Gamification Works, but How and to Whom? An Experimental Study in the Context of Programming Lessons

Luiz Rodrigues
lalrodrigues@usp.com
University of São Paulo, Brazil

Armando M. Toda
armando.toda@usp.br
University of São Paulo, Brazil

Wilk Oliveira
wilk.oliveira@usp.br
University of São Paulo, Brazil

Paula T. Palomino
paulatpalomino@usp.br
University of São Paulo, Brazil

Anderson Paulo Avila-Santos
anderson.avila@usp.br
University of São Paulo, Brazil

Seiji Isotani
sisotani@icmc.usp.br
University of São Paulo, Brazil

ABSTRACT

Programming is a complex, not trivial to learn and teach task, which gamification can facilitate. However, how gamification affects learning and the influence of context-related aspects on that effect demand research to better understand how and to whom gamification enhances programming learning. Therefore, we conducted an experimental study analyzing how gamification worked and the role of context-related aspects in terms of intervention duration and learners' familiarity with programming (i.e., the task's topic). It was a six-week study with 19 undergraduate students from an *Algorithms* class that measured their learning gains, intrinsic motivation, and number of completed quizzes. Mainly, we found gamification affected learning via intrinsic motivation, effect that depended on intervention duration and learners' familiarity with programming. That is, intrinsic motivation strongly predicted learning gains and gamification's effect on intrinsic motivation changed over time, decreasing from positive to negative as learners had less familiarity with programming. Thus, showing gamification can positively impact programming learning by improving students' intrinsic motivation, although that effect changes over time depending on one's previous familiarity with programming.

CCS CONCEPTS

• **Human-centered computing → Empirical studies in HCI**; • **Applied computing → Interactive learning environments**.

KEYWORDS

gamification, learning, motivation, moderators, longitudinal

ACM Reference Format:
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 (SIGCSE '21), March 13–20, 2021, Virtual Event, USA.* ACM, New York, NY, USA, 7 pages. https://doi.org/10.1145/3408877.3432419

SIGCSE '21, March 13–20, 2021, Virtual Event, USA

ACM ISBN 978-1-4503-8062-1/21/03…$15.00
https://doi.org/10.1145/3408877.3432419

1 INTRODUCTION

Learning to program is challenging. Students present difficulties in syntactic, conceptual, and strategic knowledge [28], and often lack motivation to learn, leading to low grades and high rates of dropout [20, 24]. To address motivational concerns aiming to improve learning, recent research started to explore gamified learning [2, 25] (i.e., adding game elements to change the learning process [19]), which is associated to overall positive effects on learning outcomes [34], including applications to Computer Science (CS) [6].

However, studies on CS applications often focus on gamification's impact on behavioral learning outcomes, such as student performance (e.g., [5, 8, 20, 22]), neglecting learners' motivation and context of use, despite those are inherently connected to learning [1, 36, 38, 42]. Thus, there is a need to understand how gamification affects programming learning while also considering motivation, as well as the role of context, corroborating recent calls for better understanding how gamification works and the context's role aiming to improve gamification's positive outcomes [15, 31, 35].

To advance this understanding, we conducted an experimental study. The experiment context is a Brazilian undergraduate *Algorithms* class focused on programming lessons. Students were randomly assigned to complete quizzes in one of two Moodle versions (gamified | non-gamified) during half semester. Then, we analyzed how gamification affected programming learning based on cognitive (learning gains), motivational (intrinsic motivation), and behavioral (number of tasks completed) learning outcomes [34]; the last two were weekly measured. Moreover, as context involves someone, with their mental state, performing an activity in a given environment for a given period (e.g. a learner, with their current knowledge, completing quizzes in a gamified system) [36], we analyzed its role based on intervention duration and students' familiarity with general course topics. Accordingly, the purpose of this paper was to answer *how does using gamification for half semester affect Brazilian undergraduates programming learning?*

We found gamification affected programming learning by influencing learners' intrinsic motivation. Additionally, we found intervention duration and familiarity with programming together moderated gamification's effect: at the intervention beginning, it was positive but then decreased for students with little familiarity with programming; whereas it started negative and then became positive for those with high familiarity with programming. Thus, we contribute with empirical evidence revealing through which construct gamification affected programming learning and when/to whom

that effect was positive or negative, besides a theory-grounded gamification design likely to contribute to undergraduates' learning depending on their familiarity with programming and usage time.

2 BACKGROUND AND RELATED WORK

Often, students consider learning to program difficult and of low motivation and, thus, perform poorly [1, 24]. A possible explanation for this phenomenon is found in Self-determination theory (SDT) [3], a well-know theory often related to learning that considers three motivation types: intrinsic (i.e., a desire/interest that comes from within), external (i.e., desire/interest in the rewards/outcomes), and amotivation (i.e., no motivation). The literature suggests the first type is ideal for learning contexts [42], and that intrinsically motivated learners are more engaged and retain information better [3]. Accordingly, evidence that learners with higher intrinsic motivation levels achieve higher exam scores has been found (e.g., [11]). Hence, suggesting intrinsically motivated people will achieve better learning than amotivated and extrinsically ones.

Gamification might aid with the motivation issue as it is often concerned with improving it [43]. However, between gamification impacting motivation, there are the moderator factors: those that increase/decrease - or pre-determinate - gamification's effect, possibly even changing it from positive to negative [19]. Studies have investigated moderators related to users' profiles (e.g., [26, 29, 37]), whereas more attention to the intervention duration and context is needed [10, 14, 31, 35]. Here, we consider the context involves users performing some action (e.g., activity) in a given environment [36]. Accordingly, we might expect that users' familiarity with topics related to actions will affect how they perceive the environment.

Moreover, by affecting motivation, gamification can help in influencing learners' behaviors [23]. Studies have shown that completing quizzes (self-testing) also relates to higher exam scores (e.g., [4, 35]). This relationship relies on the *testing effect*, which has been supported in numerous experimental settings [33]. Thereby, demonstrating gamification's potential to indirectly affect learners' behavior is of value to their learning.

In light of this context, we review empirical research on the effects of gamification on learning outcomes within the context of programming concepts. For studies selection, we analyzed two recent secondary studies [2, 34] as they were published less than one year before our time of writing and map empirical gamification research found in a broad range of databases.

Hakulinen et al. [8] evaluated the effect of badges on behavioral learning outcomes. Based on data from 281 students from a *Data Structures and Algorithms* course (around eight weeks long), both positive (e.g., badges earned and time in the system) and null (e.g., completed exercises) results were found when log data from learners' interactions with the gamified educational system were compared to log data from those who interacted with the non-gamified version of the same system.

Krause et al. [17] compared the impacts of gamification and social gamification to a non-gamified condition in the context of a course for learning Python as a statistical analysis tool (four weeks long). Considering retention, in-system quizzes' performance (n=206), and a post-test score (n=101), they found gamification versions overcame the control condition in all three measures; social gamification overcame the simpler gamification version. No moderator effect of age neither sex was found.

Fotaris et al. [5] gamified a Python programming course (12 weeks long) using Kahoot! and Codeacademy. They longitudinally compared attendance, late arrivals, and number of material downloads. Also, they compared the students' academic performance in this course to that of a non-gamified version of the same course, offered in the previous semester. Overall results were positive in favor of gamification, but only descriptive analyses were performed.

Moreno and Pineda [22] analyzed the impact of using a gamified educational system featuring automatic code judging compared to traditional workshops. Participants (n=43) were split into two groups (one for each condition) and had their learning compared in terms of performance on programming tests (e.g., conditionals and loops). Findings show those who used the gamified system achieved higher scores compared to the remaining, suggesting the benefits of the system as no performance difference was found between groups in the pre-test. However, it is unclear whether the effect emerged from gamification or other system features because gamification was not the only difference between conditions.

Marin et al. [20] analyzed data from two semesters (n=817) of a C programming course (four weeks long) to assess gamification's impact on students' performances. Despite they measured learning performance from two exams, the first one was administered after the intervention began (middle semester), not characterizing a pre-test. Overall results are positive for gamification, but the decrease from exam one to exam two was similar in both semesters and the change from one to another was not considered.

Table 1 summarizes this paper and main related works' characteristics, allowing the identification of the gaps this study faces. First, not using pre-tests, which opens the possibility of not acknowledging when one group has, for instance, an initial motivation higher than the other, which might mislead conclusions. Additionally, using pre-tests has been acknowledged as a characteristic needed for gamification studies to present high methodological rigor [34]. To address this issue, our experiment employed pre-tests right before the intervention begin. Second, the lack of longitudinal studies, which disables the possibility of understanding gamification's effects over time. To tackle this lack, we studied learners' psychological states and behavior at each experiment's week.

Table 1: Related research characterization.

Ref.	ECG	PT	LA	IA	MA	ID	SC	TB	LO
[8]	Y	N	N	Y	N	8	Y	N	B
[17]	Y	N	N	Y	Y	4	Y	N	B,C
[5]	Y	N	Y	N	N	12	N	N	B
[22]	N	Y	N	Y	N	?	Y	N	B
[20]	Y	N	N	Y	N	16	N	N	B
This	Y	Y	Y	Y	Y	6	Y	Y	M, B, C

ECG = equivalent control group; PT = pre-test; LA = longitudinal analysis; IA = inferential analysis; MA = moderation analysis; ID = intervention duration in weeks; SC = same class; TB = theoretical background; LO = learning outcomes; Y = yes; N = no; ? = undefined; B = behavioral; C = cognitive; M = motivational.

Third, the little attention to moderators that, otherwise, would advance the understanding of when/to whom gamification is more or less suited. To expand this understanding, we analyzed the impact of possible moderators that have been suggested in the literature, such as intervention time and contextual characteristics [14, 30, 35]. Fourth, not grounding gamification designs on learning-related theories neither exploring motivational or cognitive learning outcomes, which would better explain the process through which gamification affects learning, rather than just showing if it does. We approached this need grounding the gamification design on SDT, one of the most used in gamification studies and considered relevant to learning [11, 42, 43], as well as evaluated cognitive, behavioral and motivational learning outcomes.

Given this context, we tested the following hypotheses to answer our research question: (**H1**) Intrinsic motivation and completing quizzes positively affect learning gains; (**H2**) Gamification improves intrinsic motivation, effect moderated by intervention duration and users' familiarity with class' topics; and (**H3**) The more the learners' intrinsic motivation, the more quizzes they complete.

3 EXPERIMENT

To achieve the goal of understanding *how using gamification for half semester affects Brazilian undergraduates programming learning*, this experiment was performed in an undergraduate Software Engineering course of a private institution in Londrina - Brazil, with approval of the university's ethical committee. We conducted the experiment from March to June, 2020, on the *Algorithms* discipline, which explores pseudo-language to introduce first-term students to programming. The discipline instructor is male, holds a MsC. degree in CS, and had taught for six years at that time. Topics taught during the experiment included conditionals, loops, and arrays (initialization and manipulation). Our participants provided informed consent and represent 68% of those initially enrolled in the discipline. Inclusion criteria was being enrolled in the discipline and completing pre- and post-tests. From the 28 possible participants (all males), 19 met the criteria: all males with an average age of 20.32 years (±3.64).

The materials related to this experiment are the educational system, experimental task, gamification design, measures, and moderators. The educational system used to accomplish the experiment's tasks was Moodle because it is the standard educational system in the university, enabling the intervention to be within the environment students and course instructor regularly use.

The task participants had to accomplish was completing quizzes, an optional task that added extra points in the course. We explored extra activities to reduce the influence on the original course program, and along with the instructor, defined quizzes would give extra points to encourage students to engage with the tasks. Each quiz featured from three to five items, and six new quizzes were provided each week. The rationale for six quizzes was that students could complete one quiz per non-class day, aiming not to overcharge them with extra work. To align quizzes with the course design, we followed the revision of the cognitive process dimensions of Bloom's Taxonomy, since it is a renowned structure to support the development of learning outcomes [16]. Then, at each week, quizzes complexity increased according to those dimensions. That is, quizzes' highest dimension was *remember* in the first week, *understand* in the second, and so on. Multiple knowledge dimensions were intentionally explored within each week.

For gamification design, we implemented gamification heuristics focused on SDT, aiming to affect intrinsic motivation. Next, we introduce the heuristics, taken verbatim from [40], and how we implemented them.

#1 Avoid obligatory uses: Completing quizzes was optional and students could solve the week quizzes in their preferred order.

#2 Provide a moderate amount of meaningful options: Quizzes were aligned to each week's class topic, based on the instructor's schedule, offering six of those per week (i.e., learners could do one per day, excluding the class day).

#3 Set challenging but manageable goals: The discipline instructor revised all quizzes' items and provided feedback for adapting them when necessary.

#4 Provide positive, competence-related feedback: We added weekly, unannounced badges that acknowledged students according to the cognitive dimension of the quiz.

#5 Facilitate social interaction: Two out of the six weekly quizzes were team/group activities. In those, students could see peers' answers after completing the quizzes, analyze and discuss the answers, and update their owns.

#6 When supporting a particular psychological need, wary to not thwart the other needs: We i) used Cooperation to support relationships feelings, rather than Competition, which could make users feel incompetent, and ii) offered unannounced badges to prevent, for instance, feelings of *needing* to receive it (e.g., anxiety).

#7 Align gamification with the goal of the activity in question: We transformed quizzes into missions that encouraged learners to complete the quizzes.

#8 Create a need-supporting context: Implemented by attending user needs, autonomy, competence, and relatedness, through heuristics #1, #4, and #5, respectively.

#9 Make the system flexible: Not implemented because personalizing/adapting gamification is a recent, open research field [9, 32].

Nevertheless, the implementation of many of those heuristics were available in the regular Moodle version as well. As not all aspects are directly related to adding game elements to change the learning process (i.e., gamifying learning), we intentionally allowed the non-gamified Moodle version to feature them. Compared to the regular version, the gamified one adds unannounced badges and a more gameful experience by presenting quizzes as graphic-enriched missions, similar to [41], and working groups as teams, resembling a game rather than a regular group learning activity. The missions and badges used can be seen at: shorturl.at/gkrDN.

As measures, we captured cognitive, behavioral, and motivational learning outcomes [34]. To measure cognitive learning outcomes (i.e., learning gains), we designed a test featuring 15 multiple choice items to assess *remembering* and *understanding* domain processes on three programming topics: conditionals, loops, and arrays (five items per topic). The test was revised by the discipline instructor, being considered a suitable formative evaluation instrument. The behavioral outcome was operationalized as the number of completed quizzes, following previous similar research (e.g., [35]), which Moodle automatically collected. Intrinsic motivation (motivational outcome) was measured with Portuguese version of the

interest/enjoyment sub-scale of the Intrinsic Motivation Inventory as validated in [27].

As moderators, we captured intervention time and contextual characteristics. For intervention time, we considered the experiment week (*week*; 0 for the pre-test, 6 for the last week). For contextual characteristics, we captured learners' self-reports of their familiarity with topics related to the course in five-point Likert-scales: familiarity with algorithms (FAlg), C programming language (FC), programming (FProg), and pseudo-language (FPseu). These aspects concern the context because they are directly related to the action (activity) users performed in the experiment [36].

As experimental design, we employed a 2x6 mixed factorial design with random assignment to the between-subject independent variable Condition: control (i.e., used non-gamified Moodle; N = 10) and experimental (i.e., used gamified Moodle; N = 9). The within-subject independent variable concerned six repeated measures of motivational and behavioral data collected weekly. Moodle automatically collected the former, whereas the professor asked students to complete the measures of the latter during the weekly classes.

The data collection procedure followed three steps. First, pre-tests were administered to serve as baseline comparisons for cognitive and motivational learning outcomes. Then, the intervention started, which lasted for six weeks. During this phase, participants were offered a new set of extra activities related to the class' topics each week. Activities were the same for all participants, with the only difference being the Moodle version to be used. Lastly, after the sixth experiment week, the post-test was administered, generating the learning gain measure (post - pre). Completing the motivation scales and quizzes was optional. Consequently, some participants completed no quiz as well as the number of motivational measures completed varies per week. The number of completions of control and experimental groups is, respectively: W0, 7 and 7; W1, 6 and 6; W2, 3 and 6; W3, 7 and 7; W4, 6 and 3; W5, 6 AND 3; W6, 9 and 5. Completions reliability was good in all weeks ($\alpha > 0.8$) . Summarizing, we captured cognitive (learning gains; pre- and post-tests), behavioral (Moodle's logs; one per week), and motivational (self-reports; pre-test plus one per week) learning outcomes.

For data analysis, we explored regression methods. We used multiple linear regression to test **H1**, as this approach enables understanding whether and how various independent variables predict a dependent variable [7]. In **H1** case, intrinsic motivation and completed quizzes (independent) and learning gains (dependent) are the variables. Learning gains were measured based on the difference between post- and test-tests. Consequently, there is one measure per participant. However, this analysis' independent variables were repeatedly measured. Therefore, we aggregated them (average and sum), creating one measure of each variable per participant. Furthermore, we excluded outliers based on standard deviation ($|x| > 2 * SD$; two SD due to the small sample size, i.e., 19), after Shapiro-wilk tests suggested both independent variables follow a normal distribution, as regression analyses are sensitive to outliers [18]; a single item was excluded.

Testing **H2** involves dependent data (i.e., multiple answers from the same participants), which violates the independence assumption of classical regression methods [7]. Multilevel models are regression-based models that properly account for dependent data, and can be seen as a hierarchical system of regression equations (one for each level/group) [12]. This is achieved by allowing each group (e.g., of subjects) to have its own intercept and slope coefficients, which are often referred to as *random* coefficients. Additionally, these models contain *fixed* coefficients, which do not vary across groups. By doing so, multilevel methods model the groups' variance as well as find estimates applicable to the whole sample. Measuring random coefficients, however, requires sample sizes larger than that of this study. Therefore, we focus on analyzing fixed effects, which are reproducible properties of the overall data [21]. Moreover, compared to repeated-measures ANOVA, multilevel analysis has more power and handles data with varied numbers of answers per subject, besides accounting for dependencies [7]. Therefore, we test **H2** using multilevel analysis.

To test **H2**, we followed guidelines [7, 21] for properly identifying and defining multilevel models. The recommended approach to evaluate the relevance of a specific parameter (independent variable; e.g., Condition) is to test whether the parameter significantly increases model fit compared to the model without it. In the context of multilevel analyses, this is often accomplished through the Likelihood Ratio Test (LRT). Furthermore, multilevel models might be developed from bottom-up (starts simple and increasingly complexify) or top-down (starts complex and removes irrelevant parameters). The former requires multiple steps, whereas the latter allows creating a model with all parameters to be tested and, then, remove those that do not decrease model fit [7, 21].

We tested **H2** following the top-down approach to reduce the number of tests in model development. Accordingly, we defined a model that accounts for relationships between our dataset's three-level hierarchy: repeated measures (first; e.g., intrinsic motivation and experiment week), nested within students (second; e.g., FAlg) nested within a condition (third; control or experimental). That is, a model accounting for interactions between repeated measures-level and both student- and condition-level variables, as well as interactions from student- to condition-level variables. Centering variables is recommended for models with interactions [7], therefore, with did so to all independent variables before fitting the model.

To determine relevant variables, we removed each of the three-level interactions (e.g., week, FAlg and gamification) at a time, testing if some removal significantly changed model fit based on the LRT. We only tested removing the three-way interactions because terms involved in a significant interaction should be kept in the model even when the term itself is nonsignificant [7]. **H3** is similar to **H2** (involves dependent data), but no moderators were involved in this step. Therefore, the testing procedure was similar, but we only compared whether removing the single independent variable significantly decreased model fit. We adopted a more lenient alpha level (0.1) for all LRTs due to the exploratory testing of moderators of gamification's effects [12]. P-values were adjusted with the False Discovery Rate approach due to multiple comparisons [13].

4 RESULTS

H1 predicted intrinsic motivation and the number of completed quizzes would positively affect participants' learning gains. Overall regression results (N = 18; $R^2 = 0.76$; R^2-adj = 0.73; F(2,15) = 23.61; $p < 0.01$) and individual predictors (intrinsic motivation average: B = 1.24; SE = 0.24; $\beta = 0.66$; $p < 0.01$; sum of completed quizzes:

B = 0.07; SE = 0.02; β = 0.60; p < 0.01) were significant. Thus, suggesting strong positive effects of the predictors on learning gains, supporting **H1**.

H2 concerns gamification affecting intrinsic motivation while being moderated by context-related factors. Results from the LRTs (Table 2) demonstrate the only term to significantly affect model fit is the three-way interaction between gamification, week, and FAlg. Therefore, we followed literature recommendation [12] and fitted a new model adding only the variables and interactions involved in that significant term (Table 3). Figure 1 helps understanding the model. It shows how the intrinsic motivation (Y-axis) of those with (right) and without (left) gamification changed over time (X-axis) depending on their previous FAlg. The figure shows intrinsic motivation changed over time for all participants, and that gamification's impact was mixed, increasing from negative to positive inasmuch learners had more familiarity with algorithms at the beginning of the experiment. Therefore, suggesting that gamification affected intrinsic motivation and that this impact was moderated by intervention time and FAlg but not by other contextual factors analyzed. Thus, partially supporting **H2**.

H3 predicted intrinsic motivation would positively affect the number of completed quizzes. The LRT showed removing intrinsic motivation insignificantly changes the model fit (F(1, 67.425) = 2.01; p = 0.16), indicating this model is no better than an intercept-only one. Thus, we have no evidence intrinsic motivation affected the number of completed quizzes, failing to support **H3**.

In summary, our findings indicate that students who completed more quizzes and were more intrinsically motivated achieved higher learning gains (**H1**), that completing quizzes was unlikely driven by intrinsic motivation (**H3**), and that gamification's effects depended on intervention duration and learners' previous familiarity with algorithms (**H2**).

5 DISCUSSION

This section discusses our results from four perspectives. First, our research question. In terms of how gamification affected Brazilian undergraduates programming learning in a six-week period, we found that was accomplished by influencing their motivation. That is, gamification influenced learners' motivation (**H2**) that, in turn, positively predicted learning gains (**H1**). We also predicted gamification would indirectly affect learning via users' behaviors, which would be affected by intrinsic motivation (**H3**), however, we found no support for that. Furthermore, **H2** also predicted intervention duration and context-related factors would moderate gamification's effect on intrinsic motivation. Intervention duration and only one (FAlg) out of the four context-related factors were significant moderators. Besides changing the effect's strength, these moderators affected its direction: over time, gamification was positive when participants had at least some familiarity to algorithms but negative otherwise. Hence, the way gamification affected participants learning was through intrinsic motivation, positively or negatively affecting it over time according to learners' previous FAlg.

The second perspective is findings' relationship to previous work. Most studies applying gamification in programming education reported positive outcomes [17, 20, 22], with few null results [5, 8]. Differently, our results were mixed. A possible rationale is that most reviewed studies focused on analyzing behavioral outcomes, whereas we analyzed motivational and cognitive ones as well. Moreover, despite gamification's overall effect is positive, multiple factors moderate its success [34], likely leading to cases where results are null/negative [39]. As we analyzed moderators, we were able to

Table 2: Likelihood ratio tests assessing significant terms when modeling gamification's effect on intrinsic motivation while accounting for moderators. *gamified* dummy coded.

Terms	F (df_n, df_d)	p	p-adj
gamified:week:FAlg	7.669 (1, 59.920)	0.007	0.090
gamified:week:FC	3.236 (1, 64.686)	0.077	0.307
gamified:week:FPseu	2.161 (1, 63.285)	0.146	0.352
gamified:week:FPprog	0.001 (1, 64.439)	0.973	0.973
gamified:FAlg	2.603 (1, 15.450)	0.127	0.352
gamified:FC	0.778 (1, 17.923)	0.389	0.673
gamified:FPseu	0.779 (1, 13.652)	0.393	0.673
gamified:FProg	0.010 (1, 15.638)	0.923	0.973
week:FAlg	0.316 (1, 59.810)	0.576	0.739
week:FC	0.285 (1, 63.364)	0.595	0.739
week:FPseu	3.873 (1, 60.173)	0.054	0.307
week:FProg	0.255 (1, 59.716)	0.616	0.739

F = familiarity to; Alg = algorithms; Prog = programming; Pseu = Pseudo-language; C = C programming language.

Table 3: Multilevel model of gamification's effect on intrinsic motivation controlling for intervention duration (week) and learners' previous familiarity with algorithms (FAlg; Likert-scale). *gamified* dummy coded; * p < 0.1.

Coefficient	Est. (SE)	Coefficient	Est. (SE)
Intercept	5.64 (0.90)	gamified:week	-0.32 (0.21)
gamified	0.41 (1.24)	gamified:FAlg	-0.26 (0.47)
week	0.06 (0.15)	week:FAlg	-0.04 (0.06)
FAlg	-0.02 (0.35)	gamified:week:FAlg	0.17 (0.08)*

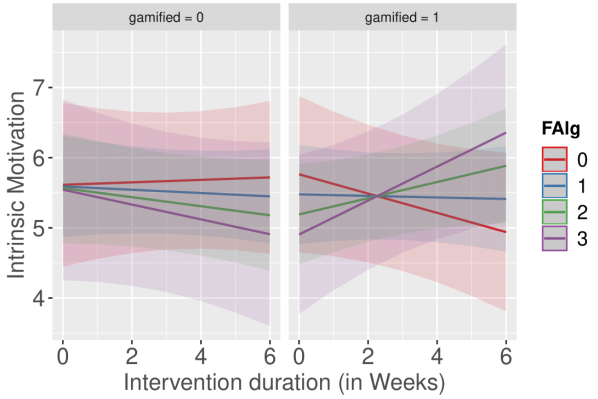

Figure 1: Effects of gamification on intrinsic motivation (Y-axis), moderated by intervention duration (X-axis) and previous familiarity with algorithms (FAlg; Likert-scale).

understand which factors affected gamification's impact, whereas a single related study performed a similar analysis [17]. Based on this context, our findings corroborate the overall gamification literature by achieving mixed outcomes [15], provide evidence for research claiming the moderator effect of context and intervention duration [10, 14, 31], and suggest the predominance of positive reports might be due to not considering moderators' impact.

The third perspective concerns our findings' implications to programming teaching. From testing our hypotheses, we found evidence towards two main directions, which poses two implications for programming learning. First, we found intrinsic motivation and the number of competed quizzes positively predicted learning gains (**H1**), further grounding previous discussions (e.g., [33, 42]) with evidence in the context of programming learning (conditionals, loops, and arrays). Therefore, the implication is instructors should seek to improve this psychological state of learners, as well as having them self-tested (e.g., completing quizzes), as this likely enhances their learning. Second, we found intervention duration moderated the impact of gamification, which was positive for those with previous familiarity to the task's topic (another moderator) but null or negative for those with no previous familiarity (**H2**). Therefore, the implication is that the decision to gamify an educational system must be made with caution, considering not only users' demographics and profiles (c.f. [23]), but also for how long it will be used as well as users' previous familiarity with the system's topics.

Lastly, we discuss three implications from our findings to future research. The first concerns results from **H1**. Despite those corroborate previous research [11, 35], further research are still needed to ground whether intrinsic motivation and the testing effect (e.g., completing quizzes) hold within the context of programming learning. The second concerns results from **H2**. The fact that the same gamification is unlikely to work for all users has been recently discussed, calling for the need of tailored gamification [40]. Within this context, recent studies have called for considering aspects related to the context and learning activities when tailoring gamification [9, 10, 30, 32, 38]. Our findings' implication to this vein is that learners' previous familiarity with the learning activity, which relates to the context, moderates how gamification impacts their intrinsic motivation. This implies research on tailored gamification should investigate these factors as tailoring gamification to such familiarity might be crucial to improve its effectiveness. The third implication relates to intrinsic motivation not driving users' behavior, unlike our expectations (**H3**). In this research, completing quizzes worth extra points (external rewards), then, students possibly were motivated to complete them due to extrinsic motivation [3]. The implication for future research, therefore, is that studies should seek to motivate learners participation through intrinsic rather than extrinsic approaches.

5.1 Limitations and Threats to Validity

First, our sample is restricted in size (19), limiting our findings' generalization. We opted for this approach to perform the study within a real class, which is costly and hard to perform with large samples. Additionally, the sample concerns a single class, which might lead to groups' contamination (information from one group leaking to another). We chose to study a single class to increase the study internal validity, guaranteeing all participants would learn from the same instructor and lessons, increasing the chances that differences are due to gamification. Second, there were missing data (38%) in the motivational outcomes, possibly because completing the scales did not worth extra points, unlike completing quizzes. Multilevel analysis handle such missings on the dependent variable, but in independent variables the common approach is deletion [12]. Hence, threatening our conclusion validity only with regards to **H3**. Moreover, the limited sample size also impacted the data analyses (e.g., low statistical power, possibly inaccurate estimates). To address this, we focused the analyses on fixed effects, as they can be estimated with smaller samples than random effects [12]. Third, there is the learning gains measuring. We opted for measuring this construct through pre- and post-tests, as recommended in the literature [34], in which we used the same test in both occasions. As there was a 42 days interval between the tests, we believe threats related to memorizing test's items were mitigated. Additionally, the test was validated by the discipline instructor, guaranteeing it was aligned to and measured the topics approached during the experiment. Fourth, there was no control over how/where/when participants completed the experiment's task. This approach increases external validity, as this freedom is similar to when students are given homework or optional tasks. However, as participants are likely to have distinct routines and livings, which might have affected our results. Fifth, the intervention only lasted for half semester due to restrictions from the university and the effort needed by the discipline instructor. Although our results suggest how gamification's impact would change in a longer intervention, this can only be ensured with more research. Lastly, during the intervention, the class changed from face-to-face to online due to covid-19 quarantine. The between-subject design mitigated this threat as participants continued using the same condition regardless of the change.

6 CONCLUSIONS

Students often lack motivation to learn to program, jeopardizing their learning. Gamification can aid with this issue, but the understanding of how it contributes to programming learning, and which aspects moderate that contribution, is scarce. Thus, we conducted a longitudinal experiment with Brazilian undergraduate students, comparing the motivation, behavior, and learning gains of those interacting with gamified quizzes to those engaged with non-gamified ones. Mainly, we found that gamification contributed to students' programming learning via intrinsic motivation and that this effect changed as intervention duration time increased, decreasing from positive to negative inasmuch learners had less familiarity with programming.

In summary, our main contributions are i) empirical evidence revealing through which construct gamification affected programming learning, ii) when/to whom that effect was positive or negative, and iii) a theory-grounded gamification design likely to improve the learning of undergraduates with previous familiarity to programming after a six-week use. As future works, we recommend conducting similar experiments with different samples to ground our findings, developing/testing other gamification designs aiming to mitigate cases of negative effects, and advancing the understanding of moderators of gamification's success.

ACKNOWLEDGMENTS

This research was partially funded by CNPq, CAPES, and FAPESP (Projects 2016/02765-2; 2018/11180-3; 2018/15917-0; 2018/07688-1).

REFERENCES

[1] Raad A Alturki et al. 2016. Measuring and improving student performance in an introductory programming course. *Informatics in Education-An International Journal* 15, 2 (2016), 183–204.

[2] 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.

[3] Edward L. Deci and Richard M. Ryan. 2000. The "What" and "Why" of Goal Pursuits: Human Needs and the Self-Determination of Behavior. *Psychological Inquiry* 11, 4 (2000), 227–268. https://doi.org/10.1207/S15327965PLI1104_01 arXiv:https://doi.org/10.1207/S15327965PLI1104_01

[4] 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* (Montreal QC, Canada) *(CHI '18)*. Association for Computing Machinery, New York, NY, USA, 1–13. https://doi.org/10.1145/3173574.3173885

[5] 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.

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

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

[8] 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).

[9] Stuart Hallifax, Audrey Serna, Jean-Charles Marty, and Élise Lavoué. 2019. Adaptive Gamification in Education: A Literature Review of Current Trends and Developments. In *European Conference on Technology Enhanced Learning*. 294–307.

[10] 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* (Barcelona, Spain) *(CHI PLAY '19)*. 559–572. https://doi.org/10.1145/3311350.3347167

[11] Michael D. Hanus and Jesse Fox. 2015. Assessing the effects of gamification in the classroom: A longitudinal study on intrinsic motivation, social comparison, satisfaction, effort, and academic performance. *Computers & Education* 80 (2015), 152 – 161. https://doi.org/10.1016/j.compedu.2014.08.019

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

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

[14] 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

[15] 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

[16] David R Krathwohl. 2002. A revision of Bloom's taxonomy: An overview. *Theory into practice* 41, 4 (2002), 212–218. https://doi.org/10.1207/s15430421tip4104_2

[17] 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* (Vancouver, BC, Canada) *(L@S '15)*. 95–102.

[18] Max Kuhn and Kjell Johnson. 2013. *Applied predictive modeling*. Vol. 26. Springer.

[19] 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.

[20] 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

[21] Daniel Mirman. 2016. *Growth curve analysis and visualization using R*. CRC press.

[22] 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).

[23] 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.

[24] IT Chan Mow. 2008. Issues and difficulties in teaching novice computer programming. In *Innovative techniques in instruction technology, e-learning, e-assessment, and education*. Springer, 199–204.

[25] Paula Palomino, Armando Toda, Wilk Oliveira, Luiz Rodrigues, and Seiji Isotani. 2020. From the Lack of Engagement to Motivation: Gamification Strategies to Enhance Users Learning Experiences. In *2020 19th Brazilian Symposium on Computer Games and Digital Entertainment (SBGames) - GranDGamesBR Forum*.

[26] Paula T Palomino, Armando M Toda, Wilk Oliveira, Luiz Rodrigues, and Seiji Isotani. 2019. Gamification journey: A Novel approach for classifying gamer types for gamified educational systems. *2019 18th Brazilian Symposium on Computer Games and Digital Entertainment (SBGames)* (2019), 165–173.

[27] Laís Zagatti Pedro. 2016. *Uso de gamificação em ambientes virtuais de aprendizagem para reduzir o problema da externalização de comportamentos indesejáveis*. Ph.D. Dissertation. Universidade de São Paulo.

[28] 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.

[29] Luiz Rodrigues and Jacques Duílio Brancher. 2019. Playing an Educational Game Featuring Procedural Content Generation: Which Attributes Impact Players' Curiosity? *Journal New Technologies on Education (Revista Novas Tecnologias na Educação - RENOTE)* (2019).

[30] 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)*.

[31] Luiz Rodrigues, Armando Toda, Wilk Oliveira, Paula Palomino, and Seiji Isotani. 2020. Just beat it: Exploring the influences of competition and task-related factors in gamified learning environments. In *Proceedings of the Brazilian Symposium on Computers in Education (Simpósio Brasileiro de Informática na Educação-SBIE)*, Vol. 31.

[32] 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 *Proceedings of the 8th International Conference on Technological Ecosystems for Enhancing Multiculturality (TEEM 2020) (Salamanca, Spain, October 21-23, 2020)*, F. J. García-Peñalvo (Ed.).

[33] 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.

[34] Michael Sailer and Lisa Homner. 2019. The Gamification of Learning: a Meta-analysis. *Educational Psychology Review* (15 Aug 2019).

[35] 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.

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

[37] Armando Toda, Wilk Oliveira, Lei Shi, Ig Ibert Bittencourt, Seiji Isotani, and Alexandra I. Cristea. 2019. Planning gamification strategies based on user characteristics and DM : a gender-based case study. In *Proceedings of the 12th International Conference on Educational Data Mining*., Michel Desmarais, Collin F. Lynch, Agathe Merceron, and Roger Nkambou (Eds.). Educational Data Mining 2019, Montréal, Canada, 438–443. http://dro.dur.ac.uk/28609/

[38] Armando Toda, Filipe Dwan Pereira, Ana Carolina Tomé Klock, Luiz Rodrigues, Paula Palomino, Wilk Oliveira, and Elaine Oliveira. 2020. For whom should we gamify? Insights on the users intentions and context towards gamification in education. In *Proceedings of the Brazilian Symposium on Computers in Education (Simpósio Brasileiro de Informática na Educação-SBIE)*, Vol. 31.

[39] 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.

[40] Rob van Roy and Bieke Zaman. 2017. Why gamification fails in education and how to make it successful: introducing nine gamification heuristics based on self-determination theory. In *Serious Games and edutainment applications*. Springer, 485–509.

[41] Rob Van Roy and Bieke Zaman. 2018. Need-supporting gamification in education: An assessment of motivational effects over time. *Computers & Education* 127 (2018), 283–297.

[42] Maarten Vansteenkiste, Eline Sierens, Bart Soenens, Koen Luyckx, and Willy Lens. 2009. Motivational profiles from a self-determination perspective: The quality of motivation matters. *Journal of educational psychology* 101, 3 (2009), 671.

[43] 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.