Analyzing the Impact of Leaderboards in Introductory Programming Courses' Short-Length Activities

1st Rafael G. de Pontes*, 2nd Kláudio H. M. Medeiros†,
3rd Dalton D. S. Guerrero‡, 4th Jorge C. A. de Figueiredo§
*‡§*Department of Systems and Computing*, *Federal University of Campina Grande*
†*Unifacisa Centro Universitário*
Campina Grande, Brazil
Email: *rafaelpontes@copin.ufcg.edu.br, †klaudiohenrique.medeiros@gmail.com,
‡dalton@computacao.ufcg.edu.br, §abrantes@computacao.ufcg.edu.br

***Abstract*—This Research Full Paper presents results from gamified in-class interventions with leaderboards. Generally, it is assumed that student expertise is directly related to deliberate and repeated practice. However, finding techniques that foster student engagement in problem-solving may not be the easiest of tasks. To that end, gamification, also known as the use of game design elements in non-game contexts, has proved to be effective in educational scenarios. Concerned with stimulating student engagement in Computer Science, this research aimed to observe the effect of the insertion of leaderboards in short-length in-class experiments performed in two different terms of undergraduate IP courses. $N = 84$ students were randomly allocated into one of two types of group: experimental groups with leaderboards; and control groups without knowledge of nor access to leaderboards. Both experiments showed that students in the experimental groups tried and completed more programming problems, on average, than those from their respective control groups. P-values below $\alpha = 5\%$ indicated a real impact from the leaderboard in students' performances. This research may benefit educators responsible for IP courses that have an online platform at their disposal, by introducing a design of leaderboard that incites student engagement, at least for short-length activities.**

***Index Terms*—Gamification, Introductory Programming, Education, Leaderboard**

I. Introduction

Novice students are more likely to develop expertise and problem-solving skills through deliberate and repeated practice [1] [2]. In the context of introductory programming (IP) education, practice is usually associated with the implementation of programs that process inputs within a given pattern, and generate the desired corresponding outputs. However, finding techniques that foster students' engagement in problem-solving may not be the easiest of tasks.

In general, people enjoy playing games [3]. Recent researches show that video games have become increasingly popular among both genders and age ranges [4] [5]. In its 2017 annual report, the Entertainment Software Association found that 65% of United States' households had at least one person who played 3 or more hours of video games per week. Over 4000 households were surveyed. From this sample, 41% of self-declared gamers were women [4] and in 2018, this number raised to 45% [5].

In Brazil, where the experiments reported in this paper were held, a survey conducted with nearly 3000 people, in 2018, showed that 58.9% of gamers were women. This presents a gradual increase of female gamers since 2013, when this survey mapped that 41% of self-declared gamers were women in this country [6].

This notable game phenomenon inspired the creation of a new concept, known as *gamification*, which was defined by Deterding *et al.* as *the use of game design elements in non-game contexts* [7]. Even globally well-known companies such as Oracle, Google, Microsoft, Adobe and IBM started implementing full-scale games and/or gamified applications in support of their business activities, with the intent of increasing the efficiency in areas such as marketing, management and administration [8] [9].

Taking that into consideration, educators and researchers have noticed they could insert game elements in the teaching/learning process. Hence, many have studied the effects of their use in engaging students [10] [11] [12] [13]. Despite all the positive results, some studies suggest that gamifying learning may not always be beneficial, advising educators to use it with caution [14].

This research attempted to isolate and measure the impact of one specific game element (leaderboard) on student engagement (completed assignments) in short-length in-class programming activities. A total of $N = 84$ students were asked to complete programming assignments that covered the content of some of the modules of a Computer Science undergraduate IP course that implemented the mastery learning teaching method[1]. Among the results from such experiments, it was verified that intermediate students tend to solve more programming problems because of the fact that a leaderboard is present.

II. Background

Many researchers have studied the effects of the use of game elements, or gamification, in a variety of non-game contexts, among which, education. Here, we cite some of the experiments and theoretical studies that have been done in recent years, and relate them with this research.

In their paper, Fotaris *et al.* [16] present a quasi-experimental study with a combination of instructor feedback, real time sequence of scored quizzes, and live coding. It uses the *Kahoot!* Classroom Response System and Codecademy's interactive platform in an entry-level Python programming course. Results from experimental gamified group were better than the control group. One disadvantage was that it was a small-scale study, such as this one.

Ibáñez *et al.* [11] evaluated the learning effectiveness and engagement appeal of a gamified learning activity with focus on the C programming language. Their study inquired which gamified learning activities were more appealing to students. The results of the evaluation showed positive effects on student engagement toward the gamified learning activities and a moderate improvement in learning outcomes. Students reported different motivations for continuing and stopping activities once they completed the mandatory assignment.

Landers & Landers [13] studied the effect of some aspects inherent in leaderboards in the context of education, focusing on learner behaviour, time-on-task, competition and goal-setting theory. Their method consisted of assigning learners to complete an online wiki-based project that had a gamified version with a leaderboard and another without it. The results showed that students with leaderboards interacted 29.61 times more, on average, with their projects. Bootstrapping was used to analyze the academic achievement with the amount of time invested on the task. Hence, it was demonstrated that leaderboards can be used to improve learners' performances, within certain circumstances. Our study relates the the aforementioned paper, but with focus on measuring in-class performance and a mastery learning curriculum.

Besides that, Hanus & Fox [12] developed a longitudinal study with students from two 16-week courses, one of which received a gamified curriculum and the other didn't. They measured students' satisfaction, motivation, effort, social comparison, learner empowerment and academic performance at four points during the courses. It was observed that students from the gamified course presented less motivation and satisfaction over time, aside from worse final grades in comparison with the non-gamified course. The paper demonstrates that it is imperative that gamification be applied with caution

[1]The mastery learning teaching method states that each student must learn the current topic before moving forward to the next. Hence, each student advances in the course at his/her own pace. For more information on this educational approach, see Bloom *et al*'s work. [15]

in education for it may present a negative impact.

Cheong et al [10] developed a gamified quiz with leaderboards that were used by students in Information Technology courses. A survey was also applied, whose feedback was mostly positive. As for the leaderboards, some students felt that the competition was motivating, but there were students that felt embarrassed for being in the last positions. Therefore, this paper is important because it brings awareness of good points and drawbacks of using leaderboards in class.

Some of the aforementioned studies and others concluded that the effects of gamification may not be long-term [17] [18] [19] [20] [12]. Taking that into consideration, this study tried to measure its impact as a one-time approach. Furthermore, most of the studies above mixed several game elements [11] [12] [16]. As rich an experience that might promote, this research had a different intent: to isolate and measure the effect of a single game element – the leaderboard – in a one-time in-class activity over the term. Therefore, its contribution is very specific and show whether the leaderboard may be used as a sporadic tool to engage students in-class, given a similar educational context.

III. Methodology

The research procedures aimed to collect data that supports an answer to the following research question:

RQ1: Can the introduction of a leaderboard lead to engaging students to complete more programming assignments in an introductory programming class?

From which the following hypotheses were proposed:

Hypothesis$_0 1$: The number of programming assignments completed by a student in class is not affected by the introduction of a leaderboard.

Hypothesis$_1 1$: The number of programming assignments completed by a student in class increases with the introduction of a leaderboard.

A. Subjects and the undergraduate courses

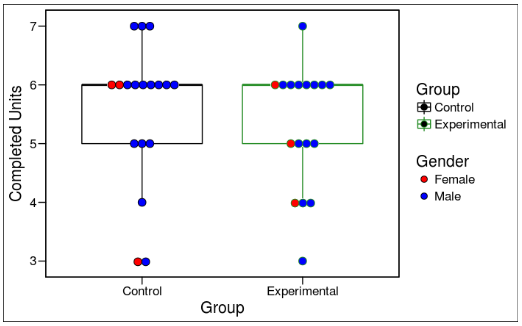
The subjects that took part in the experiments were portions of students enrolled in Computer Science undergraduate IP courses, at a Brazilian Public University. The first trial involved a total of $N_1 = 35$ students, whereas the second had $N_2 = 49$. These students were from different terms and did not overlap. Table I shows number of voluntary participants, per gender, per group.

TABLE I: Samples' Demographics

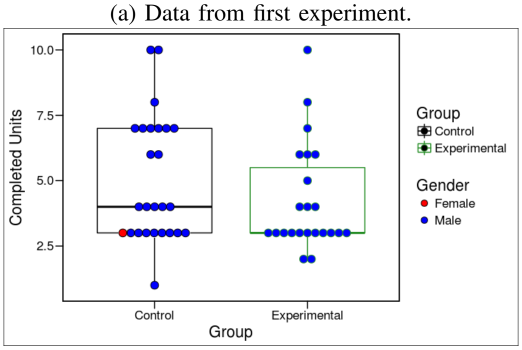
Trial	Group	Male	Female
1	Control	15	3
	Experimental	14	3
2	Control	25	1
	Experimental	23	0

Because of the mastery learning approach, at the time of each experiment, there were students that ranged from being in the first unit to those who had already completed all the ten modules (also called *units* here). The first 3 units covered basic programming concepts (simple programs, conditionals), units 4-6 covered intermediate concepts (for/while loops and functions), units 7-10 covered more advanced topics (data structures such as arrays, lists, matrices, and maps). The adopted programming language was Python 2.7. In the IP courses at hand, students were submitted to a weekly mini-test in order to be able to advance in the units. Therefore, students' partial grades were directly related to which unit they were at. Figure VII shows the distribution of students' units per group before the first and second experiments.

In order to make groups equal, students were randomly assigned to both groups taking into consideration each student's latest completed unit a week before the experiment. A script was developed in order to randomly assign students to their groups. The sample was divided into $k = 1 + 3,322(log10n)$ classes (Sturges formula) [21]. In the formula, n = total number of students expected to participate. The script made sure there was an equal amount of students in all k classes, by randomly reassigning them until this condition was met, in the control and experimental groups, respectively. The distribution of students is presented on Table II.



(a) Data from first experiment.



(b) Data from second experiment.

Fig. 1: Boxplots of students' unit completion before experiments per group.

TABLE II: Statistics for students' completed units before both experiments

		Number of completed units		
Trial	Group	Mean	Median	Standard Deviation
1	Control	5.56	6	1.20
	Experimental	5.29	6	1.05
2	Control	5.00	4	2.38
	Experimental	4.22	3	2.04

A Wilcoxon rank sum test[2] [22] [23] with continuity correction was performed for each trial to compare the number of units completed before the experiment in each group (control and experimental). The null hypothesis stated that groups were equal in relation to unit completion before the experiments. The p-values obtained in the tests presented on Table III are above $\alpha = 0.05$ and failed to refute the null hypothesis. Therefore, the samples were technically equivalent, assuming the unit completion as a measure of technical knowledge obtained in the IP course.

[2]This is a non-parametric test alternative to the t-test that only requires that the groups being compared be independent and does not rely on a specific distribution.

TABLE III: Wilcoxon rank test comparing number of completed units per group, per experiment.

Trial	W	P-value	95% CI	
1	179	0.3681	0.00	1.00
2	366.5	0.1625	0.00	2.00

B. *The leaderboard and length of the activity*

The leaderboard consisted of a table with students' positions, photos, names, followed by each one's number of problems solved and was projected publicly visibly on the laboratory's wall. The students were ranked by the number of problems they had solved. The positioning was cumulative, meaning that the if K students had solved the same amount of problems, they would all be on the same i-th position; whereas, the students that solved the number of questions immediately below, would be in $i+K$-th position. This rule held for all positions. The subjects were asked to remain in the laboratory for at least 90 minutes, trying to solve the programming problems. After this period, they could leave if they wanted to. The extra time granted to problem-solving was extended on demand. The first experiment lasted 120 minutes, whereas the second lasted 160 minutes. Students would be blocked from the system if they decided to leave early, or when the experiment reached its end.

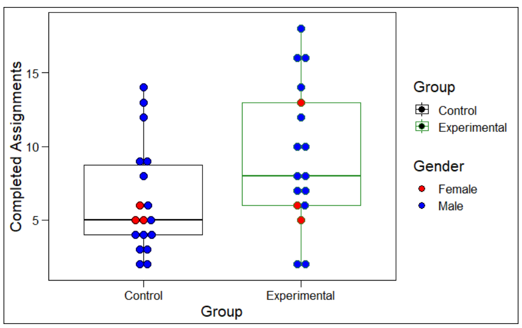
In order to mitigate the unpleasantness of being in the last positions, the leaderboard only showed students in the first positions for most of the time. However, if a specific student was curious about his/her position, the leaderboard would be scrolled down briefly so he/she could see his/her status. Furthermore, the only criteria for positioning was the amount of assignments they completed. Since there were many assignments available, even beginner students could try and reach fairly high positions by solving the easiest problems over the limited time they had.

C. Metrics

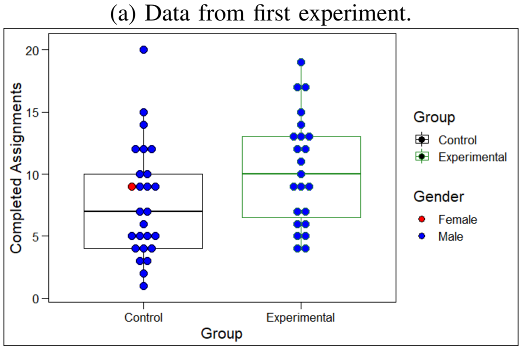
To evaluate the effectiveness of the insertion of the leaderboard in the students' performances, the following metrics were taken into consideration: number of problems solved in the experiments; student latest unit completion log before experiments; categorical variable that indicates the insertion of a leaderboard in the activity.

IV. Results

The experimental group completed statistically more coding tasks than the control group. The box-plots of Figure 2 show the distribution of number of problems solved in both experiments. They visually indicate that a good portion of both experimental groups solved more programming problems than students from control groups.



(a) Data from first experiment.



(b) Data from second experiment.

Fig. 2: Box-plots of number of completed assignments per group for each experiment.

On Table V, the means and medians of number of solved problems give a first numerical hint that the experimental group performed better on both trials. The first experimental group solved 3.08 (48.7%) more programming problems than its respective control group. On the second experiment, the experimental group solved 2.53 (32.56%) more programming problems than the control group, on average.

To verify that the number of completed units before the experiment was a good measure of technical skills for the task given, a Pearson Correlation Coefficient was calculated to compare unit completion and number of assignments completed in the experiment.

TABLE IV: Variances, Pearson Covariances and Correlation Coefficients in relation to number of Completed Units before experiments and Completed Assignments in the experiments.

Trial	Group	Variable	var(X)	Cov.	Cor.
1	Both	C.U.	1.252	1.634	0.325
		C.A.	20.205		
	Ctrl.	C.U.	1.438	1.627	0.366
		C.A.	13.765		
	Exp.	C.U.	1.096	2.184	0.434
		C.A.	23.132		
2	Both	C.U.	5.029	5.256	0.508
		C.A.	21.249		
	Ctrl.	C.U.	5.680	5.4	0.500
		C.A.	5.4		
	Exp.	C.U.	4.178	6.431	0.712
		C.A.	19.494		

TABLE V: Statistics for students' number of problems solved on experiments

		Number of problems solved		
Trial	Group	Mean	Median	Standard Deviation
1	Control	6.33	5	3.71
	Experimental	9.41	8	4.81
2	Control	7.77	7	4.53
	Experimental	10.30	10	4.41

A. Hypothesis Test

Knowing that the samples were technically equivalent, as demonstrated in Section III, another Wilcoxon rank sum test was performed for each group to determine whether the number of completed assignments was affected by the presence of the leaderboard. The results from this test may be

TABLE VI: Results Wilcoxon rank sum tests in relation to completed assignments in both experiments.

Trial	W	P-value	95% CI		Diff. in Location
1	91	0.042	-6.00	0.00	-3.00
2	198	0.043	0.00	2.00	-3.00

seen on Table VI. The P-value calculated on both experiments is below $\alpha = 0.05$, and refute the null hypothesis ($H_0 1$). The column with difference in location had a value of -3 for both trials and represents the median of the difference between samples from the control group and respective samples from the experimental group, indicating that the control group performed worse than the equivalent experimental group.

B. Bootstrap and blocking

In order to investigate in greater detail which type of students were mostly affected by the leadearboard, students from each trial were divided into three categories (blocks), according to unit completion data. More formally, let S be the set of all students that participated in the experiments and x be a specific student, such that $x \in S$; and let A, B, C, be subsets of S, as defined in Equations 1a$-$1c.

$$\{\, x \in A \mid 7 \leq u_i \leq 10 \} \tag{1a}$$

$$\{\, x \in B \mid 4 \leq u_i < 7 \} \tag{1b}$$

$$\{\, x \in C \mid 0 \leq u_i < 4 \} \tag{1c}$$

Where u_i represents the i-th student's latest completed unit before the experiment (value from 0 to 10). With that, students were divided into 3 categories: beginner, intermediate and advanced. Table VII shows the number of students per group of each category, for each trial, indicating that most of the categories were numerically equivalent.

Bootstrap [24] [25] was applied to each block of the collected samples, independently, being resampled 10000 times, with replacement. For each resampling, the mean of number of problems solved per student was calculated for each group (control and experimental) and subtracted, as expressed by Equation 2.

TABLE VII: Number of students per class, category and experiment.

Trial	Class	Number of students per group		
		Control	Experimental	Total
1	A	3	1	4
	B	13	15	28
	C	2	1	3
	Total	18	17	35
2	A	9	3	12
	B	7	7	14
	C	10	13	23
	Total	26	23	49

$$\overline{nSolDif}_i = \overline{nSolCtrl}_i - \overline{nSolExprtl}_i \tag{2}$$

Where, $i = 1, 2, 3, ..., 10000$, and represents the i-th iteration of resampling made on each block by the bootstrap technique, and:

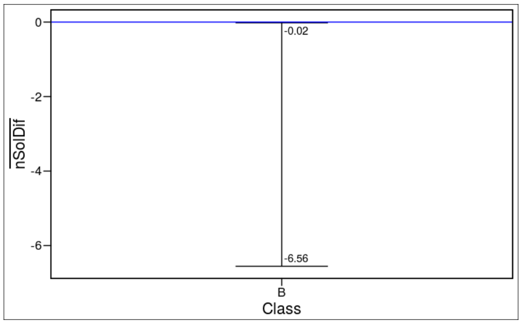
- $\overline{nSolDif}_i$ is the i-th difference of means.
- $\overline{nSolCtrl}_i$ is the i-th mean calculated for the i-th sample of the control group.
- $\overline{nSolExprtl}_i$ is the i-th mean calculated for the i-th sample of the experimental group.

After that process, a confidence interval (of 95%) was calculated and depicted in Figure 3. Since the number of students of categories A and C from the first sample were too low, bootstrap was not performed for these categories.

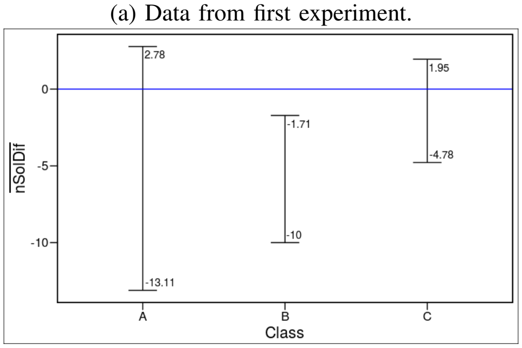
V. SURVEY

In the second experiment, when each student decided to leave after the 90-minute time-lapse, he/she would be asked to answer a survey on his/her experience and opinion on gamified educational environments. In the control group, this was done without notifying the subjects still participating in the activity. The intent here was to investigate the students' feelings towards gamified education and is some evidence of how they would take further gamified educational approaches, if they were to be reintroduced.

A total of 40 students decided to answer it. One of the questions was, "How would you evaluate the impact of the insertion of game elements in



(a) Data from first experiment.



(b) Data from second experiment.

Fig. 3: Confidence Intervals (95%) generated by bootstrap of differences between the mean of the number of problems solved by the Control and Experimental group, per class of student, for each experiment, as defined by Equation 2.

the educational context?". From those, 25 (62.5%) students chose the *Excellent* option and 15 (37.5%) chose the *Good* option. There were three other options: *There is no relevant impact*; *Bad*; and *Other(s)*, but none of the students chose any of these options.

When asked whether they preferred traditional or gamified courses: 32 (82.1%) opted for gamified courses, 2 (5.1%) preferred hybrid courses, 2 (5.1%) opted for traditional ones, 2 (5.1%) didn't opine alleging lack of experience, and 1 (2.6%) answered that depended on the course.

The great majority 29 (72.5%) indicated that had great experience with games, 5 (12.5%) had moderate experience with games, whereas the remaining 6 (15%) had little or no experience with games.

When asked about previous experience with gamified educational platforms: 11 (27.5%) said they had great experience, 15 (37.5%) said it was moderate, 11 (27.5%) had little or no experience, and 3 (7.5%) did not answer this question.

Finally, when asked about what they thought about the use of game elements in Education, some of the answers commented that gamification "can help learning", "increases interest and dedication", "motivates students to compete and learn as a consequence", "that it is more engaging than traditional methods", "that games present, in a beautiful, fun and artistic manner, the most diverse sorts of knowledge", "that it is a smart strategy", and "that competitiveness is sure the most motivating factor". Nonetheless, some answers drew attention that "some students may focus on the game itself and forget about other traditional learning methods, advising the gamified method to be used as a complement", and that "the excessive use of games for learning would rather be inefficient". Only one woman answered this survey. She said she loved games and had extensively played them over her life, but had had less time for them after the university began. She also mentioned she preferred gamified courses and highlighted that she enjoyed the challenges they present.

VI. Discussion

The data extracted from both experiments showed overall positive performance results in favor of students in the experimental groups. The Pearson Correlation Coefficients of Table IV that relates the number of Completed Units before the experiments and Assignments Completed in the experiments are above 0.325 in all scenarios, indicating a positive linear correlation. Hence, it seems like a plausible metric to verify that groups were technically similar. A curious effect that was noticed was that the coefficients obtained by the experimental groups were higher than their respective control groups, leading to an interpretation that the leaderboard increased the positive linear correlation of the student's current unit and the amount of exercises they solved, reaching a strong correlation in the case of the second experimental group (0.712).

In Section IV-B, the subtraction of solved problems' means of both intermediate categories were always negative, meaning that students from the experimental group in this category performed significantly better (95% confidence) than the ones in the control groups, implying that the leaderboard had a real effect on students from this category. However, these intervals for the beginner and advanced students (categories A and C) crossed zero, which means that we can't infer a real impact from the leaderboard on these students.

One explanation for the significance in the effect on intermediate students is hypothesized to be due to the fact that intermediate students have enough knowledge to solve a considerate number of problems, but lack interest in solving exercises voluntarily. However, the competitiveness presented by the leaderboard possibly increased their engagement to use their technical skills to actually solve problems as to fulfill their desire to reach better positions in the leaderboard; whereas intermediate students that are not in a competition don't feel that motivated to try harder.

Concerning the lack of significance of impact on engagement on higher grade students is possibly due to the fact that excellent students are normally inherently motivated to solve problems with or without competition; but it may also be due to the small sample from this category of students.

In turn, beginner students performed equivalently in both groups. We reason that even if they wanted to surpass their classmates to reach better positions on the leaderboard, they had insufficient technical knowledge to solve many problems. This fact makes one consider a leaderboard design according to which students in the last positions be hidden from public view. This might motivate them into trying to solve more problems, without demotivating them for not being in good positions.

Finally, this research assumes that these results could also be due to the one-time aspect of the approach, possibly presenting lower or even negative impact on engagement on the long term; or due to the small sample size ($N = 84$). Those concerns, however, are to be left to future studies. The main contribution of this paper is showing that a leaderboard within the proposed design may be used at least once in an undergraduate IP course with similar context in an attempt to increase student engagement in-class.

VII. Threats to Validity

Overall, herein presented results are positive, but some threats to validity are worth mentioning, such as: small sample ($N = 84$); instruction to do as many exercises as possible could be responsible for increased performance, even though activity didn't account for any grades; some students were absent or didn't want to participate, leaving certain technical unbalance; factor of novelty could be responsible for greater impact on students' performances, for the leaderboard hadn't been used before experiments; results for short length activities may not hold or be greatly diminished on the long term as other studies suggested [17] [18] [19] [20] [12]; and other ignored factors could have had influence on performance.

VIII. Future Work

There are many factors, game elements, and different contexts that could be further investigated. As a continuation of this work, we expect to implement richer gamified educational environments with other game elements such as badges and collaborative experiences and report impact on performance of Computer Science students. Aside from that, instead of short-length experiments, a longitudinal approach could assess the effect of game elements on the long-term.

IX. Acknowledgement

We thank CAPES for supporting this study. This work was also partially sponsored by the agreement No 08200.315131/2016-10 between UFCG and ePol/DPF, and by the agreement No SCON2017-01155 between UFCG, PaqTcPB and OPAS/OMS.

References

[1] K. A. Ericsson, R. T. Krampe, and C. Tesch-Römer, "The role of deliberate practice in the acquisition of expert performance." *Psychological review*, vol. 100, no. 3, p. 363, 1993.

[2] L. E. Winslow, “Programming pedagogy – a psychological overview,” *SIGCSE Bull.*, vol. 28, no. 3, pp. 17–22, Sep. 1996. [Online]. Available: http://doi.acm.org/10.1145/234867.234872

[3] K. M. Kapp, *The Gamification of Learning and Instruction: Game-based Methods and Strategies for Training and Education*. Pfeiffer, 2012.

[4] Entertainment Software Association, “Essential Facts about the Computer and Video Game Industry,” http://www.theesa.com/wp-content/uploads/2017/09/EF2017_Design_FinalDigital.pdf, 2017.

[5] ——, “Essential Facts about the Computer and Video Game Industry,” http://www.theesa.com/wp-content/uploads/2018/05/EF2018_FINAL.pdf, 2018.

[6] Sioux Group, Blend, ESPM and Go Gamers, “Pesquisa Game Brasil,” https://www.pesquisagamebrasil.com.br/, 2018, access in June 2018.

[7] S. Deterding, D. Dixon, R. Khaled, and L. Nacke, “From game design elements to gamefulness: Defining ”gamification”,” in *Proceedings of the 15th International Academic MindTrek Conference: Envisioning Future Media Environments*, ser. MindTrek ’11. New York, NY, USA: ACM, 2011, pp. 9–15. [Online]. Available: http://doi.acm.org/10.1145/2181037.2181040

[8] A. Uskov and B. Sekar, “Serious games, gamification and game engines to support framework activities in engineering: Case studies, analysis, classifications and outcomes,” in *Electro/Information Technology (EIT), 2014 IEEE International Conference on*. IEEE, 2014, pp. 618–623.

[9] K. Lakhotia, “En garde: Winning coding duels through genetic programming,” in *2013 IEEE Sixth International Conference on Software Testing, Verification and Validation Workshops*, March 2013, pp. 421–424.

[10] C. Cheong, F. Cheong, and J. Filippou, “Quick quiz: A gamified approach for enhancing learning.” in *PACIS*, 2013, p. 206.

[11] M. B. Ibanez, A. Di-Serio, and C. Delgado-Kloos, “Gamification for engaging computer science students in learning activities: A case study,” *IEEE Transactions on Learning Technologies*, vol. 7, no. 3, pp. 291–301, July 2014.

[12] M. D. Hanus and J. Fox, “Assessing the effects of gamification in the classroom: A longitudinal study on intrinsic motivation, social comparison, satisfaction, effort, and academic performance,” *Computers & Education*, vol. 80, pp. 152–161, 2015.

[13] R. N. Landers and A. K. Landers, “An empirical test of the theory of gamified learning: The effect of leaderboards on time-on-task and academic performance,” *Simulation & Gaming*, vol. 45, no. 6, pp. 769–785, 2014.

[14] A. DomíNguez, J. Saenz-De-Navarrete, L. De-Marcos, L. FernáNdez-Sanz, C. PagéS, and J.-J. MartíNez-HerráIz, “Gamifying learning experiences: Practical implications and outcomes,” *Computers & Education*, vol. 63, pp. 380–392, 2013.

[15] P. W. Airasian, B. S. Bloom, and J. B. Carroll, *Mastery Learning: Theory and Practice*. Holt Rinehart & Winston, 1971.

[16] P. Fotaris, T. Mastoras, R. Leinfellner, and Y. Rosunally, “Climbing up the leaderboard: An empirical study of applying gamification techniques to a computer programming class.” *Electronic Journal of e-Learning*, vol. 14, no. 2, pp. 94–110, 2016.

[17] J. Hamari, J. Koivisto, and H. Sarsa, “Does gamification work? – a literature review of empirical studies on gamification,” in *2014 47th Hawaii International Conference on System Sciences*, Jan 2014, pp. 3025–3034.

[18] J. Hamari, “Transforming homo economicus into homo ludens: A field experiment on gamification in a utilitarian peer-to-peer trading service,” *Electronic Commerce Research and Applications*, vol. 12, no. 4, pp. 236 – 245, 2013, social Commerce- Part 2. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S1567422313000112

[19] R. Farzan, J. M. DiMicco, D. R. Millen, B. Brownholtz, W. Geyer, and C. Dugan, “When the experiment is over: Deploying an incentive system to all the users,” in *symposium on persuasive technology*, 2008.

[20] R. Farzan, J. M. DiMicco, D. R. Millen, C. Dugan, W. Geyer, and E. A. Brownholtz, “Results from deploying a participation incentive mechanism within the enterprise,” in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, ser. CHI ’08. New York, NY, USA: ACM, 2008, pp. 563–572. [Online]. Available: http://doi.acm.org/10.1145/1357054.1357145

[21] H. A. Sturges, “The choice of a class interval,” *Journal of the american statistical association*, vol. 21, no. 153, pp. 65–66, 1926.

[22] M. Hollander and D. A. Wolfe, “Nonparametric statistical methods,” 1999.

[23] D. F. Bauer, “Constructing confidence sets using rank statistics,” *Journal of the American Statistical Association*, vol. 67, no. 339, pp. 687–690, 1972.

[24] T. C. Hesterberg, “What teachers should know about the bootstrap: Resampling in the undergraduate statistics curriculum,” *The American Statistician*, vol. 69, no. 4, pp. 371–386, oct 2015. [Online]. Available: https://doi.org/10.1080/00031305.2015.1089789

[25] B. Efron and R. Tibshirani, *An Introduction to the Bootstrap (Chapman & Hall/CRC Monographs on Statistics & Applied Probability)*. Chapman and Hall/CRC, 1993.