International Journal of Educational Technology in Higher Education

RESEARCH ARTICLE Open Access

Gamification suffers from the novelty effect but benefits from the familiarization effect: Findings from a longitudinal study

Luiz Rodrigues[1*], Filipe D. Pereira[2,3], Armando M. Toda[1,3], Paula T. Palomino[1,4], Marcela Pessoa[5,6], Leandro Silva Galvão Carvalho[6], David Fernandes[6], Elaine H. T. Oliveira[6], Alexandra I. Cristea[3] and Seiji Isotani[1]

*Correspondence: lalrodrigues@usp.br
[1] Institute of Mathematics and Computer Science, University of São Paulo, Avenida Trabalhador São Carlense, 400 - Centro, São Carlos, SP 13566-590, Brazil
Full list of author information is available at the end of the article

Abstract

There are many claims that gamification (i.e., using game elements outside games) impact decreases over time (i.e., the novelty effect). Most studies analyzing this effect focused on extrinsic game elements, while fictional and collaborative competition have been recently recommended. Additionally, to the best of our knowledge, no long-term research has been carried out with STEM learners from introductory programming courses (CS1), a context that demands encouraging practice and mitigating motivation throughout the semester. Therefore, the main goal of this work is to better understand how the impact of a gamification design, featuring fictional and competitive-collaborative elements, changes over a 14-week period of time, when applied to CS1 courses taken by STEM students (N = 756). In an ecological setting, we followed a 2x7 quasi-experimental design, where Brazilian STEM students completed assignments in either a gamified or non-gamified version of the same system, which provided the measures (number of attempts, usage time, and system access) to assess user behavior at seven points in time. Results indicate changes in gamification's impact that appear to follow a U-shaped pattern. Supporting the novelty effect, the gamification's effect started to decrease after four weeks, decrease that lasted between two to six weeks. Interestingly, the gamification's impact shifted to an uptrend between six and 10 weeks after the start of the intervention, partially recovering its contribution naturally. Thus, we found empirical evidence supporting that gamification likely suffers from the novelty effect, but also benefits from the familiarization effect, which contributes to an overall positive impact on students. These findings may provide some guidelines to inform practitioners about how long the initial contributions of gamification last, and how long they take to recover after some reduction in benefits. It can also help researchers to realize when to apply/evaluate interventions that use gamification by taking into consideration the novelty effect and, thereby, better understand the real impact of gamification on students' behavior in the long run.

Keywords: Gameful, Education, Learning, Quasi-experiment Repeated-measures

Introduction

Game elements have been widely added outside games, approach known as gamification, with the goal of affecting user behavior Deterding et al. (2011); Helmefalk (2019). The educational domain is where gamification has been researched the most Koivisto and Hamari (2019), which is in line with the view that it is an educational innovation that can enhance learning experiences Palomino et al. (2020); Hernández et al. (2021). Accordingly, empirical evidence supports its positive effect on, for instance, behavioral (e.g., class attendance) and cognitive (e.g., test performance) learning outcomes Briffa et al. (2020); Huang et al. (2020). However, researchers often discuss that gamification suffers from the *novelty effect* Hamari et al. (2014); Nacke and Deterding (2017). That is, it leads to positive outcomes when introduced and, as its novelty goes, so does the positive effect Clark (1983).

Studies have provided empirical evidence supporting the novelty effect in gamified learning. Those studies show that, for instance, users' perceived benefits from a gamified service decreased as the time using that service increased Koivisto and Hamari (2014) and that there are cases when gamification's impact on student behavior and motivation changed after an initial positive effect Sanchez et al. (2020); Rodrigues et al. (2021). Such findings question whether gamification can lead to long-lasting benefits. Meta-analytic evidence, on the other hand, suggest the moderator effect of intervention duration on behavioral outcomes is nonsignificant Sailer and Homner (2019). Accordingly, researchers have called for longitudinal studies tracking gamification's impact over long time-periods Bai et al. (2020); Alsawaier (2018); Kyewski and Krämer (2018).

This article responds to literature calls by presenting a longitudinal analysis of gamification's effects on behavioral outcomes over the course of a full semester. Specifically, we analyze data of Brazilian undergraduate students from 14 STEM[1] courses, which were enrolled at the Introductory Programming (CS1) subject. Gamification is especially valuable in that context because, oftentimes, CS1 is not motivating for STEM students, which, along with the complexity and substantial practice needed to learn the topic, leads to high drop-out and failure rates Santana and Bittencourt (2018); Fonseca et al. (2020); Pereira et al. (2020). However, while gamification might be a way to encourage positive learning behaviors for CS1, STEM learners, there is a lack of empirical evidence on gamification's long-term effects in that context. We address this gap by answering *How does gamification's effects on CS1, STEM students' behavioral outcomes change over time?*

Despite past research has studied similar questions Hanus and Fox (2015); Sanchez et al. (2020); Tsay et al. (2020), this one differs in two main perspectives, according to our best knowledge. First, research context: The literature indicates gamification's effect varies depending on the context Hamari et al. (2014); Liu et al. (2017); Toda et al. (2020), and a single research analyzed how gamification's effect changes over time when applied to CS1 students Rodrigues et al. (2021). Second, the gamification design. Meta-analyses indicate that gamified systems including *collaborative competition* and *fictional* game elements improve gamification's effect Sailer and Homner (2019); Huang et al. (2020);

[1] Science, Technology, Engineering, and Mathematics.

Toda et al. (2019), but only Putz et al. (2020) implemented both of those and Mavletova (2015) implemented one: fictional elements. However, those studies are limited to interventions lasting from a day to six weeks, while past research has called for longer analyses Bai et al. (2020); Alsawaier (2018); Nacke and Deterding (2017). This study address both of these perspectives, thus, contributing *new empirical evidence revealing how the effect of a gamification design featuring both fictional and competitive-collaborative elements changes, over a 14-week period, when applied to CS1, STEM students*.

Next, we review related work in terms of longitudinal studies assessing gamification's impact over time, discussing the main points in which this article adds to the literature ("Related work" Section). Then, we describe our study ("Method" Section) and present its results ("Results" Section). Subsequently, we discuss our findings as well as their implications and limitations ("Discussion" Section). Lastly, we draw our conclusions in Conclusions" Section.

Related work

Research has provided empirical evidence that users' perceptions about gamification usage change over time. Koivisto and Hamari found that the more users used a gamified app, the less the perceived benefit they reported was Koivisto and Hamari (2014). Similarly, Van Roy and Zaman found learners' motivations regarding a gamified course decreased throughout the semester, with indication that it then increased towards the end of the semester Van Roy and Zaman (2018). However, these studies are limited by not featuring a baseline comparison group. That is, a condition with no gamification. Hence, they do not allow the identification of whether the gamification's impact (the difference between conditions) changed overtime. This study addresses this need by comparing data from a gamified setting to a control condition with no gamification. Given that context, the remainder of this section reviews longitudinal studies focused on changes of gamification's effects over time. That is, empirical research comparing a gamified condition to a non-gamified setting (to identify gamification's effects) based on measures captured two or more times.

Some research analyzed changes on gamification's effect over time within classrooms. Hanus and Fox implemented gamification (using Badges, Coins, and a Leaderboard) to assess its effects on Communications students Hanus and Fox (2015). Measures were captured three times at every four weeks, for 14 weeks, revealing that the gamification's impact was nonsignificant at the first time point and negative for the subsequent ones for motivation, satisfaction, and empowerment. Grangeia et al. evaluated gamification's impact with final-year medical students, implemented using Levels, Medals, and Encouraging messages Grangeia et al. (2019). The authors evaluated log-data on a daily basis, longitudinally measuring student behavior and interaction with the technology used during a whole course (four weeks). Their results suggest a positive impact of gamification throughout the four weeks. Sanchez et al. added Progress bars, a Wager option, and Encouraging messages to quizzes offered in classrooms for psychology-related majors Sanchez et al. (2020). Students' performances in three tests were analyzed, which were captured from varied intervals (i.e., spacing varied between 3 and 5 weeks). Results indicated gamification was positive for the first test, but nonsignificant for the remaining ones. Tsay et al. explored gamification effects with undergraduate learners in the context

Table 1 Summary of related works and this study

Source	GF	CC	Context	#M	MS	ID
Mavletova (2015)	X		Online survey completion	2	2m	<1d
Hanus and Fox (2015)			Communication classroom	3	4w	16w
Mitchell et al. (2017)			Adults physical activities	3	1w	4w
Grangeia et al. (2019)			Medical students	28	1d	4w
Sanchez et al. (2020)			Psychology-related classrooms	3	3-5w	13w
Tsay et al. (2020)			Personal and Professional Development classroom	24	1w	24w
Putz et al. (2020)	X	X	Workshops on sustainable supply chain management	2	2w	<1d
Rodrigues et al. (2021)			CS1, Software Engineering classrooms	6	1w	6w
This study	X	X	CS1, STEM classrooms	7	2w	14w

GM Game fiction, *CC* Competitive-collaborative, *#M* Number of measures, *MS* Measures spacing, *ID* Intervention duration, *m* months, *w* weeks, *d* day

of a Personal and Professional Development module Tsay et al. (2020). Based on log data from two terms (24 weeks), the authors analyzed gamification's (Quests, Progression, Badges, and Leaderboard) impact weekly, finding that it decreased over time in the first term, but not in the second, results they attributed to changing the gamification design. Rodrigues et al. applied gamification (Badges, Objectives, and Collaboration) with CS1 undergraduates majoring in Software engineering Rodrigues et al. (2021). Measuring their motivation weekly, during six weeks, they found mixed effects that changed over time: at first, gamification was positive and negative for participants with low and high familiarity with coding, respectively. After the six weeks, the effects were the opposite.

Others performed similar research in domains apart from classrooms. Mitchell et al. experimented with a gamified app—featuring Variable difficulty levels, Player choice, and Dynamic feedback—in the context of physically active adults Mitchell et al. (2017). Authors measured participants' motivational and behavioral engagement three times over the course of four weeks, with one week of spacing between them. Unlike most research, findings indicated a positive gamification effect over time, although the decrease in the effect's magnitude. Mavletova conducted a study in the context of children completing surveys, analyzing gamification impact in two waves that had a two-month spacing Mavletova (2015). Results from using Narrative (a form of game fiction), Rules, Challenges, and Rewards were only positive in the first wave and, for some measures (e.g., survey test-retest reliability), negative in the second. Putz et al. analyzed gamification's impact on knowledge retention from workshops on sustainable supply chain management for two time points: 20 minutes and 14 days after the end of the intervention, which lasted around six hours Putz et al. (2020). The game elements used in the workshops were Time constraint, Storytelling (a form of game fiction), Rewards, Leaderboard, Collaborative Competition, Clear goals, Immediate feedback, Avatar, Points. Their findings suggest that gamification was positive for short-term knowledge retention, but roughly ineffective at the second time-point.

Summary

Compared to related work, this study advances the literature based on two main perspectives, summarized in Table 1. The first concerns the context of the study. The literature argues that gamification's success depends on the context Hamari et al. (2014); Liu

et al. (2017); Toda et al. (2020). Thus, performing similar studies in different contexts is important to advance the understanding to new contexts Seaborn and Fels (2015). This study differs from the available body of research by analyzing changes of gamification's effect over time, in the context of CS1. While one study also worked with CS1 learners Rodrigues et al. (2021), they were not STEM students and the experiment was limited to 19 participants studied for six weeks. Differently, this study features 756 participants from STEM courses that were studied over a 14-week period.

The second perspective concerns the gamification design. The gamification design employed in most studies explored extrinsic, challenge-based game elements, such as rewards and leaderboards Toda et al. (2019). However, meta-analytic evidence demonstrated that adding fictional game elements (e.g., Narrative and Storytelling Palomino et al. (2019)), as well as competitive-collaborative social interactions, enhanced the effect of gamification on behavioral learning outcomes Sailer and Homner (2019). Despite that, a single study, in which the intervention lasted less than a day and measures were taken only two times, explored both kinds of elements Putz et al. (2020). Differently, this research explores both fictional and competitive-collaborative elements in a 14-week intervention, analyzing data from seven time-points. Thereby, our study is contributing to the understanding of how the impact of gamification design changes over a semester.

Based on those premises, this article's novelty relies in analyzing how the gamification effect change over time, in a context (CS1, STEM students) it has not been previously explored in the long-run (14 weeks), with a large sample (756) and using both fictional and competitive-collaborative game elements.

Method

This study responds to calls for longitudinal gamification studies Nacke and Deterding (2017); Bai et al. (2020). Accordingly, the objective is to answer the following research question:

> *How does gamification's effects on CS1, STEM students' behavioral outcomes change over time?*

Therefore, we report results stemming from a quasi-experimental study, conducted in an ecological setting, from 2016 to 2018. The study was conducted in the context of CS1 courses of 14 STEM undergraduation programs from Federal University of Amazonas (UFAM). The experimental tasks concern activities that were already included in the course schedule, which allowed studying learners in an ecological setting. The quasi-experimental characteristic emerges because random assignment was not possible, a common feature of studies performed in ecological settings Creswell and Creswell (2017). Instead, the assignment of conditions was based on the year in which a student enrolled in a discipline, within the context of the study. This assignment procedure was necessary, because the system used for data collection was developed and deployed in 2016, without the innovation of gamification and, only since 2017, its gamification design was deployed. Hence, students from 2016 were assigned to the control condition ($N = 138$), whereas those from 2017 and 2018 were assigned to the experimental condition ($N = 659$).

Design

This study follows a 2x7 design. The first factor is *gamification*, in which levels are *yes* and *no*. The second factor is *time*, wherein levels range from one to seven. Time one refers to data captured until the end of the second week of the intervention, and subsequent times refer to data captured within two-week intervals, until week 14. Based on this design, the first factor enables the identification of the gamification impact, whilst the second provides data for analysing changes of the gamification effect over time.

Participants

Seven-hundred and fifty-six (N = 756[2]) Brazilian students participated in this research. Their average age was 22.2 years old (Standard Deviation, SD: 4.2); 62.3% were males and 37.7% were females. All of them were students of the UFAM and were enrolled in a CS1 course of one of the following STEM classes: Computer Science (2%), Materials Engineering (9%), Oil and Gas Engineering (9%), Production Engineering (9%), Electrical Engineering (13%), Mechanical Engineering (9%), Chemical Engineering (10%), Statistics (7%), Bachelor in Physics (7%), Licensure in Physics (6%), Bachelor in Mathematics (5%), Licensure in Mathematics (11%), and Applied Mathematics (3%). As this study was conducted in an ecological setting, all students enrolled in any of those courses were eligible as participants. However, we only considered those 756 learners, because of two reasons. First, they properly completed the form providing informed consent to having their data used in scientific research and basic demographic information (e.g., age and gender). That was accomplished at the beginning of the first class of the term, before students started using the system. Hence, our approach complied to ethical standards and allowed us to understand characteristics of our study sample. Second, students completed the semester without failing by attendance. This is desirable as our goal is evaluating how the effect of gamification changes over time. That led to the removal of 19.7% and 24.2% of the students, who failed by attendance in the non-gamified and gamified groups, respectively.

Materials

Educational system

According to the ecological setting of the study, data was collected through Codebench[3]. It has been developed at UFAM, Brazil, and since 2016 has been used for supporting teaching the CS1 subject of all STEM classes of the university. Figure 1 shows a snapshot of the system's main usage page, taken from Galvão et al. (2016) with copyright holder's authorization.

Codebench is an online judge, which has been used to help both instructors and students. For instructors, the system contributes by providing a list of programming problems they can use to create assignment lists, to use as activities in their classes. For students, the system offers an IDE (Integrated Development Environment) where they can develop solutions for programming problems, suggested by instructors. Furthermore, the system provides automatic feedback about solution correctness, which is

[2] Forty-one participants were in both groups due to the ecological setting of the experiment.

[3] https://codebench.icomp.ufam.edu.br/

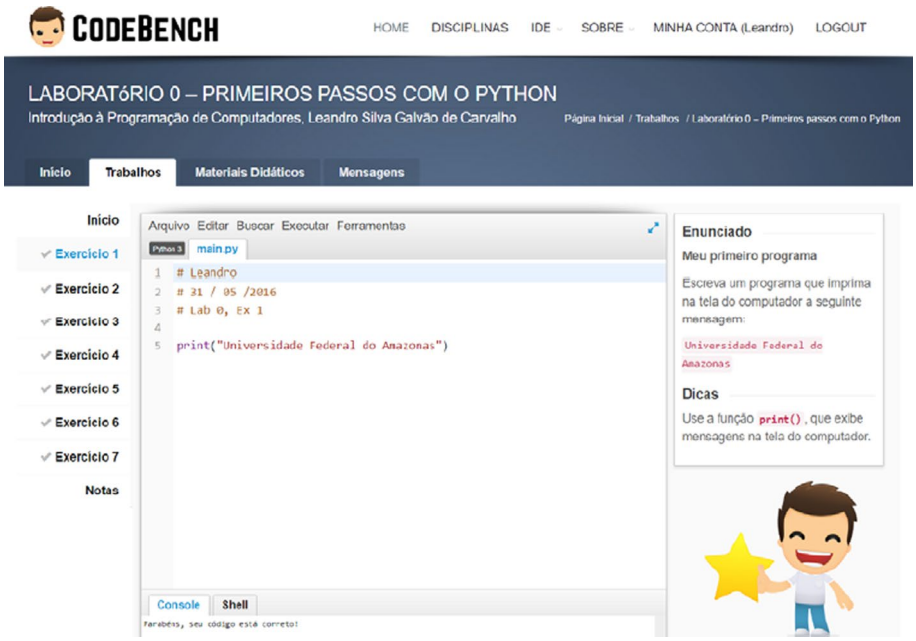

Fig. 1 Snapshot of the system participants used to complete this study's task. That image was taken from Galvão et al. (2016) with copyright holder's authorization

possible by testing each solution of a given problem against a set of hidden test cases. Hence, it helps students with instantaneous feedback, and instructors, by reducing the workload of correcting learner solutions.

Task

Given the nature of the system used for data collection, the task analyzed in this study concerns programming assignments students solved in Codebench. These assignments were offered by instructors, mainly for students to practice concepts and techniques introduced in classes, although solving them was only worth a small proportion of their final grade (around 6%). Hence, when analyzing gamification impact over time, we consider data from students solving non-compulsory programming assignments, which had a small impact on their final grade and were suggested by instructors. Moreover, we highlight that although the participants were learners from different STEM classes, all programming problems were selected from the same database, regardless of the class, and the same pedagogical materials and methodological plan were followed in the CS1 subject for all STEM courses. Thereby, this is mitigating confounding effects, which could emerge from having participants from different courses.

Experimental conditions

According to the need for first identifying gamification impact and, then, how it changes over time, this study features two experimental conditions. The first is the *control* condition, in which participants used Codebench without gamification. The second condition is the *experimental* one, in which participants used the same system, but with game elements; that is, they used the system's gamified version.

The gamified version was developed following the proposal by von Wangenheim and von Wangenheim (2012), integrating the Instructional Systematic Design model Dick et al. (2005) and the instrumental design ADDIE Branch (2009). The gamification goal is to motivate students to complete programming assignments, either during face-to-face or online classes, when more suitable to the learners, but always using Codebench. To achieve that goal, the gamification allows the students to see themselves as characters from a medieval fantasy world (narrative). The students can customize these characters (i.e., their avatar) and, if they receive a high score in a series of assignments, their avatar gets a badge that leads to a more powerful weapon. Furthermore, the gamification design also enables students to join their classmates (collaboration) to fight a monster, aiming to free the fantasy world from it. For this goal, there is a map that the students explore, as they solve programming assignments.

Note that as Codebench can be used by different subjects (i.e., not only CS1), the gamification design is not tied to any specific topic. Therefore, the integration between programming assignments and gamification happens once a student completes any assignment. When they submit a solution to the system, if the solution is correct, the student receives an in-game reward. That reward is a card selected through a draw, which can bring strength points and weapons to the student's avatar. The greater the strength, the better the student's weapon becomes, and the more hit points they take from the monster. Consequently, they have more chances of remaining in the best positions on the leaderboard, which is helpful, because the top three students receive a special reward. Nevertheless, a single student cannot defeat the monster alone, as this is a collective mission. Thus, a percentage of the class must reach the end of the map, to be able to beat the monster and, finally, win the challenge together (cooperation).

Given this context, we might summarize the two main elements of the gamification design as follows. First, the narrative aspect immerses students into a fictional story. To advance in that story, students have to complete real programming assignments—hence, connecting the learning activities to the gamified design. Second, the students have to collaborate to finish the gamified story, while they also compete with each other during that process, towards staying at the top of the leaderboard and gaining rewards to improve and customize their avatars. Thus, the competitive-collaborative aspect emerges, because students must work together; at the same time, they have to outperform others, to receive more rewards.

Measures

We analyze gamification impact based on three behavioral outcomes, which were all collected through data logs from Codebench. Those are:

- **Attempts:** The number of attempts that a students made. That is, how many times they submitted their solution to the online judge for checking the correctness of the code. Consequently, this measure accounts for both correct and incorrect submissions.
- **IDE Usage:** The time (in minutes) that a student used the IDE integrated in the online judge. Note that inactive usage does not count in this metric. Hence, we can see it as

a measure of the time a student spent working towards improving their programming skills; that is, practicing to code.

- **System Access:** How many times a student accessed the online judge. Thereby, it accounts for each time a student logged into the system. Thus, we can see it as a measure of students' intentions to practice, consult learning materials, and interact with the gamification.

Those measures were selected considering that practicing is prominent to learn to code Estey and Coady (2016); Pereira et al. (2021). We can see **attempts** as a measure of relevant learning behavior based on the testing effect: the idea that the more one is tested, the more one learns Rowland (2014); Rodrigues et al. (2021); Sanchez et al. (2020). Specifically, empirical evidence supports that the simple act of trying to complete a test is valuable for learning Goldstein (2014); Ahadi et al. (2016). Therefore, we might expect that the more attempts a student has, even if those are failed attempts, the better for their learning. For similar reasons, we might expect that **IDE usage** will benefit learning as well. For instance, empirical evidence demonstrates that time-on-task is positively related to learning outcomes Landers and Landers (2014); Pereira et al. (2020), which further supports the positive role spending time on the IDE has on learning. Lastly, **system access** reflects valuable learning behaviors, as it captures students' interest in practicing (i.e., using the IDE to submit a solution—attempt) and interacting with the gamification.

Procedure

According to the study setting, participants followed a straight-forward procedure. First, they provided informed consent for their data to be used in scientific research, and completed a brief demographic questionnaire, with information such as gender and age. Then, students were provided with assignment lists they could complete in Codebench. Note that although optional, completing these lists was worth 6% of the students' final grade. By completing the programming task-lists on the online judge, data used for analysis was generated, as the online judge automatically captured the participants' interactions, storing it onto log data. At the end of the semester, these logs were grouped, to represent students' interactions at each of the seven points in time analyzed in this study.

Data analysis process

Given the design of this study, we use a Robust ANOVA based on trimmed means Wilcox (2017) for data analysis. We chose this method because it behaves better than the traditional ANOVA for real world situations, while maintaining the flexibility of the traditional approach, as the literature discusses and recommends Cairns (2019). The rationale is that robust ANOVA using trimmed means maintains the ability to compare changes in location (e.g., how much one is bigger than the other), unlike rank-based tests (e.g., Kruskal-Wallis), while still providing results robust to violations of assumptions, such as non-normality and homogeneity of sphericity Keselman et al. (2000); Kowalchuk et al. (2003) (for a comprehensive discussion, see Cairns (2019)). Hence, we use the

Robust ANOVA[4] to test whether there is an interaction between the two factors (gamification and time). This can indicate how the effect of gamification changes over time.

In cases of significant interactions, we perform post-hoc tests using the Yuen-Welch test Yuen (1974), following literature recommendations Cairns (2019). Here, we limit the post-hoc analysis to testing the difference between conditions (gamified and non-gamified) for each time-point, aiming to identify the magnitude of the difference, as well as finding how they change (by how much they increase/decrease) over time. Accordingly, difference magnitudes are estimated using the explanatory measure of effect size introduced by Wilcox and Tian Wilcox and Tian (2011), which is a robust measure, aligned to the Yuen-Welch test. Effect sizes of 0.1, 0.3, and 0.5 are interpreted as small, moderate, and large, respectively. The alpha level is set to 0.05 for all tests and p-values are corrected using the false-discovery-rate approach, which has been recommended over the more common Bonferroni alternative Jafari and Ansari-Pour (2019). Inferential analyses were conducted using R R Core Team (2020), R studio RStudio Team (2019), and the *WRS2* package Mair and Wilcox (2018).

Results

First, this section presents preliminary analyses due to the study design. Then, it introduces analyses and results answering our research question.

Preliminary analyses

Due to the lack of random assignment, we compare the demographic characteristics of participants for each experimental condition. In terms of gender, 38% and 62% of the participants were females and males, respectively, for both experimental conditions. In terms of age, those in the control condition had an average age of 23.3 years (SD: 4.0), whereas those in the experimental setting were, on average, 22 years old (SD: 4.2). While there was no difference in gender, the difference in age might confound the results. Then, we further examined whether age is a predictor for any of the behavioral measures.

- For attempts, linear regression results were statistically significant, $F(1, 7782) = 18.42$, $p < 0.001$, $R^2 = 0.002$, adj-$R^2 = 0.002$, and age was a significant predictor, $t = -4.29$, $p < 0.001$, $B = -0.58$, $\beta = -0.05$.
- For IDE usage, linear regression results were not statistically significant, $F(1, 7782) = 2.48$, $p = 0.116$, $R^2 <$ textless 0.001, adj-$R^2 < 0.001$.
- For system access, linear regression results were statistically significant, $F(1, 7782) = 12.25$, $p < 0.001$, $R^2 = 0.002$, adj-$R^2 = 0.001$, and age was a significant predictor, $t = -3.5$, $p < 0.001$, $B = -0.38$, $\beta = -0.04$.

These findings indicate that, although statistically significant for two of the three behavioral measures, age's practical effect is negligible, based on the proportion of variance explained (all R^2 <1%) and standardized coefficients (all $\beta \leq 0.05$) Kotrlik and Williams (2003). Therefore, these results mitigate the threat of age confounding our results as the

[4] Note that we rely on a robust alternative because it handles such violations when they are present, as well as provides similar results to the standard approach, otherwise Cairns (2019); Wilcox (2017). Additionally, we do not test for assumption violations (e.g., normality) as the reliability of such tests has been critiqued Cairns (2019).

Table 2 Descriptive statistics for behavioral outcomes of participants from control (CTR; non-gamified) and experimental (EXP; gamified) conditions for seven time-points (TP) of the semester, as well as overall (1-7). Data represented as Mean±Standard Deviation

	Attempts		IDE usage		System access	
TP	CTR	EXP	CTR	EXP	CTR	EXP
1-7	40±51	47±48	80±94	119±118	29±24	64±39
1	62±53	85±52	112±64	185±103	19±19	48±27
2	35±33	57±57	56±42	126±82	22±12	52±30
3	68±73	61±53	134±134	185±159	33±16	60±32
4	42±59	36±36	89±113	108±112	42±23	66±36
5	29±37	32±35	76±90	95±107	25±26	67±39
6	22±26	28±31	48±68	74±93	30±27	73±44
7	24±35	29±37	47±81	60±83	35±31	79±47

difference in groups' overall age is unlikely to be the source of relevant differences in student behavior.

Does gamification's effect change over time?

Table 2 presents an overview of the behavioral measures, overall (1-7) and for each time-point, for both experimental conditions. The table shows that participants in the experimental condition (gamified): attempted problems, used the IDE, and accessed the system 17.5%, 48.6%, and 220% more, respectively, than those with the non-gamified setting. These magnitude differences are small ($g = 0.144$), moderate ($g = 0.341$), and large ($g = 0.948$) according to Hedge's g[5]. Similarly, Fig. 2 presents boxplots that demonstrate the distribution of each behavioral measure for each group in each session. Next, we present results from hypothesis tests, to provide statistical evidence on how the impact of gamification changes over time.

First, we analyzed student attempts. A robust two-way ANOVA on 20% trimmed means was used to evaluate the effects of *gamification* and *time* on student attempts. A statistically significant interaction between the effects of gamification and time was found, $F(6, 410.3948) = 4.5677$, $p < 0.001$. The main effects of gamification, $F(1, 429.5147) = 41.2818$, $p < 0.001$, and time, $F(6, 409.9150) = 73.1199$, $p < 0.001$, were statistically significant as well. Next, we conducted post-hoc analyses using the Yuen-Welch tests to assess differences of gamification impact over the seven time points. Table 3 presents the results from comparing the attempts of participants of both experimental conditions. It shows results from the Yuen-Welch test, comparing groups and estimates of the difference in location among groups. The results reveal moderate to large statistically significant differences in the number of attempts in time points one and two, insignificant differences in time-points three to five, and small to moderate statistically significant differences in time-points six and seven.

Second, we analyzed students' IDE usage. We repeated the previous analysis, but with students' IDE usage as dependent variable. A statistically significant interaction between

[5] Calculated using www.polyu.edu.hk/mm/effectsizefaqs/calculator/calculator.html

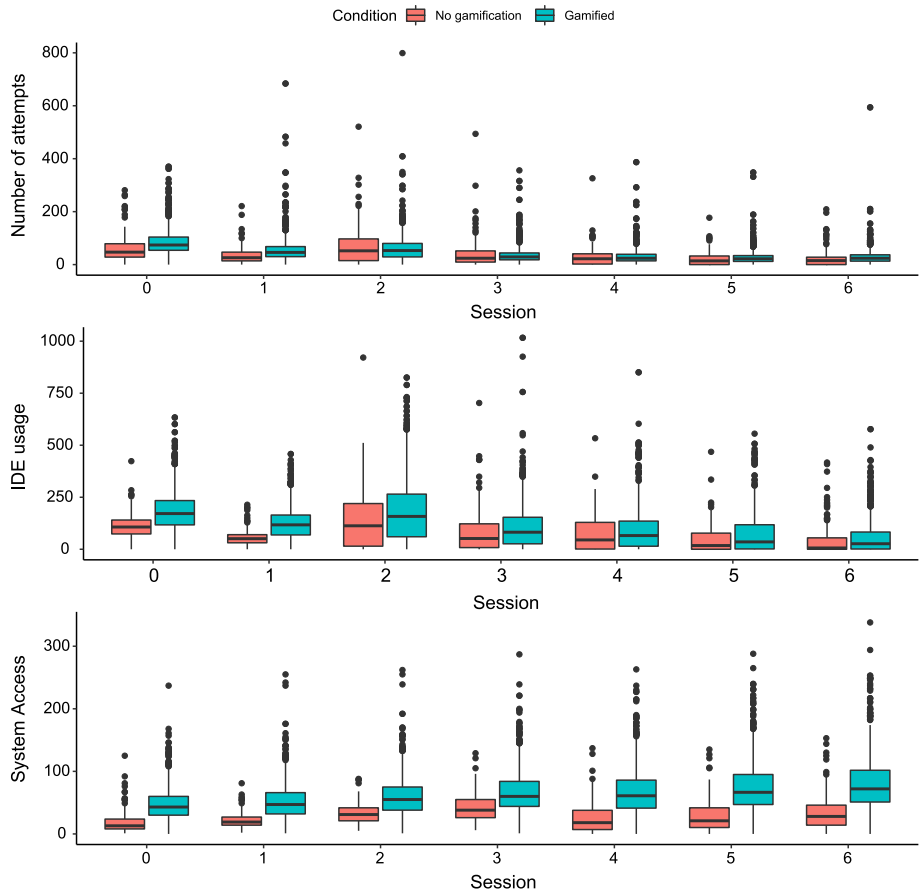

Fig. 2 Boxplots demonstrating the distribution of each behavioral measure of each condition for each session

Table 3 Comparison of experimental conditions (control, non-gamified; and experimental; gamified) in terms of participant attempts for each point in time (TP)

	Yuen-Welch test				Effect size	
TP	t(df)	p	p-adj	95% CI	ES	95% CI
1	7.872 (121.11)	0.000	0.000	− 33.123 − 19.811	0.448	0.305 0.596
2	7.781 (130.84)	0.000	0.000	− 22.894 − 13.613	0.443	0.296 0.584
3	0.023 (97.45)	0.982	0.982	− 10.969 10.717	0.033	0.000 0.194
4	0.717 (95.90)	0.475	0.555	− 8.500 3.991	0.061	0.000 0.249
5	1.468 (102.10)	0.145	0.203	− 8.285 1.236	0.100	0.000 0.260
6	3.523 (103.19)	0.001	0.001	− 11.192 − 3.130	0.240	0.070 0.378
7	5.407 (113.15)	0.000	0.000	− 13.916 − 6.453	0.331	0.178 0.467

the effects of gamification and time was found, $F(6, 412.7856) = 7.5375$, $p < 0.001$. The main effects of gamification, $F(1, 440.7169) = 91.5288$, $p < 0.001$, and time, $F(6, 412.3179) = 93.1663$, $p < 0.001$, were statistically significant as well. Next, we conducted post-hoc analyses using the Yuen-Welch tests, to assess differences in gamification impact over the seven time-points. Table 4 presents the results from comparing the IDE

Table 4 Comparison of experimental conditions (control, non-gamified; and experimental; gamified) in terms of participants' IDE usage in each time-point (TP)

	Yuen-Welch test				Effect size	
TP	t(df)	p	p-adj	95% CI	ES	95% CI
1	12.132 (186.58)	0.000	0.000	− 78.997 − 56.898	0.599	0.469 0.719
2	18.060 (282.93)	0.000	0.000	− 73.875 − 59.354	0.717	0.594 0.812
3	3.347 (120.61)	0.001	0.002	− 72.908 − 18.713	0.213	0.064 0.367
4	3.692 (126.42)	0.000	0.001	− 44.746 − 13.517	0.214	0.067 0.368
5	1.935 (113.97)	0.056	0.056	− 34.214 0.405	0.123	0.000 0.288
6	3.233 (154.58)	0.002	0.002	− 31.532 − 7.611	0.199	0.038 0.346
7	3.227 (132.07)	0.002	0.002	− 25.468 − 6.109	0.199	0.037 0.354

Table 5 Comparison of experimental conditions (control, non-gamified; and experimental; gamified) in terms of participants' system access for each time-point (TP)

	Yuen-Welch test				Effect size	
TP	t(df)	p	p-adj	95% CI	ES	95% CI
1	22.458 (199.67)	0.000	0.000	− 32.564 − 27.307	0.794	0.721 0.882
2	22.252 (303.48)	0.000	0.000	− 30.655 − 25.673	0.822	0.737 0.899
3	15.019 (193.11)	0.000	0.000	− 27.913 − 21.433	0.708	0.582 0.806
4	10.696 (152.51)	0.000	0.000	− 26.744 − 18.404	0.561	0.440 0.678
5	18.783 (152.56)	0.000	0.000	− 47.941 − 38.816	0.783	0.680 0.863
6	18.640 (163.20)	0.000	0.000	− 49.943 − 40.375	0.780	0.670 0.860
7	16.877 (153.56)	0.000	0.000	− 50.217 − 39.692	0.749	0.629 0.832

usage of participants of both experimental conditions. It shows results from the Yuen-Welch test, comparing groups and estimates of the difference in location among groups. The results reveal large, statistically significant differences in IDE usage for the first two time-points and small, statistically significant differences for all the remaining ones, with the exception of point five, in which the difference is only marginally significant.

Third, we analyzed system access, with the approach used for the other measures. A statistically significant interaction between the effects of gamification and time was found, $F(6, 422.4725) = 7.4256$, $p < 0.001$. The main effects of gamification, $F(1, 840.7586) = 672.3375$, $p < 0.001$, and time, $F(6, 422.6884) = 30.5048$, $p < 0.001$, were statistically significant, as well. Next, we conducted post-hoc analyses using the Yuen-Welch tests, to assess differences of gamification impact over the seven points in time. Table 5 presents the results from comparing the system access of participants of both experimental conditions. It shows results from the Yuen-Welch test, comparing groups and estimates of the differences among groups. The results reveal large, statistically significant differences in IDE usage for all time points, despite the decrease in time-point four, which was readily recovered in time-point five.

How does the effect of gamification change over time?

The way how the effect of gamification changes over time, based on the effect sizes estimated in previous analyses, is shown in Fig. 3. In this figure, time-points are

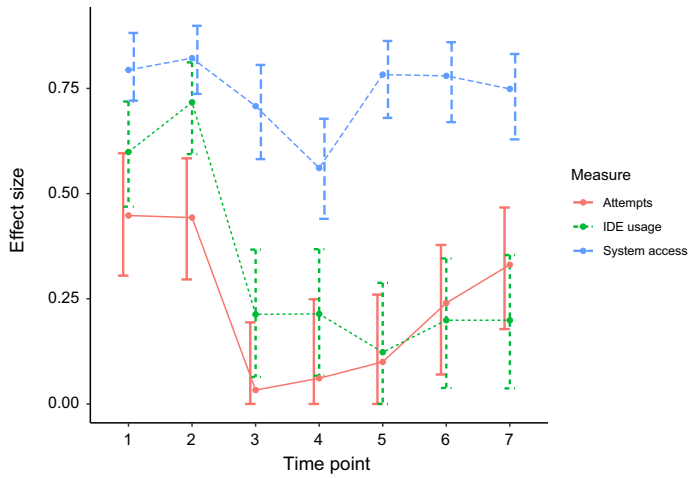

Fig. 3 Gamification's effect (compared to no gamification) over time, based on effect sizes estimated from differences in behavioral outcomes. Error bars represent the 95% confidence intervals estimated from effect sizes

shown on the x-axis and the magnitude of the effect size is shown on the y-axis, with each measure represented by a different line type and color and presenting error bars, based on the confidence intervals estimated in the analyses presented previously. The figure demonstrates that, in terms of attempts, the effect of gamification substantially decreased after the second time-point but, thereafter, progressively increased, until time-point seven. For IDE usage, the figure demonstrates that the effect of gamification increased from time one to time two but, subsequently, incrementally decreased, until time-point five and, in the end, suffered a small increase. In addition, the figure shows that the effect of gamification on system access also increased from time point one to time-point two, then decreased until time-point four, subsequently shifting to an increase in time-point five and, lastly, roughly decreasing, until time-point seven.

Overall, these results suggest that, over time, the effect of gamification on behavioral learning outcomes change, following, to some extent, a U-curve. We must note, however, that this conclusion is the product of an exploratory analysis. That is, while we expected that the effect of gamification would vary over time, we had no assumption on the shape of such a variation, prior to the data analysis. Accordingly, we intentionally chose to first analyze variations in gamification effect based on effect sizes estimated during post-hoc testing. Then, we analyzed such results by inspecting plots featuring confidence intervals. Importantly, our approach differs from the standard, quantitative analyses that mainly rely on p-values (e.g., a polynomial regression). Our justification is that the literature discourages using such a quantitative approach in the context of exploratory studies Vornhagen et al. (2020); Cairns (2019); Dragicevic (2016). Therefore, according to the exploratory nature of our analysis, we followed a more qualitative, visual interpretation, based on confidence intervals, as the above literature recommends. Thus, summarizing such results, we have i) empirical evidence that the gamification effect does change over time and ii) a model (i.e., the U-shaped

function) of how such variation occurs during the course of a 14-week period, which must be further explored and validated in future research.

Discussion

Based on our results, this article's novelty lies in providing the following contributions, which have not been addressed in prior research, to the best of our knowledge:

- Empirical evidence showing the long-term effectiveness of an innovative gamification design, which features fictional and competitive-collaborative game elements, in improving positive learning behaviors;
- Empirical evidence revealing how the effect of gamification decreases (novelty effect) and naturally increases, with no intervention (familiarization effect) over time;
- A heuristic on a 'safe period' in which educators can use gamification before its effect starts to decrease;
- A heuristic on how long gamification studies should last to ensure their findings are roughly stabilized.

To further support these contributions, the remainder of this section discusses our findings, their implications, this article's limitations, and future research recommendations.

Findings

First, our findings revealed that the effect of gamification effect decreased after four weeks, for all behavioral measures, although the magnitude of that change differed between measures. THe effect of gamification either was maintained from the first time-point, or it increased between time-points one and two. However, it decreased from moderate-to-large to negligible, large to small, and had it size diminished for attempts, IDE usage, and system access, respectively. These findings support the novelty effect as, during the first four weeks, the effects were positive, after which they decreased. These findings corroborate most previous research, showing gamification impact decreasing over a period of time (e.g., Mitchell et al. (2017); Putz et al. (2020); Sanchez et al. (2020)). Compared to these, our study expands the literature, by providing evidence that the novelty effect holds in a different context, as well as for a distinct gamification design. In addition, this study expands the literature by demonstrating the extent to which the novelty effect influences the gamification impact, in terms of changes on effect size magnitudes. We also show that changes in magnitudes differ among similar measures, similar to what Mavletova (2015) found in the context of online surveys, yet extended to our context and gamification design. On the other hand, results by Grangeia et al. Grangeia et al. (2019) suggested that gamification impact was positive throughout the whole intervention. Whereas we show that this impact decreased after four weeks, their intervention lasted only four weeks. Therefore, considering the design we employed features elements likely to enhance gamification potential Sailer and Homner (2019), their results might be due to the limited intervention duration, or the different gamification design.

Second, our results indicate that after the decrease, the gamification effect showed a trend of increasingly positive impacts. While the decrease first appeared from time-point two to three, it promptly ended for attempts, but lasted until time-points four

and five for system access and IDE usage, respectively. After the end of the decline, our results show an increase in gamification impact for all measures, which either continued to increase (attempts) or achieved a rough plateau (IDE usage and system access). These findings support the *familiarization effect*, in which users need some time to fully familiarize themselves with gamification. This effect was previously discussed by Van Roy and Zaman Van Roy and Zaman (2018), when they found that learner motivation appeared to follow a polynomial behavior over time. Similarly, Tsay et al. Tsay et al. (2020) noted that, over time, the impact of gamification decreased in the first term, but not in the second one, which also suggests a familiarization effect. However, neither of these studies performed such analyses by comparing data from the gamification intervention to a control, non-gamified group. Therefore, our contribution here is providing empirical support for the familiarization effect, based on a quasi-experimental study, comparing data from gamified and non-gamified interventions, besides extending it to the context of Brazilian students from CS1 classes from STEM courses, as well as to our gamification design. Additionally, it is worth noting other long-term studies (e.g., Hanus and Fox (2015); Sanchez et al. (2020)) are limited in the number of time-points considered (three) and data analysis method (linear modeling), which prevented them from finding such a non-linear pattern. Thus, our findings suggest that the decrease in gamification impact, likely due to the novelty effect, might happen because learners need some time to familiarize themselves with it.

Third, our findings revealed that, overall, the gamification effect was positive, and, importantly, that it was not negative at any point in time. Considering the whole intervention period, all behavioral measures of participants in the gamified condition outperformed those in the control setting. Considering each time-point separately, effect sizes estimated in the analyses corroborate that positive impact, with the exception of a few points in which gamification impact was nonsignificant. This finding is aligned to the overall gamification literature, which shows most studies report positive with null results Koivisto and Hamari (2019). Additionally, these findings are aligned to the gamified learning literature Sailer and Homner (2019), revealing average small-to-moderate positive impacts on behavioral learning outcomes. The literature has demonstrated cases in which gamification applied to education is associated with negative outcomes Toda et al. (2018); Hyrynsalmi et al. (2017), which is harmful to students' learning experience and reinforces the need to carefully designing gamification, to prevent such cases. A possible reason for avoiding these pitfalls might be the inclusion of fictional and competitive-collaborative elements, which have been suggested as positive moderators of gamification's impact on behavioral outcomes Sailer and Homner (2019). In this sense, this study contributes by scrutinizing the effects from the employed gamification design over several time periods, finding that besides being positive on average, it was not harmful to learners at any point in time.

Finally, these findings additionally demonstrate the effectiveness of adding an innovative design to the educational process of teaching programming. On one hand, gamification has been seen as an educational innovation, with the potential to improve learning experiences Hernández et al. (2021); Palomino et al. (2020). However, most research on gamification has been limited to standard designs, featuring elements such as points, badges, and leaderboards Bai et al. (2020); Koivisto and Hamari (2014);

prior gamification research also lacked long-term studies Nacke and Deterding (2017); Alsawaier (2018); Kyewski and Krämer (2018). On the other hand, unlike most prior research, this paper analyzed a gamification design featuring fictional game elements Toda et al. (2019). Additionally, that design also features competitive-collaborative elements, while most research is often limited to leaderboard-based competition Mustafa and Karimi (2021). Such distinction is important, because, despite being rarely used, such fictional and competitive-collaborative game elements are considered maximizers of gamification's effects Sailer and Homner (2019). Therefore, Codebench provides to students an innovative educational design, not only in terms of being gamified, but also because its gamification design itself is innovative, compared to the prior works. Thus, this article contributes with evidence demonstrating that adding an innovative educational design to programming lessons, implemented through gamification featuring both fictional and competitive-collaborative game elements, is able to improve learning behaviors in the long run.

Nevertheless, in addition to the gamification design, this study differs from previous research in terms of the context. Compared to Mavlova Mavletova (2015), Mitchell et al. Mitchell et al. (2017), and Putz et al. Putz et al. (2020), contexts differ, because we focused on the learning domain and they did not. Compared to Hanus and Fox Hanus and Fox (2015), Grangeia et al. Grangeia et al. (2019), Sanchez et al. Sanchez et al. (2020), and Tsay et al. Tsay et al. (2020), contexts differ because neither of them involved CS1. Consequently, our study context differs from theirs in terms of the task: we gamified programming assignments, whereas coding is not part of the classes involved in their research (e.g., psychology, medicine, and communication). Compared to Rodrigues et al. Rodrigues et al. (2021), who also gamified CS1 activities, most of those activities were not coding assignments. Additionally, their participants were Software Engineering students, a course highly linked to coding. In our study, most participants were students of STEM courses, such as Electrical Engineering and Mathematics. That difference is key, as such STEM students often struggle to see the value of coding for their future occupations Fonseca et al. (2020); Pereira et al. (2020). Therefore, we can summarize that our study context differs from those of prior research in terms of domain, learning task, and major. Note that acknowledging such differences is important because research suggests that one's experience with gamification might change, depending on the task Liu et al. (2017); and that different designs should be adopted depending on the learning activity Rodrigues et al. (2019); Hallifax et al. (2019). Consequently, the difference in contexts might have affected our findings when compared to similar studies. Thus, we also contribute to the discussions on the role of contextual aspects (e.g., domain, task type, and major) in gamification, raising the question of how those factors affect gamification effectiveness.

Implications

Our findings have major implications for higher education. That is, *they provide consistent support for the effectiveness of an increasingly popular approach, which has, nevertheless, been rarely analyzed in the long term: gamification*. From our findings, we have empirical evidence that gamifying programming assignments has an overall positive effect on students' submissions (attempts), IDE usage, and system access. Such behaviors

are valuable for learning, as empirical evidence shows that the more one solves questions and spends time on a task, the better their learning is Rowland (2014); Rodrigues et al. (2021); Sanchez et al. (2020). Importantly, our study demonstrates that the innovative design was able to improve positive learning behaviors throughout the course of a whole semester. Thus, our findings' educational implication is that if practitioners deploy a gamification design similar to ours, they are likely to be adopting an innovative educational design that will improve their students' learning.

From that main contribution, we have derived five additional implications that contribute to educational practice and gamification research. First, *despite the fact that the gamification effect is likely to decrease over time, it is unlikely to start decreasing before four weeks.* Beyond supporting the *novelty effect*, our findings open the debate in terms of *when it starts to act*. Our results suggest that the effect od gamification only started to decrease after four weeks of use. Despite several discussions in the literature that gamification effect suffers from the novelty effect, little has been explored and discussed in terms of when the decrease starts. Therefore, this finding has both practical and theoretical implications, informing practitioners on a *safe period* in which gamification can be used without losing its power, as well as providing researchers with a threshold that can be assessed in further research, to ground the extent to which the novelty effect starts acting.

Second, *whereas the novelty effect is often present, the familiarization effect seems to naturally address it.* Our findings demonstrate that the impact of gamification on all measures started to decrease at some point in time. However, we also showed that such impact then shifted back to an increasing trend, without any intervention (e.g., changing the gamification design). This behavior, which resembles a U-shaped curve, has been called the *familiarization effect*: after some (familiarization) time, gamification's effect is enhanced. Additionally, it should be noted that such effect appeared after the downtrend of the novelty effect. Providing empirical support for the familiarization has practical implications as, despite the fact that the impact of gamification might decrease after a period of time, a recovery is likely to happen, after users familiarize themselves with the game elements. Furthermore, empirically supporting the familiarization effect has theoretical implications. From our findings, the familiarization effect tackles the drawback from the novelty effect within a few weeks (two to six) after the end of the downtrend. To our best knowledge, however, such analysis, based on a comparison to a control group, has not been performed before in previous research. Hence, our findings call for future long-term, longitudinal gamification studies, to better understand the familiarization effect.

Third, *gamification likely suffers from the novelty effect, but benefits from the familiarization effect, which contributes to an overall positive impact.* Our findings corroborate gamification literature on three perspectives. They demonstrate that i) using gamification positively impacted on learner behavior, ii) gamification impact suffered from the novelty effect, and iii) some time after the end of the novelty effect, gamification's effect was enhanced via the familiarization effect. Therefore, from a practical point of view, despite a gamified intervention seemingly loosing power or failing to work after some time, it is likely to gain power again in the future, although the effect might not go back to its initial values. From a theoretical point of view, more research is needed to

understand the magnitude of both the novelty and the familiarization effects, as well as for shedding light into how long the uptrend of the familiarization effect lasts.

Fourth, *studies lasting less than 12 weeks are likely to be insufficient for truly revealing gamification's impact*. Our findings indicate that the novelty effect starts to act only after four weeks. In addition, our findings show that the final uptrend of the familiarization might start only after 10 or 12 weeks. Therefore, studies lasting less than this period of time might reveal unreliable findings, as they might fail to consider such changes that happen during learners' experiences with a gamified intervention. Nevertheless, it might be that even longer studies are necessary, given that, for attempts, the increase continued until the end of the intervention (14 weeks). These findings have practical and theoretical implications as well, informing researchers on the duration of the experiments they will conduct, and practitioners on how long they should use gamification before assessing its impacts.

Lastly, *our article has implications on how researchers can deploy similar studies in other contexts, learning activities, countries, and for other student types, further underpinning our findings*. As our evidence is limited to the context of STEM students completing programming assignments, our results should be seen in this context. Then, as the literature will benefit from understanding how the gamification effect varies in other settings, future research needs to generalize our study design to that end. For instance, one could use gamification to motivate completing different learning activities (e.g., multiple-choice quizzes) or following other behaviors (e.g., class attendance). In exploring other behaviors/learning activities, researchers can develop similar studies outside the STEM context. Thus, we contribute to future research with directions on how to ground our findings based on other samples/populations.

Limitations

This study generated several limitations that must be acknowledged when interpreting its findings. First, participants were not randomly assigned to experimental conditions, because the gamified version of the system used in this research was only available after 2016. To mitigate that, we compared participants in terms of demographic characteristics and, through preliminary analyses, showed the roughly one-year difference among groups is unlikely to confound our results. Additionally, meta-analytic evidence suggests the lack of random assignment does not affect the effect of gamification on behavioral learning outcomes Sailer and Homner (2019), further mitigating the impact of this limitation.

Second, the disciplines involved in this study did not have the same instructors throughout the three years of data collection. This limitation especially emerges because UFAM has a policy that professors take turns with the disciplines they teach. Despite that, pedagogical plans, learning materials, and the problems within the tasks participants completed were similar. In addition, meta-analytic evidence suggests that experimental conditions having different instructors does not affect gamification impact Bai et al. (2020), further mitigating this limitation's impact.

Third, group sizes were highly unbalanced. From the perspective of sample sizes, this limitation is handled by the number of participants for each condition ($\geq$ 138) being above the average of the total sample of gamification studies, as reported by secondary

studies (see Koivisto and Hamari (2019) and Sailer and Homner (2019), for instance). From the perspective of conclusion validity, we addressed this limitation, by answering our research question using robust statistical analyses that, among other advantages, handle groups with different sizes Wilcox (2017).

Fourth, our discussions regarding the shape (change) of the effect of gamification over time are based on visual inference. Although we selected robust statistical methods to guarantee conclusion validity, we did not perform, for instance, growth curve analysis. We chose this approach to enable comparing gamification effect at each time-point, as well as determining the effect's magnitude, rather than understanding the effect's curve. Consequently, we provided evidence on each time-point's difference and inferred the curve shape from visual analysis, upon that evidence. While that is a limitation, we opted for it instead of more conclusive approaches (such as regression analysis based on p-values), due to our exploratory analysis goal, as the literature recommends Vornhagen et al. (2020); Cairns (2019); Dragicevic (2016).

Fifth, regarding our study design, as this was a quasi-experimental study conducted over three academic years, its internal validity is likely affected. Indeed, gamification studies are often criticized, due to lack of methodological rigor Hamari et al. (2014); Dichev and Dicheva (2017). However, we made this choice in exchange of having a higher external validity, as it was conducted using a real system, within the context of real classrooms. Another aspect that could be discussed is the lack of a pre-test, in which pre-existent differences could be the source of differences rather than the experimental manipulation, especially given the lack of random assignment. Nevertheless, it must be noted that we analyzed gamification effect in terms of student behavior. Consequently, this is mitigating the lack of a pre-test, which would have to be based on learners' intentions (e.g., to access the system) rather than on actual behavior. Additional positive points are that this study features a control group and that the dependent variables reflect participants' behavior.

Lastly, in terms of study context, this research concerned STEM students completing programming assignments in either a gamified or non-gamified setting. That is, the study is limited to a single learning activity type. Additionally, we only deployed a single gamification design throughout the data collection period. In contrast, research suggests that one's experience with gamification might change depending on the task Liu et al. (2017) and that different designs should be adopted, depending on the learning activity Rodrigues et al. (2019); Hallifax et al. (2019). Therefore, we cannot rule out the possibility that we observed an effect due to the combination of the gamification design and the task type (i.e., programming assignments). However, given that the task type was invariant in our study, the effect observed could only be attributed to the gamification. Nevertheless, we opted for a single-factor study (i.e., only manipulating designs—gamified or not—within a single task) to maximize the findings' interval validity, following literature recommendations Cairns (2019).

Future work

Based on our results and their implications, we call for further similar research (i.e., long-term, longitudinal gamification studies featuring a control group) in different contexts, based on other measures, to further ground our findings as well as determine

whether they generalize to new contexts. Future research is also needed to confirm for how long gamification positively acts, until the novelty effect starts to act—a four-week period according to this study. Also, the literature demands such longitudinal research to ground the familiarization effect. We found that it takes between six and 10 weeks to start the uptrend. However, the lack of longitudinal studies and appropriate data analysis methods lead to little evidence on whether this period differs in other contexts. Additionally, future research should seek for further evidence on how much of the initial effect of gamification the familiarization can recover, for how long its uptrend lasts, and, mainly, ground whether there is a point in which the gamification impact actually achieves a plateau. In doing so, researchers could explore mixed-methods approaches. Specifically, qualitative data would allow understanding what maintains user motivation and persistence over time, based on their subjective experiences. Then, researchers could triangulate quantitative and qualitative data to advance the overall understanding of gamification's effect over time.

Furthermore, once researchers are aware that these changes in gamification effect happen over time, interventions to mitigate them should be sought for. Although the familiarization effect appears to naturally address the novelty effect's negative impact, it seems not to recover gamification's initial benefit. Nevertheless, the period in which the gamification impact decreases should be tackled, to maximize its contributions. To that end, a promising research direction is tailored gamification, which can be accomplished through personalization or adaptation Klock et al. (2020); Rodrigues et al. (2020). In this approach, game elements are tailored based on user and/or contextual information, with the goal of enhancing gamification potential (e.g., Rodrigues et al. (2021); Lopez and Tucker (2021)). The rationale is that the same gamification design is unlikely to work for all users (i.e., one size does not fit all) because people have different preferences, perceptions, and experiences, even under the same conditions Van Roy and Zaman (2018); Rodrigues et al. (2020). Accordingly, in future studies, tailoring can be triggered once a user logs into the system (i.e., personalization), providing game elements that better suit the user, whilst expecting to mitigate the magnitude of the novelty effect, due to this choice, or during system usage (i.e., adaptation), changing the gamification design when its effect starts to decrease Tondello (2019). As personalized gamification is a recent research field Klock et al. (2020); Tondello et al. (2017), long-term experimental studies assessing its effects compared to general gamification approaches would be beneficial.

Conclusions

Gamification is an educational innovation that has been widely used in several domains. Researchers often advocate that gamification's impact is positive due to its novelty and that it consequently vanishes as the novelty passes (i.e., the novelty effect). However, little is known about how the gamification effect changes over time, due to the lack of longitudinal studies in this field. Accordingly, recent literature has called for research to address this gap. This study responded to such calls with a 14-week longitudinal study, in which we analyzed the impact of gamification on students' behavioral outcomes at seven points in time. Specifically, our findings are threefold. First, they confirm the effectiveness of the gamification design we proposed to innovate the educational process of teaching programming. Second, concerning the novelty effect, we found that novelty: i)

starts to act after four weeks of intervention, ii) lasts for two to six weeks, and iii) diminishes a moderate effect to null in the worst case. Third, concerning the familiarization effect, we found that i) its positive trend starts between six and ten weeks after the start of the intervention and that ii) it seems to naturally tackle the negative effects of novelty expiration, although it did not restore the positive effect to the level of the initial contribution of the gamification.

These findings have several implications. Mainly, they reveal that innovating the educational practice with a gamification strategy similar to the one we used is likely to improve students' behavioral learning outcomes, even in the long-run. Additional implications are: informing practitioners on how long it takes for the novelty effect to start acting; revealing that although gamification's impact might decrease over time, it will likely recover after a familiarization period; and making the case that assessing gamification effect in less than 12 weeks is likely to yield unreliable results, due to the joint effects of novelty and familiarization. Furthermore, our findings have raised concerns that demand further research, such as confirming for how long the novelty effect decreases gamification impact and for how long (and to what extent) the familiarization effect recovers that effectiveness lost. Hence, our main contribution is advancing the understanding of how gamification impacts students' behavior over time, by providing empirical evidence on how the novelty effect acts, as well as supporting the familiarization effect. Ultimately, our conclusion is that gamification's impact on students' behaviors suffers from the novelty effect, but benefits from the familiarization effect, a counter-balance that is likely responsible for the overall positive effect gamification often demonstrates in maximizing behavioral learning outcomes.

Abbreviations

STEM: Science, Technology, Engineering, and Mathematics; CS1: Introductory Programming; IDE: IntegratedDevelopment Environment; ANOVA: Analysis of Variance; SD: Standard Deviation.

Acknowledgements

The authors are thankful for all organizations that provided funding to make this research possible.

Authors' contributions

LR: conception, design, data analysis, data interpretation, drafting the work; FP: conception, design, data analysis, data interpretation, drafting the work; AT: conception, design, drafting the work; PP: data interpretation, revised the work; MP: data acquisition, software creation, revised the work; LO: data acquisition, revised the work; DF: data acquisition, software creation, revised the work; EO: data acquisition, revised the work; AC: revised the work; SI: revised the work. All authors read and approved the final manuscript.

Funding

This research was partially funded by the following organizations: Brazilian National Council for Scientific and Technological Development (CNPq)—processes 141859/2019-9, 163932/2020-4, 308458/2020-6, and 308513/2020-7; Coordination for the Improvement of Higher Education Personnel (CAPES)—Finance Code 001; and São Paulo State Research Support Foundation (FAPESP)—processes 2018/15917- 0 and 2013/07375-0. Additionally, this research was carried out within the scope of the Samsung-UFAM Project for Education and Research (SUPER), according to Article 48 of Decree n° 6.008/2006 (SUFRAMA), and partially funded by Samsung Electronics of Amazonia Ltda., under the terms of Federal Law n° 8.387/1991, through agreements 001/2020 and 003/2019, signed with Federal University of Amazonas and FAEPI, Brazil.

Availability of data and materials

The datasets used and/or analyzed during the current study are available from the corresponding author on reasonable request

Declarations

Ethics approval and consent to participate

All participants provided consent allowing collection and usage of their data, in an anonymized way, for scientific purposes.

Consent for publication
All authors and institutions involved in this research are aware of and agree with the publication of this document.

Competing interests
The authors declare that they have no competing interests.

Author details
[1]Institute of Mathematics and Computer Science, University of São Paulo, Avenida Trabalhador São Carlense, 400 - Centro, São Carlos, SP 13566-590, Brazil. [2]Department of Computer Science, Federal University of Roraima, Boa Vista, Brazil. [3]Durham University, Durham, UK. [4]HCI Games Group, University of Waterloo, Stratford, Canada. [5]Amazonas State University, Manaus, Brazil. [6]Institute of Computing, Federal University of Amazonas, Manaus, Brazil.

Received: 2 September 2021 Accepted: 21 December 2021
Published online: 15 February 2022

References

Ahadi, A., Lister, R., Vihavainen, A. (2016). On the number of attempts students made on some online programming exercises during semester and their subsequent performance on final exam questions. In: Proceedings of the 2016 ACM Conference on Innovation and Technology in Computer Science Education, pp. 218–223
Alsawaier, R. S. (2018). The effect of gamification on motivation and engagement. *The International Journal of Information and Learning Technology*. https://doi.org/10.1108/IJILT-02-2017-0009.
Bai, S., Hew, K. F., & Huang, B. (2020). Does gamification improve student learning outcome? evidence from a meta-analysis and synthesis of qualitative data in educational contexts. *Educational Research Review, 30,* 100322.
Branch, R. M. (2009). *Instructional design: The ADDIE approach* (Vol. 722). Berlin: Springer.
Briffa, M., Jaftha, N., Loreto, G., Pinto, F. C. M., Chircop, T., & Hill, C. (2020). Improved students' performance within gamified learning environment: A meta-analysis study. *International Journal of Education and Research, 8*(1), 223–244.
Cairns, P. (2019). *Doing better statistics in human-computer interaction*. Cambridge: Cambridge University Press.
Clark, R. E. (1983). Reconsidering research on learning from media. *Review of educational research, 53*(4), 445–459.
Creswell, J. W., & Creswell, J. D. (2017). *Research design: qualitative, quantitative, and mixed methods approaches*. New York: Sage publications.
Deterding, S., Dixon, D., Khaled, R., Nacke, L. (2011). From game design elements to gamefulness: defining gamification. In: Proceedings of the 15th International Academic MindTrek Conference: Envisioning Future Media Environments, pp. 9–15 . ACM
Dichev, C., & Dicheva, D. (2017). Gamifying education: What is known, what is believed and what remains uncertain: a critical review. *International Journal of Educational Technology in Higher Education, 14*(1), 9.
Dick, W., Carey, L., Carey, J. O. (2005). The systematic design of instruction.
Dragicevic, P. (2016). Fair statistical communication in hci. In: Modern Statistical Methods for HCI, pp. 291–330. Springer
Estey, A., Coady, Y. (2016). Can interaction patterns with supplemental study tools predict outcomes in CS1? Proceedings of the 2016 ACM Conference on Innovation and Technology in Computer Science Education—ITiCSE '16, 236–241
Fonseca, S. C., Pereira, F. D., Oliveira, E. H., Oliveira, D. B., Carvalho, L. S., Cristea, A. I. (2020). Automatic subject-based contextualisation of programming assignment lists. In *Proceedings of The 13th International Conference on Educational Data Mining (EDM 2020)*.
Galvão, L., Fernandes, D., Gadelha, B.(2016). Juiz online como ferramenta de apoio a uma metodologia de ensino híbrido em programação. In: Brazilian Symposium on Computers in Education (Simpósio Brasileiro de Informática na Educação-SBIE), vol.27, p. 140
Goldstein, E. B. (2014). *Cognitive Psychology: Connecting Mind*. Nelson Education: Research and Everyday Experience.
Grangeia, T.d.A.G., de Jorge, B., Cecílio-Fernandes, D., Tio, R.A., de Carvalho-Filho, M.A. (2019). Learn+ fun! social media and gamification sum up to foster a community of practice during an emergency medicine rotation. *Health Professions Education, 5*(4), 321–335.
Hallifax, S., Serna, A., Marty, J.-C., & Lavoué, É. (2019). Adaptive gamification in education: A literature review of current trends and developments. In M. Scheffel, J. Broisin, V. Pammer-Schindler, A. Ioannou, & J. Schneider (Eds.), *Transforming learning with meaningful technologies* (pp. 294–307). Cham: Springer.
Hamari, J., Koivisto, J., Sarsa, H. (2014). Does gamification work?–a literature review of empirical studies on gamification. In: 2014 47th Hawaii International Conference on System Sciences, pp. 3025–3034 . Ieee
Hanus, M. D., & Fox, J. (2015). Assessing the effects of gamification in the classroom: A longitudinal study on intrinsic motivation, social comparison, satisfaction, effort, and academic performance. *Computers& Education, 80,* 152–161. https://doi.org/10.1016/j.compedu.2014.08.019.
Helmefalk, M. (2019). An interdisciplinary perspective on gamification: Mechanics, psychological mediators and outcomes. *International Journal of Serious Games, 6*(1), 3–26.
Hernández, M. I. O., Lezama, R. M., Gómez, S. M.(2021). Work-in-progress: The road to learning, using gamification. In: 2021 IEEE Global Engineering Education Conference (EDUCON), pp. 1393–1397 . IEEE
Huang, R., Ritzhaupt, A. D., Sommer, M., Zhu, J., Stephen, A., Valle, N., et al. (2020). The impact of gamification in educational settings on student learning outcomes: A meta-analysis. *Educational Technology Research and Development*. https://doi.org/10.1007/s11423-020-09807-z.
Hyrynsalmi, S., Smed, J., Kimppa, K. (2017). The dark side of gamification: How we should stop worrying and study also the negative impacts of bringing game design elements to everywhere. In: GamiFIN, pp. 96–104
Jafari, M., & Ansari-Pour, N. (2019). Why, when and how to adjust your p values? *Cell Journal (Yakhteh), 20*(4), 604.

Keselman, H., Algina, J., Wilcox, R. R., & Kowa, R. K. (2000). Testing repeated measures hypotheses when covariance matrices are heterogeneous: Revisiting the robustness of the welch-james test again. *Educational and Psychological Measurement, 60*(6), 925–938.
Klock, A. C. T., Gasparini, I., Pimenta, M. S., & Hamari, J. (2020). Tailored gamification: A review of literature. *International Journal of Human-Computer Studies*. https://doi.org/10.1016/j.ijhcs.2020.102495.
Koivisto, J., & Hamari, J. (2014). Demographic differences in perceived benefits from gamification. *Computers in Human Behavior, 35,* 179–188. https://doi.org/10.1016/j.chb.2014.03.007.
Koivisto, J., & Hamari, J. (2019). The rise of motivational information systems: A review of gamification research. *International Journal of Information Management, 45,* 191–210. https://doi.org/10.1016/j.ijinfomgt.2018.10.013.
Kotrlik, J., & Williams, H. (2003). The incorporation of effect size in information technology, learning, information technology, learning, and performance research and performance research. *Information Technology, Learning, and Performance Journal, 21*(1), 1.
Kowalchuk, R. K., Keselman, H., & Algina, J. (2003). Repeated measures interaction test with aligned ranks. *Multivariate Behavioral Research, 38*(4), 433–461.
Kyewski, E., & Krämer, N. C. (2018). To gamify or not to gamify? an experimental field study of the influence of badges on motivation, activity, and performance in an online learning course. *Computers& Education, 118,* 25–37.
Landers, R. N., & Landers, A. K. (2014). An empirical test of the theory of gamified learning: The effect of leaderboards on time-on-task and academic performance. *Simulation& Gaming, 45*(6), 769–785.
Liu, D., Santhanam, R., & Webster, J. (2017). Toward meaningful engagement: A framework for design and research of gamified information systems. *MIS quarterly, 41*(4), 1011–1034.
Lopez, C. E., Tucker, C. S. (2021). Adaptive gamification and its impact on performance. In: International Conference on Human-Computer Interaction, pp. 327–341 . Springer
Mair, P., Wilcox, R. (2018). Robust statistical methods using wrs2. The WRS2 Package.
Mavletova, A. (2015). A gamification effect in longitudinal web surveys among children and adolescents. *International Journal of Market Research, 57*(3), 413–438.
Mitchell, R., Schuster, L., & Drennan, J. (2017). Understanding how gamification influences behaviour in social marketing. *Australasian Marketing Journal (AMJ), 25*(1), 12–19.
Mustafa, A. S., Karimi, K. (2021). Enhancing gamified online learning user experience (ux): A systematic literature review of recent trends. Human-Computer Interaction and Beyond-Part I, 74–99
Nacke, L. E., & Deterding, C. S. (2017). The maturing of gamification research. *Computers in Human Behaviour*. https://doi.org/10.1016/j.chb.2016.11.062.
Palomino, P.T., Toda, A.M., Oliveira, W., Cristea, A.I., Isotani, S. (2019). Narrative for gamification in education: why should you care? In: 2019 IEEE 19th International Conference on Advanced Learning Technologies (ICALT), vol. 2161, pp. 97–99 . IEEE
Palomino, P., Toda, A., Rodrigues, L., Oliveira, W., Isotani, S.(2020). From the lack of engagement to motivation: Gamification strategies to enhance users learning experiences. In: 2020 19th Brazilian Symposium on Computer Games and Digital Entertainment (SBGames)-GranDGames BR Forum, pp. 1127–1130
Pereira, F. D., Fonseca, S. C., Oliveira, E. H., Cristea, A. I., Bellhäuser, H., Rodrigues, L., Oliveira, D. B., Isotani, S., Carvalho, L. S. (2021). Explaining individual and collective programming students' behaviour by interpreting a black-box predictive model. IEEE Access
Pereira, F. D., Oliveira, E. H., Oliveira, D. B., Cristea, A. I., Carvalho, L. S., Fonseca, S. C., Toda, A., Isotani, S. (2020). Using learning analytics in the amazonas: Understanding students' behaviour in introductory programming. *British Journal of Educational Technology*. https://doi.org/10.1111/bjet.12953.
Pereira, F. D., Toda, A., Oliveira, E. H., Cristea, A. I., Isotani, S., Laranjeira, D., Almeida, A., Mendonça, J. (2020). Can we use gamification to predict students' performance? a case study supported by an online judge. In: International Conference on Intelligent Tutoring Systems, pp. 259–269. Springer
Putz, L.-M., Hofbauer, F., & Treiblmaier, H. (2020). Can gamification help to improve education? Findings from a longitudinal study. *Computers in Human Behavior*. https://doi.org/10.1016/j.chb.2020.106392.
R Core Team. (2020). R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria. R Foundation for Statistical Computing. Retrieved from https://www.R-project.org/
Rodrigues, L., Oliveira, W., Toda, A., Palomino, P., Isotani, S. (2019). Thinking inside the box: How to tailor gamified educational systems based on learning activities types. In: Proceedings of the Brazilian Symposium on Computers in Education (Simpósio Brasileiro de Informática na Educação—SBIE)
Rodrigues, L., Palomino, P.T., Toda, A.M., Klock, A.C., Oliveira, W., Avila-Santos, A.P., Gasparini, I., Isotani, S. (2021). Personalization improves gamification: Evidence from a mixed-methods study. Proceedings of the ACM on Human-Computer Interaction 5(CHI PLAY), 1–25
Rodrigues, L., Toda, A. M., Oliveira, W., Palomino, P. T., Avila-Santos, A. P., Isotani, S. (2021). Gamification works, but how and to whom? an experimental study in the context of programming lessons. In: Proceedings of the 52nd ACM Technical Symposium on Computer Science Education, pp. 184–190
Rodrigues, L., Toda, A. M., Oliveira, W., Palomino, P. T., Isotani, S. (2020). Just beat it: Exploring the influences of competition and task-related factors in gamified learning environments. In: Anais do XXXI Simpósio Brasileiro de Informática na Educação, pp. 461–470 . SBC
Rodrigues, L., Toda, A. M., Palomino, P. T., Oliveira, W., Isotani, S. (2020). Personalized gamification: A literature review of outcomes, experiments, and approaches. In: Eighth International Conference on Technological Ecosystems for Enhancing Multiculturality, pp. 699–706
Rowland, C. A. (2014). The effect of testing versus restudy on retention: A meta-analytic review of the testing effect. *Psychological Bulletin, 140*(6), 1432.
RStudio Team. (2019). RStudio: Integrated Development Environment for R. RStudio, Inc., Boston, MA . RStudio, Inc. Retrieved from http://www.rstudio.com/
Sailer, M., & Homner, L. (2019). The gamification of learning: A meta-analysis. *Educational Psychology Review*. https://doi.org/10.1007/s10648-019-09498-w.

Sanchez, D. R., Langer, M., & Kaur, R. (2020). Gamification in the classroom: Examining the impact of gamified quizzes on student learning. *Computers& Education, 144,* 103666. https://doi.org/10.1016/j.compedu.2019.103666.
Santana, B. L., Bittencourt, R. A. (2018). Increasing motivation of cs1 non-majors through an approach contextualized by games and media. In: 2018 IEEE Frontiers in Education Conference (FIE), pp. 1–9 . IEEE
Seaborn, K., & Fels, D. I. (2015). Gamification in theory and action: A survey. *International Journal of human-computer studies, 74,* 14–31.
Toda, A. M., Klock, A. C., Oliveira, W., Palomino, P. T., Rodrigues, L., Shi, L., Bittencourt, I., Gasparini, I., Isotani, S., Cristea, A. I. (2019). Analysing gamification elements in educational environments using an existing gamification taxonomy. *Smart Learning Environments, 6*(1), 16.
Toda, A., Pereira, F. D., Klock, A. C. T., Rodrigues, L., Palomino, P., Oliveira, W., Oliveira, E. H. T., Gasparini, I., Cristea, A. I., Isotani, S. (2020). For whom should we gamify? insights on the users intentions and context towards gamification in education. In: Anais do XXXI Simpósio Brasileiro de Informática na Educação, pp. 471–480 . SBC
Toda, A. M., Valle, P. H. D., & Isotani, S. (2018). The dark side of gamification: An overview of negative effects of gamification in education. In A. I. Cristea, I. I. Bittencourt, & F. Lima (Eds.), *Higher education for all. From challenges to novel technology-enhanced solutions* (pp. 143–156). Cham: Springer.
Tondello, G. F. (2019). Dynamic personalization of gameful interactive systems. PhD thesis, University of Waterloo
Tondello, G. F., Orji, R., Nacke, L. E. (2017). Recommender systems for personalized gamification. In: Adjunct Publication of the 25th Conference on User Modeling, Adaptation and Personalization, pp. 425–430 . https://doi.org/10.1145/3099023.3099114.ACM
Tsay, C.H.-H., Kofinas, A. K., Trivedi, S. K., & Yang, Y. (2020). Overcoming the novelty effect in online gamified learning systems: An empirical evaluation of student engagement and performance. *Journal of Computer Assisted Learning, 36*(2), 128–146.
Van Roy, R., & Zaman, B. (2018). Need-supporting gamification in education: An assessment of motivational effects over time. *Computers& Education, 127,* 283–297.
von Wangenheim, C. G., & von Wangenheim, A. (2012). *Ensinando computação com jogos*. Florianópolis, SC, Brasil: Bookess Editora.
Vornhagen, J. B., Tyack, A., Mekler, E. D. (2020). Statistical significance testing at chi play: Challenges and opportunities for more transparency. In: Proceedings of the Annual Symposium on Computer-Human Interaction in Play, pp. 4–18
Wilcox, R. (2017). *Introduction to robust estimation and hypothesis testing* (4th ed.). Amsterdam: Elsevier.
Wilcox, R. R., & Tian, T. S. (2011). Measuring effect size: A robust heteroscedastic approach for two or more groups. *Journal of Applied Statistics, 38*(7), 1359–1368.
Yuen, K. K. (1974). The two-sample trimmed T for unequal population variances. *Biometrika, 61*(1), 165–170.

Publisher's Note