Are they learning or playing? Moderator conditions of gamification's success in programming classrooms

LUIZ RODRIGUES, University of São Paulo, Brazil
FILIPE PEREIRA, Federal University of Roraima, Brazil and Durham University, United Kingdom
ARMANDO TODA, University of São Paulo, Brazil and Durham University, United Kingdom
PAULA PALOMINO, University of São Paulo, Brazil and University of Waterloo, Canada
WILK OLIVEIRA, University of São Paulo, Brazil and Tampere University, Finland
MARCELA PESSOA, Amazonas State University, Brazil, and Federal University of Amazonas, Brazil
LEANDRO CARVALHO, Federal University of Amazonas, Brazil
DAVID OLIVEIRA, Federal University of Amazonas, Brazil
ELAINE OLIVEIRA, Federal University of Amazonas, Brazil
ALEXANDRA CRISTEA, Durham University, United Kingdom
SEIJI ISOTANI, University of São Paulo, Brazil

Students face several difficulties in introductory programming courses (CS1), often leading to high dropout rates, student demotivation, and lack of interest. The literature has indicated that the adequate use of gamification might improve learning in several domains, including CS1. However, the understanding of which (and how) factors influence gamification's success, especially for CS1 education, is lacking. Thus, there is a clear need to shed light on *pre-determinants of gamification's impact*. To tackle this gap, we investigate how user and contextual factors influence gamification's effect on CS1 students through a quasi-experimental retrospective study (N = 399), based on a between-subject design (conditions: gamified or non-gamified) in terms of final grade (academic achievement) and the number of programming assignments completed in an educational system (i.e., how much they practised). Then, we evaluate whether and how user and contextual characteristics (such as age, gender, major, programming experience, working situation, internet access, and computer access/sharing) moderate that effect. Our findings indicate that gamification amplified to some extent the impact of practising. Overall, students practising in the gamified version presented higher academic achievement than those practising the same amount in the non-gamified version. Intriguingly, those in the gamified version that practised much more extensively than the average showed lower academic achievements than those who practised comparable amounts in the non-gamified version. Furthermore, our results reveal gender as the only statistically significant moderator of gamification's effect: in our data, it was positive for

Authors' addresses: Luiz Rodrigues, lalrodrigues@usp.br, University of São Paulo, Brazil; Filipe Pereira, filipe.dwan@ufrr.br, Federal University of Roraima, Brazil and Durham University, United Kingdom; Armando Toda, armando.toda@usp.com, University of São Paulo, Brazil and Durham University, United Kingdom; Paula Palomino, paulatpalomino@usp.br, University of São Paulo, Brazil and University of Waterloo, Canada; Wilk Oliveira, wilk.oliveira@usp.br, University of São Paulo, Brazil and Tampere University, Finland; Marcela Pessoa, msppessoa@uea.edu.br, Amazonas State University, Brazil, and Federal University of Amazonas, Brazil; Leandro Carvalho, galvao@icomp.ufam.edu.br, Federal University of Amazonas, Brazil; David Oliveira, david@icomp.ufam.edu.br, Federal University of Amazonas, Brazil; Elaine Oliveira, elaine@icomp.ufam.edu.br, Federal University of Amazonas, Brazil; Alexandra Cristea, alexandra.i.cristea@durham.ac.uk, Durham University, United Kingdom; Seiji Isotani, sisotani@icmc.usp.br, University of São Paulo, Brazil.

XXXX-XXXX/2023/9-ART $15.00
https://doi.org/10.1145/nnnnnnn.nnnnnnn

females, but nonsignificant for males. These findings suggest which (and how) personal and contextual factors moderate gamification's effects, indicate the need to further understand and examine context's role, and show gamification must be cautiously designed to prevent students from playing instead of learning.

CCS Concepts: • **General and reference** → **Empirical studies**; • **Social and professional topics** → **Computing education**.

Additional Key Words and Phrases: Gamified Learning, Testing Effect, Moderation, Gaming the system, Context

ACM Reference Format:
Luiz Rodrigues, Filipe Pereira, Armando Toda, Paula Palomino, Wilk Oliveira, Marcela Pessoa, Leandro Carvalho, David Oliveira, Elaine Oliveira, Alexandra Cristea, and Seiji Isotani. 2023. Are they learning or playing? Moderator conditions of gamification's success in programming classrooms. 1, 1 (September 2023), 27 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 INTRODUCTION

Learning to code is challenging and demands significant effort from the learners [75]. Research on introductory programming (i.e., CS1 [33]) has been conducted for several decades and, up to date, findings report high failure and dropout rates in these courses [8–10, 45]. Furthermore, CS1 is compulsory not only for computing-related majors but for STEM[1] courses as well [22, 69]. For students of the latter majors, the problem is enhanced, as students often lack affinity with programming and even fail to see its value for their professional lives [21, 87]. As such, students' lack of interest and effort negatively contribute to their achievement, given that learning to code requires significant amounts of practice [38, 75, 103].

Recently, there is increasing evidence that the need for motivating students to practice programming might be addressed with gamification: adding game elements into non-gaming contexts [17]. It has been widely used in the educational domain [52], with empirical results reporting overall positive outcomes in, for instance, academic performance [101]. Reasons for such positive effects include allowing goal-setting, providing performance feedback and recognition, and fostering enthusiasm [4]. To that end, it is important to design gamification to drive the expected motivational and behavioural outcomes [96], as well as enhance learning in the educational context [46]. Otherwise, the game elements might jeopardise students' learning, such as in cases when too much engagement with the system leads to behaviours such as gaming the system, distraction, or lack of utility [5, 86, 94].

1.1 Problem

Gamification's success is assumed to depend on multiple factors, such as who will interact with it (i.e., its users) and the context[2] in which it will be used [32, 50, 90]. Advancing the understanding regarding such moderators is important, to shed light on which aspects predetermine gamification's success, as well as how each one acts (i.e., maximising or minimising it) [43]. However, the precise factors that moderate gamification's effectiveness are not well known [85], especially in the context of CS1, where empirical studies assessing gamification's impact often lack moderator analyses [24, 53, 58]. Therefore, despite some of the existing literature arguing that gamification depends on factors related to the user (e.g., gender and age) and context (e.g., the environment or circumstance) [4, 31, 77], there is a gap in the understanding of which factors moderate gamification's success

[1] Science, Technology, Engineering, and Mathematics
[2] Resources, methods, people's mental representations, environment, and circumstance involved in an activity (see Section 1.3.3).

when applied to CS1 learning contexts. Thus, we tackle this gap by answering the following research question:

- **RQ:** *How do user and contextual factors influence the effect of gamification on the academic achievement of CS1 students?*

Given the current literature on gamification applied to CS1 education, we expand it, by evaluating *how* gamification improves CS1 students' academic achievement, assuming this happens by influencing their behaviours, along with an analysis of which contextual and demographic factors moderate that effect. Additionally, for the gamification design, we used fictional, social, and challenge-based game elements [93], whereas previous similar studies only focused on the last two kinds (e.g., [19, 44, 53]). Featuring fictional game elements is valuable, as a recent meta-analysis found those to maximise gamification's impact on learning outcomes [85]. Thus, the overall contribution of this study is *to offer empirical evidence revealing which user and contextual characteristics moderate the impact of a gamification design featuring fictional game elements on CS1 students' academic achievement, as well as how those moderators act, that is, maximising or minimising that impact.*

1.2 Literature Review

Most empirical research applying gamification to programming learning focuses on *whether* gamification had a positive effect on learning or not, compared to no gamification. In that context, [30] added badges to an eight-week-long Data Structures and Algorithms course, which resulted in a positive effect on students' time-on-system, but not on the number of completed exercises, when compared to the condition with no badges. Similarly, [23] used Kahoot! and Codeacademy to gamify a 12-week-long programming course. Based on descriptive analyses, they found positive results, compared to the course's previous run that was not gamified, mainly in terms of final grades and attendance. [53] used the UDPiler to gamify a four-week-long C programming course. By analysing data collected for two semesters, they found positive results on learning performance from gamification usage. [19] used the OneUp platform to gamify a Data Structures (semester-long) course. Their findings suggested gamification was positive in terms of the number of challenge attempts and students' final grades. [57] gamified QueryCompetition by adding points and leaderboards and compared it to a nongamified version in terms of students' performance, motivation, and user experience. Based on a five-week intervention featuring pre- and post-tests, they found positive results favouring gamification, especially in performance. These studies support the claim that gamification has the potential to positively influence learner behaviour. However, they do not contribute to understanding what factors moderate the impact of gamification.

In contrast, few studies analyse moderators of the impact of gamified programming learning. For instance, [29] analysed the role of two factors: achievement goal orientation and motivation towards badges. They found a relationship among those factors, indicating that students with different goal orientations have distinct motivations towards badges. However, their findings suggested that those factors did not moderate students' behaviour in the gamified system. Two points of this study that must be noted are the intervention duration (half-semester) and the use of a single game element (badges). In another research, [44] evaluated whether learners' gender (male or female) and major (computer science or psychology) moderated gamification's impact on student retention, quiz accuracy, and test performance. They found no significant interactions between conditions (e.g., gamified and non-gamified) and moderators, suggesting neither gender nor major moderated gamification's effect.

Similarly, [64] assessed the role of three possible moderators: gender, major, and gaming experience. Again, their empirical findings suggested that the gamification's impact was not moderated by any of the three factors analysed. [2] studied the role of two moderators - group size and time - on students' performance and satisfaction. Based on a 16-week data collection involving 229 participants, they found both factors moderating the gamification's effect, alone and together, on both outcomes. [79] also evaluated the influence of two moderators on gamification's effect on intrinsic motivation: usage time and previous affinity to the content to be learned. They found that both moderators affected the gamification's impact only when considered together. Two limitations of [44], [64], and [79] must be acknowledged, though: the limited sample sizes - 71, 102, and 19, respectively - and intervention duration - four, four, and six weeks, respectively. On the other hand, [2] contributed to understanding a moderator's effect, but did not use fictional game elements and focused on a moderator related to the gamification design. Hence, while it advances the understanding of gamification design, it does not provide evidence on how contextual and situational factors predeterminate the effectiveness of gamification.

We showed thus that most previous studies failed to analyse moderators of gamification's success and that, those that did, are limited in terms of sample size, intervention duration, moderator's nature (i.e., user or context-related), and/or gamification design. In this study, we tackle these limitations, by presenting a 15-week empirical study analysing moderators of gamification's success, based on a wide sample of 399 learners and a well-planned gamification design. Furthermore, all of the reviewed studies explored social (e.g., collaboration) and challenge-based (e.g., challenges and rewards) gamification. Besides game elements similar to those, the gamification design employed in this study also features fictional elements [93]. Meta-analytic evidence indicates that this kind of game element is likely to improve gamification's effectiveness [85]. Hence, we expand the literature, by evaluating moderators of gamification's effectiveness based on a different design.

Lastly, studies performing moderator analysis [2, 44, 64] indicate gamification's impact does not depend on user characteristics (e.g., gender and goal orientation) and contextual factors (e.g., student major). These results contradict empirical findings from non-programming contexts. For instance, [67] found that gamification only worked for boys; that is, a moderator effect of gender. In contrast, [72] found that gamification influences women and older people more. Similarly, results from [47] suggest gamification only works for people with good attitudes towards it. Accordingly, it has been noted that studies must control contextual characteristics, such as the educational level of the learners, when analysing the gamification's effect [43]. Empirical findings suggest the relevance of other contextual factors, such as the moderator role of the geographic location [4, 81] and the motivational improvements from considering the learning task when designing gamification [78]. Some literature reviews also discuss contextual factors in general terms, such as usage domain (e.g., [31, 32]). However, because of the broadness of those factors, they are likely products of more specific ones [49]. Then, the contradiction might be attributed, for instance, to not finding and studying the most relevant moderators. Thus, this demonstrates the need for empirical studies investigating other moderators, to reveal which factors affect gamification's success [43, 85]. Table 1 contrasts the present study to those reviewed in this section, summarising the main points discussed here.

1.3 Hypothesis

As our RQ concerns understanding moderators of gamification's effect on academic achievement, we first need to test for such effects. For that, we rely on the Theory of Gamified learning [46], which is considered a framework suitable to understand how gamification acts according to recent research [85]. That framework advocates that for gamification to affect outcomes such as academic achievement, it affects user behaviour. Therefore, to answer our RQ, we first needed to consider the behavioural source of its effect, which we hypothesise to be *practising* (H2). Similarly, we need to ensure that practising is working as expected (H1). Lastly, we needed to test the user and contextual

Table 1. Summary of related work on gamification applied to computing education.

Study, year	Moderator analysis?	Uses game fiction?	Intervention duration	Sample size
[30], 2015			8 weeks	281
[23], 2016			12 weeks	106
[53], 2018			16 weeks	817
[19], 2019			16 weeks	27
[57], 2020			5 weeks	139
[29], 2014	X		8 weeks	278
[44], 2015	X		8 weeks	71
[64], 2019	X		4 weeks	102
[2], 2020	X		16 weeks	229
[79], 2021	X		6 weeks	19
Our study	X	X	15 weeks	399

moderators (H3) to answer our RQ. Based on that, we discuss the relevant literature that supports our research model next.

1.3.1 Testing effect. The claim that learning to code requires practising can be supported by the theory of the testing effect. Also referred to as test-enhanced learning, it is concerned with the fact that long-term memory is often improved when learners dedicate some of their studying time to retrieve the information they expect to be remembered [27]. For instance, that might be achieved by completing quizzes or, in a more elaborated way, by retrieving programming information presented in lectures, to elaborate when performing problem-solving activities. Overall results for the testing effect are positive (e.g., [54, 66]), as the literature shows that learners who study and are tested (like in a school test) present higher long-term knowledge retention compared to those that studied but were not tested [83]. Furthermore, in a meta-analysis comparing restudy (i.e., reading again) to testing, Rowland [84] found the testing effect on recalling tasks (e.g., short answers) is much larger. Despite the fact that this effect is commonly studied in simple tasks, such as quizzes and short answers, the literature supports the effectiveness of test-enhanced learning for contents that are highly related to others (e.g., you need to know conditionals to understand loops) as well [40].

As programming presents learners with difficulties in aspects such as natural language, syntax, and abstraction [73], the need for practising can be related to improving long-term memory about these aspects, as well as receiving feedback on their programs, that is, testing themselves concerning their ability to code. Additionally, it has been claimed that students learn to code by doing (e.g., [99]). Accordingly, empirical evidence supports that need. For instance, [79] shows that the more students engaged with quizzes, the higher were their learning gains. Similarly, [70] demonstrated that the more students practised, the higher were their performances. Nevertheless, such effects are less known for complex materials, showing the need for empirical research to further examine it [74]. Consequently, we might expect that the more students practice, or test themselves, the more they will be successful in programming. Thus, our first hypothesis H1 is:

H1: Practising positively affects academic achievement in CS1.

1.3.2 Gamified Learning. Gamification has been widely explored within the educational domain, based on beliefs that it can improve, for instance, learners' engagement, motivation, and learning [18, 52, 65]. Consequently, several empirical studies assessing its effectiveness emerged, some of which have been recently summarised in secondary studies. Sailer and Homner [85] presents a meta-analysis of gamification's impact on three kinds of learning outcomes, in which they found it has, overall, small positive effects for motivational (e.g., intrinsic motivation), behavioural (e.g., performance), and cognitive (e.g., conceptual knowledge) learning outcomes. Results of the meta-analysis by [35] corroborate those findings, demonstrating a positive impact from gamification on cognitive learning outcomes. In another recent meta-analysis, Bai et al. [4] focused on gamification's effect on academic performance, finding it has a small-to-moderate positive influence.

Additionally, Bai et al. [4] summarised reasons for students enjoying gamification. These include inciting enthusiasm, providing performance feedback and means to be recognised (e.g., badges), and goal-setting. While enthusiasm is likely to motivate people to use the gamified system, goal-setting is valuable to drive performance [97], providing feedback is highly important for learning programming [68, 75, 103], and recognition is likely to incite feelings of self-efficacy and fulfil competence needs, aspects that are also positively related to academic achievement and performance [36, 71, 91, 102]. This demonstrates the importance of defining gamification designs that support the desired behaviours/outcomes (e.g., enhancing learning by fulfilling competence needs) and minimises side effects (e.g., gaming the system rather than using it to study). Hence, the overall literature provides evidence of the potential of well-designed gamification to contribute to learning, demonstrating it can affect cognition, motivation, and behaviour and indicating reasons for those effects to incite self-efficacy and fulfil competence needs as well as goal-setting.

Thereby, we might expect using gamification will enhance programming practice, consequently improving academic achievement, and formulate hypothesis H2 as:

> **H2**: Gamification enhances the testing effect and, consequently, academic achievement.

1.3.3 Moderators of gamification's success. Although gamification has overall positive impacts on learning outcomes, there are cases in which results are null or negative [37, 94]. A recurrent justification for those outcomes is the quality of the gamification design [15, 51]. According to many discussions, what leads to such low-quality gamification designs is the lack of consideration of user and contextual characteristics (e.g., [18, 32, 60, 89, 90]). That is, many of such characteristics moderate whether gamification will be effective for a given user in a given context, hence, when those are not taken into account, it is likely that effectiveness will not be achieved for some users [48].

According to those discussions, studies have shown empirical support for the influence of users' characteristics regarding their experiences, perceptions, and preferences. For instance, research has demonstrated that people with different gender, age, socio-economic conditions, behavioural profiles, and personality traits have different preferences, perceptions, and experiences, even when doing the same task [12, 29, 49, 56, 62, 63, 80, 86, 92, 95]. Moreover, context-related aspects are acknowledged as relevant moderators of user experience as well [4, 32]. However, studies often understand context differently and define it in general terms (e.g., based on usage domain/aim [31, 43]). In contrast, in the scope of this paper, we interpret the context to involve resources (e.g., gamified system) and methods involving human activity (e.g., task to be done in a gamified system) [100]. Further, context might be seen as internal and external: the former relates to users' mental representations (e.g., user characteristics that influence the learning experience) while the latter relates to the environment/circumstance (e.g., where the activity takes place) [88]. Thus, by relying on context-specific definitions (i.e., from Human-Computer Interaction and Educational Technology literature), we can study and understand contextual factors in a fine-grained, comprehensive way.

However, empirical analyses of the role of contextual aspects are rare, although recent research highlights their importance and lack of studies in this regard [31, 41, 43, 50, 77]. For instance, research has explored major's moderator effect [44, 64], which fits within the external context, but found no significant effects. On the other hand, [79] found initial evidence on the moderator effect of a factor related to the internal context: one's previous affinity to the content to be learned, when considered together with intervention duration. Hence, whereas the literature acknowledges there are multiple factors likely to moderate gamification's effectiveness, evidence from related work is limited and suggests we need to explore new factors that could explain the role of context on the effects of gamification. Thus, highlighting the need for studies to confirm, discover, and better understand the role of those factors [36, 85]. Moreover, understanding whether these factors' moderator effect will be positive or negative, as well as its magnitude, is even more uncertain. Therefore, in a more exploratory hypothesis, we expect that user and contextual characteristics will moderate the effects of gamification, as predicted by **H3**, with no assumption on the direction (positive or negative) and the magnitude of these moderators:

H3: Gamification's effects will be moderated by a) user and b) contextual characteristics.

The research model illustrating this study's hypotheses is shown in Figure 1. It shows the testing effect prediction (**H1**), the assumption that gamification will enhance academic achievement by improving the testing effect (**H2**), and the expectation that user and contextual characteristics will moderate the gamification's impact (**H3**).

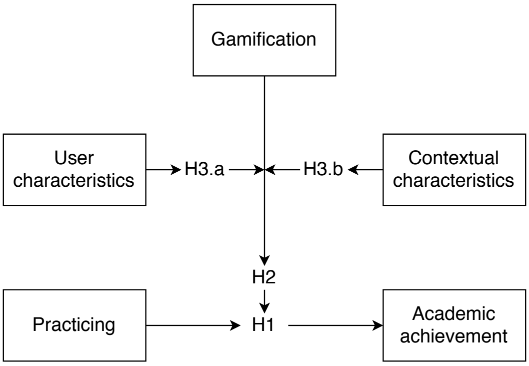

Fig. 1. Study research model on Gamification in CS1, hypothesising that i) academic achievement is positively affected by practice ii) and by gamification, iii) effect that is moderated by user and contextual characteristics.

2 METHOD

This is a retrospective study, as we examine data captured in the past that was made available for analysis by Federal University of Amazonas (UFAM in Portuguese) from Brazil.

2.1 Inclusion and Exclusion

As this is a retrospective study, we prepared the dataset (N = 1309) in four steps. First, we removed participants (325) that did not provide consent to participating in the research. Second, we removed participants (30) that completed the characterisation survey with unrealistic values (e.g., age of 4). Third, we removed participants (198) that dropped out due to two reasons. First, to ensure analysing data from subjects that participated in the whole semester, which is necessary because gamification's effect might decline with time [4, 85]. Therefore, analysing it based on long usage periods is imperative to achieve reliable findings. Second, our dependent variable (see Section 2.3) depends on several assignments completed from semester's week two to week 14. Importantly, those assignments' weights increase progressively. Therefore, we had to remove those who dropped out because their final grades' measures would be misleading. Lastly, we removed data from majors in which the number of subjects in any condition (control or experimental) was five or less (357)[3].

2.2 Participant Characteristics

Our analysis concerns data from 399 CS1 students of seven majors of the UFAM. Majors are Materials Engineering, Electrical Engineering, Mechanical Engineering, Statistics, Physics, Mathematics, and Applied Mathematics. Although different, all majors followed the same methodological plan and pedagogical materials, and all activities offered for participants to practice were selected from the same database. The participants are 399 learners (64.2% males, 35.8% females) with an average age of 22.1 years (±3.6) who self-reported whether they had i) previous experience with any programming language (yes: 37.3%; no: 62.7%), ii) worked/interned before (yes: 20.1%; no: 79.9%), iii) had internet at home (yes: 82%; no: 18%), iv) a computer at home (yes: 87.7%; no: 12.3%), and whether they shared their home computer with someone else (yes: 27.3%; no: 72.7%).

Table 2 presents a comparison between control and experimental groups regarding their categorical demographic characteristics. Two significant differences were found: control group was more experienced (p-adj < 0.05) and worked/interned before the degree less (Worked; p-adj < 0.01) than the experimental one. Also, the experimental group (M = 21.84; ± 3.87) was younger than the control one (M = 22.78; ± 2.80), W = 23073; p < 0.001; CI = 1.000-2.000. While comparing groups with different characteristics might affect the results, we handled this limitation by inserting all these variables as covariates during the data analysis process. Then, if some difference was to be found due to a demographic characteristic (e.g., experience) rather than condition (i.e., gamification), the analysis would reveal it.

Table 2. Comparison of demographic data from study groups (i.e., control - Ctr, no gamification; experimental - Exp, gamification). Data represented as percentages; p-values adjusted using the False Discovery Rate approach [39]; all comparisons' degree of freedom was one.

	Gender	Has PC?	Shares PC?	Int?	Exp?	Worked?
	Mal/Fem.	No/Yes	No/Yes	No/Yes	No/Yes	No/Yes
Ctr	56/44	12/88	70/30	18/82	47/53	73/23
Exp	58/42	13/87	74/26	18/82	69/31	83/17
χ^2	0.927	0.000	0.311	0.000	18.439	5.222
P-val	0.336	1.000	0.577	1.000	0.000	0.030
P-adj	0.672	1.000	0.865	1.000	0.001	0.089

Has/Shares PC = whether the participant has/shares a PC at home; Int = whether the participant has internet at home; Exp = whether the participant has experience with any programming language; Worked = whether the participants worked/internet before the degree.

2.3 Measures and Covariates

We analysed two measures that concern a whole semester: academic achievement and practising to code. The former is the study dependent variable, measured as a student's final grade (Equation 1).

$$FG_s = \frac{(Ex_{t1} + Ex_{t2}).1 + (Ex_{t1} + Ex_{t2}).2 + (Ex_{t1} + Ex_{t2}).3 + \frac{PA_{t1} + ... + PA_{t7}}{7}}{16} \qquad (1)$$

[3] The number of students removed in this step is large, because many majors were highly unbalanced among conditions (i.e., many students in one condition, very few in the other) because we analysed majors in **H3**.

A final grade (*FG*) is calculated based on the student's (*s*) scores on the programming assignment (*PA*) and their exams' marks (at each two weeks, the students were required to take an exam – Ex – about the same topic of the PA.). As the content of programming is cumulative, the weights of the seven exams increased over the topics. The students' groups (control or experimental) did not affect how their final grades were calculated.

The latter – practising to code – acted as a proxy to the former, measuring the extent to which learners practised programming. We operationalised it based on how many times they submitted their codes during the semester, that is, the number of all attempts that a student made in the educational system throughout the whole semester (sum attempts hereafter).

The characterisation questionnaire captured the moderators analysed in this study, which were selected by convenience, due to this study's retrospective nature. Moderators consider user characteristics as well as data related to the internal and external context. Table 3 presents each moderator, along with a brief description that indicates its possible values, an alias that will be used hereafter, and the category it fits in. We consider age and gender as user characteristics because they are basic user information. Because the circumstance and environment wherein participants completed the activities differ depending on their major, we see it as part of the external context. The others are classified as internal context, as they are characteristics that influence users' learning experiences, such as previous experience, currently working/interning, or having internet access.

Table 3. Possible moderators of gamification's success analysed.

Description	Alias	Category
Student's age (numeric)	Age	User characteristic
Whether one is male (1) or female (0)	Male*	User characteristic
Whether one has a PC at home (1) or not (0)	PC	Internal context
Whether one share a PC at home (1) or not (0)	SharesPC	Internal context
Whether one has internet at home (1) or not (0)	Internet	Internal context
Whether one has previous experience with any programming language (1) or not (0)	Exp	Internal context
Whether one has worked/interned (1) or not (0)	Worked	Internal context
Which major one is enrolled at	Major	External context

* Because this information was dichotomous (male or female), we dummy coded it as 1 for Male and 0 for Female to facilitate the interpretation of the regression coefficient.

Note that despite covariates being selected by convenience, selecting them addresses literature limitations. On one hand, the frameworks recommended in the literature, which explain how gamification works, do not define what are the possible moderators of its effects [46, 48], possibly because that question remains open [43, 85]. On the other hand, evidence from empirical studies is unclear in terms of what those moderators are, especially in terms of contextual factors (see Section 1.2). Therefore, while the literature often indicates moderators, what are these factors remain undefined. Thus, we approach it through an exploratory perspective, based on new factors.

2.4 Data Collection

First, all students completed a characterisation questionnaire that captured demographics and internal context-related information presented in Table 3. This was accomplished in the first week of the semester. *Second*, students were offered several programming assignments that they could complete to practice their programming skills. Completing these assignments was optional and had a very small impact (about 6 %) on the final grade of both groups. There were PA available during the whole term, that is, for 15 weeks. In 2016, the system did not feature gamification yet, whereas it was present in that system in 2017 and 2018. Therefore, students from 2016 feature the control group and students from 2017 and 2018 belong to the experimental group. *Third*, students had to complete a programming exam that worth part of their final grade every two weeks, starting in the term's second week. Thus, their final grade (the measure of academic achievement) was based on scores from exams collected throughout the whole semester.

2.5 Instrumentation

We used to instruments for data collections: an educational system and programming assignments implemented within the system.

2.5.1 Educational System. All participants used the CodeBench[4] system, which is a home-made online judge created by one of the authors. Through this system, instructors/monitors select problems to create assignment lists for programming classes. The system features an embedded Integrated Development Environment - IDE - where students develop solutions and submit them at the same place. When the learner submits a solution for a given problem (i.e., attempts), the system provides instantaneous feedback on whether the solution is correct, partially correct, or incorrect. Such automatic assessment system is both convenient for instructors due to the reduction of workload related to correcting students' attempts and for learners, as they receive instantaneous feedback about the correctness of their solutions.

2.5.2 Programming Assignments. The task participants solved are programming assignments (PA) offered for them to practice programming concepts/techniques introduced in classes. In total they were required to solve seven PA, each one related to one of the programming topics (sequentially taught): (t1) sequential, (t2) conditional (if-then-else), (t3) nested conditionals, (t4) repetition by condition (using loops with *while*), (t5) vectors and strings, (t6) repetition by counting (using loops with *for*), (t7) matrices. Biweekly, students learned one of the seven topics and could solve the PA of the current topic.

2.6 Masking

No masking took place in this study because we collected data in a natural context.

2.7 Conditions and Design

The study follows a quasi-experimental between-subject design to evaluate whether gamification improves students' academic achievement by influencing their behaviours, as well as which factors moderate that effect. The quasi-experiment is characterised due to the lack of random assignment, a common feature for maintaining a natural setting [13]. Instead of random assignment, condition assignment was done based on the year a student was enrolled at a CS1 class of the majors considered in this study.

This study concerns data from five semesters from three different years (2016 to 2018). Data from the first year represent our control condition (control group; N = 118), in which subjects used a non-gamified educational system. Data from the subsequent years represent our experimental condition (experimental group; N = 281), in which subjects used the same system but with gamification. Hence, characterising the between-subject design as participants interacted with a single condition.

While students of the control group used the standard, non-gamified version of the system, those in the experimental group used it with gamification. The gamification design was planned according to the proposal of [98], which combined aspects of the ADDIE (Analyse, Design, Develop,

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

Implement, and Evaluate) approach [11] with the Instructional Systematic Design model [20]. ADDIE is a product development concept used in educational environments to build student-centered learning [11]. Similarly, the Instructional Systematic Design model [20] is an instructional design process based on learning theories and research, and practical experience. According to those, the gamification design resulted in the following:

- *Goal*: Motivating students to solve the PA's problems.
- *Media type*: Digital and online.
- *Context*: Programming classes or when appropriate to the learner.
- *Interaction*: Single user.
- *Narrative*: A medieval fantasy world where characters (students from the same class) must face a monster (Chimera) to free their lands from domination.
- *Description*: The student chooses their avatar among several options available. When they solve a problem, they progress in the map towards the Chimera, winning strength points and weapons. The greater the strength and the better the avatar's weapon, the more hit points the student can take from the monster when facing it.
- *Results*: A percentage of the class must reach the end of the map to find and kill the Chimera. Although the interaction is individual, this collective objective aims to minimise undesired competition among students. Killing the Chimera leads to the "winning state".
- *Feedback*: While using the gamified online judge, students receive additional feedback about their performance according to their characters' positions, weapons and strength.

Next, we further describe the system's gamified version, relating its main aspects to the game elements of a recent taxonomy of game elements for education [93]. In the gamified version, deployed since 2017, the students see themselves as one of the characters in a fictional world, where they can walk through maps, overcome an enemy, investigate stories and explore environments (Storytelling and Narrative). The gamification does not influence the content of the course, nor completing its assignments is mandatory for students. However, when the students correctly solve the problems from the assignment lists created by the instructors/monitors, they receive rewards (Acknowledgement) that allow them to advance within the story (Progression), unlock new items and new interactions with the environment (Novelty). Thus, the reward is given to those who are more dedicated to solving the programming problems. Furthermore, the students can compete to each other (Competition) based on the rewards received after completing assignments. Nevertheless, they can only finish the fictional world's story if a large proportion of the class completes the assignments. That is, they have to work together (Cooperation). Summarising, the gamification design is a mix of immersive-, social-, and challenge-based gamification: it is mostly concerned with providing an immersive experience through fictional game elements (i.e., narrative and storytelling), while also presenting performance feedback (e.g., points and progression). Figure 2 shows the map in which learners explore the story.

2.8 Analytic Strategy

Because our data follows a hierarchical structure (e.g., students grouped by their majors), we used multilevel regression for data analysis. Multilevel regression is designed for statistically analysing, at the same time, multiple variables from different levels of a hierarchical structure, as well as taking into account their dependencies [34]. As we need to simultaneously analyse the relationship between variables of distinct hierarchical levels (e.g., users' data, level 1, and their classes, level 2) to achieve our goal, multilevel regression is adequate for testing our hypotheses and answering our research question.

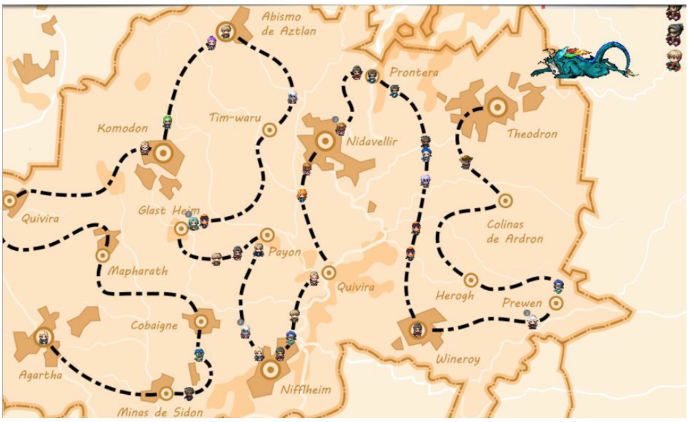

Fig. 2. Map in which learners' avatars advance within the story provided by the gamification design as they complete assignments.

To account for such group differences, multilevel models allow each one to have its intercept and regression coefficients, which are known as *random coefficients*. In the case of this study, students are grouped by majors, thus, each major will have an intercept. Additionally, we want to evaluate whether gamification's effect differs depending on the student's major (i.e., do majors moderate gamification's effect?). Therefore, our model will allow the gamification coefficient to vary across majors to estimate these possible differences. Complementary, multilevel models also estimate coefficients that apply to the overall model, which are known as *fixed coefficients*. Fixed coefficients estimate a predictor's overall effect, whereas random coefficients capture how much a group differs from the sample's average. As our hypotheses assume a single effect depends on the grouping (i.e., gamification's influence depends on students' majors), the remaining variables we analyse are all considered as fixed coefficients only as including many random coefficients substantially increase model complexity and sample size needed [14].

For the analysis procedure, we followed a set of literature recommendations concerning data preparation and model development. For data preparation, we applied three transformations. First, we transformed the continuous variables using the squared root transformation to guarantee model validity as, without these transformations, some multilevel regression assumptions were being violated (e.g., heteroscedasticity) [25]. Second, we scaled these square-root-transformed continuous variables, which is recommended for models featuring interactions. Scaling is important to guarantee variables from an interaction are in similar scales, as well as for coefficients interpretation [25]. Third, we transformed nominal variables (e.g., gamified or not, is male or not) into factors. Scaling was not applied to non-numerical variables because they are all dummy coded (e.g., gamified or not, has previous experience or not), thereby, their values and interactions are meaningful without scaling.

For model development, we followed a top-down approach [34]. That is, starting with a *full* model (i.e., all interactions between all variables examined) and iteratively removing predictors that do not affect the model fit. In the context of this study, this was accomplished by starting with a model that can be represented as:

$$finalgrade = gamified \times sumattempts \times (usercharacteristics + internalcontextdata)$$

where the equation aims to model a student's final grade (the academic achievement measure) based on interactions between the condition (gamified or not; dummy coded), the measure of practice (*sumattempts*), and moderators under analysis (e.g., gender and previous experience with some programming language). Thus, this model allows us to analyse all influences and moderations predicted by our hypotheses. Note that the measure of external context (i.e., students' major) is the grouping factor, considered as a random coefficient. Then, to identify which predictors to remove, we used the Likelihood Ratio Test (LRT). The LRT is recommended as it calculates a predictor's impact based on the change in model fit when it is removed, indicating whether it is significant or not [25]. Hence, we iteratively removed all predictors that insignificantly affect model fit, starting from the three-way interactions (interactions between three predictors), then the two-way, and, lastly, the single terms; that is, from the most complex ones to the simplest.

Given the exploratory nature of investigating these moderations, we adopted a 10% alpha level, following literature suggestions [28, 34]. Furthermore, due to the multiple comparisons in this procedure, we adjusted p values using the False Discovery Rate approach as is has been recommended over the Bonferroni approach [39]. All analyses were conducted using R[5], R studio[6], and the *lme4* package [7].

3 RESULTS

Table 4 presents descriptive statistics of the main measures analysed in this study: students' final grade (academic achievement) and how much they practised programming (sumattempts). The table describes these values in the raw form as well as after the transformations applied for data analysis for the reader's reference when interpreting our results. Next, this section presents the development process for modelling students' academic achievement based on the predictors previously introduced. Then, we present how the multilevel model developed answers our hypotheses, as well as additional insights it reveals.

Table 4. Measures' descriptive statistics.

	Academic Achievement			Sum of attempts		
Statistic	Raw	Root squared	S+S	Raw	Root squared	S+S
Mean	5.40	2.15	0	394.7	18.21	0
SD	3.22	0.87	1	349.70	7.95	1
Q1	2.39	1.55	-0.70	189.00	13.75	-0.56
Q3	8.20	2.86	0.82	507.50	22.53	0.54

S+S = Root squared then Scaled.

3.1 Modelling Students' Academic Achievement

Following the top-down approach [34], we fitted a full model as previously defined. We found no three-way interaction significantly affected model fit (p-adj > 0.1), as shown in Table 5. Therefore, we removed predictors corresponding to those interactions and moved to the next step: testing whether any two-way interaction affects the model fit (Table 6). Results show five two-way interactions significantly affect model fit: gamification and sumattempts, gamification and Male, sumattempts and Age, sumattempts and Male, and sumattempts and Internet. Therefore, we kept these interactions in the model. Then, we tested whether the predictors not involved in any significant interaction affect the model fit alone (Table 7), finding that Worked was the only non-significant predictor. Thus, we removed it and fitted the final model, which is summarised in Table 8.

[5] https://www.r-project.org/

[6] https://rstudio.com/

Table 5. Likelihood ratio tests assessing three-way interactions in modelling students' academic achievement.

Predictors	F(num_df, den_df)	p-val	p-adj
gamified:sumattempts:Age	0.076(1, 380.38)	0.783	0.818
gamified:sumattempts:Male	0.053(1, 382.61)	0.818	0.818
gamified:sumattempts:PC	3.393(1, 326.67)	0.066*	0.465
gamified:sumattempts:SharesPC	0.856(1, 367.27)	0.355	0.718
gamified:sumattempts:Internet	0.680(1, 339.46)	0.410	0.718
gamified:sumattempts:Exp	0.380(1, 392.20)	0.538	0.753
gamified:sumattempts:Worked	0.876(1, 371.97)	0.350	0.718

* p < 0.1

Table 6. Likelihood ratio tests assessing two-way interactions in modelling students' academic achievement.

Predictors	F(num_df, den_df)	p-val	p-adj
gamified:sumattempts	6.623(1, 396.95)	0.010***	0.052**
gamified:Age	2.522(1, 385.17)	0.113	0.242
gamified:Male	10.476(1, 396.56)	0.001***	0.010**
gamified:PC	0.189(1, 392.07)	0.664	0.824
gamified:SharesPC	1.146(1, 391.44)	0.285	0.428
gamified:Internet	0.632(1, 391.69)	0.427	0.582
gamified:Exp	0.001(1, 371.16)	0.979	0.979
gamified:Worked	0.021(1, 389.76)	0.884	0.947
sumattempts:Age	20.543(1, 393.96)	0.000***	0.000***
sumattempts:Male	5.719(1, 396.58)	0.017**	0.065*
sumattempts:PC	2.703(1, 393.00)	0.101	0.242
sumattempts:SharesPC	1.673(1, 396.74)	0.197	0.369
sumattempts:Internet	4.683(1, 392.47)	0.031**	0.093*
sumattempts:Exp	1.256(1, 394.37)	0.263	0.428
sumattempts:Worked	0.135(1, 393.81)	0.714	0.824

* p < 0.1; ** p < 0.05; *** p < 0.01

Table 7. Likelihood ratio tests assessing single predictors in modelling students' academic achievement in the presence of significant interactions.

Predictors	F(num_df, den_df)	p-val	p-adj
PC	4.459(1, 392.58)	0.035**	0.052*
SharesPC	4.211(1, 392.23)	0.041**	0.052*
Exp	8.093(1, 393.18)	0.005***	0.011**
Worked	0.121(1, 392.14)	0.728	0.728
gamified:sumattempts	8.005(1, 395.10)	0.005***	0.011**
gamified:Male	9.756(1, 395.28)	0.002***	0.009***
sumattempts:Age	20.510(1, 393.89)	0.000***	0.000***
sumattempts:Male	5.276(1, 396.28)	0.022**	0.040**
sumattempts:Internet	3.390(1, 392.65)	0.066*	0.075*

* p < 0.1; ** p < 0.05; *** p < 0.01

Table 8. Multilevel model predicting students' academic achievement.

Predictors	Est.(SE)	CI	p-val
Fixed Effects			
(Intercept)	-0.55 (0.17)	-0.84 – -0.27	0.001***
gamified	0.62 (0.14)	0.38 – 0.86	0.000***
Male	0.40 (0.14)	0.17 – 0.64	0.004***
sumattempts	0.58 (0.12)	0.38 – 0.77	0.000***
Age	0.04 (0.04)	-0.02 – 0.11	0.276
Internet	0.50 (0.15)	0.25 – 0.76	0.001***
PC	-0.39 (0.18)	-0.69 – -0.09	0.035**
SharesPC	-0.17 (0.08)	-0.31 – -0.03	0.041**
Exp	0.22 (0.08)	0.09 – 0.34	0.005***
gamified * sumattempts	-0.28 (0.10)	-0.44 – -0.12	0.005***
gamified * Male	-0.54 (0.17)	-0.82 – -0.25	0.002***
sumattempts * Age	0.17 (0.04)	0.11 – 0.23	0.000***
sumattempts * Male	0.19 (0.08)	0.05 – 0.33	0.022**
sumattempts * Internet	0.15 (0.08)	0.02 – 0.28	0.065*
Random Effects			
	Var(SD)		
Residual	0.50 (0.71)		
Major	0.04 (0.19)		
gamified * Major	0.00 (0.01)		
Model fit			
Marginal R^2	0.46		
Conditional R^2	0.50		

* p < 0.1; ** p < 0.05; *** p < 0.01

Furthermore, to test whether the external context moderator (students' major) affects the model fit, we performed an LRT test comparing the full model shown in Table 8 to a version without the random coefficient that allows gamification's effect to be moderated by a student's major. This test yielded non-significant results (LRT(2) = 0.02; p = 0.989), indicating the random effect does not affect model fit. We highlight that adding or removing a random coefficient does not change a model's estimates, however, removing it increases the chances of finding false-positive fixed coefficients (i.e., inflating type I errors) [34]. Therefore, we left the random effect in our final model, although it does not significantly improve model fit. In the last step of this process, we tested the final model validity concerning aspects such as normality of random effects and heteroscedasticity; none was violated. The normality of residuals was not met, but due to our sample size, this is unlikely to affect model validity [55]. Assumptions' testing and a summary of all models developed in this process are available in the supplementary material.

3.2 Study Hypotheses

From our final model (Table 8), we can discuss our hypotheses as well as answer our research question. **H1** predicted that practising to code would positively influence students' academic achievement. According to the model, the extent to which students practised (sumattempts) has a positive, highly significant effect on academic achievement, suggesting the more students practised, the higher was their academic achievement. Therefore, supporting **H1**.

H2 predicted gamification would positively influence learners' academic achievement, by maximising the testing effect. Our model suggests gamification (gamified) has a positive, highly significant effect on academic achievement. However, its interaction with how much students practised (sumattempts) is negative and highly significant. This indicates gamification had a positive effect on academic achievement, as expected, but the more students practised, the more this effect decreased (see Figure 3), unlike per our expectations. Thus, this only partially supports **H2**.

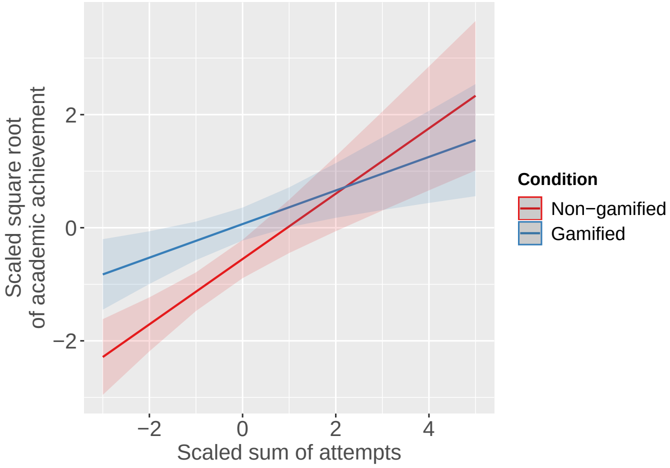

Fig. 3. Predicted effects of gamification on the impact that practising to code (sum of attempts) has on academic achievement.

H3 predicted that gamification's effects would depend on user and contextual characteristics. The only significant moderator of gamification's effect was learners' gender (Figure 4). The moderator effect of the remaining user and contextual characteristics were non-significant. For data from the internal context and the user, this was found during the model development process as those interactions were found to not affect the model fit. For the external context moderator that we evaluated, this was shown by testing the random coefficient effect, but can also be seen by the small variances in the random part shown in Table 8, as well as the negligible improvement of the Conditional R^2 (4%; Conditional R^2 = 0.50), which considers the random and the fixed model parts, compared to the marginal R^2 (46%), which only considers the fixed part. Thus, only partially supporting **H3**.

3.3 Additional Findings

Moreover, our model revealed additional insights that do not directly relate to this study's hypotheses and research questions. Despite our focus on gamification's effect, we found insights concerning direct moderators of academic achievement as well as factors that moderate the impact of practising. Surprisingly, the developed model indicates having a PC at home has a negative, highly significant impact on academic achievement. The model also indicates that sharing the PC at home has a negative, significant effect on academic achievement. On the other hand, the model indicates previous experience with any programming language and having internet have positive, highly significant impacts on academic achievement. The model also indicates males presented higher academic achievements than females. Additionally, the final model indicates the testing

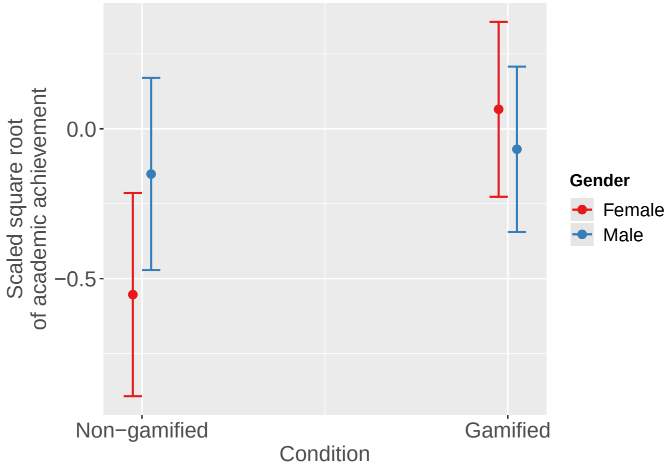

Fig. 4. Gamification's effects on academic achievement as moderated by students' gender.

effect (i.e., practising to code to improve learning via self-testing) is moderated by students' age, gender, and whether they have internet at home. The indications are that i) compared to younger students, older ones accomplish higher academic achievement as they practice more; ii) the more males practice, the higher is the difference in their academic achievement compared to females, and iii) the more one with internet practices, the larger is their academic achievement compared to those with no internet.

3.4 Summary

From our findings, the answer to *how user and contextual factors influence gamification's effect on the academic achievement of CS1 students* is that gender significantly moderated that influence, positively for females and null for males, whereas user age and contextual factors (i.e., having a PC at home, sharing a PC at home, having internet, having previous experience with some programming language, having already worked/interned, and major enrolled at) had null (non-significant) influences on that effect. Furthermore, we found that for those who practised more than two standards deviations above the average (i.e., 1094 attempts), gamification's effect changed from positive to negative. Our results also revealed additional insights concerning moderators of academic achievement as well as of the testing effect.

4 DISCUSSION

Our results support the testing effect in the study context and show that students in the gamified version yielded higher academic achievement. The difference was moderated by participants' gender. We noted that some learners self-tested to a point in which their academic achievement was below that of those who practised less. In contrast to this study, previous work on gamification applied to CS1 education rarely performed moderator analysis and results were often based on small usage periods and relatively small samples. Differently, we analysed data from 399 learners, evaluated gamification's effects after they used it for a whole semester (15 weeks), and analysed the moderator role of eight factors, including user (e.g., age and gender) and contextual (major) information. Thus, our contribution is revealing user and contextual characteristics that moderate gamification's impact on CS1 students' academic achievement, as well as whether those maximise or minimise that impact, based on empirical evidence build upon a sample of substantial size and long-term usage of the intervention. Next, we discuss our results related to relevant literature.

First, our findings provide empirical support for the testing effect [84] in the context of programming learning. The results demonstrate that the more the students practised programming, the higher their academic achievements were at the end of the semester. On one hand, previous research has explored and demonstrated the testing effect's benefits for programming learning, but either with experienced programmers or in a non-gamified environment (e.g., [75, 103]). On the other hand, studies have also shown evidence on the testing effect's positive impact on learning from using gamified systems (e.g., [16, 79, 86]) but in other contexts or based on small samples. Whereas our findings corroborate the results of those studies, we studied the testing effect based on data from beginner programmers (CS1 learners) using a gamified educational system. Therefore, we contribute with support for the testing effect within the context of CS1, providing evidence that practising and submitting programming assignments is positively related to higher academic achievement.

Second, our findings demonstrate gamification's contribution to students' academic achievement was positive. The analyses suggested gamification had a positive impact on learners' academic achievement. This finding corroborates the general gamification literature [43] as well as gamified learning effects, in which meta-analytic results indicate that gamified interventions lead to small positive learning outcomes, compared to non-gamified conditions [85]. This positive result is aligned to the positive outcomes of gamified programming learning as well (e.g., [23, 53]; Section 1.2). It should be noted, however, that empirical studies' methodological rigour has been a limitation of the field, which raises questions on findings' validity [18, 32, 90]. Differently, our findings are based on an examination featuring a control group, a substantial sample size and a long usage period, besides controlling for covariates in data analysis. Therefore, by corroborating the overall literature, our findings contribute robust empirical evidence on gamification's benefits to learning. Thus, we expand the literature both in terms of gamification's impact, with evidence on its effectiveness in the context of programming learning based on data from a significant sample size who used gamification for an entire semester, as well as to computing education, supporting the potential from gamification to support programming learning.

We also found that gamification's impact was influenced by two factors: the amount of practice and gender. Concerning the relationship between gamification and the testing effect, our analysis revealed that gamification's contribution was mixed. The results for this analysis indicate gamification improved the testing effect; however, such improvement not only vanished but became negative for students who practised substantially more than the average (see Figure 3). We interpret this finding from two perspectives. One possible explanation might lie behind the support gamification provides to learning. In a meta-synthesis of gamification literature, Bai et al. [4] found learners enjoy gamification because it provides performance feedback and means to be recognised (e.g., badges), as well as goal-setting. Feedback is important for programming learning [75, 103], being recognised is likely to fulfil competence needs, which also contributes to meaningful learning experiences, as well as goal-setting [91, 97]. Similarly, the game elements might have contributed to learners' feelings of self-efficacy, which is positively related to learning performance as well [36, 102].

On the other hand, Bai et al. [4] found gamification might cause anxiety or jealousy. Those feelings might lead students to do whatever it takes to not feel this way, such as substantially interacting with the system aiming to receive rewards and climb up the leaderboard, without paying attention to the learning task. For instance, students might start submitting several attempts, in a "desperate" effort to solve the question, wherein most submissions are likely to be partially or completely wrong. Specifically, the system we used does not check whether a new submission differs from previous ones. Then, in the gamified version, students might engage in behaviours such as adding small incremental changes (or even simple resubmissions) due to the anxiety to climb the leaderboard and/or their jealousy of those at the top [4], instead of putting the adequate effort to understand and correctly solve the question. Differently, in the version without gamification, students would have no motivation for engaging in such behaviours. Such potentially desperate, reward-driven behaviours indicate students were just gaming the system: seeking game-like rewards instead of properly using it to practice and learn [5]. Hence, explaining why some of our participants ended up yielding lower academic achievement than those who submitted fewer or the same number of attempts in the non-gamified version. A similar outcome was reported in Ghaban and Hendley [26], wherein learners of the gamified version dropped out less, but showed worse learning gains. Thus, we expand the literature with insights about gaming the system behaviour in such context, empirically demonstrating that, although gamification is of value, it might lead to outcomes opposed to the expected, which also contributes to the literature by responding to the need of analysing gamification’s negative effects [43].

Concerning gender, our analysis revealed that the academic achievement of male learners was almost the same, regardless of using the gamified system or not, while showing that female learners’ were positively affected by gamification (see Figure 4). This finding corroborates claims that user characteristics affect their experiences with gamified systems (e.g., [18, 60]), as well as has been shown in the context of games (e.g., [61, 76]). On the other hand, this finding is contrary to studies that found no moderator effect of gender [44, 72, 86]. A reason for that contradiction might be the gamification design. Research acknowledges that the gamification design is determinant for users’ experience with gamified systems [51, 59]. Whereas we used a mix of immersion, challenge, and social gamification in this study, previous research [44, 72, 86] mostly focused on challenge-based designs. As different designs were used, it might be that one of them was equally good, bad, or null for all learners, whereas the other was different for females and males. Hence, explaining the fact that gender was a significant moderator in one case but not in others.

Despite gender being the only factor we found to moderate gamification’s impact, we analysed another seven aspects, related to both the user and the context. Those are based on arguments that for gamification to yield positive outcomes, it needs to be designed according to the user and the context (e.g., [43, 59, 93]). Therefore, one might expect that user and contextual characteristics will moderate gamification’s effect. There are studies exploring these claims, showing whether gamification usage is successful or not depends on aspects such as age [44], major [72], attitudes towards game-based learning [3], goal orientation [29], performance [1], and the user and context in general [18, 31]. Nevertheless, recent research still highlights the need for studying moderators, with calls to advance the understanding of pre-determinants and occasions in which learners take advantage of gamification [43, 49, 85]. Therefore, this study contributes by responding to such calls, analysing moderators not only related to user demographics (e.g., gender and age) but also exploring those related to internal (e.g., previous experience with programming) and external (e.g., the major a student is enrolled at) context. Nevertheless, our findings are mostly contradictory to the overall discussion in related work, as ours indicate the impact of gamification only depended on learners’ gender. In contrast, what we found mostly corroborates moderator analysis in the context of programming learning, except for the moderator role of gender (see Section 1.2). Thus, reinforcing the need for advancing the understanding of moderators of gamification’s success, especially in varied contexts.

Furthermore, we also found some factors moderated the testing effect as well as students’ academic achievement. Despite the focus of this study is on gamification and its moderators, our analyses revealed interesting findings that contribute with evidence on factors likely to affect programming learning. On one hand, we found some characteristics related to internal context

[88] moderated academic achievement. As expected, users with previous experience in some programming languages achieved higher academic achievement than those with no previous experience. Differently and surprisingly, students with at least one computer at home presented lower academic achievement compared to those with no computer. Similarly, having to share the home computer negatively affects academic achievement whereas the academic achievement of learners with internet at home was higher than that of learners without it. On the other hand, we also found three factors moderated the testing effect (i.e., age, gender, and internet), in which all of those maximised it. These findings do not directly relate to the objective of this study, they rather emerged as a consequence of our data analysis process. Then, we briefly presented and interpreted them so that the interested reader can understand what our analysis revealed. Nevertheless, we believe these findings are of value for those interested in moderators of learning as well as of the testing effect. Those provide insights on the characteristics of learners that are more likely to take more or less advantage from the testing effect, as well as factors that play a role in CS1 students' learning. Thus, opening directions for future research to investigate these insights.

4.1 Implications

We highlight five main implications of our findings. First, we have shown that students who practice more are likely to yield higher academic achievement at the end of the semester. Instructors can explore this finding by providing their students with opportunities to practice programming as much as possible, which can be accomplished by making several programming assignments available during the course. Second, we demonstrated gamification can enhance as well as mitigate the testing effect. On one hand, this suggests that instructors can rely on gamified systems to improve the effectiveness of the testing effect. On the other hand, this implies that gamification must be designed cautiously, aiming to prevent behaviours such as gaming the system or feelings of jealousy and anxiety, as for some users gamification might end up diminishing the testing effect. Third, our finding concerning students who gamed the system contributes to the design of online judges. The tool our participants used did not impose restrictions on the extent to which one submission should differ from previous ones, nor decreased gamification outcomes upon repeated attempts. Hence, designers of online judges might consider imposing such restrictions, to avoid the submission of (almost) identical solutions, aiming to push students to work on improving their code instead of just trying the correction system until getting it right. Fourth, we found gamification's impact was different depending on whether users were males or females. This further supports the claim that one size does not fit all [60], indicating the need for developing and providing gamification designs tailored to specific audiences. Fifth, we found no support for the moderator effect of various user and contextual characteristics, which is contrary to the overall discussion from previous studies (e.g., [31, 81, 93]) and our initial expectations. Hence, pointing that more research is needed to understand what moderates gamification's success.

4.2 Limitations

This study has some limitations that must be considered when interpreting its findings. First, as we explored data from three years, we could not guarantee majors had the same instructor in both the control and experimental condition. This was mitigated with all other aspects being the same (e.g., class program, handbooks, exam structure). Nevertheless, this is likely to not affect our findings as a recent secondary study suggests gamification's effect is the same regardless of instructors being different or not [4].

Second, because this is a retrospective study, we could not use random assignment, and groups significantly differed in some demographic characteristics. Considering covariates (e.g., age, gender, previous experience) helps to handle this limitation as the analysis would reveal if some difference is due to the covariate itself and not due to conditions. Additionally, the meta-analysis by Sailer and Homner [85] found randomisation did not affect gamification's impact on behavioural outcomes, suggesting our second limitation is unlikely to threaten our results. Also, most covariates are self-reported and, in some cases, based on binary answers, which respectively inserts subjectivity and limits the information they provide. While that limits statistical results, the successful use of similar data in prior research (e.g., [29, 64, 79]) suggests it represents a mitigated threat to the findings.

Third, we did not conduct a pre-test, which opens the possibility for students from one condition to have more previous knowledge than those in the other. Consequently, this could lead to misleading conclusions regarding a condition's contribution to students' academic achievement. To mitigate this threat, we analysed covariates related to possible prior knowledge (i.e., previous experience and having worked/interned before), hence, accounting for those possible differences in the data analysis. Meta-analytic evidence further supports the reduced role of this limitation (the lack of a pre-test) in our findings: it demonstrates gamification's effect on cognitive outcomes (e.g., academic achievement) does not significantly change, depending on whether only post-test or pre- and post-tests were used [85].

Fourth, the number of participants in each condition was unbalanced ($N_{control}$ = 118; $N_{experimental}$ = 281). This limitation emerged because data from a single year (2016) composed the control group, whereas data from two years (2017 and 2018) composed the experimental group. The reason is that gamification was deployed to the system used for data collection in 2017. To cope with this limitation, we performed statistical analysis to shed light on whether differences were not by mere chance. Nevertheless, the number of participants per group in our study is above the average of participants per study in similar research (e.g., 107, 95, and 75 according to Bai et al. [4], Sailer and Homner [85], and Koivisto and Hamari [42], respectively), which demonstrates our analysis is based on a representative sample size (N = 399) in terms of gamification studies for both study groups.

Lastly, we note our sample consists of STEM, not computing-related majors, such as Computer Science and Software Engineering. While studying STEM-related majors is important, because their students face difficulties with CS1 often [21, 87], we cannot ensure that our findings will hold with computing-related majors. However, we note that the effects did not vary from one major to another within our sample, so it is reasonable to expect similar outcomes for other majors.

4.3 Future Work

As future work, we suggest some lines of research. First, we call for more research to understand when gamification does not work. We found mixed effects, in which we discussed those in terms of gaming the system, possibly motivated by participants' feeling anxious or jealous. Further experiments to understand when and why such effects happen would contribute to the design of more effective gamified systems, consequently, contributing to students' learning, as has been suggested recently [43]. Towards preventing these undesired behaviours, game elements exploring game fiction and socialisation [93] are promising [85], as they do not rely on external motivators. Thus, we encourage future studies to evaluate the impacts of game fiction and socialisation on CS1 learning.

We also call for research revealing factors that moderate gamification's effectiveness. In this study, we found evidence gender was one of such moderators, while our results indicated other factors (e.g., age, previous experience, and student's major) were not. Such indications add to the literature, confirming the need to determine which are those factors, as well as understanding when and in which occasion gamification works [6, 85]. Accordingly, experiments to identify which are those moderators would contribute to the understanding of factors to consider when designing gamified interventions.

Based on that need, we call for research on designing more tailored gamified interventions. This also emerges from our finding that the gamification design we used was effective for females but null for males, highlighting the need for developing gamification designs that are tailored to the target sample [35]. That is, once we are aware there are males and females in the target population, gamification should be planned accordingly, such as in personalised gamification [82]. Consequently, this calls for research to understand how to gamify an educational environment to a target population while considering all of its relevant characteristics [41], including the users, the task, and the context [31, 50, 77].

5 CONCLUSIONS

Through a quasi-experimental retrospective study (N = 399), we showed that gamification positively affected CS1 students' academic achievement, by enhancing the testing effect. However, the results also suggested that some learners were gaming the system rather than studying, which led to negative learning outcomes. This leads us to conclude that gamification can contribute to CS1 education, by enhancing behaviours valuable to learning (e.g., self-testing). However, it should be planned and deployed with caution, to prevent inciting undesired behaviours. Furthermore, we examined the moderator role of several user and contextual characteristics, finding only gender as a significant moderator. Surprisingly, this contradicts previous research advocating that gamification's success depends on multiple user and contextual characteristics, which was not supported by our analyses. Nevertheless, more research is needed to better understand and ground which factors moderate the impact of gamification, given that, currently, there are many theoretical discussions and few empirical examinations in this direction, especially those concerning the role of context, as in this study.

6 ACKNOWLEDGMENTS

The authors are thankful to the financial support of the following agencies: Brazilian National Council for Scientific and Technological Development (CNPq) - processes 141859/2019-9, 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, 2018/07688-1, 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 nº 6.008/2006 (SUFRAMA), and partially funded by Samsung Electronics of Amazonia Ltda., under the terms of Federal Law nº 8.387/1991, through agreements 001/2020 and 003/2019, signed with Federal University of Amazonas and FAEPI, Brazil.

REFERENCES

[1] Samuel Abramovich, Christian Schunn, and Ross Mitsuo Higashi. 2013. Are badges useful in education?: It depends upon the type of badge and expertise of learner. *Educational Technology Research and Development* 61, 2 (2013), 217–232.

[2] Adnan Ahmad, Furkh Zeshan, Muhammad Salman Khan, Rutab Marriam, Amjad Ali, and Alia Samreen. 2020. The impact of gamification on learning outcomes of computer science majors. *ACM Transactions on Computing Education (TOCE)* 20, 2 (2020), 1–25.

[3] Michael B. Armstrong and Richard N. Landers. 2017. An Evaluation of Gamified Training: Using Narrative to Improve Reactions and Learning. *Simulation & Gaming* 48, 4 (2017), 513–538. https://doi.org/10.1177/1046878117703749 arXiv:https://doi.org/10.1177/1046878117703749

[4] Shurui Bai, Khe Foon Hew, and Biyun Huang. 2020. Is gamification "bullshit"? Evidence from a meta-analysis and synthesis of qualitative data in educational contexts. *Educational Research Review* (2020), 100322.

[5] Ryan SJd Baker, Albert T Corbett, Ido Roll, and Kenneth R Koedinger. 2008. Developing a generalizable detector of when students game the system. *User Modeling and User-Adapted Interaction* 18, 3 (2008), 287–314.

[6] Gonçalo Baptista and Tiago Oliveira. 2019. Gamification and serious games: A literature meta-analysis and integrative model. *Computers in Human Behavior* 92 (2019), 306–315.

[7] Douglas Bates, Martin Mächler, Ben Bolker, and Steve Walker. 2014. Fitting linear mixed-effects models using lme4. *arXiv preprint arXiv:1406.5823* (2014).

[8] Brett A Becker and Keith Quille. 2019. 50 years of cs1 at sigcse: A review of the evolution of introductory programming education research. In *Proceedings of the 50th acm technical symposium on computer science education*. 338–344.

[9] Jens Bennedsen and Michael E Caspersen. 2019. Failure rates in introductory programming: 12 years later. *ACM Inroads* 10, 2 (2019), 30–36.

[10] Paulo Blikstein. 2011. Using learning analytics to assess students' behavior in open-ended programming tasks. In *Proceedings of the 1st international conference on learning analytics and knowledge*. 110–116.

[11] Robert Maribe Branch. 2009. *Instructional design: The ADDIE approach*. Vol. 722. Springer Science & Business Media.

[12] Patrick Buckley and Elaine Doyle. 2017. Individualising gamification: An investigation of the impact of learning styles and personality traits on the efficacy of gamification using a prediction market. *Computers & Education* 106 (2017), 43–55.

[13] John W Creswell and J David Creswell. 2017. *Research design: Qualitative, quantitative, and mixed methods approaches*. Sage publications.

[14] Jan De Leeuw, Erik Meijer, and Harvey Goldstein. 2008. *Handbook of multilevel analysis*. Springer.

[15] Luis De-Marcos, Adrián Domínguez, Joseba Saenz-de Navarrete, and Carmen Pagés. 2014. An empirical study comparing gamification and social networking on e-learning. *Computers & education* 75 (2014), 82–91.

[16] Paul Denny, Fiona McDonald, Ruth Empson, Philip Kelly, and Andrew Petersen. 2018. Empirical Support for a Causal Relationship Between Gamification and Learning Outcomes. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems (CHI '18)*. Association for Computing Machinery, New York, NY, USA, 1–13. https://doi.org/10.1145/3173574.3173885

[17] Sebastian Deterding, Dan Dixon, Rilla Khaled, and Lennart Nacke. 2011. From game design elements to gamefulness: defining gamification. In *Proceedings of the 15th international academic MindTrek conference: Envisioning future media environments*. ACM, 9–15.

[18] Christo Dichev and Darina Dicheva. 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 (2017), 9.

[19] Darina Dicheva, Keith Irwin, and Christo Dichev. 2019. OneUp: Engaging Students in a Gamified Data Structures Course. In *Proceedings of the 50th ACM Technical Symposium on Computer Science Education*. 386–392.

[20] Walter Dick, Lou Carey, and James O Carey. 2005. The systematic design of instruction. (2005).

[21] Leovy Echeverría, Ruth Cobos, Liliana Machuca, and Ivan Claros. 2017. Using collaborative learning scenarios to teach programming to non-CS majors. *Computer applications in engineering education* 25, 5 (2017), 719–731.

[22] Samuel C Fonseca, Filipe Dwan Pereira, Elaine HT Oliveira, David BF Oliveira, Leandro SG Carvalho, and Alexandra I Cristea. 2020. Automatic Subject-based Contextualisation of Programming Assignment Lists. EDM.

[23] Panagiotis Fotaris, Theodoros Mastoras, Richard Leinfellner, and Yasmine Rosunally. 2016. Climbing up the Leaderboard: An Empirical Study of Applying Gamification Techniques to a Computer Programming Class. *Electronic Journal of e-learning* 14, 2 (2016), 94–110.

[24] MN Gari, GS Walia, and AD Radermacher. 2018. Gamification in computer science education: A systematic literature review. In *American Society for Engineering Education*.

[25] Andrew Gelman and Jennifer Hill. 2006. *Data analysis using regression and multilevel/hierarchical models*. Cambridge university press.

[26] Wad Ghaban and Robert Hendley. 2020. Can We Predict the Best Gamification Elements for a User Based on Their Personal Attributes?. In *International Conference on Human-Computer Interaction*. Springer, 58–75.

[27] E Bruce Goldstein. 2014. *Cognitive psychology: Connecting mind, research and everyday experience*. Nelson Education.

[28] Joseph F Hair Jr, G Tomas M Hult, Christian Ringle, and Marko Sarstedt. 2016. *A primer on partial least squares structural equation modeling (PLS-SEM)*. Sage publications.

[29] Lasse Hakulinen and Tapio Auvinen. 2014. The effect of gamification on students with different achievement goal orientations. In *2014 international conference on teaching and learning in computing and engineering*. IEEE, 9–16.

[30] Lasse Hakulinen, Tapio Auvinen, and Ari Korhonen. 2015. The Effect of Achievement Badges on Students' Behavior: An Empirical Study in a University-Level Computer Science Course. *International Journal of Emerging Technologies in Learning* 10, 1 (2015).

[31] Stuart Hallifax, Audrey Serna, Jean-Charles Marty, Guillaume Lavoué, and Elise Lavoué. 2019. Factors to Consider for Tailored Gamification. In *Proceedings of the Annual Symposium on Computer-Human Interaction in Play (CHI PLAY '19)*. Association for Computing Machinery, New York, NY, USA, 559–572. https://doi.org/10.1145/3311350.3347167

[32] Juho Hamari, Jonna Koivisto, and Harri Sarsa. 2014. Does gamification work?–a literature review of empirical studies on gamification. In *2014 47th Hawaii international conference on system sciences*. Ieee, 3025–3034.

[33] Matthew Hertz. 2010. What do" CS1" and" CS2" mean? Investigating differences in the early courses. In *Proceedings of the 41st ACM technical symposium on Computer science education*. 199–203.

[34] Joop J Hox, Mirjam Moerbeek, and Rens Van de Schoot. 2010. *Multilevel analysis: Techniques and applications*. Routledge.

[35] Rui Huang, Albert D Ritzhaupt, Max Sommer, Jiawen Zhu, Anita Stephen, Natercia Valle, John Hampton, and Jingwei Li. 2020. The impact of gamification in educational settings on student learning outcomes: a meta-analysis. *Educational Technology Research and Development* (2020), 1–27.

[36] Xiaoxia Huang and Richard E Mayer. 2019. Adding self-efficacy features to an online statistics lesson. *Journal of Educational Computing Research* 57, 4 (2019), 1003–1037.

[37] Sami Hyrynsalmi, Jouni Smed, and Kai Kimppa. 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*. 96–104.

[38] Petri Ihantola, Arto Vihavainen, Alireza Ahadi, Matthew Butler, Jürgen Börstler, Stephen H Edwards, Essi Isohanni, Ari Korhonen, Andrew Petersen, Kelly Rivers, et al. 2015. Educational data mining and learning analytics in programming: Literature review and case studies. In *Proceedings of the 2015 ITiCSE on Working Group Reports*. 41–63.

[39] Mohieddin Jafari and Naser Ansari-Pour. 2019. Why, when and how to adjust your P values? *Cell Journal (Yakhteh)* 20, 4 (2019), 604.

[40] Jeffrey D Karpicke and William R Aue. 2015. The testing effect is alive and well with complex materials. *Educational Psychology Review* 27, 2 (2015), 317–326.

[41] Ana Carolina Tomé Klock, Isabela Gasparini, Marcelo Soares Pimenta, and Juho Hamari. 2020. Tailored gamification: A review of literature. *International Journal of Human-Computer Studies* (2020), 102495.

[42] Jonna Koivisto and Juho Hamari. 2014. Demographic differences in perceived benefits from gamification. *Computers in Human Behavior* 35 (2014), 179 – 188. https://doi.org/10.1016/j.chb.2014.03.007

[43] Jonna Koivisto and Juho Hamari. 2019. The rise of motivational information systems: A review of gamification research. *International Journal of Information Management* 45 (2019), 191 – 210. https://doi.org/10.1016/j.ijinfomgt.2018.10.013

[44] Markus Krause, Marc Mogalle, Henning Pohl, and Joseph Jay Williams. 2015. A Playful Game Changer: Fostering Student Retention in Online Education with Social Gamification. In *Proceedings of the Second (2015) ACM Conference on Learning @ Scale (L@S '15)*. Association for Computing Machinery, New York, NY, USA, 95–102. https://doi.org/10.1145/2724660.2724665

[45] Carmen Lacave, Ana I Molina, and José A Cruz-Lemus. 2018. Learning Analytics to identify dropout factors of Computer Science studies through Bayesian networks. *Behaviour & Information Technology* 37, 10-11 (2018), 993–1007.

[46] Richard N Landers. 2014. Developing a theory of gamified learning: Linking serious games and gamification of learning. *Simulation & gaming* 45, 6 (2014), 752–768.

[47] Richard N Landers and Michael B Armstrong. 2017. Enhancing instructional outcomes with gamification: An empirical test of the Technology-Enhanced Training Effectiveness Model. *Computers in human behavior* 71 (2017), 499–507.

[48] Richard N Landers, Elena M Auer, Andrew B Collmus, and Michael B Armstrong. 2018. Gamification science, its history and future: Definitions and a research agenda. *Simulation & Gaming* 49, 3 (2018), 315–337.

[49] Richard N Landers, Gustavo F Tondello, Dennis L Kappen, Andrew B Collmus, Elisa D Mekler, and Lennart E Nacke. 2019. Defining gameful experience as a psychological state caused by gameplay: Replacing the term 'Gamefulness' with three distinct constructs. *International Journal of Human-Computer Studies* 127 (2019), 81–94.

[50] De Liu, Radhika Santhanam, and Jane Webster. 2017. Toward Meaningful Engagement: A Framework for Design and Research of Gamified Information Systems. *MIS quarterly* 41, 4 (2017), 1011–1034.

[51] Kevin Loughrey and Daire Broin. 2018. Are We Having Fun Yet? Misapplying Motivation to Gamification. In *2018 IEEE Games, Entertainment, Media Conference (GEM)*. IEEE, 1–9.

[52] Jenni Majuri, Jonna Koivisto, and Juho Hamari. 2018. Gamification of education and learning: A review of empirical literature. In *Proceedings of the 2nd International GamiFIN Conference, GamiFIN 2018*. CEUR-WS.

[53] B. Marín, J. Frez, J. Cruz-Lemus, and M. Genero. 2018. An Empirical Investigation on the Benefits of Gamification in Programming Courses. *ACM Trans. Comput. Educ.* 19, 1, Article 4 (Nov. 2018), 22 pages. https://doi.org/10.1145/3231709

[54] KB McDermott, S Kang, and HL Roediger III. 2005. Test format and its modulation of the testing effect. In *biennial meeting of the Society for Applied Research in Memory and Cognition, Wellington, New Zealand*.

[55] Douglas C Montgomery, Elizabeth A Peck, and G Geoffrey Vining. 2012. *Introduction to linear regression analysis*. Vol. 821. John Wiley & Sons.

[56] Alberto Mora, Gustavo F Tondello, Laura Calvet, Carina González, Joan Arnedo-Moreno, and Lennart E Nacke. 2019. The quest for a better tailoring of gameful design: An analysis of player type preferences. In *Proceedings of the XX International Conference on Human Computer Interaction*. ACM, 1.

[57] Miguel Ehécatl Morales-Trujillo and Gabriel Alberto García-Mireles. 2020. Gamification and SQL: An Empirical Study on Student Performance in a Database Course. *ACM Transactions on Computing Education (TOCE)* 21, 1 (2020), 1–29.

[58] Julian MORENO and Andres F PINEDA. 2018. Competitive programming and gamification as strategy to engage students in computer science courses. *Revista ESPACIOS* 39, 35 (2018).

[59] Benedikt Morschheuser, Lobna Hassan, Karl Werder, and Juho Hamari. 2018. How to design gamification? A method for engineering gamified software. *Information and Software Technology* 95 (2018), 219–237. https://doi.org/10.1016/j.infsof.2017.10.015

[60] Lennart E Nacke and Christoph Sebastian Deterding. 2017. The maturing of gamification research. *Computers in Human Behaviour* (2017), 450–454.

[61] Rita Orji, Regan L Mandryk, and Julita Vassileva. 2017. Improving the efficacy of games for change using personalization models. *ACM Transactions on Computer-Human Interaction (TOCHI)* 24, 5 (2017), 1–22.

[62] Rita Orji, Kiemute Oyibo, and Gustavo F. Tondello. 2017. A Comparison of System-Controlled and User-Controlled Personalization Approaches. In *Adjunct Publication of the 25th Conference on User Modeling, Adaptation and Personalization (UMAP '17).* ACM, New York, NY, USA, 413–418. https://doi.org/10.1145/3099023.3099116

[63] Rita Orji, Gustavo F Tondello, and Lennart E Nacke. 2018. Personalizing persuasive strategies in gameful systems to gamification user types. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems.* ACM, 435. https://doi.org/10.1145/3173574.3174009

[64] Margarita Ortiz-Rojas, Katherine Chiluiza, and Martin Valcke. 2019. Gamification through leaderboards: An empirical study in engineering education. *Computer Applications in Engineering Education* 27, 4 (2019), 777–788.

[65] Paula T Palomino, Armando M Toda, Luiz Rodrigues, Wilk Oliveira, and Seiji Isotani. 2020. From the Lack of Engagement to Motivation: Gamification Strategies to Enhance Users Learning Experiences. In *Brazilian Symposium on Computer Games and Digital Entertainment.* 1127–1130.

[66] Harold Pashler, Nicholas J Cepeda, John T Wixted, and Doug Rohrer. 2005. When does feedback facilitate learning of words? *Journal of Experimental Psychology: Learning, Memory, and Cognition* 31, 1 (2005), 3.

[67] Lais Z Pedro, Aparecida MZ Lopes, Bruno G Prates, Julita Vassileva, and Seiji Isotani. 2015. Does gamification work for boys and girls? An exploratory study with a virtual learning environment. In *Proceedings of the 30th annual ACM symposium on applied computing.* 214–219.

[68] Filipe Pereira, Elaine Oliveira, David Fernandes, Leandro Silva Galvão de Carvalho, and Hermino Junior. 2019. Otimização e automação da predição precoce do desempenho de alunos que utilizam juízes online: uma abordagem com algoritmo genético. In *Brazilian Symposium on Computers in Education (Simpósio Brasileiro de Informática na Educação-SBIE)*, Vol. 30. 1451.

[69] Filipe D Pereira, Elaine HT Oliveira, David BF Oliveira, Alexandra I Cristea, Leandro SG Carvalho, Samuel C Fonseca, Armando Toda, and Seiji Isotani. 2020. Using learning analytics in the Amazonas: understanding students' behaviour in introductory programming. *British Journal of Educational Technology* (2020).

[70] Filipe D Pereira, Armando Toda, Elaine HT Oliveira, Alexandra I Cristea, Seiji Isotani, Dion Laranjeira, Adriano Almeida, and Jonas Mendonça. 2020. Can we use gamification to predict students' performance? A case study supported by an online judge. In *International Conference on Intelligent Tutoring Systems.* Springer, 259–269.

[71] Daniel H Pink. 2011. *Drive: The surprising truth about what motivates us.* Penguin.

[72] Ana Isabel Polo-Peña, Dolores María Frías-Jamilena, and María Lina Fernández-Ruano. 2020. Influence of gamification on perceived self-efficacy: gender and age moderator effect. *International Journal of Sports Marketing and Sponsorship* (2020).

[73] Yizhou Qian and James Lehman. 2017. Students' Misconceptions and Other Difficulties in Introductory Programming: A Literature Review. *ACM Trans. Comput. Educ.* 18, 1, Article 1 (Oct. 2017), 24 pages. https://doi.org/10.1145/3077618

[74] Katherine A Rawson. 2015. The status of the testing effect for complex materials: Still a winner. *Educational Psychology Review* 27, 2 (2015), 327–331.

[75] Anthony V Robins. 2019. Novice programmers and introductory programming. *The Cambridge Handbook of Computing Education Research, Cambridge Handbooks in Psychology* (2019), 327–376.

[76] Luiz Rodrigues, Robson Bonidia, and Jacques Brancher. 2020. Procedural versus human level generation: Two sides of the same coin? *International Journal of Human-Computer Studies* 141 (2020), 102465.

[77] Luiz Rodrigues, Wilk Oliveira, Armando Toda, Paula Palomino, and Seiji Isotani. 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).*

[78] Luiz Rodrigues, Paula T. Palomino, Armando M. Toda, Ana C. T. Klock, Wilk Oliveira, Anderson P. Avila-Santos, Isabela Gasparini, , and Seiji Isotani. 2021. Personalization Improves Gamification: Evidence from a Mixed-methods Study. *Proc. ACM Hum.-Comput. Interact.* 5, CHI PLAY, Article 287 (sep 2021), 25 pages. https://doi.org/10.1145/3474714

[79] Luiz Rodrigues, Armando M Toda, Wilk Oliveira, Paula T Palomino, Anderson Paulo Avila-Santos, and Seiji Isotani. 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*. 184–190.

[80] Luiz Rodrigues, Armando M Toda, Wilk Oliveira, Paula T Palomino, and Seiji Isotani. 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*. SBC, 461–470.

[81] Luiz Rodrigues, Armando M. Toda, Wilk Oliveira, Paula T. Palomino, Julita Vassileva, and Seiji Isotani. 2021. Automating Gamification Personalization: To the User and Beyond. arXiv:cs.HC/2101.05718

[82] Luiz Rodrigues, Armando M Toda, Paula T Palomino, Wilk Oliveira, and Seiji Isotani. 2020. Personalized gamification: A literature review of outcomes, experiments, and approaches. In *Eighth International Conference on Technological Ecosystems for Enhancing Multiculturality*. 699–706.

[83] Henry L Roediger-III and Jeffrey D Karpicke. 2006. Test-enhanced learning: Taking memory tests improves long-term retention. *Psychological science* 17, 3 (2006), 249–255.

[84] Christopher A Rowland. 2014. The effect of testing versus restudy on retention: a meta-analytic review of the testing effect. *Psychological Bulletin* 140, 6 (2014), 1432.

[85] Michael Sailer and Lisa Homner. 2019. The Gamification of Learning: a Meta-analysis. *Educational Psychology Review* (15 Aug 2019). https://doi.org/10.1007/s10648-019-09498-w

[86] Diana R. Sanchez, Markus Langer, and Rupinder Kaur. 2020. Gamification in the classroom: Examining the impact of gamified quizzes on student learning. *Computers & Education* 144 (2020), 103666. https://doi.org/10.1016/j.compedu.2019.103666

[87] Bianca L Santana and Roberto A Bittencourt. 2018. Increasing motivation of cs1 non-majors through an approach contextualized by games and media. In *2018 IEEE Frontiers in Education Conference (FIE)*. IEEE, 1–9.

[88] Isabelle Savard and Riichiro Mizoguchi. 2019. Context or culture: what is the difference? *Research and Practice in Technology Enhanced Learning* 14, 1 (2019), 1–12.

[89] Sofia Schöbel and Matthias Söllner. 2016. How to Gamify Information Systems - Adapting Gamification to Individual User Preferences. In *24th European Conference on Information Systems*.

[90] Katie Seaborn and Deborah I Fels. 2015. Gamification in theory and action: A survey. *International Journal of human-computer studies* 74 (2015), 14–31.

[91] Geneviève Taylor, Tomas Jungert, Geneviève A Mageau, Kaspar Schattke, Helena Dedic, Steven Rosenfield, and Richard Koestner. 2014. A self-determination theory approach to predicting school achievement over time: The unique role of intrinsic motivation. *Contemporary Educational Psychology* 39, 4 (2014), 342–358.

[92] Armando Toda, Filipe Dwan Pereira, Ana Carolina Tomé Klock, Luiz Rodrigues, Paula Palomino, Wilk Oliveira, Elaine Harada Teixeira Oliveira, Isabela Gasparini, Alexandra Ioana Cristea, and Seiji Isotani. 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*. SBC, 471–480.

[93] Armando M Toda, Ana CT Klock, Wilk Oliveira, Paula T Palomino, Luiz Rodrigues, Lei Shi, Ig Bittencourt, Isabela Gasparini, Seiji Isotani, and Alexandra I Cristea. 2019. Analysing gamification elements in educational environments using an existing Gamification taxonomy. *Smart Learning Environments* 6, 1 (2019), 16.

[94] Armando M. Toda, Pedro H. D. Valle, and Seiji Isotani. 2018. The Dark Side of Gamification: An Overview of Negative Effects of Gamification in Education. In *Higher Education for All. From Challenges to Novel Technology-Enhanced Solutions*, Alexandra Ioana Cristea, Ig Ibert Bittencourt, and Fernanda Lima (Eds.). Springer International Publishing, Cham, 143–156.

[95] Gustavo F Tondello, Alberto Mora, and Lennart E Nacke. 2017. Elements of gameful design emerging from user preferences. In *Proceedings of the Annual Symposium on Computer-Human Interaction in Play*. ACM, 129–142.

[96] Gustavo F Tondello and Lennart E Nacke. 2020. Validation of User Preferences and Effects of Personalized Gamification on Task Performance. *Frontiers in Computer Science* 2 (2020), 29.

[97] Gustavo F Tondello, Hardy Premsukh, and Lennart Nacke. 2018. A theory of gamification principles through goal-setting theory. Hawaii International Conference on System Sciences.

[98] Christiane Gresse von Wangenheim and Aldo von Wangenheim. 2012. Ensinando computação com jogos. *Bookess Editora, Florianópolis, SC, Brasil* (2012).

[99] Szymon Wasik, Maciej Antczak, Jan Badura, Artur Laskowski, and Tomasz Sternal. 2018. A survey on online judge systems and their applications. *ACM Computing Surveys (CSUR)* 51, 1 (2018), 1–34.

[100] Robert Wellington. 2015. Context to Culture for Gamification HCI Requirements: Familiarity and Enculturement. In *Gamification in Education and Business*. Springer, 151–163.

[101] Zamzami Zainuddin, Samuel Kai Wah Chu, Muhammad Shujahat, and Corinne Jacqueline Perera. 2020. The impact of gamification on learning and instruction: A systematic review of empirical evidence. *Educational Research Review* (2020), 100326.

[102] Amy L Zeldin, Shari L Britner, and Frank Pajares. 2008. A comparative study of the self-efficacy beliefs of successful men and women in mathematics, science, and technology careers. *Journal of Research in Science Teaching: The Official*

Journal of the National Association for Research in Science Teaching 45, 9 (2008), 1036–1058.

[103] Wayne Xin Zhao, Wenhui Zhang, Yulan He, Xing Xie, and Ji-Rong Wen. 2018. Automatically learning topics and difficulty levels of problems in online judge systems. *ACM Transactions on Information Systems (TOIS)* 36, 3 (2018), 1–33.