] J. Gerring, What is a case study and what is it good for? *Am. Polit. Sci. Rev.* 98 (2) (2004) 341–354.
- [17] E.A. Plant, K.A. Ericsson, L. Hill, K. Asberg, Why study time does not predict grade point average across college students: implications of deliberate practice for academic performance, *Contemp. Educ. Psychol.* 30 (1) (2005) 96–116.
- [18] S. Bai, K.F. Hew, B. Huang, Does gamification improve student learning outcome? evidence from a meta-analysis and synthesis of qualitative data in educational contexts, *Educ. Res. Rev.* 30 (2020) 100322.
- [19] R. Huang, A.D. Ritzhaupt, M. Sommer, J. Zhu, A. Stephen, N. Valle, J. Hampton, J. Li, The impact of gamification in educational settings on student learning outcomes: a meta-analysis, *Educ. Technol. Res. Dev.* 68 (4) (2020) 1875–1901.
- [20] M. Sailer, L. Hommer, The gamification of learning: a meta-analysis, *Educ. Psychol. Rev.* 32 (1) (2020) 77–112.
- [21] J. Hamari, Do badges increase user activity? a field experiment on the effects of gamification, *Comput. Human. Behav.* 71 (2017) 469–478.
- [22] L. Hakulinen, T. Auvinen, A. Korhonen, The effect of achievement badges on students' behavior: an empirical study in a university-level computer science course, *Int. J. Emerg. Technol. Learn.* 10 (1) (2015) 18–29.
- [23] B. Huang, K.F. Hew, Do points, badges and leaderboard increase learning and activity: A quasi-experiment on the effects of gamification, in: *Proceedings of the 23rd International Conference on Computers in Education*, Society for Computer in Education Hangzhou, China, 2015, pp. 275–280.
- [24] P. Denny, The effect of virtual achievements on student engagement, in: *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, ACM, 2013, pp. 763–772.

- [25] E.A. Linnenbrink, P.R. Pintrich, Motivation as an enabler for academic success, *School Psych. Rev.* 31 (3) (2002) 313–327.
- [26] J.A.C. Hattie, G.M. Donoghue, Learning strategies: a synthesis and conceptual model, *npj Sci. Learn.* 1 (1) (2016) 1–13.
- [27] V. Beattie IV, B. Collins, B. McInnes, Deep and surface learning: a simple or simplistic dichotomy? *Account. Educ.* 6 (1) (1997) 1–12.
- [28] A.-M. Bliuc, R. Ellis, P. Goodyear, L. Piggott, Learning through face-to-face and online discussions: associations between students' conceptions, approaches and academic performance in political science, *Brit. J. Educ. Technol.* 41 (3) (2010) 512–524.
- [29] R.A. Ellis, G. Marcus, R. Taylor, Learning through inquiry: student difficulties with online course-based material, *J. Comput. Assist. Learn.* 21 (4) (2005) 239–252.
- [30] D. Easley, A. Ghosh, Incentives, gamification, and game theory: an economic approach to badge design, *ACM Trans. Econ. Comput.(TEAC)* 4 (3) (2016) 1–26.
- [31] L. Facey-Shaw, M. Specht, P. Van Rosmalen, D. Brner, J. Bartley-Bryan, Educational functions and design of badge systems: a conceptual literature review, *IEEE Trans. Learn. Technol.* 11 (4) (2017) 536–544.
- [32] G.H. Seijts, G.P. Latham, The effect of distal learning, outcome, and proximal goals on a moderately complex task, *J. Org. Behav.: Int. J. Ind. Occupat. Org. Psychol. Behav.* 22 (3) (2001) 291–307.
- [33] G.P. Latham, G.H. Seijts, The effects of proximal and distal goals on performance on a moderately complex task, *J. Org. Behav.* 20 (4) (1999) 421–429.
- [34] L.D. Ordóñez, M.E. Schweitzer, A.D. Galinsky, M.H. Bazerman, Goals gone wild: the systematic side effects of overprescribing goal setting, *Acad. Manag. Perspect.* 23 (1) (2009) 6–16.
- [35] M.T. Damgaard, H.S. Nielsen, Nudging in education, *Econ. Educ. Rev.* 64 (2018) 313–342.
- [36] C. Groening, C. Binnewies, "Achievement unlocked!"-the impact of digital achievements as a gamification element on motivation and performance, *Comput. Human. Behav.* 97 (2019) 151–166.
- [37] D.J. Simons, C.F. Chabris, Gorillas in our midst: sustained inattention blindness for dynamic events, *Perception* 28 (9) (1999) 1059–1074.
- [38] M. Saqr, J. Jovanovic, O. Viberg, D. Gašević, Is there order in the mess? a single paper meta-analysis approach to identification of predictors of success in learning analytics, *Stud. Higher Educ.* (2022) 1–22.
- [39] G. Lust, J. Elen, G. Clarebout, Students' Tool-use within a web enhanced course: explanatory mechanisms of students' tool-use pattern, *Comput. Human. Behav.* 29 (5) (2013) 2013–2021.
- [40] V. Kovanović, D. Gašević, S. Joksimović, M. Hatala, O. Adesope, Analytics of communities of inquiry: effects of learning technology use on cognitive presence in asynchronous online discussions, *Internet Higher Educ.* 27 (2015) 74–89.
- [41] D. Gašević, J. Jovanovic, A. Pardo, S. Dawson, Detecting learning strategies with analytics: links with self-reported measures and academic performance, *J. Learn. Anal.* 4 (2) (2017) 113–128.
- [42] M.G. Eley, Differential adoption of study approaches within individual students, *Higher Educ.* 23 (3) (1992) 231–254.
- [43] A. Duff, Understanding academic performance and progression of first-year accounting and business economics undergraduates: the role of approaches to learning and prior academic achievement, *Account. Educ.* 13 (4) (2004) 409–430.
- [44] J. Jovanović, D. Gašević, S. Dawson, A. Pardo, N. Mirriahi, Learning analytics to unveil learning strategies in a flipped classroom, *Internet Higher Educ.* 33 (4) (2017) 74–85.
- [45] N. Ahmad Uzir, D. Gašević, W. Matcha, J. Jovanović, A. Pardo, Analytics of time management strategies in a flipped classroom, *J. Comput. Assist. Learn.* 36 (1) (2020) 70–88.
- [46] M. Bannert, P. Reimann, C. Sonnenberg, Process mining techniques for analysing patterns and strategies in students' self-regulated learning, *Metacogn. Learn.* 9 (2) (2014) 161–185.
- [47] J. López-Zambrano, J.A.L. Torralbo, C. Romero, Early prediction of student learning performance through data mining: a systematic review, *Psicothema* 33 (3) (2021) 456–465.
- [48] Y. Cui, F. Chen, A. Shiri, Y. Fan, Predictive analytic models of student success in higher education: a review of methodology, *Inf. Learn. Sci.* 120 (3/4) (2019) 208–227.
- [49] M. Burgermaster, J.H. Son, P.G. Davidson, A.M. Smaldone, G. Kuperman, D.J. Feller, K.G. Burt, M.E. Levine, D.J. Albers, C. Weng, A new approach to integrating patient-generated data with expert knowledge for personalized goal setting: a pilot study, *Int. J. Med. Inform.* 139 (2020) 104158.
- [50] J. Langenhagen, The use of a gamified learning app in accounting education: Exploring the impact of COVID-19, *Higher Education Learning Methodologies and Technologies Online. HELMeTO 2021, Comm. Com. Inf. Sci.* 1542 (2022) 156–169.
- [51] S.B. Kotsiantis, Decision trees: a recent overview, *Artif. Intell. Rev.* 39 (4) (2013) 261–283.
- [52] A. Priyam, G.R. Abhijeeta, A. Rathee, S. Srivastava, Comparative analysis of decision tree classification algorithms, *Int. J. Curr. Eng. Technol.* 3 (2) (2013) 334–337.
- [53] S. Singh, P. Gupta, Comparative study ID3, CART and C4.5 decision tree algorithm: a survey, *Int. J. Adv. Inf. Sci. Technol.(IJAIST)* 27 (27) (2014) 97–103.
- [54] H. Sharma, S. Kumar, A survey on decision tree algorithms of classification in data mining, *Int. J. Sci. Res.(IJSR)* 5 (4) (2016) 2094–2097.
- [55] S. Piramuthu, Input data for decision trees, *Expert. Syst. Appl.* 34 (2) (2008) 1220–1226.
- [56] G. James, D. Witten, T. Hastie, R. Tibshirani, *An introduction to statistical learning*, Springer, New York, NY, 2021.
- [57] S. Rizvi, B. Rienties, S.A. Khoja, The role of demographics in online learning: a decision tree based approach, *Comput. Educ.* 137 (2019) 32–47.
- [58] V. Matzavela, E. Alepis, Decision tree learning through a predictive model for student academic performance in intelligent m-learning environments, *Comput. Educ.: Artif. Intell.* 2 (2021) 100035.
- [59] A.J. Myles, R.N. Feudale, Y. Liu, N.A. Woody, S.D. Brown, An introduction to decision tree modeling, *J. Chemometric. Soc.* 18 (6) (2004) 275–285.
- [60] S. Einig, Supporting students' learning: the use of formative online assessments, *Account. Educ.* 22 (5) (2013) 425–444.
- [61] D. Massoudi, S. Koh, P.J. Hancock, L. Fung, The effectiveness of usage of online multiple choice questions on student performance in introductory accounting, *Iss. Account. Educ.* 32 (4) (2017) 1–17.
- [62] D. Malekian, J. Bailey, G. Kennedy, Prediction of students' assessment readiness in online learning environments: The sequence matters, in: *Proceedings of the 10th International Conference on Learning Analytics & Knowledge*, 2020, pp. 382–391.
- [63] R. Conijn, C. Snijders, A. Kleingeld, U. Matzat, Predicting student performance from LMS data: a comparison of 17 blended courses using moodle LMS, *IEEE Trans. Learn. Technol.* 10 (1) (2017) 17–29.
- [64] D. Gašević, S. Dawson, T. Rogers, D. Gasevic, Learning analytics should not promote one size fits all: the effects of instructional conditions in predicting academic success, *Internet Higher Educ.* 28 (2016) 68–84.
- [65] B.S. Bloom, Time and learning, *Am. Psychol.* 29 (9) (1974) 682–688.
- [66] J. Stallings, Allocated academic learning time revisited, or beyond time on task, *Educ. Res.* 9 (11) (1980) 11–16.
- [67] N.E. Cagiltay, E. Ozcelik, N.S. Ozcelik, The effect of competition on learning in games, *Comput. Educ.* 87 (2015) 35–41.
- [68] C.-H. Chen, C.-C. Shih, V. Law, The effects of competition in digital game-based learning (DGBL): a meta-analysis, *Educ. Technol. Res. Dev.* 68 (2020) 1855–1873.
- [69] S.H. McIntyre, J.M. Munson, Exploring cramming: student behaviors, beliefs, and learning retention in the principles of marketing course, *J. Market. Educ.* 30 (3) (2008) 226–243.
- [70] S.K. Carpenter, N.J. Cepeda, D. Rohrer, S.H.K. Kang, H. Pashler, Using spacing to enhance diverse forms of learning: review of recent research and implications for instruction, *Educ. Psychol. Rev.* 24 (3) (2012) 369–378.
- [71] N. Kornell, Optimising learning using flashcards: spacing is more effective than cramming, *Appl. Cognit. Psychol.* 23 (9) (2009) 1297–1317.
- [72] I. YeckehZaare, V. Mulligan, G. Ramstad, P. Resnick, Semester-level Spacing but Not Procrastination Affected Student Exam Performance, in: *Proceedings of the 12th International Conference on Learning Analytics & Knowledge*, 2022, pp. 304–314.
- [73] A.J. Wile, G.A. Shouppe, Does time-of-day of instruction impact class achievement? *Perspect. Learn.* 12 (1) (2011) 21–25.
- [74] G. Winocur, L. Hasher, Age and time-of-day effects on learning and memory in a non-matching-to-sample test, *Neurobiol. Aging* 25 (8) (2004) 1107–1115.
- [75] I. Khan, A. Pardo, Data2U: Scalable real time student feedback in active learning environments, in: *Proceedings of the 6th International Conference on Learning Analytics & Knowledge*, 2016, pp. 249–253.
- [76] A. Gabadinho, G. Ritschard, N.S. Mueller, M. Studer, Analyzing and visualizing state sequences in r with traminer, *J. Stat. Softw.* 40 (4) (2011) 1–37.
- [77] F. Chen, Y. Cui, Utilizing student time series behaviour in learning management systems for early prediction of course performance, *J. Learn. Anal.* 7 (2) (2020) 1–17.
- [78] B. Harkin, T.L. Webb, B.P.I. Chang, A. Prestwich, M. Conner, I. Kellar, Y. Benn, P. Sheeran, Does monitoring goal progress promote goal attainment? a meta-analysis of the experimental evidence, *Psychol. Bull.* 142 (2) (2016) 198.
- [79] L.J. Rawsthorne, A.J. Elliot, Achievement goals and intrinsic motivation: a meta-analytic review, *Personal. Soc. Psychol. Rev.* 3 (4) (1999) 326–344.
- [80] C. Cruz, M.D. Hanus, J. Fox, The need to achieve: players' perceptions and uses of extrinsic meta-game reward systems for video game consoles, *Comput. Human. Behav.* 71 (2017) 516–524.
- [81] C. Tran, K. Schenke, D.T. Hickey, Design Principles for Motivating Learning with Digital Badges: Consideration of Contextual Factors of Recognition and Assessment, in: *Proceeding of the 11th International Conference of the Learning Sciences*, 2014, pp. 1027–1031. Boulder, Colorado, USA
- [82] E.A. Locke, G.P. Latham, New directions in goal-setting theory, *Curr. Dir. Psychol. Sci.* 15 (5) (2006) 265–268.
- [83] D. Huynh, L. Zuo, H. Iida, An assessment of game elements in language-learning platform Duolingo, in: *2018 4th International Conference on Computer and Information Sciences (ICCOINS)*, IEEE, 2018, pp. 1–4.
- [84] T. Epton, S. Currie, C.J. Armitage, Unique effects of setting goals on behavior change: systematic review and meta-analysis, *J. Consult. Clin. Psychol.* 85 (12) (2017) 1182.
- [85] M. Burke, B. Settles, Plugged in to the community: Social motivators in online goal-setting groups, in: *Proceedings of the 5th International Conference on Communities and Technologies*, 2011, pp. 1–10.