Studying Student Differentiation in Gamified Education: A Long-Term Study[1]

Gabriel Barata[2], Sandra Gama, Joaquim Jorge, Daniel Gonçalves
Dept. of Computer Science and Engineering
INESC-ID / Instituto Superior Técnico, Universidade de Lisboa
Lisbon, Portugal

ABSTRACT

Gamified learning is a novel concept that according to recent studies, can increase student activity and improve learning outcomes. However, little is known about how different students experience and are engaged by it. We present a long-term study that identified distinct behavioral and performance patterns in participants taking a gamified college course. Our study lasted for three years, during which we deployed three consecutive instances of the course, each featuring improvements based on student feedback from the previous instances. To understand how different students behaved in our gamified experience, according to their daily performance, we performed cluster analysis and assessed student engagement in the last year using a formal instrument. We then did a cluster-wise analysis using different performance and behavioral measures, to further assess and characterize every cluster. To wit, we identified six different student clusters, each featuring different behaviors and performance levels. However, only four were present in the last year, which differed in terms of engagement with the course. In this paper we carefully describe each student cluster, explain how they evolved, and derive meaningful design lessons.

Keywords: Gamification, Engagement, Gamified Learning, Student Types, Cluster Analysis, Student Performance

1. INTRODUCTION

Gamification is a recent concept that adopts game elements in non-gaming contexts [1, 2, 3] to engage users and encourage them to adopt specific behaviors, such as being more ecofriendly [4], becoming loyal to a brand [5], raising health awareness [6], improving productivity [7], or learning how to drive [8]. Gamification draws on the many motivational qualities of good games, which make them good behavior drivers and powerful learning tools [9, 10, 11]. As opposed to traditional educational materials, games can deliver information on demand and within context, and are designed to maximize choice and ease the impact of failure [12]. Good games aim at preventing players from getting either bored or frustrated, thus allowing them to experience flow [13, 14] and endure. Researchers and educators have for long been studying the effort and resilience of gamers when playing games, and how these can be put to use to help in learning [15]. Games have been used to educate with success [16, 17], with documented improvements in learning outcomes, motivation and diligence in different academic fields.

Gamification of education is a recent subject, and research shows promising results. Pioneer studies already demonstrate that gamification may potentially increase student activity [18] and performance [19], although replicating these results over several iterations of the same course seems to be

[1] Accepted for Publication in Computers and Human Behavior (Elsevier), Special Issue on Gamification, Feb. 2016
[2] Corresponding author. Email: gabriel.barata@ist.utl.pt. Address: INESC-ID, Rua Alves Redol, 9, 1000-029, Lisboa, Portugal. Phone: +351 213 100 205

unexplored. Furthermore, little information exists regarding how different students adapt to a gamified course and what kind of strategies they adopt. Recent works show that in a gamified setting, only users that bother to monitor their progress and that of others seem be to be significantly affected by gamification [20]. Therefore, it is paramount to understand how different students play and learn in a gamified course and how they are engaged by it, to develop new gamification approaches that can adapt to their needs and engage more learners.

In previous work, we fully gamified a college course and observed that students participated more when compared to the previous non-gamified version of the course, and felt both more motivated and interested as compared to other "regular" courses [21, 22, 23]. We have now collected student data from consecutive terms of our gamified course, over a three-year period, where iterative changes took place to address students' needs. We analyzed how they acquired experience points over time, via cluster analysis, and then made a cluster-wise assessment of different behavior and performance measures, including online participation, lecture attendance, evaluation results, among others. Furthermore, we also used a validated instrument to assess how each student was engaged with the course in the third year, and studied how this was portrayed to each cluster. Our study identifies six different student clusters, observed over the years, each exhibiting different behavior and performance traits. However, only four were present in the third instance of the course, where two of them were more engaged than the others. In this paper we carefully characterize each student cluster and study how they relate to each other. We also discuss how changes made to the course over the years may have affected their composition and how students from the third year differed in terms of engagement. We finalize by describing the most important lessons learned from this experiment and present guidelines for designers of gamified learning experiences.

2. GAMIFICATION IN EDUCATION

Games have long been considered good learning tools and their usage in education has been studied for more than a decade. Research shows that games can both be used to engage students and increase their activity and learning outcomes, at diverse academic levels, ranging from grade school [24], through high school [25], to college [26], and in diverse fields of learning, such as numerical methods [26], biology [27], programing [28], or electromagnetism [29]. Drawing on these pedagogical benefits, gamification was soon adopted in education to engage learners, with prominent examples being Khan Academy[3] and Codecademy[4]. In these services, students learn by watching videos online and performing exercises, while their progress is tracked via points and collectible badges.

Gamification applies game design elements to non-gamified processes [1, 2]. Even though there is not a formal list of elements to use, some of the most common are [30, 31, 32, 5, 33, 34]: experience points and levels, serving the main purpose of transmitting feedback and progress; challenges or quests, which provide tasks with clear goals, progress assessment and train users for more complex tasks; badges, collectible artifacts that aim at boosting the user's motivation by appealing to her natural desire to collect; and leaderboards, which spur competitiveness and encourage users to continually strive to achieve their desired ranking. Using leaderboards in gamification is controversial, given that users at the bottom usually become demotivated. However, research on leaderboards in gamified settings did not find any significantly harmful effect on participant motivation [35, 36].

Measuring success of gamified learning has become a great concern. A major review on empirical studies on gamification suggests that effectiveness greatly depends on both context of application and participant's characteristics [34]. Another study tried to assess how effective gamification might be at motivating students, but formal measurements of intrinsic motivation do not support a correlation [37]. Still, gamification's potential to shape student behavior is hard to overlook. We differentiate gamification applied to education into two different phenomena: 1) partially-gamified and 2) fully-gamified experiences. Whereas the former consists of typically gamifying a single evaluation

[3] https://www.khanacademy.org/
[4] http://www.codecademy.com/

component of a course (or other unit of teaching), the latter comprises the gamification of the entire course, changing how it is evaluated as whole.

There has been a lot of research in partially gamified learning. In their study, Cheong et al. used a gamified quiz to evaluate IT undergrad students [38], whereby they received points for answering questions and could then compare their scores with those of other students, in a leaderboard. Students self-reported that the quiz improved both their learning effectiveness and their grades, and also their enjoyment and engagement. However, this study presented no empirical results other than self-reports. Domínguez et al. also proposed a gameful approach to an e-learning ICT course [19], where students could take optional exercises, either by reading a PDF file or via a gamified system. In the latter, students were awarded with badges and medals by completing the exercises. They found that students using the gamified approach had better exam grades and reported higher engagement with the course.

Another work [18] studied the effect of adding badges to an online repository of student-generated multiple-choice questions. This was used to evaluate the students of a course featuring frameworks and tools to understand and control the impact of disease in populations. Students using gamification answered significantly more questions and were more active than those not using it. In a similar study [39], achievement badges were added to an online learning environment where students completed interactive automatically assessed exercises on data structures and algorithms. According to data collected from the system logs, achievement badges had a significant impact on student behavior, with more of them getting more perfect scores. However, only a small group of students was especially motivated to collect badges.

Fully-gamified learning focuses not on changing one evaluation component but in creating a whole new learning experience, where typically most evaluation components have to be adapted. In his book, Lee Sheldon [40] explains how a conventional course can be turned into a game, without using technology, where students start with an F grade and go all the way up to an A+, by completing challenges and gaining experience points. Several reported case studies using this method showed improvements that covered student attendance, willingness to participate and work, and grade performance. Following this approach, several other studies have reported encouraging and diverse findings. Aguilar, Fishman and Holman conducted a series of experiments with several gamified college courses [41, 36, 42], where they studied correlations between student perceptions of the gamified grading systems and adaptive outcomes associated with gameful course designs. The courses had comparable grading systems, where students had the freedom to specify the type of assignments and their respective weight covered by 60% of the grade. The remaining 40% respected to traditional criteria, like attending classes and discussion sessions. A gamified gradebook named GradeCraft was also used by both students and instructors alike. Students were presented with a dashboard which displayed their current score and the level currently achieved. Also displayed were badges which were earned by students when they met specific skills defined by the instructor. The interface allowed students to compare their score to the class' average and check how it would translate to final grade. It was also possible for them to predict their grade by specifying scores to assignments they plan to complete.

The authors found that "affordances of gameful grading systems lead to positive perceptions of the grading system" and that "gameful mechanics were positively predictive of students' assessment of various aspects of the course, which in turn predicted positive non-cognitive motivational outcomes" [36]. They also observed equally in two sequential iterations of the same course that "whether students 'like' the grading system is positively related to whether they feel encouraged to work harder; their perceptions of control over their final grade; whether students completed more assignments; and the ease with which students feel they can earn the grade they want" [42]. GradeCraft was viewed as a capable tool that allowed instructors to have access to information that would be otherwise unavailable in a traditional learning environment [41].

In another work, Schutter [37, 43] presented a gamified undergrad college course on game design for education purposes and compared formal measurements of student intrinsic motivation and engagement with those of another non-gamified course, on principles of game design. The gamified course included avatars, guilds (groups of students), quests, a backstory, experience points and levels,

skills (powers gained through progression), and a leaderboard. Students had both mandatory and optional quests that could be of several types, which would earn students experience points on completion. As students leveled up, they would be able to choose and unlock a skill, which had direct consequences in class. The authors performed formal evaluation of student motivation using the Situational Motivation Scale (SiMS) [44] and student engagement using the Game Experience Questionnaire (GEQ) [45], on both courses. The authors concluded that "gameful instruction did not necessary lead to higher levels of intrinsic motivation or engagement in comparison to traditional teaching methods, and that further improvements to the design and documentation of the course are necessary."

Despite the considerable body of knowledge on gamified learning, empirical evidence of its effects over student motivation and performance are still scarce, which provides evidence that further research in the area is needed. Furthermore, to our knowledge, there are no other studies that try to understand how different students experience a gamification course and how they are engaged by it. The study here presented aims to fill that gap.

3. THE COURSE

Multimedia Content Production (MCP) is a semester-long MSc course in Information Systems and Computer Engineering, taught yearly at *Instituto Superior Técnico*, University of Lisboa. The course follows a blended learning method, whereby students attend both theoretical lectures and lab classes, but also engage in discussions and complete online assignments using the Moodle[5] platform. Theoretical lectures cover multimedia concepts such as capture, editing and production techniques, file formats and multimedia standards, as well as Copyright and Digital Rights Management. Laboratory classes explore concepts and tools on image, audio and video manipulation, via a series of regular assignments.

The student population comprised 35, 52 and 54 students in the first, second and third years respectively. Students attended parallel lectures in two sites, *Alameda* and *Taguspark*, some 35km apart, but the course ran synchronized across campuses and both instances shared the same Moodle platform. There were no distinctions between students from either campus besides their physical location. The faculty included four people, two professors and two lab assistants in the first year, and three people, the same two professors and a different lab assistant, during both the second and third years. The professors carried out most of the face-to-face contact (two times a week) and online interaction with students, while lab assistants had more limited face-to-face interactions (once a week) and fewer online responsibilities. Although most of our students are Portuguese, between two and three foreign exchange students took the course each year. Therefore, lectures and all content on Moodle were delivered in English (including the final exam). This is a requirement in our University, in order to make content accessible to non-Portuguese students.

Before the gamification experiment, MCP students were graded based on regular quizzes and lab assignments, online participation, a multimedia presentation about a particular subject, and a final exam. Students were scored using a 20-point scale, where 10 is the minimum passing grade. We noticed, however, that students often focused solely on the major evaluation components and overlooked the online participation. To further captivate students and engage them with the course we decided to adopt a gamified approach to study and analyze how student behavior and performance were affected by it.

We adopted an analysis method based on design-based research. This technique relies on iterative design cycles in real-world learning contexts, where participants interact with each another within the defined settings rather than in laboratory context, isolated from everyday practice [46, 47, 48]. The resulting principles are considered to have greater external validity than those arising from laboratory settings [49] and are viewed as more informative regarding systematic issues in education [50]. Accordingly, our experiment covered three academic terms, where three gamified course instances were deployed. Each instance featured improvements based on observation and student feedback collected by the end of the preceding year. In this section we describe the changes we made to the course on each

[5] www.moodle.org

iteration, provide an overview of the observed effects, and present the design rationale underlying our implementation.

3.1. The First Year

We started to gamify MCP by shaping course activities into meaningful challenges and achievements, using experience points (XP), levels, leaderboards, and badges as means to communicate progress and provide feedback to students. Most course activities took place on the online forums, where students earned XP for undertaking traditional activities (i.e. quizzes, multimedia presentation, lab classes and exam). We adopted an approach similar to that of Lee Sheldon [40] who translated grades to XP. started the course with 500 XP and earned more by completing achievements, which would also earn them badges. Examples of achievements include attending lectures, finding resources related to class subjects, finding bugs in class materials and completing challenges. Achievements could either be single-level or multi-level, according to the amount of iterations required to complete them. Each level (or iteration) earned students a level-specific amount of XP and a badge. Detailed information about every achievement and their usage throughout the years can be found in Table A.1. The amount of XP allocated to every level of every achievement can be found in

Joaqui
Delet

Table A.2

While most achievements did not have a time limit associated, a few had specific deadlines. Examples of these were achievements that required students to participate in Challenges and the Online Quests, both requiring participants to complete time-limited assignments. Challenges came in two flavors: Theoretical Challenges required students to solve problems related to the topics taught in a theoretical lecture. Lab Challenges involved producing creative content using tools and techniques taught in lab classes. Online Quests encouraged students to share knowledge and tools to solve specific problems, unrelated to any particular class. To mitigate failure, we decided to create varied achievements, some of them granting extra XP (the final grade was still limited at 20). This would allow students to make up for other achievements they could not or did not want to complete.

For each accumulated 900 XP students would increase their experience level, with each level being represented by a unique honorary title. Levels maxed out at 20 (18000XP), to match our traditional 20-point grading system, and students had to make level 10 to pass the course (see Figure 1). Their grade in the 20-point grading system corresponded to the current student level, which means that a student finishing the course at level 18 would receive a final grade of 18 out of 20. The leaderboard was the main entry point to the gamified experience by allowing students to assess both their progress and their peers' (see Figure 2). Students were displayed by row, sorted by descending order of amount of XP. By clicking on a row in the leaderboard, a dedicated page for that student would be displayed, where her achievement history and list of collected badges was displayed (see Figure 3). This made game progression transparent, by showing what had been accomplished so far and what could still be done.

XP (Level)	Current Grade
900 (lvl. **1**)	**1**
1800 (lvl. **2**)	**2**
2700 (lvl. **3**)	**3**
...	...
9000 (lvl. **10**)	**10** (minimum passing grade)
...	...
16200 (lvl. **18**)	**18**
17100 (lvl. **19**)	**19**
18000 (lvl. **20**)	**20**

Figure 1. XP to Grade matching.

Figure 2. The MCP leaderboard.

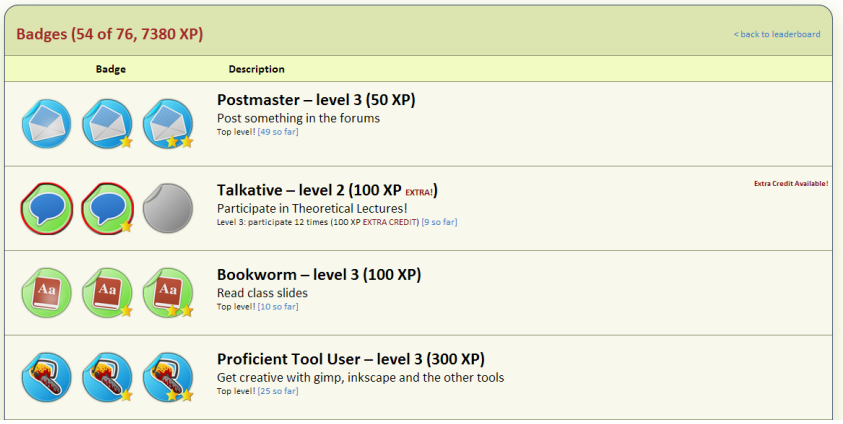

Figure 3. A student's badge list.

In the first year, student evaluation was processed as follows: 20% of maximum XP was allocated to regular quizzes, 20% to the multimedia presentation, 15% to lab class assessments, 35% to the final exam, and 10% (plus 5% extra) to the achievements. In lab classes students had assignments where they developed plugins for the PCM Media Mixer, a prototype media player developed exclusively for the course. These assignments were legacy from previous years and were replaced in the subsequent versions of the course by similar exercises using Processing[6].

By the end of the first year, we observed a significant improvement in terms of online participation (more student posts), lecture attendance and downloads of reference materials, as compared to the previous non-gamified year, and students considered the course as being more motivating and interesting than other regular courses [21]. However, by analyzing data collected at the end of the term, we concluded that our gamified course had a few problems. For example, 10% XP for the achievements was not enough to captivate all students, and many seem to have ignored that component altogether. Some students actually complained about the achievements being too much work for too few XP. Furthermore, challenges and quests were not evenly distributed over the term, as the majority of them occurred during the first half of the course. This reduced the students' opportunities to participate and recover from other challenges they could not or did not want to attend to. Informal student feedback as well as their responses to open questions in a final satisfaction questionnaire suggested that oral participation in class should be rewarded and there should be more room for collaboration, even though

[6]https://processing.org/

several people asked for more ways to compete. They also felt lack of autonomy in game and did not have many choices to make. We proceeded to fix these issues in the second iteration of our gamified course.

3.2. The Second Year

In the second year we tried to solve the problems identified at the end of the previous installment. We increased the amount of XP allocated to the achievements, by assigning them an additional 10% and taking 10% away from the quizzes. This aimed to make the game aspect more appealing and worth the students' time. To grant students with additional opportunities to participate and recover lost XP, we created more challenges and distributed them more evenly over the semester. For a change, we modified the amount of XP per level, from 900 to 1200. We also added the student dashboard, which would be displayed at the beginning of each student's page, right before the badges (see Figure 4). The dashboard contained line charts, histograms and statistics about their performance, allowing students to better assess their progress in the course. We created two new achievements to encourage collaboration in the labs, where one was awarded to the whole class if all students in that lab class had above 80% XP in an assessment (the Guild Warrior achievement), and the other one was awarded to students groups with the best score in lab assessments (the Guild Master). We also added new achievements to reward oral participation in class (the Talkative achievement), prompt responses to challenges (the Proactive), and compiling challenge results (the Archivist), so students could have more materials to study from. Each of these had three levels to complete. We removed two achievements that were present in the first year: The Good Host achievement, which rewarded students for attending invited lecturers on both campuses, because it was hard to ensure that there would be an equal amount of those lectures on both campuses; and the Bug Squasher, which required students to fix bugs on our media player prototype, which would no longer be used in our course.

We also added a new game element in the second year: the Skill Tree (see Figure 5). It consisted of a precedence tree where each node represented a thematic task, which would earn students XP upon completion. There were five initial nodes and subsequent nodes could be unlocked when two preceding ones were completed. The main goal of the Skill Tree was to give students more autonomy to make choices, by allowing them to choose different paths to the top nodes, according to their liking. However, due to its early stage of development, the Skill Tree was only added after two weeks of classes. Given its provisional nature, the Skill Tree was not mandatory and only 5% of extra XP could be earned from this element. It could be used, for example, as an alternative to some of the achievements. However, the initial

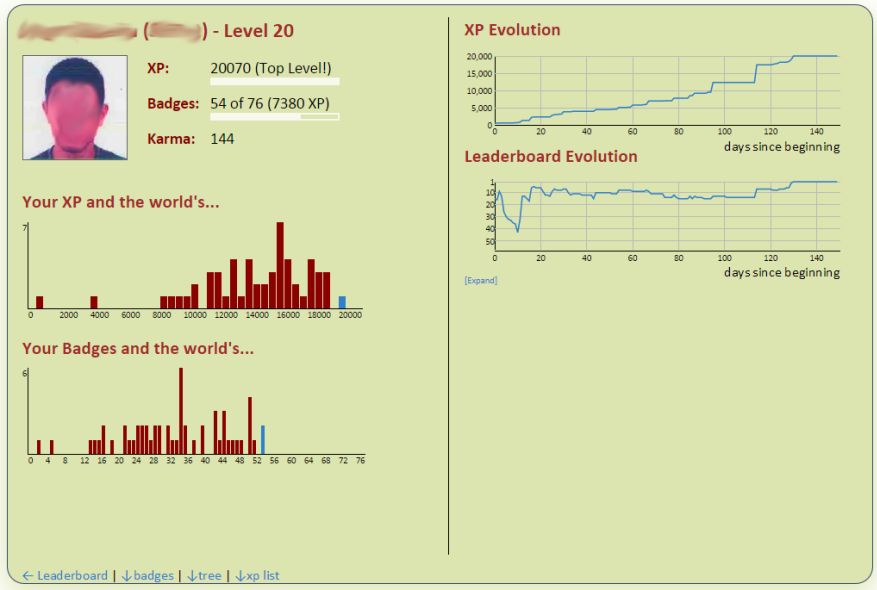

Figure 4. The student dashboard.

Figure 5. The MCP Skill Tree.

nodes did not earn XP, but subsequent ones did. The Skill Tree's largest problem though, was that unlocking nodes was cumbersome. Not only did students have to unlock the preceding nodes, but they also had to accumulate a particular amount of meta-points, which could be of the following elements: Text, Audio, Image, Video and Design. These could be earned by completing tasks (i.e. challenges and quests) related to one of those topics. The main goal here was to make the Skill Tree require more strategy, bud it ended up making the unlocking process harder.

We did all these changes to improve the students' autonomy and make the gamified experience more engaging. Ultimately, we wanted to reach out to more students and as we will see in the Results section, we believe we did. But, we also observed that some additions did not work as well as expected. Among the new achievements, only the one promoting oral participation worked as envisioned. Those promoting collaboration were mostly ignored, and the one encouraging prompt replies to challenges ended up getting students to post a lot in detriment of quality. This felt unfair to students that really wanted to make quality contributions, but it only became more evident in the second year because the game component was now worth more XP. The achievement that required students to compile challenge results was considered too much work for what it was worth and only a single badge was acquired. Because the Skill Tree was not well rewarded and was hard to participate in, only a few students actually engaged in it. However, they still felt they lacked autonomy, they could not be creative, and while some praised the competition, others found the experience to be too competitive. Once again, we aimed to get these issues fixed in the following year.

Despite the observed limitations, we believe the second deployment made for a more engaging gamified learning experience. The number of student posts almost doubled as compared to the first gamified year and they considered the course as more motivating, interesting and easier to learn from than other courses [22].

3.3. The Third Year

In the third year we performed additional modifications to overcome the problems identified in the previous course instance. We excluded the achievement that required students to compile the results from challenges and the one that encouraged prompt replies. Because students affirmed, in the satisfaction questionnaire, that the course required more work than the other regular courses, we removed the regular

Table 1. Comparison of the evaluation components across years.

	1st Year	2nd Year	3rd Year
Quizzes	20%	10%	N/A
Presentation	20%	20%	15%
Labs	15%	15%	15%
Exam	35%	35%	30%
Achievements	10% (+5% extra)	20% (+5% extra)	30% (+5% extra)
Skill Tree	N/A	5% extra	10%

quizzes and added the remaining XP to the Achievements. The final multimedia presentation was now worth less 5% of total course XP and the Skill Tree was worth 5% more.

Posts started to be graded based on a rating from 0 to 4 assigned by faculty, to promote quality over quantity. Therefore, a student making poor quality posts would have to make four times more posts to have the same grade as students making quality posts. To encourage them to learn from failure and further increase the amount of participation occasions, we changed the participation procedure on Challenges. Students were now allowed to post up to three times in the same challenge, to make up for lack of participation in other less appealing challenges. We also improved the Skill Tree, abolishing the meta-points system, rewarding the initial nodes, increasing the amount of XP earned from it, and adding 10 new nodes. Playing with the Skill Tree was now easier: the first nodes were unlocked at the beginning of the game and further unlocks depended solely on the completion of the preceding nodes. Again, we changed the amount of XP per level from 1200 to 1000. A comparison of the evaluation components across the years is depicted by Table 1.

To improve student autonomy and allow them to exercise their creativity, we also added a new element named AvatarWorld (see Figure 6). It consists of a 2.5D virtual world that evolves and grows as students earn XP, with new buildings and characters emerging. Students are represented by an avatar that they can use to explore the world. Avatars can be customized with equipment items and handheld objects, which can be unlocked by acquiring certain course badges. Students can also create custom content for the game, such as buildings and equipment items, using tools and techniques introduced in class.

Figure 6. The AvatarWorld.

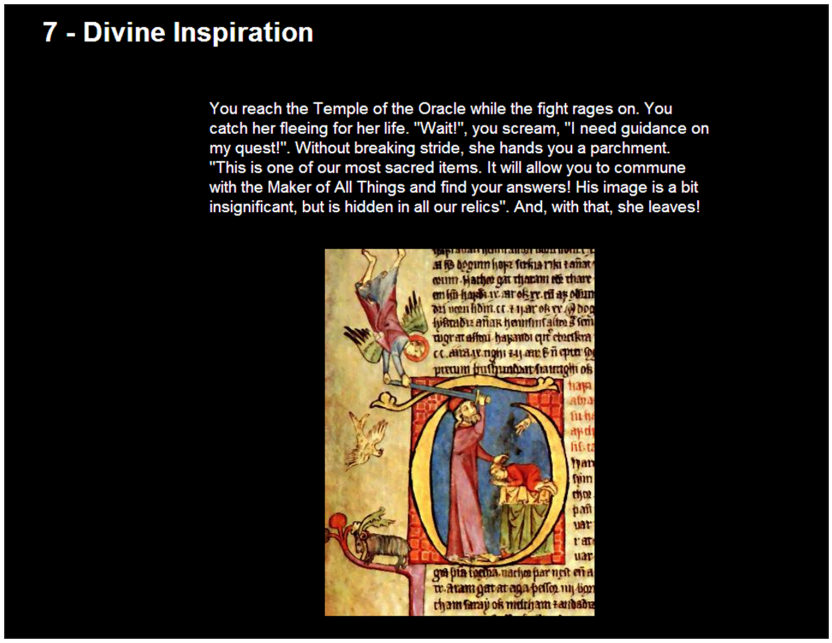

Figure 7. Level 7 of the MCP Quest.

Submissions were made via posts and graded by faculty, based on their creativity and technical correctness. Up to 3% of total XP could be earned from it by completing two related 3-level achievements (the Blacksmith and the Master Builder). The main goal here was to allow students to develop a sense of identity, to customize their learning experience and to be creative, while still using what they learned in class. However, owing to its early stage of development, AvatarWorld was not fully integrated with the course and students had little to do there besides creating custom content.

To fill the collaboration gap, we introduced yet another component named MCP Quest (see Figure 7), which replaced the former Online Quests, given their similarity to the Challenges. It consisted of an online-riddle where students started from a webpage with some sort of multimedia content, which they had to manipulate to find the URL for the next clue of the riddle, or next level. The amount of earned XP was proportional to the quest level reached (from a total of 20). To encourage every student to participate, they had to contribute at least once to earn the XP, but an individual could never post twice in a row. Contributions were posted in the forums and rated by faculty. This element was designed to encourage students to collaborate towards a common goal. The more they participated and the further ahead they got, the more XP everybody would earn.

Even though we observed some meaningful behavioral changes over the years, we knew little about how different students reacted to the gamified course and how engaged they were. In section four we describe the procedure we used to identify different types of student based on their progression on the course. In section five we formally assess how each student type was engaged with the course.

3.4. Design Rationale

Student performance is tightly related to how intrinsically motivated they are [51]. The Self-Determination theory (SDT) identifies three innate needs of intrinsic motivation [52]: Autonomy, a sense of control over the learning environment; Competence, referring to a need for challenge and feelings of effectance; and Relatedness, experienced when a person feels connected to others. The core motivation strategies that served as a basis for the choice and integration of game-design elements in our learning experience were drawn from the SDT. Our experience was inherently multiplayer and together with the leaderboard, the online forums, the cooperative achievements and the MCP Quest – a major collaborative effort – we aimed at boosting the student's feeling of Relatedness. To appeal to the students' feelings of

competence we provided several forms of positive feedback [52] and progress assessment, which include the achievement badges, the experience points and leveling system, and the student dashboard.

Autonomy is both a key feature of good games and good learning [53, 54]. We tried to give students a sense of autonomy and control over their learning experience by providing several meaningful choices and alternatives. Examples of these were the Skill Tree, which had several combinations of paths to achieve the maximum allocated XP, the achievements which had extra XP creating alternative routes towards completion, and the challenges starting from the third year, which allowed students to do up to three of the same kind. Providing alternative paths, allowing students to repeat challenges and allocating extra XP to achievements was an effort to allow students to learn from failure and control the pace of their learning, allowing them to do more of what they like or that interests them on a particular moment [55]. Similar approaches have been used on other studies on gamified learning [42, 43, 56].

4. STUDENT CLUSTERING

During course instruction we felt that not all students were reacting in the same way to the gamified experience. This was observable in different types of behavior, feedback and levels of performance. There were students getting really hyped about the course and doing everything they could to get just a few XP. This effort was usually transversal to other aspects of the course and often translate to better grades. Other students seemed to enjoy it at first and put forth effort, but would eventually grow tired of it and start to neglect several aspects of the course, ending up relying mostly on major evaluation checkpoints. There were also students that would simply ignore most of the gaming aspect from the beginning, and just did enough on the traditional evaluation components to pass. Therefore, we felt the need to understand what was going on; were there significant differences regarding how our students progressed in the course? Were these enough to differentiate them into different categories? This was the main motivation behind the study here presented, were we aimed at differentiating students by the way they perform over time.

In order to be able to classify students into different types, according to their progression in the course, we had to find a measurement capable of transmitting student progress over time. Leaderboard rank was the most straightforward, but two students with equal performance could never be at the same ranking at a given time. Therefore, we chose accumulated XP over time, which allows students with similar performance levels to be equally represented. Informally printing accumulated XP over time for each student revealed a few interesting patterns, which backed up our initial premise regarding the existence of different students types. Therefore, for a given instance of the course, we used as attributes of clustering analysis the amount of accumulated XP per day, in order to group students by similarity of XP acquisition. Expectation-Maximization (EM) [57] was the selected clustering algorithm; this is reasonable given our small sample size on both years and that the number of clusters was not known beforehand [58]. The course lasted for 139, 142 and 156 days in the first, second and third years respectively.

Usually, during the first days of class, most students have residual XP, either because they are not fully enrolled in the course or because there was still no significant activity. This makes all students' activity look alike, which might mislead the clustering algorithm to group all students into the same cluster. To overcome this, we excluded from the analysis the first days that satisfied at least one of the following criteria: a) there were still students whose enrollment process was not completed and, thus, were not yet playing the game, and b) students were tied up at zero score due to lack of initial activity. As a result of this filtering process, 4, 15 and 16 days were excluded from the first, second and third years respectively. In the first year there was no significant student activity during the first four days. In the second year, professors were attending a conference between the first and the second week, which limited student activity. In the third, some of the students only became fully enrolled in the course in the second week. We also excluded those who dropped the class mid-term and were not able to complete the course.

Three clusters were identified in the first year and all of them seem to describe different XP acquisition patterns. However, four clusters were observed both in the second and in the third years, but their characteristics diverged between the two instances. This suggests that different students adopted different strategies and adapted differently to the changes made to the game. We observed that some clusters seem to have a representation in more than one year. However, because each term is different

from another (i.e., different amount of days, assignments occurring in different dates, etc.), clusters from different years were generated using different sets of rules, which encapsulate different XP accrual patterns. What is more, it makes it impossible to use a trained model with data from another term so long as the number of days differs. Therefore, we had to compare clusters based on their accumulation curve and performance metrics, and consider those exhibiting the same patterns as instances of the same cluster. We acknowledge this is a limited approach, which does no guarantee that two representations of the same cluster, in two different years, characterize in fact the same entity. Nonetheless, we believe this is a viable method, providing fruitful discussion around what differentiates students in a gamified learning environment.

In this section we will describe all the identified clusters, taking into account several student performance and participation measures, and also qualitative feedback collected by the end of the term, via a satisfaction questionnaire. Given that normality could not be assumed due the clusters' small size, we checked for differences between the data using a Kruskal-Wallis test, with post hoc Mann-Whitney's U tests and Bonferroni correction. The results will be discussed ahead in the Discussion section.

4.1. The First Year

In the first year our course had 35 students and three distinct clusters were identified. Figure 8 shows each cluster's average ranking per day and Figure 9 the average accumulated XP. In the latter, we presented both the students' average XP accumulation curve, but also one for what we call the "Ideal" student, a hypothetical ideal student who gradually earns XP and achieves the maximum by the end of the course.

The Ideal student's XP accrual was included in Figure 9 to give a comparison basis relative to what would ideally be expected of our students. Such a student would eventually get the maximum XP of the course. To this end, we assumed that the Ideal student would complete all achievements and score a perfect grade on all evaluation components. However, she should also "rationality" distributed the workload throughout the semester, whenever possible. We believe the Ideal student's behavior would be best modeled using the Rational Choice Theory, where rational choices would be compared based on the costs and benefits of different courses of action [59]. However, in the absence of data on how students managed their time, using the Rational Choice Theory was impractical. To overcome this problem, we compromised and assumed the Ideal student would manage his time using the following criteria: 1) for achievements that had a deadline (usually a couple of weeks), the Ideal student would do it soon, by

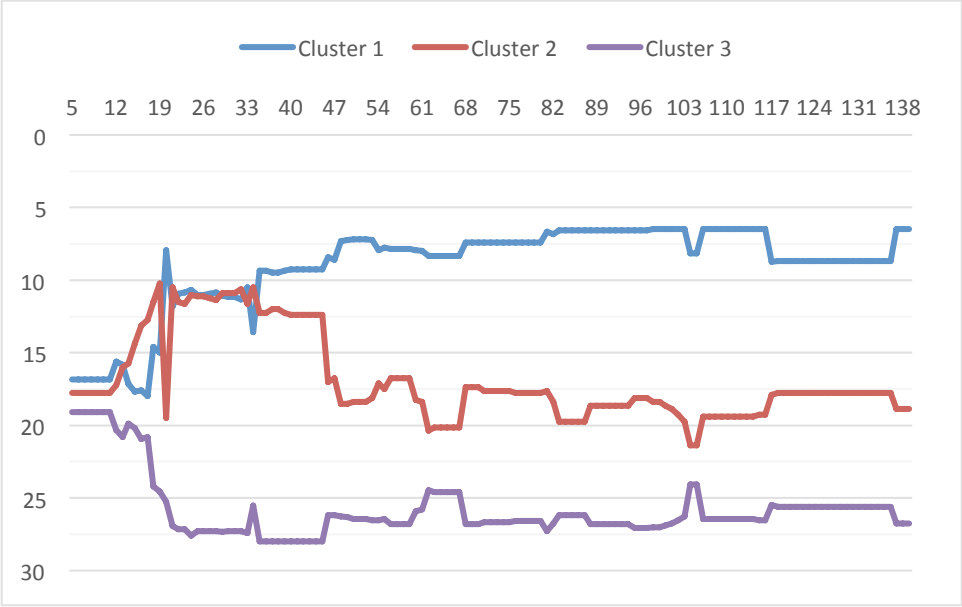

Figure 8. Average student leaderboard rank in the first year.

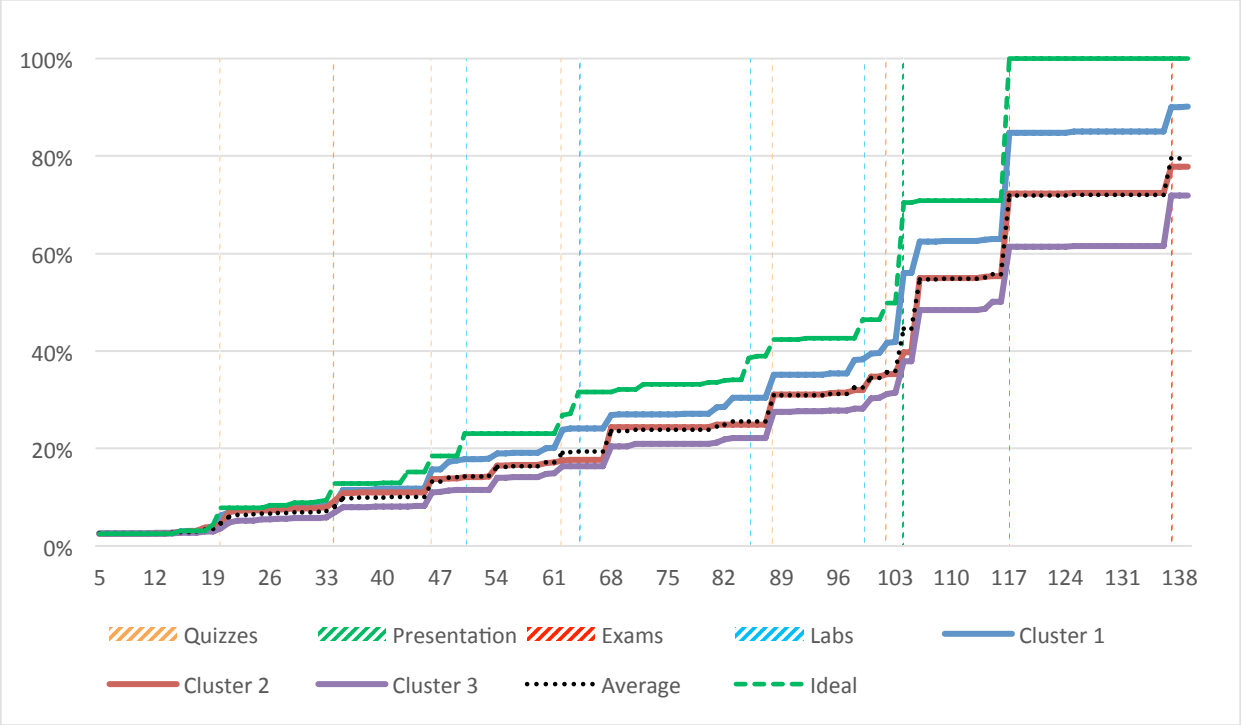

Figure 9. Average accumulated XP by students in the first year.

Friday of the week in which that achievement was issued; 2) for achievements that did not have a deadline and could be completed at the beginning, in middle or at the end of the course, we assumed that the Ideal student would earn one badge per month (1st of March, April and May); 3) evaluation checkpoints like quizzes, the presentation and the exams were done in the respective day.

We provide detailed information about every cluster's performance in Table B.1, where each row presents the mean value of a metric, for every cluster and for the whole student population. Colors reflect a scale from green to red, where green represents the cluster with the highest value, red the one with the lowest, and yellow the one in between. The last column shows between which clusters there were statistically significant differences, with green reflecting metrics where differences were significant among all clusters, yellow representing those were there were only differences between some of the clusters, and red when no significant differences were observed. The last rows whose name begin with a "[A]" denote the mean amount of badges earned for the respective achievement.

The first cluster consisted of 12 students that are characterized by a larger and steeper average XP accumulation curve, way above the average (see Figure 9), which suggests that these students were most of the time ahead of others. This is corroborated by their mean ranking in the leaderboard, which shows they clearly dominated the top (see Figure 8). Every slope in the curve represents a moment where a lot of XP was gathered, and these students seem to have tried to grab every XP available to them. They were closer to the ideal student than any other cluster. Performance-wise, they also seem to have scored better than the other clusters on all measures (see Table B.1). We named these students the *Achievers*.

The second cluster consisted of eight students that started the course collecting XP at rates, similar to Achievers, but soon fell behind and stabilized with a lower XP acquisition rate. This caused a sharp drop in leaderboard rank, thus failing to keep up with the Achievers. For these reasons we named them the *Disheartened* students. Indeed, this cluster's average XP accumulation curve was the closest to the observed mean. Regarding measured performance variables, these students also performed worse than the achievers, and closer to the mean. The third cluster contained 14 students, typically with the lowest XP accumulation curve and with the fewest slopes, who spent most of the time at the bottom of the leaderboard. These students also scored lowest on all measurements. This suggests that most of these

Table 2. Performance measures with significant differences in their mean value, between one cluster and all other, in the first year.

Achievers	Disheartened	Underachievers
Final Grade (%)	Final Grade (%)	Final Grade (%)
Labs Grade (%)		
Presentation Grade (%)		
Theoretical Challenge Posts (#)		
Quest Posts (#)		
XP from Quests (%)		
XP from Achievements (%)		
Explored Achievements (#)		
[A] Challenger of the Unknown		

students must have had a lower level of interest and engagement with the course, which made us name them the *Underachievers*. We will discuss these results in more detail in the Discussion section.

Table 2 summarizes performance variables for which significant differences in their mean value were observed between one cluster and all others. We conclude that the final grade is the only feature whose differences were significant among all clusters, making it the best differentiating factor. Other measures show significant performance differences between Achievers and other clusters. These include grades from the Labs, Multimedia Presentation and Exam, posts in Theoretical Challenges, total posts and XP from the Online Quests, XP earned from achievements, explored achievements, and badges collected in the Challenger of the Unknown achievement.

By the end of the first year (and all other too), students replied to a satisfaction questionnaire where they had to rate statements using a five-point Likert scale. Table C.1 shows all the items in the questionnaires and in which years each question was used. All but seven students replied to a satisfaction questionnaire in the first year and their responses are depicted in Table 3. Here, "x̃" denotes the response's median, "Mo" the mode, and "Mo %" the amount of students in the mode. The first column contains a question code that can be cross-checked with Table C.1 for more details about the question. Of the 28 respondents, 11 were Achievers (92%), 5 were Disheartened (63%) and 12 were Underachievers (80%).

Overall, all students seem to have equally found that the gamified experiment performed well. The Achievers were the ones considering the course as being more motivating as compared to other regular courses, but all students seem to have shared this overall opinion. They also considered the course as being more interesting and as requiring more work than other non-gamified courses, but as being as difficult. They also found the course as not being easier nor harder to learn from, with only 35.7% of students rating this item above 3, where 1 was much less [easier], 5 much more [easier] and 3 was a neutral response between both. All clusters considered that the study performed in this course was more continuous than that performed in other courses, and they also considered it to be made in more quantity, with the exception being the Disheartened students, who presented a more neutral view. In general, our students had a mild feeling they were actually playing a game instead of just attending a course, but Achievers seem to have had the strongest game-like experience.

When asked if the achievements should account for a greater part of the grade, overall student opinion seems to have been neutral. While the Underachievers appear to have been more opposed to this idea, Disheartened students were more in favor, which suggests that for some the achievements were not worth the effort. This encouraged us to rethink some of the achievements and improve their balance in the next year. Students were also asked if, when faced with non-mandatory tasks that would earn them an achievement (those earning extra XP), they would do them more for the grade or for the game. Most students had a neutral position regarding this matter, considering they would do them as much for the

Table 3. Student responses for first year's satisfaction questionnaire.

Questions (code)	Achievers			Disheartened			Underachievers			All			Significant Differences ($p < 0.016$)
	$\tilde{x}$	Mo	Mo %	$\tilde{x}$	Mo	Mo %	$\tilde{x}$	Mo	Mo %	$\tilde{x}$	Mo	Mo %	
Overall	4.0	4	55%	4.0	4	80%	4.0	4	50%	4.0	4	57%	none
Motivation	5.0	5	91%	4.0	4	80%	4.0	4	50%	5.0	5	57%	(A, B), (A, C)
Interest	4.0	4	64%	4.0	4	100%	4.0	4, 5	42%	4.0	4	61%	none
Work	4.0	4	64%	3.0	3	60%	4.0	4	50%	4.0	4	54%	none
Difficul	3.0	3	55%	3.0	3	80%	3.0	3	67%	3.0	3	64%	none
Easiness	4.0	3, 4	36%	3.0	3	60%	3.0	3	50%	3.5	3	46%	none
Quantity	4.0	4	45%	3.0	3	80%	3.0	3	50%	3.0	3	46%	none
Continuity	4.0	4	64%	4.0	4	80%	3.5	4	50%	4.0	4	61%	(A, C)
Game Feeling	4.0	4	45%	3.0	3	60%	3.0	3	42%	3.0	3	39%	none
Achievement Grade	3.0	2	36%	3.0	3, 4	40%	3.0	1	33%	3.0	3	29%	none
Non-mandatory	3.0	4	36%	2.0	2	80%	3.0	4	42%	3.0	4	32%	none
Level Distribution		1	73%		3	60%		1	42%		1	54%	none
Extra Actions	4.0	4	45%	4.0	4	60%	3.5	4	33%	4.0	4	43%	none
Extend	5.0	5	55%	4.0	4	40%	4.5	5	50%	4.0	5	46%	none

grade as they would for the game. However, while the Disheartened students seemed to have a slight tendency towards the grade, the other two clusters seem to have one towards the game.

In our course, XP distribution per achievement level had a decreasing factor, where the first levels were worth more than the subsequent levels. When asked if XP distribution should be changed to the same amount of XP per level or to use an incremental factor, most students affirmed they would rather leave it as it was. The exception were the Disheartened, who affirmed that the last levels should be worth more XP. The majority of students also considered that achievements that required extra actions, such as the Theoretical challenges and the Quests, contributed to their learning experience, and that it would be a good idea to extend gamification to other courses. Of all questions, only the one regarding how motivating the course is had significant differences between one cluster (the Achievers) and the other two.

4.2. The Second Year

In the second year, 52 students completed the course and four student clusters were identified. Details about their ranking over time are provided in Figure 10, and about the average XP acquisition patterns in Figure 11. An Ideal student was again computed, following the same approach as the previous analysis. Table B.2 provides detailed information about performance differences between clusters.

The first cluster presented the highest XP accumulation curve, with the steepest slopes, which suggests that these students took advantage of the most chances to get any additional XP (see Figure 11). Consequently, this had a direct reflex in their ranking. These students dominated the first half of the leaderboard since the beginning of the game. They also had the best performance on all measured performance variables, with the exception being the number of posts on the Online Quests. However, they still managed to get the most XP out of this component. Given the several resemblances to the Achievers from the first year, we kept the name *Achievers* in the second year for this cluster. Interestingly, the Achievers were now much closer to the Ideal student, especially at the beginning and at the ending of the course (see Figure 11).

A cluster with similar features to the Underachievers of the previous experiment was observed this year as well (Cluster 4). It consisted of 13 students, which were characterized by having the lowest XP accumulation curves, with the fewest and smallest slopes, which made them typically occupy the lower positions on the leaderboard. They also scored the poorest on most measured performance variables. An exception was the mean presentation grade where cluster #3 seems to have performed

Joaqui
Delet

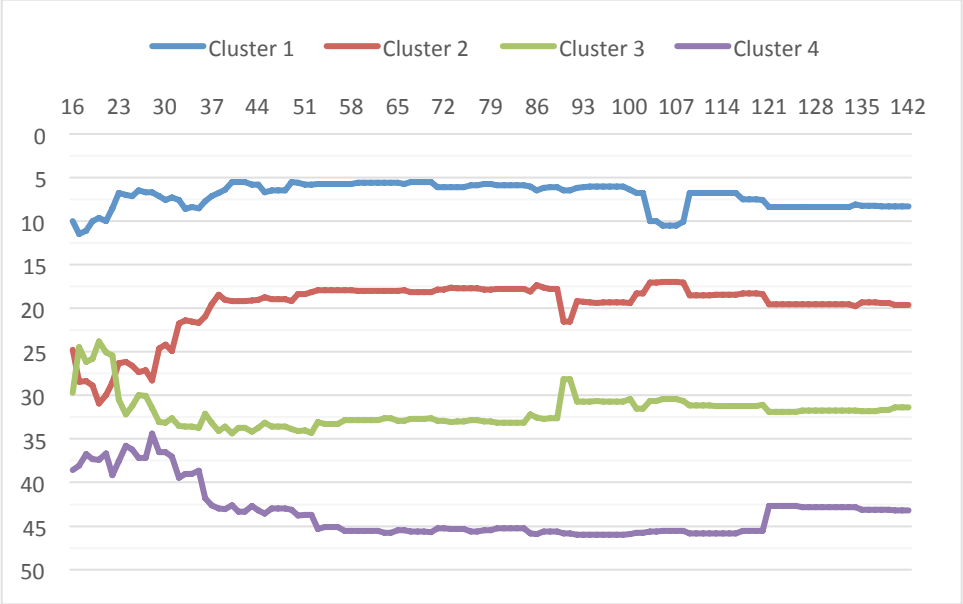

Figure 10. Average student leaderboard rank in the second year.

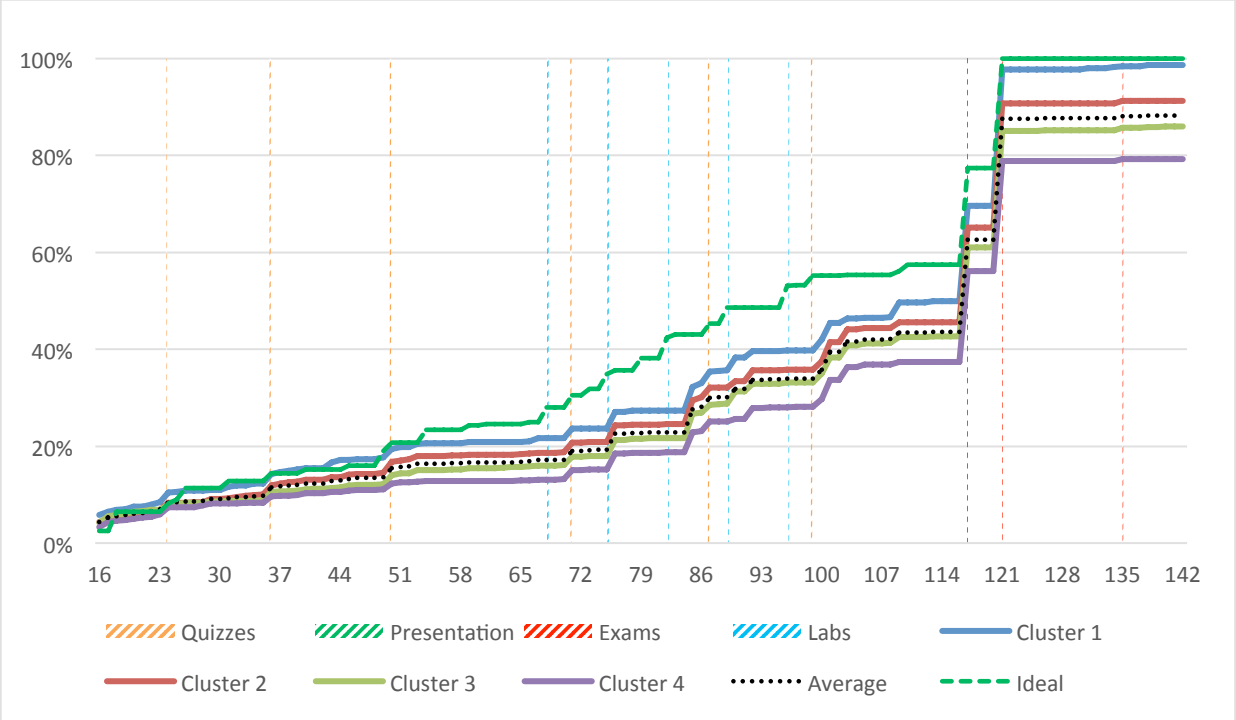

Figure 11. Average accumulated XP by students in the second year.

marginally worse. However, this difference was not significant. Given the remarkable similarities, these students kept the name *Underachievers*.

Fourteen students composing Cluster 3 resembled the Disheartened students, by presenting an XP accumulation curve closer to that of the achievers during the first weeks and then assuming a middle-ground between the Achievers and the Underachievers. Moreover, their positioning in the leaderboard

was also very similar over the term. The resemblances made us name these the *Disheartened* students. These students presented below average performance on most measurements.

Cluster 2 was first observed in the second year, which consisted of 15 students. These students were characterized by an XP accumulation curve that shared some similarities with that of the Achievers and the Disheartened students from the previous year. However, these students accumulated in average more XP than the latter student type, but less than the former. The average leaderboard position of this cluster is similar to the inverse of that of the Disheartened students, which is translated in what seems to be like a fast rank loss in the beginning of the course followed by a progressive recovery (see Figure 10). This led us to name this new student type *Late Awakeners*. We hypothesize the emergence of this cluster might be related to the existence of additional opportunities for students to participate and recover lost XP, but we will further address this issue in the Discussion section. These students were also characterized by scoring above average on most performance measures.

Table 4 presents all performance features for which significant differences in their means were observed between one cluster and all other. Significant differences between all clusters were observed for final grade, the amount of earned badges, XP from achievements, and the number of explored achievements, which suggests these are the traits that best differentiate them. It is also possible to observe that the Achievers have significant differences to all other clusters for the number of posts, the amount of explored Skill Tree nodes and the amount of XP earned from the Skill Tree. Disheartened students also significantly differed from other students in the amount of badges collected from the post Postmaster achievement (do plenty of posts). Several features significantly differed the Underachievers from the other clusters, like the number of posts, reply posts, Challenge posts and Theoretical Challenge posts, the amount of XP earned from Challenges and Theoretical Challenges, and the number of badges acquired in the Postmaster and Rise to the Challenge achievements (complete Theoretical Challenges).

Out of 52 students, 45 answered the satisfaction questionnaire, where 8 where Achievers (80%), 13 where Late Awakeners (87%), 12 were Disheartened (86%) and 12 were Underachievers (92%). in Table 5 shows response statistics. In general, all students considered that gamification applied to the course went very well, with Disheartened students presenting a more neutral opinion on this subject. They also affirmed the course was more motivating an interesting than other regular courses, with Achievers finding it the most motivating and Disheartened considering it the least interesting. Students found the course as requiring a larger amount of work in comparison to other courses, but they considered it as being as difficult. Students also found it easier to learn from, with 53.3% of the students rating it above 3. This time, most students considered that they studied more for this course in comparison to other courses, except for the Achievers. Nonetheless, all students agreed that the study was more continuous. Students found that the distribution of the amount of work asked from them during the semester to be even, and they also considered even the amount of work they put in the course.

Table 4. Performance measures with significant differences in their mean value, between one cluster and all other, in the second year.

Achievers	Late-Awakers	Disheartened	Underachievers
Final Grade (%)	Final Grade (%)	Final Grade (%)	Final Grade (%)
Skill Tree Posts (#)	Badges (#)	Badges (#)	Posts (#)
XP from Skill Tree (%)	XP from Achievements (%)	XP from Achievements (%)	Reply Posts (#)
Explored Skill Tree Nodes (#)	Explored Achievements (#)	Explored Achievements (#)	Challenge Posts (#)
Badges (#)		[A] Postmaster	XP from Challenges (%)
XP from Achievements (%)			Theoretical Challenge Posts (#)
Explored Achievements (#)			XP from Theoretical Challenges (%)
			Badges (#)
			XP from Achievements (%)
			Explored Achievements (#)
			[A] Postmaster
			[A] Rise to the Challenge

Just like in the previous year, students mildly felt they were playing a game, with Achievers having the most game-like experience and Underachievers having the least. When asked if the achievements should account for a greater part of the grade, overall student opinion was positive. While the Underachievers seem to have been more opposed to this idea, the Achievers were more in favor. When asked if they would do non-mandatory tasks more for the grade or for the game's sake, most students affirmed they would do it for the grade, with Achievers having a more neutral opinion between both. When asked if XP distribution should be changed, 40% considered the first level should be worth less XP and the later level worth more. However, 50% of the Achievers and 42% of the Disheartened students considered that XP distribution should remain the same. Most students found that achievements that required extra actions contribute to their learning experience and also that it would be a good idea to extend gamification to other courses. However, here both the Late Awakeners and the Underachievers seem to have had a rather negative opinion. The general impression of our students was that the course was more competitive rather than collaborative, but it still allowed them to be rewarded for things the enjoyed doing, to be creative, and to learn useful skills for the future. The Achievers considered the Skill Tree to be great, whereas the Late Awakeners did not like it. The other two clusters gave it a neutral rating. Rating the Skill Tree was the only item in the questionnaire whose differences between a cluster's response (the Achievers in this case) and that of the other three were significant.

Regarding the newly introduced achievements, students had mixed feelings towards them. The achievements promoting collaboration were equally and moderately adopted by all student types (avg. 2 badges). However, instead of collaborating, students often blamed others for not getting the respective badges instead of helping them. It turns out helping others at the expense of one's own time did not sound a good idea to our students (even when they would end up benefiting from it themselves). The achievement promoting prompt challenge response was mostly ignored, with Achievers earning the most

Table 5. Student responses for second year's satisfaction questionnaire.

Questions (code)	Achievers			Late Awakeners			Disheartened			Underachievers			All			Significant Differences
	$\tilde{x}$	Mo	Mo %	$\tilde{x}$	Mo	Mo %	$\tilde{x}$	Mo	Mo %	$\tilde{x}$	Mo	Mo %	$\tilde{x}$	Mo	Mo %	**(p < 0.008)**
Overall	4.0	4	63%	4.0	3, 4	31%	3.0	3	42%	3.5	4	42%	4.0	4	40%	none
Motivation	5.0	5	63%	4.0	4, 5	31%	4.0	3, 4	33%	3.5	4	50%	4.0	4	38%	(A, D)
Interest	4.0	4	63%	4.0	4	38%	3.5	3	50%	3.5	4	33%	4.0	4	42%	none
Work	5.0	5	75%	4.0	4, 5	38%	4.0	4, 5	33%	4.0	4	50%	4.0	5	40%	none
Difficul	3.0	3	75%	3.0	2, 3	31%	3.0	3	58%	3.0	3	50%	3.0	3	51%	none
Easiness	4.0	4	75%	3.0	3	46%	4.0	4	50%	3.5	4	50%	4.0	4	44%	none
Quantity	4.0	4, 5	38%	3.0	3	54%	3.5	2, 4	33%	3.0	3	58%	3.0	3	40%	none
Continuity	5.0	5	75%	4.0	5	38%	4.0	4	58%	4.0	4	42%	4.0	4	38%	(A, C), (A, D)
Workload	4.0	4	63%	3.0	3	38%	3.5	4	50%	3.0	1, 3	25%	3.0	4	33%	none
Workload MCP	4.0	4	38%	3.0	3	31%	4.0	4	50%	3.0	2	42%	3.0	4	36%	none
Game Feeling	4.0	4	50%	3.0	3	54%	3.0	3	33%	2.0	3	42%	3.0	3	40%	(A, D)
Achievement Grade	4.5	5	50%	4.0	5	46%	4.0	4	42%	3.0	3	33%	4.0	5	29%	none
Non-mandatory	3.0	3	38%	2.0	1	38%	2.0	1, 2	33%	2.0	1	42%	2.0	1	36%	none
Level Distribution		1	50%		3	54%		1, 2	42%		3	50%		3	40%	none
Extra Actions	4.5	4, 5	50%	3.0	3, 4	31%	4.0	4, 5	33%	3.0	3, 5	25%	4.0	4	31%	none
Extend	4.0	4	75%	3.0	1	38%	4.0	4	50%	2.0	2	33%	3.0	4	33%	none
Competitive	4.0	4	63%	4.0	4	38%	5.0	5	58%	4.5	5	50%	4.0	5	36%	none
Rate Skill Tree	4.5	5	50%	2.0	2	54%	3.0	3	50%	3.0	3	42%	3.0	3	29%	(A, B)(A, C), (A, D), (B, C), (B, D)
Things I Like	5.0	5	63%	4.0	3	38%	3.5	4	42%	4.0	4	42%	4.0	4	38%	(A, C)
Creative	5.0	5	75%	5.0	5	62%	4.0	4	83%	4.5	5	50%	4.0	5	47%	none
Useful	5.0	5	63%	4.0	4	54%	4.0	4	50%	4.0	4	42%	4.0	4	44%	none

badges (avg. 1.1), while other clusters earned only 0.5 or less. This achievement was highly criticized for encouraging students to post fast in detriment of post quality. The achievement for compiling challenge results was the least popular and only one Disheartened student participated. Students considered this achievement to be poorly rewarded and unappealing.

The achievement promoting oral participation was well received but had little adoption, with the Achievers earning the most badges (avg. 1.6) and the Late Awakeners earning the second most (avg. 0.93). The other two student types earned only 0.2 badges. Even though this achievement encouraged students to participate in class, thus adding value to the learning experience, a few claimed they did not enjoy it because they did not like to talk in class. On the other hand, the recently added Skill Tree had little success, given its optional character and late introduction into the game. Only eight Achievers, four Late Awakeners, one Disheartened and one Underachiever made Skill Tree posts and from these, only seven Achievers and one Disheartened earned any XP.

4.3. The Third Year

In the third year we had 54 students and four student types were again identified. Details about their average XP acquisition can be found in Figure 13 and about their average ranking in Figure 12. Table B.3 shows how each cluster scored on all measured performance variables. Two of the clusters presented characteristics closer to two student types seen in the second year, which were the Achievers (Cluster 1) and the Underachievers (Cluster 4). The Achievers of the third year were composed by 7 students, typically had a high and steep XP accumulation curve (see Figure 13) and also led the leaderboard (see Figure 12). They had the best performance on most performance features, with exceptions being the XP earned from the Multimedia Presentation and from achievements like the Talkative (oral participation in class) and Lab Master (having top grades in lab classes), where the second cluster appears to have performed better. However, these differences were not significant. The Underachievers, composed by 11 students, were still characterized by the lowest XP accumulation curve of the four clusters, by occupying the bottom of the leaderboard, and by having the lowest performance on all measured variables. Even though we still identified two other clusters in the third year, they did not share the same XP accumulation pattern with the Late Awakeners and the Disheartened students, identified in the second year.

One of the newly identified clusters contained 23 students (Cluster 2). It was the largest of all

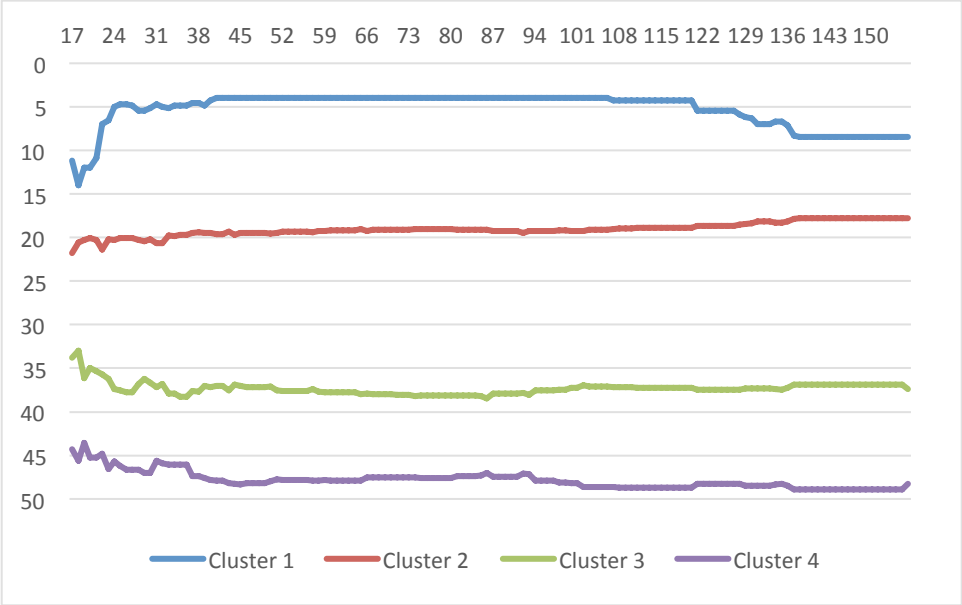

Figure 12. Average student leaderboard rank in the third year.

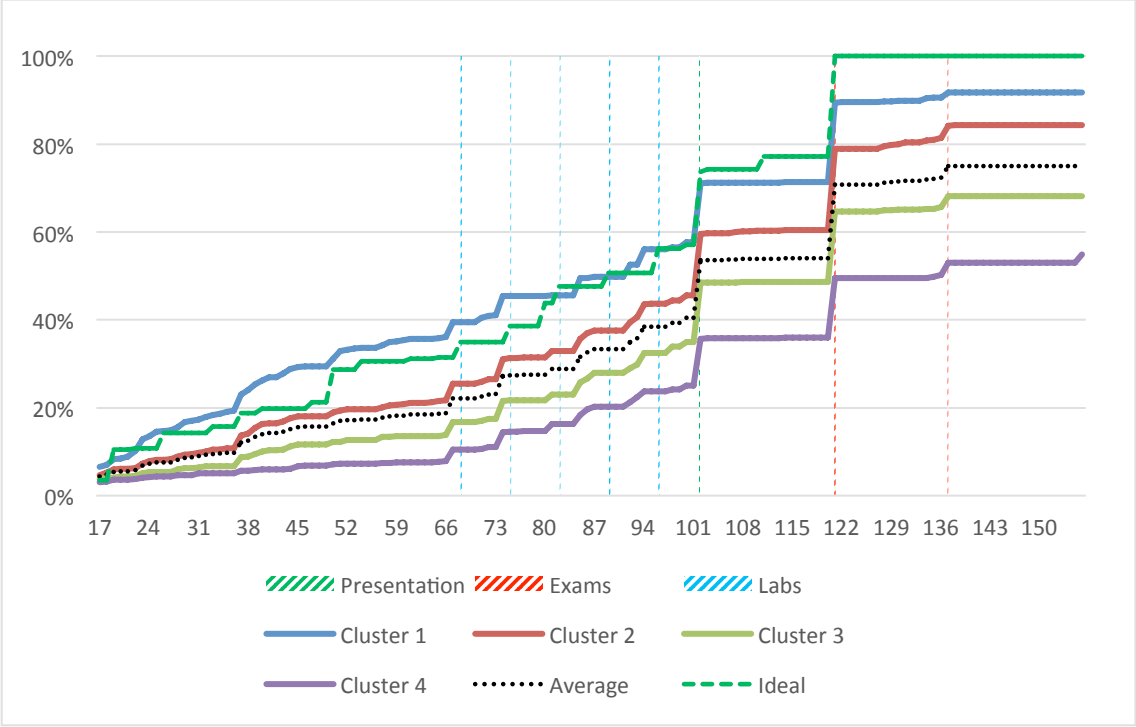

Figure 13. Average accumulated XP by student in the third year.

clusters seen so far, containing as much as 43% of our student population. It was characterized by an XP accumulation curve closer to that of the Achievers, resultant of an above average score on most performance variables. Because this cluster was so large, it comprised students that had grades that were both lower and much closer to those of the Achievers. This is corroborated by the lack of significant between the final grades of those two clusters (see Table B.3) and by how close their mean ranking became by the end of the course. Because of these traits we named these the *Regular students*.

The other new cluster was composed by 13 students, who presented an XP accrual between that of the Regular students and the Underachievers. These students present below average scores on most performance features, and seem to have avoided some aspects of the course that went beyond Challenges and the MCP Quest, such as the Skill Tree. They were typically positioned in the leaderboard right above the Underachievers. We named these the *Halfhearted students*. The emergence of these two clusters and the conditions that led to the fall of the Late Awakeners and Disheartened students will be addressed in the Discussion section.

Figure 13 shows a different relationship between the clusters and the Ideal students, as compared to the previous years. The Achievers were ahead of the Ideal student for half of the semester, which suggests they went after some of the achievements earlier than we expected. However, they lagged behind the Ideal student once they reached major evaluation checkpoints like the Multimedia Presentation and the Exams. On the other hand, the Regular students were able to take advantage of the second Exam to pull their grades up. Surprisingly, this year we observed what seems to be a wider discrepancy between the clusters' XP accumulation curves, which suggests that the clusters might have differed more and earlier. We will address this issue in the discussion section.

Table 6 shows that there were significant differences between all clusters for the number of badges, completed and explored achievements. It is also noticeable that there are a couple of them whose differences are not significant between the Achievers and the Regular students but are significant between these two and the Halfhearted and the Underachievers (and between these two as well), such as the final

Table 6. Performance measures with significant differences in their mean value, between one cluster and all other, in the third year.

Achievers	Regular	Halfhearted	Underachievers
Final Grade (%)*	Final Grade (%)*	Final Grade (%)	Final Grade (%)
Posts (#)	Posts (#)	Badges (#)	Badges (#)
Reply Posts (#)	Reply Posts (#)	XP from Achievements (%)	XP from Achievements (%)
Rated Posts (#)	Rated Posts (#)	Completed Achievements (#)	Completed Achievements (#)
Skill Tree Posts (#)	Skill Tree Posts (#)	Explored Achievements (#)	Explored Achievements (#)
XP from Skill Tree (%)	XP from Skill Tree (%)	Posts (#)**	Posts (#)**
Explored Skill Tree Nodes (#)	Explored Skill Tree Nodes (#)	Reply Posts (#)**	Reply Posts (#)**
AvatarWorld Posts (#)	Badges (#)	Rated Posts (#)**	Rated Posts (#)**
XP from AvatarWorld (%)	XP from Achievements (%)*	Skill Tree Posts (#)**	Skill Tree Posts (#)**
AvatarWorld Submissions (#)	Completed Achievements (#)	XP from Skill Tree (%)**	XP from Skill Tree (%)**
Badges (#)	Explored Achievements (#)	Explored Skill Tree Nodes (#)**	Explored Skill Tree Nodes (#)**
XP from Achievements (%)*	[A] Postmaster	[A] Postmaster**	[A] Postmaster**
Completed Achievements (#)			
Explored Achievements (#)			
[A] Postmaster			
[A] Blacksmith			
[A] Master Builder			
[A] Squire			

* Not significat between the Achievers and the Reguar students.

** Not significant between Halfhearted and Underachievers.

grade and the amount of XP earned from achievements. This suggests that the Achievers and the Regular have comparable performance levels on these two measures. There are other measures that can discriminate the Achievers and the Regular students from all the other clusters, but that cannot differentiate the Halfhearted from the Underachievers, such as the total number of posts, reply posts and rated posts, the amount of posts, explored nodes and XP earned from the Skill Tree, and the number of badges earned from the Postmaster achievement. We also found that features like the amount of posts, submissions and XP earned from AvatarWorld, and the amount of badges gathered from the Blacksmith, the Master Build and the Squire achievements presented significant differences between the Achievers and all of the other clusters. Interestingly, a closer look at Table B.3 shows that the amount of badges earned from the Artist achievement seems to separate the Achievers and the Regular from the Halfhearted and the Underachievers. We will discuss this later.

Joaqui
Delet

Forty-seven students replied to the satisfaction questionnaire, of which 7 were Achievers (100%), 23 were Regular students (100%), 11 where Halfhearted (85%) and 6 where Underachievers (55%), and the results can be found in Table 7. In general, all students considered the gamified experiment performed very well, with the Underachievers having a more neutral position here. Most students considered the course to be more motivating and interesting than other courses, although three of the Achievers considered it to be less interesting. Students found that our course required more work than other courses, but also that it was not more difficult. They also found our course as being easier to learn from, with 61.7% of them rating it above 3. Students considered that the distribution of the amount of work asked from them during the semester to be slightly uneven, and they also considered somehow uneven the amount of work they put in the course. Just like the other years, students had a mild game-like feeling in our learning experience. However, contrary to previous years, Achievers reported the experience as feeling less game-like, while Regular students seem to have a more positive opinion. When faced with non-mandatory tasks, students presented a neutral position between doing it more for the grade or for the game's sake. However, Achievers seem to have done it more for the grade and Underachievers more for the game. Most students considered that achievements that required extra actions contributed to their learning experience and also that it was a good idea to extend gamification to other courses.

Table 7. Student responses for third year's satisfaction questionnaire.

Questions (code)	Achievers			Regular			Halfhearted			Underachievers			All			Significant Differences ($p < 0.008$)
	$\tilde{x}$	Mo	Mo %	$\tilde{x}$	Mo	Mo %	$\tilde{x}$	Mo	Mo %	$\tilde{x}$	Mo	Mo %	$\tilde{x}$	Mo	Mo %	
Overall	4.0	4	43%	4.0	4	57%	4.0	4	55%	3.5	3	50%	4.0	4	51%	none
Motivation	4.0	4	43%	4.0	4, 5	48%	4.0	4	45%	4.0	4	67%	4.0	4	49%	none
Interest	4.0	2	43%	4.0	4	52%	4.0	4	36%	4.0	4	67%	4.0	4	47%	none
Work	5.0	5	57%	4.0	4	52%	4.0	4	64%	4.0	4, 5	33%	4.0	4	49%	none
Difficul	3.0	3	43%	3.0	3	48%	3.0	3	55%	3.5	3, 4	50%	3.0	3	49%	none
Easiness	3.0	3	71%	4.0	4	70%	3.0	3	45%	4.0	4	67%	4.0	4	51%	(A, B)
Quantity	4.0	3, 4, 5	29%	3.0	3	39%	3.0	3	73%	4.0	4	50%	3.0	3	45%	none
Continuity	3.0	3	57%	4.0	4	57%	3.0	4	45%	3.0	3	67%	4.0	4	47%	none
Workload	3.0	1, 5	29%	4.0	4	52%	4.0	4	64%	3.0	3	33%	4.0	4	45%	none
Workload MCP	3.0	1, 4	29%	4.0	4	52%	3.0	4	36%	2.0	2	50%	3.0	4	38%	(B, D)
Game Feeling	2.0	1, 2, 3	29%	3.0	4	39%	3.0	3	45%	3.0	3	50%	3.0	3	36%	none
Learn Traditional		2, 3	43%		2	83%		2	64%		2	67%		2	70%	none
Grade Traditional		2	71%		2	78%		2	45%		2	67%		2	68%	none
Exam vs. Quizzes	5.0	5	100%	5.0	5	65%	4.0	5	45%	4.5	5	50%	5.0	5	64%	none
Non-mandatory	2.0	1, 2, 3	29%	3.0	3	48%	3.0	3	55%	3.5	4	50%	3.0	3	43%	none
Extra Actions	4.0	4	57%	4.0	4	43%	3.0	3	45%	3.5	1, 5	33%	4.0	4	40%	none
Extend	4.0	4	43%	4.0	4	39%	3.0	3, 5	27%	4.0	4	67%	4.0	4	36%	none
Competitive	5.0	5	71%	3.0	3	30%	3.0	3	55%	4.0	3, 5	50%	3.0	3	36%	(A, C)
Things I Like	4.0	5	43%	4.0	5	43%	3.0	2, 3, 4	27%	4.0	4	50%	4.0	4, 5	34%	none
Creative	4.0	4	43%	5.0	5	57%	4.0	5	45%	4.0	4	50%	4.0	5	47%	none
Useful	3.0	1, 3, 4	29%	4.0	4, 5	43%	3.0	3	45%	4.0	4	50%	4.0	4	38%	(A, B)
The Quest	2.0	4	43%	5.0	5	52%	4.0	5	45%	3.0	1, 3, 5	33%	4.0	5	40%	(A, B)
Engaged AW	1.0	1	71%	2.0	1	43%	1.0	1	64%	1.0	1	67%	1.0	1	55%	none
Engaged Badges	3.0	1, 4	29%	4.0	4	70%	3.0	3	36%	4.0	4	33%	4.0	4	45%	none
Engaged Challenges	3.0	1, 3, 4	29%	5.0	5	52%	4.0	4	45%	5.0	5	50%	4.0	5	38%	none
Engaged XP	5.0	5	57%	4.0	5	48%	4.0	4	55%	4.0	4, 5	33%	4.0	4, 5	40%	none
Engaged Labs	3.0	3	43%	5.0	5	57%	4.0	4	45%	4.5	5	50%	4.0	5	47%	none
Engaged Leaderboard	5.0	5	57%	4.0	4	39%	3.0	3	45%	4.0	4	50%	4.0	4	34%	(A, C)
Engaged Levels	5.0	5	57%	4.0	4	43%	3.0	3	55%	4.0	4	50%	4.0	4	36%	(B, C)
Engaged Quest	3.0	1, 5	29%	4.0	5	43%	4.0	4	36%	3.5	1, 4	33%	4.0	5	34%	none
Engaged Skill Tree	4.0	5	43%	4.0	4	39%	3.0	4	36%	4.0	4, 5	33%	4.0	4	36%	none
Engaged Lectures	2.0	1, 3	43%	4.0	4	43%	2.5	1	27%	3.0	3	50%	3.0	3	32%	(A, B)

Most students mentioned they would have not learned more from a more traditional course. The Achievers, however, were more divided between "no" and "about the same". Notwithstanding, most of the students affirmed that they would not get a better grade on a more traditional course. We asked students what they would prefer: to have just one exam or several quizzes throughout the course, and most preferred the quizzes. Most of the students found the MCP Quest to be interesting, although the Achievers and the Underachievers seem to only consider it somewhat interesting. In general, students found the course to be as much collaborative as it was competitive, and they considered it allowed them to get rewards for things they like to do and to be creative, and also that it taught them useful skills for the future. When asked how several of our course elements made them feel engaged, they considered the MCP Quest, the XP, the Challenges and the lab classes to be the most engaging elements, but badges, the leaderboard, the levels and the Skill Tree were also regarded as engaging. The theoretical classes were considered as somewhat engaging and the AvatarWorld was considered the least engaging.

Student feedback is in line with our perception of how students played the MCP game. The Skill Tree, now in a more mature form, was explored by all student types, especially the Achievers and the Regular students. The MCP Quest was a great success. We designed the quest with twenty levels and planed that it would take several weeks for students to figure it out. It turns out that 40 students

collaborated to solve the quest in a single week, with some of them mentioning that it should have lasted longer so that they could have more opportunities to participate. The AvatarWorld seem to have highly engaged some students, especially at the beginning, but they soon grew tired of it [60]. We believe the main cause was the dissociation with the rest of the course, given its early stage of development.

We saw some fluctuation on the average final grades between the three years, which were 79.5% (stdev: 9.3), 88.3% (8.1), and 75.4% (14.4) respectively, with the highest value observed in the second year. A Kruskal-Wallis test revealed differences between the final grades on the three years ($p < .001$), and post hoc Mann-Whitney's U tests with Bonferroni correction showed that these differences were only observed between the second year and both the first ($p < 0.001$) and the third years ($p < .001$), but not between the first and the third.

5. ENGAGEMENT ASSESSMENT

In the first and second years we assessed student engagement indirectly, via their response to questions like whether they felt more motivated or interested by our course in comparison to others. In our third year we evaluated student engagement via a formal validated instrument named Student Course Engagement Questionnaire (SCEQ), which assesses student engagement with a college course [61]. The

Table 8. Descriptive statistics for SCEQ scores.

		N	Mean	Std. Deviation	Std. Error
SCEQ #1	Achievers	7	82.29	10.128	3.828
	Regular	22	81.59	11.316	2.413
	Halfhearted	11	70.09	7.752	2.337
	Underachievers	6	72.83	9.786	3.995
	Total	46	77.80	11.198	1.651
SCEQ #2	Achievers	7	78.86	13.886	5.248
	Regular	22	80.86	14.538	3.100
	Halfhearted	11	65.91	11.537	3.478
	Underachievers	6	70.83	7.441	3.038
	Total	46	75.67	14.186	2.092

Table 9. Tests for Normality for SCEQ scores.

		Kolmogorov-Smirnov[a]			Shapiro-Wilk		
Cluster		Statistic	df	Sig.	Statistic	df	Sig.
SCEQ #1	Achievers	.199	7	,200*	.903	7	.347
	Regular	.143	22	,200*	.955	22	.403
	Halfhearted	.183	11	,200*	.953	11	.686
	Underachievers	.257	6	,200*	.880	6	.271
SCEQ #2	Achievers	.181	7	,200*	.872	7	.191
	Regular	.125	22	,200*	.939	22	.192
	Halfhearted	.111	11	,200*	.977	11	.945
	Underachievers	.315	6	.063	.855	6	.173

*. This is a lower bound of the true significance.

a. Lilliefors Significance Correction

instrument includes 23 statements and students had to classify how characteristic of them each statement was using a five-point Likert scale, where one corresponded to “not at all characteristic of me” and five to “very characteristic of me”. A score could be calculated for an individual by summing all the responses, with the maximum being 115. We made six small adaptations to the questionnaire. Four of them aimed at simplifying the English given that our students are not native speakers. The other two replaced evaluation components not present in the same format in our course. The adapted questionnaire can be found in APPENDIX D and the list of adaptations in Table D.1.

We made SCEQ available on the course forums and performed two measurements of student engagement, one in the middle of the semester, during Easter break, and the other one by the end of the semester, along with our satisfaction questionnaire. The main intent of these two measurements was to understand to what extent student engagement would decrease as the course approached the end. Students had around one week to reply to both instances of the questionnaire. Forty-six students replied to both questionnaires, of which seven were Achievers (100%), 22 were Regular students (95.65%), 11 were Halfhearted students (84.62%) and six were Underachievers (54.55%). Underachievers had the largest abstention (45.45%), followed by the Halfhearted (15.38%), the Regular students (4.35%) and the Achievers (0%). We used SPSS to perform our data analysis, in order to investigate differences between student types and between measurements.

Descriptive statistics for SCEQ responses are depicted in Table 8, where “SCEQ #1” denotes the score for the mid-term SCEQ measurement and “SCEQ #2” represents the score for the SCEQ issued by the end of the course. We checked for outliers by probing for cases scoring not inside the open interval $(Q_1 - (Q_3 - Q_1) \times g, Q3 + (Q_3 - Q_1) \times g)$, where $g = 2.2$ and Q_1 and Q_3 represent the first and the third quartiles respectively [62]. No outliers were found.

5.1. Analysis of Variance

Before we could compare the means of the student responses using a parametric test, we verified the normality of the responses of all clusters to both instances of the questionnaire using both Kolmogorov-Smirnov and Shapiro-Wilk tests (see Table 9). We could not reject the null hypothesis that the population is normally distributed for any cluster ($p > .05$). We then performed Repeated Measures Analysis of Variance (ANOVA), where the responses to the first and second instances of the questionnaire were the within-subjects variables (coded as the Time factor) and the Student Type was the between-subjects factor. Besides being supplied with normal data, Repeated Measures ANOVA also assumes the homogeneity of variance and homogeneity of variance-covariance matrices. The first assumption was verified with the Levene's Test of Equality of Error Variances, and neither the first measurement nor the second violated the assumption (SCEQ #1: $F = .368$, $p > .05$; SCEQ #2: $F = .739$, $p > .05$). The second one was verified with the Box's Test of Equality of Covariance Matrices (Box’s $M = 13.016$, $F = 1.271$, $p > .05$).

By looking to the results of the multivariate tests, there was no apparent interaction effect between Time and Student Type, as none of the multivariate tests were statistically significant. These findings were confirmed by the tests of within-subjects effects where, again, no significant effect was

Table 10. Tests of Between-subjects Effects in Repeated Measures ANOVA.

Measure: MEASURE_1
Transformed Variable: Average

Source	Type III Sum of Squares	df	Mean Square	F	Sig.	Partial Eta Squared	Eta Squared
Intercept	408093.501	1	408093.501	1961.579	.000	.979	
Student Type	3060.917	3	1020.306	4.904	.005	.259	.259
Error	8737.823	42	208.043				

observed. However, between-subjects tests revealed that the differences between the mean values of the student types were significant, as shown by Table 10 ($p < .01$). Furthermore, the Eta Squared is greater than 0.14 which represents a large effect size [63].

In order to have a more discriminative view on the differences between student types for both SCEQ measurements, we performed a One-way ANOVA where SCEQ results were the two dependent variables and the factor was the student type. The results show that mean scores among student types were different for SCEQ #1 ($F = 4.019$, $p < .05$) and for SCEQ #2 ($F = 3.601$, $p < .05$). Post hoc tests using the Fishers Least Significant Difference (LSD) found significant differences between the Achievers and Halfhearted ($p < .05$) and between Regular and Halfhearted ($p < .01$). However, this test does not account for familywise errors, which are type I errors that are more likely to occur when performing multiple hypotheses tests. To overcome this problem we performed post hoc tests using Bonferroni correction, which is a conservative method to control the familywise error rate. These tests revealed that differences were only significant between the Regular and the Healfhearted students ($p < .05$). No differences were found in comparison with the Underachievers. We believe this happened because we only had 6 cases on this group, given that only around 50% of the Underachievers replied.

5.2. Discriminant Function Analysis

The Repeated Measures ANOVA showed that the mean score of the different student types was indeed significantly different, and that the effect was strong. However, a One-way ANOVA showed that these differences were only significant between the Halfhearted and the Regular students, which raised the need to further study the relationship between the SCEQ scores and the different student types. To this end, we performed a Discriminant Function Analysis (DFA), using the SCEQ scores as independent variables and the student type as the grouping variable. DFA has an additional assumption not covered by the previous ANOVAs which is multivariate normality (independent variables are normal for each level of the grouping variable). Because SPSS does not perform this kind of test, we performed the Henze-Zirkler's Multivariate Normality Test in R, which confirmed the assumption ($HZ = 0.74$, $p > .05$).

The DFA produced two canonical discriminant functions. A Wilks' Lambda test shows that the prediction of function number one is statistically significant ($p < .05$) but the same is not true for function number two. Function number one explains 97.7% of the variance and has a canonical correlation of 0.512. We then checked case-wise classification statistics to verify if there were any multivariate outliers, which could plague this kind of analysis. We achieved this by verifying if there were any cases whose squared Mahalanobis distance to their group's centroid was above a critical value following a χ^2 distribution [64]. We computed the critical value in SPSS, with 99% significance level and 2 degrees of freedom, one for each independent variable, with the result being 9.21. We found that case number 36 had a squared Mahalanobis distance to the group's centroid of 22.514, which was above the critical value. It turns out that this student had an abnormal score drop from the first to the second SCEQ measurement, from 96 to 40, which was affecting the analysis. We excluded this case and re-ran the DFA.

Like the first analysis, the second DFA produced two canonical discriminant functions. Again, a Wilks' Lambda test shows function number one's differentiation between groups is statistically significant ($p < .05$), but the same is not true for the second function (see Table 11). Function number one explains 89.5% of the variance and has a canonical correlation of 0.545. The function coefficients for SCEQ #1 and SCEQ #2 are 0.063 and 0.952 respectively, which suggests that the second SCEQ measurement has a higher predicting capability that the first one. However, the resulting structure matrix shows that both SCEQ measurements highly correlate with the function (see Table 12). Table 13 shows the mean discriminant scores for each grouping variable. The large difference between Achievers and Regular students and Halfhearted and Underachievers suggests two things: 1) the function can separate with smaller error the first two, who will likely have a higher score, from the other two, who will likely present a lower score; 2) the function will present more error distinguishing Achievers from Regular and Halfhearted from Underachievers. The resulting model could only correctly classify 80.95% of the Regular students and 63.64% of the Halfhearted students (see Table 14), which means it correctly classified 53.3% of all students. We believe these results were greatly affected by the small sample size of

Table 11. Wilks' Lambda test of the 2nd Discriminant Function Analysis.

Test of Function(s)	Wilks' Lambda	Chi-square	df	Sig.
1 through 2	.670	16.437	6	.012
2	.953	1.984	2	.371

Table 12. Structure Matrix correlating discriminant functions and independent variables.

	Function	
	1	2
SCEQ #2	,999	-.042
SCEQ #1	,774	.633

Table 13.Discriminant functions at group centroids.

	Function	
Type	1	2
Achievers	.228	.487
Regular	.546	-.118
Halfhearted	-.917	-.024
Underachievers	-.494	-.112

Table 14. DFA classification results.

	Predicted					
	Achievers	Regular	Halfhearted	Underachievers	Total	Correct (%)
Achievers	0	5	2	0	7	0.00%
Regular	0	17	4	0	21	80.95%
Halfhearted	0	4	7	0	11	63.64%
Underachievers	0	2	4	0	6	0.00%

both the Achievers' and the Underachievers' samples, which had only 7 and 6 cases respectively, but also by the large differences between group sizes.

These findings seem to complement those from our ANOVA analyses, which suggest Achievers and Regular students have similar levels of engagement, superior to that of the Halfhearted and Underachievers. However, these differences are only significant between Regular and Halfhearted.

6. DISCUSSION

Our results show that different student types can be identified in our gamified course, each representing a different performance level and approach towards the course. In this section we will describe each student type and explain what tells them apart. We will also explore how they are characterized in terms of engagement, based on data from the third experiment, and also address their evolution and adaptation to the changes the course underwent over the years.

6.1. Student Type Overview

Our experiments revealed a total of six different types of students, each with distinct behavioral patterns. In the last year, only four types were observed, differentiated by performance, post quality and engagement.

6.1.1. The Achievers

In the three years, the Achievers were characterized by having the largest and steepest XP accumulation curve, which reflects their struggle to be the best and collect every badge they can get their hands on. These were the students that were closer to the Ideal student in all years, although they were much closer in the second than in the other terms. We believe that one of the main reasons for this is that in the second year there were an additional 5% of extra grade allocated to the Skill Tree, on top of the usual 5% extra, which granted them more opportunities to score. This is corroborated by two observations: 1) from the eight students that actually got any XP from the Skill Tree, seven were Achievers; 2) once these additional 5% extra XP were removed in the third year, the distance between XP accumulation curves from the Achievers and Ideal student grew. However, we believe that unassessed population traits might have played an important role here.

We observed that the Achievers had not only the best mean final grade, but also made more posts, explored more nodes and earned more XP from the Skill Tree, made more posts, submissions and earned more XP from AvatarWorld, and earned more badges, explored more achievements and earned more XP from the achievements as well, as compared to other clusters. These differences were significant across years, with the exception of the final grade and the amount of XP received from achievements, whose mean was closer to Regular students. Although differences in terms of the number of posts were not significant, these students seem to have made more first and reply post on all years, which suggests they were the most proactive and participative.

These students' behavior remained relatively constant over the years, with significant improvements regarding their participation on the game aspect of the course. In the first year these students scored only around 78% of all XP reserved to achievements. In the second year, we created more opportunities for students to participate and to do more of what they like, which seem to have encouraged them to pursue achievements. As result, the Achievers got a staggering 97% of the XP allocated to the achievements in the second year, significantly more than the other clusters. Furthermore, they also earned 100% of the total XP allocated to the Challenges and the Online Quest. They were also the only student type that seem to have contributed substantially for the Skill Tree, having earned 30% of the allocated XP, which was significantly more than the other clusters. In the third year, we diversified the game component and created further opportunities for participation and collaboration. In this year, the Achiever student type became smaller, composed by an elite of students that excelled on all achievements. They earned all the XP from the Challenges, Skill Tree, the MCP Quest, and most of the XP from AvatarWorld (97.6%). This allowed them to receive the maximum XP allocated to achievements. Given that posts started to be rated in this year, these students also presented the highest mean rate per post, which implies that their contributions were of higher quality, even though differences were not significant.

While in the first year these students considered that the achievements should not account for a larger part of the grade, this changed in the second year, probably due to the increased weight they had in the course. As the amount of work and respective rewards increased, the achievements became more important and students started taking them more seriously. This might also explain why their opinion seems to have gradually shifted from doing non-mandatory tasks more for the game's sake in the first

year, to more for the grade's sake in the third. However, this did not come as a surprise. Ultimately students are still taking a course and they want to have the best grade possible. The Achievers also considered the course as being more motivating and interesting than other regular courses, even though they seem to have considered it less interesting in the third year. However, the average response was 3.43, which leads us to believe that the small size of the sample (N = 7) makes it hard to interpret this rubric in the third year. Data also revealed that these students seem to have been those enjoying the leaderboard and the levels the most, and that they considered the course as being competitive.

Our engagement study suggests that this group and the Regular students were the two most engaged clusters, which seems to be corroborated by the findings here reported. Even though differences between the first and the second measurements were not significant, the means suggest that these students were more engaged with the course (they were actually the most engaged), but became slightly less engaged as the course came to an end (thus becoming the second most engaged group). We have no data to further explore this issue, and we believe this situation should be verified in an upcoming iteration, in order to understand whether or not his was an isolated effect.

6.1.2. The Underachievers

Underachiever students featured the lowest XP accumulation curve and occupied the bottom of the leaderboard, which made them the furthest from the Ideal student. Although the only performance measure that can tell this group apart from the other student types consistently across years is the final grade, which was always the lowest, other discriminant features emerged in the second year. These students are also characterized by having the lowest amount of badges, the fewest XP earned from achievements, and by exploring the fewest achievements. Although these differences were significant in the second and third years, the amount of explored achievements and earned XP from achievements was not significant only between them and the Halfhearted students, which suggest both groups had similar performance in these components. These students made the fewest posts in general and the fewest first posts as well. While not consistently significant, these differences suggest that these students were the least proactive and participative.

Like the Achievers, these students remained mostly constant throughout the three years, with some noticeable changes in the third term. Between the first and the second years there were minor improvements regarding most performance and participation measures. In fact, these students presented an XP accumulation curve closer to that of the Disheartened students during the first five weeks, but this seems to be resultant of the latter performing poorly in the second year. In the third year we observed the lowest values recorded for this group. We believe this drop might be related to the emergence of the Halfhearted students, which might have included students that would otherwise have the best performance within the Underachievers. In the third year, Underachievers presented the lowest mean rate per post, which suggests their work was indeed of lesser quality, possibly a reflex of a poor engagement with a course. They also seem to have ignored several components of the course, like the MCP Quest, the Skill Tree and AvatarWorld, and only earned 44% of all XP allocated to the achievements.

Even though their performance was rather low, no direct manifestation of their possible low engagement with the course was observed in the final questionnaire, as their responses were similar to those of the other student types. They considered the course as being more motivating and interesting than other courses, and that the inherent experience was mildly game-like and competitive. However, the formal engagement assessment tells another story. Like the Halfhearted students, they seem to be less engaged with the course than the Achievers and the Regular students, which seems to be supported by their lower performance on most aspects. It appears that these students were not engaged with the course like the other students and that they did just enough to pass the course. However, our experience tells that some students enroll on additional courses to meet a particular amount of credits, which they already expect to fail or marginally pass by chance or with help of other students. We believe these students might be a significant part of the Underachievers.

6.1.3. The Disheartened students

Disheartened students were identified in the first year and persisted until the second. They exhibited an XP accumulation behavior less linear than any of the student types previously described, by performing similarly to the Achievers during the first weeks on both years, but then falling into a tier of their own, with performance levels between those of the Achievers and the Underachievers. These students presented slightly below average performance, participation levels and attendance, which lead them to typically occupy the middle of the leaderboard. In both years, their slightly below average final grade, was significantly different from the other student types. In the second year, they also significantly differed regarding the number of badges, amount of XP earned from the achievements, number of explored achievements, and number of badges earned from the Postmaster achievement, all presenting below average scores. While the differences were not significant, the low amount of reply posts and first posts suggest that these students were not very proactive and participative in general.

This student type existed only during the first two years, but it underwent some changes during that timespan. For example, the initial period during which they performed closer to the Achievers lasted for around 45 days in the first year, whilst it lasted only 25 in the second, which implies their performance drop occurred earlier in the second year. These students also experienced some performance improvements derived from the additional participation opportunities created in the second year. The students slightly felt they were playing a game more than just attending a course and thought it was competitive. However, while in the first year they found the course as being more motivating and interesting than other courses, their opinion in the second year was more neutral. Together with the previous findings, this suggests that the Disheartened students were less engaged by our second gamified instance of the course. We have two hypotheses that may explain this effect: 1) with the additional weight of the game component in the grade and the additional participation opportunities also came more competition and a heavier workload, which might have had a demotivating impact over these students; 2) A new cluster emerged – the Late Awakeners – which might have included some students that would be formerly classified as Disheartened, thus possibly changing nature of this cluster. Both hypotheses are plausible, but we cannot prove neither of them.

6.1.4. The Late Awakeners

In the second year a fourth cluster emerged, which we named Late Awakeners. These students shared similarities with both the Achievers and the Disheartened students, presenting a slightly above average performance and participation. These students presented an XP accumulation curve situated between that of the Achievers and the Disheartened students, and they shared the middle of the leaderboard with the latter student type. These students presented significant differences in terms of final grade, number of badges, and amount of explored achievements and XP earned from achievements, and these features presented slightly above average scores. These students also presented above average number of first and reply posts, which suggests that they were more proactive than the Disheartened and the Underachievers, but less than the Achievers, even though the differences were not significant. Late Awakeners received 0 XP from the Skill Tree and had a similar performance regarding challenges to the Disheartened students, but they earned the most XP from Quests, possibly to recover from some other component. Just like the Disheartened students, Late Awakeners did not feel they were playing a game more than attending a course, but they considered they did non-mandatory tasks for the grade, and also found the experience as being competitive. However, they still found the course more motivating and interesting than "regular" courses.

Both Late Awakeners and Disheartened students have similar XP accumulation curves. However, by analyzing how their leaderboard ranks evolved over time (see Figure 10), we can see that the progression curve of the Late Awakeners is similar to the inverse of that of the Disheartened students, which implies that both student groups were directly competing for the same ranks. Between days 20 and 37, Late Awakeners started to grab larger chunks of XP and Disheartened students seem to have fallen behind, which caused an average ranking tradeoff between the two clusters and led Late Awakeners to

occupy higher ranks. A closer individual look at these students' progress shows that they usually "woke up" around the first quiz, on day 24.

We believe that the emergence of the Late Awakeners in the second year might be in part responsible for the changes observed between the two first years on the Disheartened students. We hypothesize that a great part of the Late Awakeners were originally Disheartened students and Underachievers that were able to take advantage of the extra participation chances to turn the game around in the second year, such as additional challenges and achievements. Even though we lack data to statistically prove it, this might in part explained the differences observed on the Disheartened students.

Understanding the appearance of this cluster is not an easy task. The sheer fact that there were 17 additional students in the second year might have triggered the emergence of a fourth cluster. However, we hypothesize that the increased amount of challenges and their better distribution on the term, as well as the additional value of the achievements, allowed students to do more of what they liked and created further opportunities for them to recover lost XP. This might have reengaged some of the students that would otherwise be included in the Disheartened students or even the Underachievers.

6.1.5. The Regular Students

Regular Students were observed in the third year only and, together with Halfhearted students, replaced both Late Awakeners and Disheartened students. We believe that including more diversified game elements, promoting post quality over quantity, and creating additional opportunities for students to collaborate, be creative and doing more of what they enjoy, allowed the gamified experience to reach out to more students. Furthermore, Achievements and the Skill Tree earned students more XP than in previous years, thus becoming more attractive and harder to overlook. Consequently, our experience attracted more students, who were able to fit in and find something they enjoyed doing. Regular students composed the largest student type observed (43%).

Regular students are characterized by an XP accumulation curve situated right below that of the Achievers, and also by steadily placing on the top half of the leaderboard over the term,. Regulars scored above average and significantly differ from other groups in terms of number of posts, reply posts, rated posts, Skill Tree posts, XP from Skill Tree, explored Skill Tree nodes, acquired badges, completed and explored achievements, and notably, Postmaster badges since they are prolific posters. Conversely, the final grade and the amount of XP from achievements differentiated all clusters except Regular students from Achievers, which suggests that these two had similar performance in these components. Regulars made 38% less first posts and 37% fewer replies than Achievers. While they seem less participative and proactive, differences between the two groups were not significant in these categories.

These students considered our course as being more motivating and interesting than regular courses, and they showed strongest signs of playing a game instead of just attending a course, in the third year. This, combined with their positive stance towards the course and their high performance, suggests a high engagement. This is further corroborated by our assessment, which shows that Regulars were the most engaged students by the end of the course even though this was significant in comparison to Halfhearted participants only.

6.1.6. The Halfhearted Students

Halfhearted students emerged in the third year, together with Regular students. They comprise roughly one quarter of the population. The Half.-hearted accumulate XP at a rate lower than the Regular students, but higher than Underachievers, thus occupying a region in the leaderboard between these two student types. Additionally, they exhibit below-average performance on most aspects. Statistically significant examples are the final grade, accumulated badges, XP earned from achievements, and number of either completed or explored achievements. They also present a low number of reply posts and a number of first posts similar to Underachievers, which suggests that these students tend to participate little and are not proactive. They also seem to present lower performance on several aspects of the course, such as the Skill Tree and AvatarWorld, but perform closer to average on the Challenges and the MCP Quest.

Halfhearted responses to the final questionnaire were on par with those from other students. They considered the course as being more engaging and interesting, not particularly game-like, and equally collaborative as competitive. Their low performance and their feedback led us to believe they were not very engaged by the course. Surprisingly, our formal engagement study suggests that these students were actually the less engaged in the game, even less than the Underachievers. Unfortunately, we do not have enough data to further investigate this matter.

6.1.7. Summary

In order to provide a better understanding on what distinguishes clusters from one another, we summarize in this section the most significant differences observed between them.

Performance-wise, both Achievers and Regular were the best performing students, amassing similar amounts of XP. However, while the former group focused more on the lab component of the course and on the Skill Tree and the AvatarWorld, the latter seems to have invested more effort on the final presentation and the exam. Next came the Late-Awakeners, who accumulated less XP than the Achievers. However, such XP losses occurred in the achievements and in the Skill Tree but not in the more traditional evaluation components. Both the Halfhearted and the Disheartened students performed below average in most aspects of the course, which was noticeable in their final grade. They neglected the Skill Tree and, in the case of the Halfhearted, the AvatarWorld as well. Underachievers were the worst performing group, ignoring both Skill Tree and AvatarWorld, while showing poor performance in the MCP Quest and all Challenges. An anecdotal representation of how each cluster performed in comparison to others can be found in Figure 14. The farther away a cluster is from the center of the radar, the better it performed on average in comparison to other clusters, accross all years.

Participation-wise, the Achievers where the ones posting the most, especially on the Skill Tree and the AvatarWorld. They explored the most Skill Tree nodes and the most achievements, and also completed more achievements that anybody else. Consequently, they also earned more badges. The Regular students participated significantly less than the Achievers, but more than the other two student types in the same rubrics. The Late-Awakeners explored more achievements and earned more badges than the Disheartened and the Underachievers, but less than the Achievers. The Halfhearted too, explored and completed more achievements and acquired more badges than the Underachievers, but less than the Achievers and the Regular students. In the previous year, the Disheartened students followed the same pattern, gathering more badges, and exploring and completing more achievements than the Underachievers, but fewer than the Achievers and Late-Awakeners. The Underachievers had the lowest participation rates, making fewest posts, exploring and completing the fewest achievements, and earning fewer badges than anybody else. We summarize these differences in Figure 15.

Indeed, students that participated more also present better performance [22], which leads us to conjecture that those participating more are more engaged with the course. This seems to be corroborated by formal measurements of engagement, suggesting that Achievers and Regular students were in fact more engaged with the course than the Halfhearted and the Underachievers.

6.2. Cluster Dynamics and Engagement

Cluster dynamics suffered notable changes from the first to the second year. We believe the addition of new participation opportunities and the increased XP value of the achievements made the experience more attractive to all student types, thus increasing their competitiveness. This might have made some students within these clusters to stand out, which made clusters overlap. This is corroborated by their final leaderboard positioning on the second year, which is much more heterogeneous as compared to that of the first year (see Figure 16 and Figure 17). Because the game component was better rewarded and competition was fiercer, students seem to have taken the game more seriously and to have worked more for the course, particularly towards the achievements. Interestingly, between the first two years we observed a large decrease on the percentage of Underachievers (from 40% to 25%) and an increase in the percentage of students considering the course as being easier to learn from (from 35.7% to 53.3%). This

Figure 14. Radar chart comparison of student XP accrual. The further away from the center the better a cluster performed on average, in all years it was represented.

Figure 15. Radar chart comparison of several student participation measures. The further away from the center the better a cluster performed on average, in all years it was represented.

suggests that we reached out to more students and that the learning process was indeed easier in the second year, and we believe this was potentiated by the broader availability of participation opportunities.

Between the second and the third years, remarkable changes in cluster composition were also observed. We believe that promoting quality over quantity and presenting a diversified set of game elements that were well rewarded and balanced, allowed for more students to be engaged by the course, in a more continuous fashion. This created the conditions for a new type of "high-middle" class of students to emerge, the Regular students. These had above average performance on most evaluation aspects, except for AvatarWorld, which was for the most part avoided. Another cluster emerged, which we named Halfhearted students. These had below average performance, avoided the AvatarWorld and the Skill Tree,

and had average performance on the MCP Quest and the Challenges. In the third year, the Underachievers presented their lowest performance ever, but we believe this happened because some of the best performing students that would be classified as Underachievers in the previous year were now classified as Halfhearted, thus lowering the former type's average performance.

Telling Underachievers and Halfhearted students apart is fairly easy, as both present gradual loss of performance in almost all evaluation components but with different levels of decay, falling into distinct tiers of performance. This is easily graspable by taking a look at their final ranking at the end of the course (see Figure 18), where they seem to occupy distinct areas on the bottom half of the leaderboard. Nonetheless, they also shared some common traits, such as similar amounts of posts, reply posts, rated posts, Skill Tree posts, amount of XP from the Skill Tree, number of explored Skill Tree nodes, and number of badges in the Postmaster achievement. These features differed significantly between every cluster except for the Halfhearted and Underachievers.

Telling Achievers and Regular students apart is a harder task, as both equally shared the top 1/3 of the leaderboard. The differences in terms of final grade and grade earned from Achievements are not significant; what seems to change is how they achieved those scores. While Achievers seem to have reached top grades by having top scores on the achievements, the Skill Tree and the lab assignments, Regular students seem to have participated less on the Skill Tree and AvatarWorld but performed better on the multimedia presentation and the exam. Figure 13 seems to support this theory, showing that the Regular students became closer to the Achievers between the first and the second exam, and even closer in the day of the second exam. Between both exams, Regular students explored several achievements that did not have a deadline to get additional XP. Consequently, both types shared the same ranking, as they accumulated similar levels of XP through different paths. This implies that our third gamified instance of the course might have catered to different needs and learning preferences, which confers flexibility to the learning experience and allows it to reach more students, without hindering others. This is further corroborated by a decrease on the percentage of Underachievers in the third year (from 25% to 20.4%) and an increase on the percentage of students considering the course as being easier to learn (from 53.3% to 61.7%).

Our analysis also suggests that there are two traits that seem to separate the pair Achievers / Regular students from the pair Halfhearted / Underachievers. Firstly, the number of badges in the Artist achievement (make rating-four posts), was significantly different between the two pairs but not between pair elements, which implies that the first pair made more quality posts then the second pair. Secondly, the engagement assessment shows that the Regular students were the most engaged and the Halfhearted student were the least engaged, and that these differences were significant. Although the mean average SCEQ score for the Achievers and the Underachievers were closer to that of the Regular and Halfhearted students, no significant differences were observed. However, the Discriminant Function Analysis suggests that the Achievers' and Regular students' SCEQ grow similarly and the same was observed for the Halfhearted and the Underachievers, which seems to support our case.

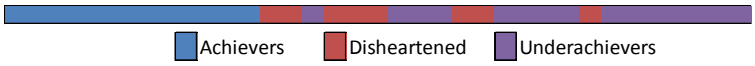

Figure 16. Final leaderboard position of different student types on the 1st year (top position on the left).

Figure 17. Final leaderboard position of different student types on the 2nd year (top position on the left).

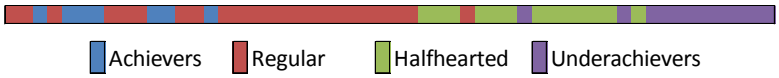

Figure 18. Final leaderboard position of different student types on the 3rd year (top position on the left).

Although we did expect to observe more differences in terms of what achievements each student type would pursue, we did observed that some components seem to have had separated students by different levels of performance and engagement with the course. For example, we observed that AvatarWorld was mostly sought by Achievers, who had the best performance and were highly engaged, and ignored by others. On the other hand, most of the participation on the Skill Tree was carried out by Achievers and Regular students only. The MCP Quest reached the Achievers, the Regular students and also the Halfhearted students. We believe that because they only had to participate once to earn the XP, Halfhearted were more encourage to participate on the MCP Quest than the AvatarWorld or the Skill Tree. The Challenges seem to have reached every student type. We believe one of the main reasons for this was that there were several of them, which constantly reminded students of their existence and encouraged them to participate. These observations suggest that we were indeed able to cater to different audiences, as we provided achievements for different tastes.

Finally, a closer look at how the XP accumulation curves evolved between years reveals two interesting effects worth discussing. The first one is the fact that the Achievers were the closer to the Ideal student in the second year in comparison to the others, which we already hypothesized that was provoked by both the additional 5% extra XP from the Skill Tree and unassessed traits from that year's population. The second effect consists of how the Achievers were ahead of the Ideal student during two thirds of the course in the third year and the got dragged down by the Multimedia Presentation and the exams. On the other hand, the Regular students took good advantage of the exams and other achievements to boost their XP. We believe that because the achievements were better rewarded, the Achievers went after them as early as possible, before we expected. However, we cannot clearly explain the reason for the decline of the Achievers upon the presentation and the exams. We believe three things might have had impact here: 1) there were no regular quizzes which might have left them less prepared for those checkpoints; 2) there were other courses that also required attention; maybe because Achievers already had a good grade, they decided to invest more time on other courses, which also had exams and project deliveries on the same period, whereas the Regular students, who had a lower grade, felt the need to work more on the final evaluation checkpoints, to get more XP; 3) unassessed and uncontrolled traits from the population, given that all traditional evaluation measures, such as the labs, the presentation and the exam also had worse scores than the previous year, but the same was not observed for the amount of XP earned from the Achievements.

6.3. All Different, All Equal

So far we have been focusing in what differentiates one cluster from another. But we also found, from their replies to the questionnaires, that our students share many opinions in common as well, which provides a broader view of what they really think about the course. For instance, most students found the gamified experience performed very well and considered the course to be more motivating, more interesting, to require more work and to be as difficult as compared to other non-gamified courses. Similarly, students in general find our course to require about as much study as other courses but they consider it to be more continuous. Also, each iteration had more and more students considering it easier to learn from (as compared to other course) than the previous one.

The majority of the students considered that they felt the course allowed them to be creative, to do things they like, and to learn useful skills for the future. Regarding whether they feel like playing a game or not, their opinion also appears to be consensual: they feel like playing a game as much as they feel like attending a course. At the end of the day, they still want to learn and pass the course with the best grade possible. If on top of that they can still experience something somewhat resembling to a game, which is interesting and motivating, and allows them to do things they like and learn useful skills, than we believe we can call this a good learning experience; one that they considered easier to learn from. As such, most students, in all years, affirmed that they would like to see gamification applied to other courses as well.

6.4. Study Overview and Design Implications

Our study consists of a series of iterative experiments, where our gamified course was progressively adapted to face many engagement challenges. The first year showed that gamification can indeed make a college course more interesting and motivating, and led students to participate and work more [21]. We learned the importance of using challenges and achievements to shape student behavior. By conceiving these elements as multi-level tasks, and by providing instant feedback via visual and collectible artifacts, students might develop a deeper sense of competence and mastery. However, by the end of the first year, we were still facing the following problems:

1. Students did not have enough opportunities to recover XP lost from achievements and challenges they could not or chose not to attend. We believe this was one of the main reasons behind the existence of the Disheartened students.
2. Students lacked autonomy, which hindered the game-feeling of the experience.
3. Students wanted to be rewarded for activities they performed in class, like oral participation.
4. Students considered the experience to be too competitive and wanted collaboration to be promoted as well.
5. Our gamified experience did not reach every student. Some seem to overlook it because it was not adequately rewarded.

We tried to overcome these problems by issuing a few changes in the second year. In response to the first problem, we included new achievements and added more challenges in the game. Additionally, we also made an effort to make them more evenly distributed over the term, so that students could recover XP whenever they saw fit. We believe this problem was partially solved, given the increased participation observed in the second year and the emergence of the Late Awakeners, who were able to take advantage of the additional participation opportunities.

To tackle the second problem, we created additional achievements and added the Skill Tree, which provided students with more choices and more control over the experience. Given its early stage of development and late introduction to the course, the Skill Tree was ignored by most except the Achievers.

We tried to solve the third and the fourth problems by creating new achievements, e.g. Talkative, Guild Master and Guild Warrior. The first earned students XP and badges when they made meaningful oral comments in class, which pleased those who enjoyed talking and displeased those who did not. The other two achievements rewarded students with both XP and badges when their lab section performed the best, or if everybody in that section had grades above 80%. The collaboration achievements had little success, as students would rather blame others for not getting the grade instead of directly helping them. It turns out that, in the particular case of our course and academic culture, students are rather competitive and seem not to enjoy helping others do their work. Therefore, while we believe we tackled the third problem, we still had a lot to do to encourage students to collaborate.

Last, to make the game component more appealing and encourage students to participate, we decided to allocate a larger chunk of the course grade to achievements, which resulted in a significant improvement of student participation. One may argue that students participated more only because there were more XP at stake, but a previous study shows that students actually made more posts per challenge, even though they would not be rewarded for doing that [22]. While this suggests that students were more engaged than the previous gamified version of the course, the progression of Late Awakeners and Disheartened students was neither stable nor gradual in the first half of the term. This led us to believe that these students were still overlooking game aspects of the course.

By the end of the second year, problems #2 and #4 were still far from controlled, problem #5 could still use some work, and we identified the following additional problems:

6. Posts were graded based on quantity, which encouraged students to make a lot of low-quality posts. Students who really cared about the course and produced quality content perceived this as unfair. Although a few students also pointed this out in the first year, the problem only became evident in the second year, when the grade associated to posting increased.
7. The Skill Tree did not earn students any XP for the first nodes, which made this initial effort seem wasteful. Furthermore, it did not seem very appealing by accounting for only 5% of total grade.

In the third year, we tried to further improve the student autonomy by introducing AvatarWorld and enhancing the Skill Tree. Not only was it worth more XP, but now all nodes counted, even the base ones. Furthermore, students no longer had to accumulate meta-points to participate in the Skill Tree, which made the whole process easier to understand and more convenient. The improved Skill Tree attracted more participation as compared to the previous year, with students acquiring a larger portion of the maximum XP allocated to this element and considering it as an engaging game element. Given its lack of additional features and proper integration with the whole gamified experience, AvatarWorld had limited success, but it still captivated some students and allowed them to be creative. Given that students considered the course as allowing them to be creative, do more of what they like, and that this feeling seems to be stronger in the third year, we believe we made progress concerning problems #2 and #7.

We dealt with problem #4 by introducing a new collaborative element named MCP Quest, which required all students to work together for a common good. This component replaced the old Online Quests, and it was considered one of the most engaging elements in the gamified experience, which finally got students to work together. The MCP Quest was different from previous collaborative attempts in the sense that students now where encourage to work on a single task together and did not have to spend any of their precious time solely on others. Indeed, students considered the course as being as much collaborative as it was competitive, whereas they considered it to be solely competitive in the previous year. This suggests we hit a balance between the two, which we consider to be a great outcome and improvement in our approach.

To further mitigate problem #5, we allocated larger grades to the achievements and the Skill Tree, and added new elements like the MCP Quest and AvatarWorld. This allowed the game component to look more of an integral part of the course and made it harder to be overlooked. In the third year, Disheartened students and Late Awakeners were replaced by Regular and Halfhearted students, which presented a more stable and gradual progression over the term. Furthermore, the Regular students, with an above average performance, became the largest student type, and the size of the Underachievers gradually shrunk over the years. On the other hand, student perception that the course was easier to learn (than other courses) also grew. Thus, we believe students find it increasingly easier to learn from our course and that we have been able to reach out to more students, which seems to greatly reduce problem #5 and suggests we are in the right direction. We solved problem #6 by grading posts not by quantity but by rating (given by faculty).

As take-away messages, we point the importance of increasing student agency, which has a high impact in their motivation to learn, of balancing out competition and collaboration, and providing an untangled gamified experience. Agency can be promoted by, for example, evenly distributing gamified assignments, such as challenges and quests, throughout the semester and allowing students to make multiple contributions in those that they like the most. This will provide them with several opportunities to be creative, to do more of what they like and to recover lost XP. Adding other elements that enable meaningful choices, like the Skill Tree, is also a viable approach. A good gamified experience should allow behaviors of both competition and collaboration to emerge naturally among students. While competition is easier to promote, collaboration requires students to have goals in common that require them to work together towards mutual benefit. Creating passive achievements where they can simply wait for others to perform well enough so that everybody will earn more XP will not work. Finally, like a good game, a gamified learning experience should be easy to play and rules should be clear an uncomplicated. The first version of our Skill Tree was the opposite of this: students had a cumbersome unlock system which ultimately led them to avoid this element altogether. Dropping the meta-point unlock system seems to have rendered the Skill Tree more appealing and easier to use.

Our study reveals that there are different types of students in our course, with different levels of performance and engagement. We consider that studying these in an important steppingstone to understand how different students perform and react to gamified setting. This is paramount to develop new and advanced learning environments where content and interventions may be devised on a per-cluster level.

6.5. Study Limitations

Our long-term study has a few limitations that should be considered, most of which related to the presence of uncontrolled variables, transversal to all years. For example, we could not control the number of students enrolled nor the nature of the student population, which might have had impact on the observed differences among years. Moreover, some of the course materials had to change in order to keep the course updated, according to the university's policy, and the composition of the faculty staff also differed from one year to the other. Student engagement might also have been affected by these factors, differently in the three years. A within subjects study on a single year could attenuate these limitations. However, a gamified course like ours involves a different and demanding evaluation method and having two groups evaluated with different criteria would neither be fair nor ethical. Furthermore, this approach will not work when performing multi-year comparisons.

The perceived nature of the identified clusters could also have been influenced by the size of the population. Our small student population originated even smaller clusters, which limited the amount of significant differences observable among them. An example of this is the engagement study with the SCEQ, where a high level of absence in some clusters associated to a small sample size led to results with weak explanatory power for the Achievers and Underachievers. Another factor that might have had impact on the results is the number of first days excluded from the cluster analysis. We decided to specify criteria that would exclude the first days where either there was no activity or some students were not yet playing the game. Including these in the analysis would lead the clustering algorithm to consider that there were more students with zero XP for a long initial period of time then there actually were, because they were not playing the game yet. We assumed this compromise and we accept that it might have had impact over the results. We acknowledge the limitations of this study and consider that our results must be interpreted with due caution. However, this exploratory research can very well establish groundwork on gamified learning, by investigating how different students perform and are engaged in a gamified setting.

7. CONCLUSION

In a previous study we described our experience at gamifying a graduate-level course on Multimedia Content Production. A comparison with the previous non-gamified versions of the course shows increased student participation with the gamified version. Furthermore, learners also considered the course as both more interesting and motivating than "regular" courses. However, despite these advantages, we knew very little about how different students experienced and reacted to gamified learning.

In this paper we presented a long-term study to identify how different students played our gamified course and how engaged they felt. To this end, we collected student behavior data over three academic years, while we iteratively improved the course, based on student feedback. We analyzed how different students progressed through the course by performing cluster analysis on XP accumulated per day, for every term. From this process different types of students emerged, each with different traits. In this study we carefully characterize every student type and extensively explain how these changed over the years, as the course evolved, and what may have caused these changes. Furthermore, in the third year we used a validated instrument to assess how different students were engaged by the course.

We identified six student types, all representative of different performance levels and approaches to playing the course. However, these types did not all co-exist in a single instance. While two types were consistently observed –Achievers and Underachievers – others emerged as the course evolved. In the last version of the course, we observed four types: 1) Achievers, who focused on the achievements and strived to acquire all the XP they could get their hands on; 2) Regular students – the largest group – who had above average performance and balanced game achievements with more traditional evaluation components; 3) Halfhearted students, who presented below-average performance and seem to have neglected some aspects of the course; and 4) Underachievers, who had the lowest performance, neglected several gamified evaluation components and seem to have done barely enough to score a passing grade.

Our research also suggests that both Achievers and Regular students were more engaged with the game and made more quality posts than either Halfhearted or Underachievers.

From our experiments we learned several lessons. Just like games, it is of utmost importance to allow players to learn from trial and error and to be able to pick their own path. This can be achieved by creating different opportunities for students to learn and score, during the full span of the course, and by providing alternative paths. As the course evolved over time, we learned that as Achievements become more prominent and are better rewarded, the experience gets more competitive. Hence, it becomes important to create new occasions for students to collaborate and balance the competitive pressure. Just like in a game, students often try to exploit game rules to get ahead of others. Performance Assessment rules emphasizing quantity over quality become easy targets for this "over-gaming". Thus it is healthy to adapt grading methods to account for quality instead.

Reaching out to every student has been one of the hardest challenges in our gamified experiment. This study provides a means to better understanding what different kinds of behaviors may emerge in gamified learning settings and how different students are engaged by the game. We thus believe our findings contribute to enrich the body of knowledge on gamified education and pose valuable hypotheses to guide future research in the field. We hope our results can be used to guide other gamified learning endeavors so that every student can learn better while enjoying the experience. As for future work we would like to use our findings about student differentiation together with machine learning techniques to develop adaptive gamified learning environments. We envision such approaches would allow us to detect a student's profile very early in the game, so that both content and delivery could be automatically adapted to participants to make their learning experience enjoyable, enriching and productive.

8. ACKNOWLEDGMENTS

This work was partially supported by national funds through Fundação para a Ciência e a Tecnologia (FCT) with references UID/CEC/50021/2013 and PTDC/EEI-SII/3154/2012. Gabriel Barata was supported by FCT grant SFRH/BD/72735/2010.

9. REFERENCES

[1] S. Deterding, D. Dixon, R. Khaled, L. Nacke, From game design elements to gamefulness: defining "gamification", in: Proceedings of the 15th International Academic MindTrek Conference Envisioning Future Media Environments, Vol. Tampere, F, ACM, 2011, pp. 9–15.

[2] S. Deterding, M. Sicart, L. Nacke, K. O'Hara, D. Dixon, Gamification. using game-design elements in non-gaming contexts, in: Proceedings of the 2011 annual conference extended abstracts on Human factors in computing systems, CHI EA'11, ACM, New York, NY, USA, 2011, pp. 2425–2428.

[3] K. Huotari, J. Hamari, Defining gamification: a service marketing perspective, in: Proceeding of the 16th International Academic MindTrek Conference, ACM, 2012, pp. 17–22.

[4] O. Inbar, N. Tractinsky, O. Tsimhoni, T. Seder, Driving the scoreboard: Motivating eco-driving through in-car gaming, in: Proceedings of the CHI 2011 Workshop Gamification: Using Game Design Elements in Non-Game Contexts, ACM, 2011.

[5] G. Zichermann, C. Cunningham, Gamification by Design: Implementing Game Mechanics in Web and Mobile Apps, O'Reilly Media, Inc., 2011.

[6] P. Brauner, A. Calero Valdez, U. Schroeder, M. Ziefle, Increase physical fitness and create health awareness through exergames and gamification, in: A. Holzinger, M. Ziefle, M. Hitz, M. Debevc (Eds.), Human Factors in Computing and Informatics, Vol. 7946 of Lecture Notes in Computer Science, Springer Berlin Heidelberg, 2013, pp. 349–362.

[7] S. Sheth, J. Bell, G. Kaiser, Halo (highly addictive, socially optimized) software engineering, in: Proceeding of the 1st international workshop on Games and Software Engineering, Vol. 11 of GAS, 2011, pp. 29–32.

[8] Z. Fitz-Walter, P. Wyeth, D. Tjondronegoro, B. Scott-Parker, Driven to drive: Designing gamification for a learner logbook smartphone application, in: Proceedings of the 2013 Symposium on Gameful Design, Research, and Applications, Gamification 2013, Stratford, ON, Canada, 2013, pp. 42–49.

[9] S. Bennett, K. Maton, L. Kervin, The 'digital natives' debate: A critical review of the evidence, British Journal of Educational Technology 39 (5) (2008) 775–786.

[10] H. F. O'Neil, R. Wainess, E. L. Baker, Classification of learning outcomes: evidence from the computer games literature, Curriculum Journal 16 (4) (2005) 455–474.

[11] K. Squire, Video Games and Learning: Teaching and Participatory Culture in the Digital Age. Technology, Education–Connections (the TEC Series), ERIC, 2011.

[12] J. P. Gee, What video games have to teach us about learning and literacy, ACM Computers in Entertainment 1 (1) (2003) 20–20.

[13] J. Chen, Flow in games (and everything else), Communications of the ACM 50 (2007) 31–34.

[14] M. Csikszentmihalyi, Flow: The psychology of optimal experience, Harper Perennial, 1991.

[15] M. Prensky, The motivation of gameplay: The real twenty-first century learning revolution, On the Horizon 10 (1) (2002) 5–11.

[16] M. de Aguilera, A. Mendiz, Video games and education: (education in the face of a "parallel school"), ACM Computers in Entertainment 1 (1) (2003) 1:1–1:10.

[17] K. D. Squire, Video games in education, International Journal of Intelligent Games & Simulation 2 (1) (2003) 49–62.

[18] P. Denny, The effect of virtual achievements on student engagement, in: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, CHI '13, ACM, New York, NY, USA, 2013, pp. 763–772.

[19] A. Domínguez, J. Saenz-de Navarrete, L. de Marcos, L. Fernández-Sanz, C. Pagés, J.-J. Martínez-Herráiz, Gamifying learning experiences: Practical implications and outcomes, Computers & Education 63 (0) (2013) 380 – 392.

[20] 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 12 (4) (2013) 236 – 245, social Commerce- Part 2.

[21] G. Barata, S. Gama, J. Jorge, D. Gonçalves, Engaging engineering students with gamification, in: Games and Virtual Worlds for Serious Applications (VS-GAMES), 2013 5th International Conference on, 2013, pp. 1–8.

[22] G. Barata, S. Gama, J. Jorge, D. Gonçalves, Improving participation and learning with gamification, in: Proceedings of the First International Conference on Gameful Design, Research, and Applications, Gamification '13, ACM, New York, NY, USA, 2013, pp. 10–17.

[23] G. Barata, S. Gama, J. Jorge, D. Gonçalves, Gamification for smarter learning: tales from the trenches, Smart Learning Environments 2 (1) (2015) 1–23.

[24] J. Lee, K. Luchini, B. Michael, C. Norris, E. Soloway, More than just fun and games: assessing the value of educational video games in the classroom, in: CHI '04 Extended Abstracts on Human Factors in Computing Systems, CHI EA '04, ACM, New York, NY, USA, 2004, pp. 1375–1378.

[25] M. Kebritchi, A. Hirumi, H. Bai, The effects of modern math computer games on learners' math achievement and math course motivation in a public high school setting, British Journal of Educational Technology 38 (2) (2008) 49–259.

[26] B. Coller, D. Shernoff, Video game-based education in mechanical engineering: A look at student engagement, International Journal of Engineering Education 25 (2) (2009) 308–317.

[27] P. Mcclean, B. Saini-eidukat, D. Schwert, B. Slator, A. White, Virtual worlds in large enrollment science classes significantly improve authentic learning, in: Proceedings of the 12th International Conference on College Teaching and Learning, Center for the Advancement of Teaching and Learning, 2001, pp. 111–118.

[28] J. Moreno, Digital competition game to improve programming skills, Educational Technology & Society 15 (3) (2012) 288–297.

[29] K. Squire, M. Barnett, J. M. Grant, T. Higginbotham, Electromagnetism supercharged!: learning physics with digital simulation games, in: Proceedings of the 6th international conference on Learning sciences, ICLS '04, International Society of the Learning Sciences, 2004, pp. 513–520.

[30] C. Crumlish, E. Malone, Designing Social Interfaces, O'Reilly, 2009.

[31] A. J. Kim, Putting the fun in functional, http://www.slideshare.net/amyjokim/putting-the-fun-in-functiona (March 2008).

[32] K. Werbach, D. Hunter, For the Win: How Game Thinking Can Revolutionize Your Business, Wharton Digital Press, 2012.

[33] C. Lewis, N. Wardrip-Fruin, J. Whitehead, Motivational game design patterns of'ville games, in: Proceedings of the International Conference on the Foundations of Digital Games, ACM, 2012, pp. 172–179.

[34] J. Hamari, J. Koivisto, H. Sarsa, Does gamification work? – a literature review of empirical studies on gamification, in: System Sciences (HICSS), 2014 47th Hawaii International Conference on, 2014, pp. 3025–3034.

[35] E. D. Mekler, F. Brühlmann, K. Opwis, A. N. Tuch, Do points, levels and leaderboards harm intrinsic motivation?: an empirical analysis of common gamification elements, in: Proceedings of the First International Conference on Gameful Design, Research, and Applications, ACM, 2013, pp. 66–73.

[36] S. Aguilar, C. Holman, B. Fishman, Multiple paths, same goal: Exploring the motivational pathways of two distinct game-inspired university course designs, Games+ Learning+ Society, Madison, WI.

[37] B. De Schutter, V. Vanden Abeele, Gradequest—evaluating the impact of using game design techniques in an undergraduate course, in: 9th International Conference on the Foundations of Digital Games, 2014.

[38] C. Cheong, F. Cheong, J. Filippou, Quick quiz: A gamified approach for enhancing learning, in: Pacific Asia Conference on Information Systems, 2013.

[39] L. Hakulinen, T. Auvinen, A. Korhonen, Empirical study on the effect of achievement badges in trakla2 online learning environment, in: Learning and Teaching in Computing and Engineering (LaTiCE), 2013, 2013, pp. 47–54.

[40] L. Sheldon, The Multiplayer Classroom: Designing Coursework as a Game, Course Technology PTR, 2011.

[41] C. Holman, S. Aguilar, B. Fishman, Gradecraft: what can we learn from a game-inspired learning management system?, in: Proceedings of the Third International Conference on Learning Analytics and Knowledge, ACM, 2013, pp. 260–264.

[42] S. Aguilar, B. Fishman, C. Holman, Leveling-up: Evolving game-inspired university course design, in: Games+Learning+Society Conference 9.0, University of Wisconsin-Madison. Madison, WI., 2013.

[43] B. De Schutter, "the gradequest tale of scrotie mcboogerballs" evaluating the second iteration of a gameful undergraduate course., in: Meaningful Play 2014, Michigan State University, East-Lansing, Michigan, USA, 2014.

[44] F. Guay, R. J. Vallerand, C. Blanchard, On the assessment of situational intrinsic and extrinsic motivation: The situational motivation scale (sims), Motivation and emotion 24 (3) (2000) 175–213.

[45] W. IJsselsteijn, W. van den Hoogen, C. Klimmt, Y. de Kort, C. Lindley, K. Mathiak, K. Poels, N. Ravaja, M. Turpeinen, P. Vorderer, Measuring the experience of digital game enjoyment, in: Proceedings of Measuring Behavior, Maastricht Netherlands, 2008, pp. 88–89.

[46] A. L. Brown, J. C. Campione, Psychological theory and the design of innovative learning environments: On procedures, principles, and systems., Lawrence Erlbaum Associates, Inc, 1996.

[47] A. Collins, The changing infrastructure of education research, Issues in education research (1999) 289–198.

[48] F. Wang, M. J. Hannafin, Design-based research and technology-enhanced learning environments, Educational technology research and development 53 (4) (2005) 5–23.

[49] J. Greeno, A. Collins, L. Resnick, Handbook of educational psychology, New York, NY, 1996, Ch. Cognition and learning, pp. 15–46.

[50] P. Bell, C. M. Hoadley, M. C. Linn, Design-based research in education, Internet environments for science education (2004) 73–85.

[51] R. M. Ryan, E. L. Deci, Handbook on motivation at school, Routledge, 2009, Ch. romoting self-determined school engagement, pp. 171–196.

[52] E. Deci, R. Ryan, Handbook of Self-determination Research., University of Rochester Press, 2004.

[53] T. W. Malone, M. R. Lepper, Making learning fun: A taxonomy of intrinsic motivations for learning, Aptitude, learning, and instruction 3 (1987) (1987) 223–253.

[54] J. Schell, The Art of Game Design: A Book of Lenses, CRC Press, 2008.

[55] B. C. Tatum, J. C. Lenel, A comparison of self-paced and lecture/discussion methods in an accelerated learning format, Publication of National University 139.

[56] B. J. Fishman, S. Deterding, Beyond badges & points: Gameful assessment systems for engagement in formal education, University of Wisconsin, Wisconsin, Madison.

[57] A. P. Dempster, N. M. Laird, D. B. Rubin, Maximum likelihood from incomplete data via the em algorithm, Journal of the Royal Statistical Society. Series B (Methodological) (1977) 1–38.

[58] N. Sharma, A. Bajpai, M. R. Litoriya, Comparison the various clustering algorithms of weka tools, facilities 4 (2012) 7.

[59] L. E. Blume, D. Easley, rationality, in: S. N. Durlauf, L. E. Blume (Eds.), The New Palgrave Dictionary of Economics, Palgrave Macmillan, Basingstoke, 2008.

[60] G. Barata, S. Gama, M. J. Fonseca, D. Gonçalves, Improving student creativity with gamification and virtual worlds, in: Proceedings of the First International Conference on Gameful Design, Research, and Applications, Gamification '13, ACM, New York, NY, USA, 2013, pp. 95–98.

[61] M. M. Handelsman, W. L. Briggs, N. Sullivan, A. Towler, A measure of college student course engagement, The Journal of Educational Research 98 (3) (2005) 184–192.

[62] D. C. Hoaglin, B. Iglewicz, J. W. Tukey, Performance of some resistant rules for outlier labeling, Journal of the American Statistical Association 81 (396) (1986) 991–999.

[63] J. Cohen, Statistical power analysis for the behavioral sciences, Academic Press, 1977.

[64] C. Becker, U. Gather, The largest nonidentifiable outlier: a comparison of multivariate simultaneous outlier identification rules, Computational Statistics & Data Analysis 36 (1) (2001) 119 – 127.

APPENDIX A Achievement Description and Grade Allocation

Table A.1. Achievement description and comparison of requirements per level across years. The "=" sign means the rule remained the same as the previous year.

Name	Description	Year 1			Year 2			Year 3	
		Level 1	Level 2	Level 3	Level 1	Level 2	Level 3	Level 1	Level 2
Postmaster	Post in something in the forums	make twenty posts	make thirty posts	make fifty posts	=	=	=	=	=
Challenger of the Unknown	Submit contributions to the Online Quests Forum / MCP Quest	make two contributions	make three contributions	make four contributions	make one post	make two post	make three posts	participate in the quest	
Bookworm	Read class slides	read slides for 50% of lectures	read slides for 75% of lectures	read all lectures slides	=	=	=	=	=
Quizmaster	Excel at the quizzes	top grade in four quizzes	top grade in five quizes	top grade in all six quizzes	=	=	=		
Lab Master	Excel at the labs	top grade in two graded classes	top grade in three graded classes	top grade in all graded classes	=	=	=	=	=
Proficient Tool User	Get creative with gimp, inkscape and the other tools	post one type of creative result	post two types of creative results	post three (or more) types of creative results	=	=	=	get four points	get eight point
Class Annotator	Find related resources, more information, about class subjects	contribute for one class	contribute for two classes	contribute for three classes	=	=	contribute for five classes	get four points	get eight point
Rise to the Challenge	Complete theoretical class challenges	one challenge	three challenges	five challenges	=	five challenges	ten challenges	get four points	get twelve points
Wild Imagination	Suggest presentation subjects	sugest a new subject for your presentation			=			=	
Right on Time	Don't be late for class!	be on time 15 times	be on time 17 times	always be there on time	=	=	=	be on time 13 times	=
Amphitheatre Lover	Show up for theoretical lectures!	be there for 50% of lectures	be there for 75% of lectures	be there for all of the lectures	=	=	=	=	=
Lab Lover	Show up for labs!	be there for 60% of labs	be there for 80% of labs	be there for all of the labs	be there for 50% of labs	be there for 75% of labs	be there for all of the labs	be there for 50% of labs	be there for 75 of labs
Student Worker	Make summaries of the lectures of the week (student workers only!)	Do summaries for at least 50% of the lectures	Do summaries for at least 75% of the lectures	Do summaries for all the lectures	Do summaries for at least 50% of the lectures	Do summaries for at least 75% of the lectures	Do summaries for all the lectures	Do summaries for at least 50% of the lectures	Do summaries for at least 75 of the lectures
Good Host	Attend invited lectures	Attend all invited lectures at your campus	Attend all invited lectures at your campus and one in the other campus	Attend all invited lectures at both campii					
Attentive Student	Find bugs in class materials	find one bugs	find three bugs	find five bugs	=	=	=	get four points	get eight point
Bug Squasher	Correct problems in PCMMedia Mixer	correct one problem	correct two problems	correct three problems					
Popular Choice Award	Have the most liked multimedia presentation	be the third most liked	be the second most liked	be the most liked!	=	=	=	=	=
Hollywood Wannabe	Create great videos for your presentation	remixed video	created own video (single shoot)	created own video, relevant edits	=	=	=	=	=
Presentation Zen Master	Think about your presentation before opening powerpoint	hand in document about the rationalle of your presentation.			=			=	

Continuation of Table A.1

Golden Star	Be creative and do relevant things to help improve the course	perform one task	perform two tasks	perform three tasks	=	=	=	=	=
Quiz King	Take the quizzes, be the best	Have the highest grade in the quizzes			=				
Lab King	Attend the labs, be the best	Have the highest grade in the labs			=			=	
Presentation King	Present your thing, be the best	Have the highest grade in the presentations			=			=	
Exam King	Take the exams, be the best	Have the highest grade in the exams!			=			=	
Course Emperor	Take the course, be the best	Have the highest course grade!			=			=	
Archivist	Help compile the results of challenges, etc				Create the archive for 3 events	Create the archive for 5 events	Create the archive for 10 events		
Guild Master	Your Guild was the best!				The average score of your lab class was the best of all classes (1 time)	The average score of your lab class was the best of all classes (2 times)	The average score of your lab class was the best of all classes (5 times)	The average score of your lab class was the best of all classes (1 time)	The average score of your la class was the best of all classes (2 times)
Guild Warrior	You helped your Guild				All students in your lab had 80% or better grade (1 time)	All students in your lab had 80% or better grade (2 times)	All students in your lab had 80% or better grade (5 times)	=	=
Proactive	Reply to a theoretical challenge in a timely manner				Be in the first three to reply to a challenge/quest (1 time)	Be in the first three to reply to a challenge/quest (4 times)	Be in the first three to reply to a challenge/quest (10 times)		
Talkative	Participate in Theoretical Lectures!				participate 5 times	participate 10 times	participate 20 times	participate 2 times	participate 6 times
Apprentice	Give answers in the 'questions' or "Labs" forums							get four points	get eight point
Artist	Show creativity and quality							get four posts of four points	get six posts of four points
Blacksmith	Create new objects/skins for Avatar World							get four points	get eight point
Master Builder	Create new buildings for Avatar World							get four points	get eight point
Replier Extraordinaire	Respond to the gamification questionnaires							respond to first questionnaire	respond to firs two questionnaires
Squire	Help your colleagues by writting tutorials of your tree challenges							get four points	get ten points

Table A.2. Comparison of the percentage of total XP allocated to each achievement level across years. Green shaded cells denote bonus XP.

Name	Year 1			Year 2			Year 3		
	Level 1	Level 2	Level 3	Level 1	Level 2	Level 3	Level 1	Level 2	Level 3
Postmaster	0.39%	0.17%	0.06%	0.83%	0.33%	0.25%	0.50%	0.40%	0.25%
Challenger of the Unknown	0.94%	0.39%	0.06%	1.25%	0.83%	0.42%	3.00%		
Bookworm	0.56%	0.39%	0.22%	0.42%	0.33%	0.25%	0.50%	0.50%	0.50%
Quizmaster	0.56%	0.17%	0.06%	0.83%	0.63%	0.21%			
Lab Master	0.56%	0.17%	0.06%	0.83%	0.63%	0.21%	0.75%	0.75%	0.50%
Proficient Tool User	0.56%	0.33%	0.22%	1.46%	0.83%	0.42%	1.50%	1.50%	1.50%
Class Annotator	0.56%	0.33%	0.17%	0.83%	0.42%	0.21%	1.00%	1.00%	1.00%
Rise to the Challenge	0.56%	0.39%	0.06%	2.29%	1.46%	0.83%	2.00%	2.00%	2.00%
Wild Imagination	0.67%			0.83%			1.00%		
Right on Time	0.22%	0.17%	0.06%	0.33%	0.21%	0.04%	0.25%	0.25%	0.25%
Amphitheatre Lover	0.22%	0.17%	0.06%	0.33%	0.25%	0.13%	0.50%	0.50%	0.50%
Lab Lover	0.22%	0.17%	0.06%	0.33%	0.25%	0.13%	0.25%	0.25%	0.25%
Student Worker	0.67%	0.50%	0.17%	0.67%	0.46%	0.17%	1.00%	1.00%	1.00%
Good Host	0.44%	0.11%	0.11%						
Attentive Student	0.22%	0.17%	0.06%	0.33%	0.17%	0.08%	0.50%	0.50%	0.50%
Bug Squasher	0.22%	0.17%	0.06%						
Popular Choice Award	0.22%	0.17%	0.06%	0.42%	0.29%	0.21%	0.25%	0.25%	0.25%
Hollywood Wannabe	0.56%	0.39%	0.06%	0.63%	0.42%	0.29%	0.75%	0.50%	0.35%
Presentation Zen Master	0.56%			0.83%			1.00%		
Golden Star	0.22%	0.17%	0.06%	0.42%	0.42%	0.42%	0.50%	0.50%	0.50%
Quiz King	0.22%			0.33%					
Lab King	0.22%			0.33%			0.40%		
Presentation King	0.22%			0.33%			0.40%		
Exam King	0.22%			0.33%			0.40%		
Course Emperor	0.22%			0.33%			0.40%		
Archivist				0.21%	0.13%	0.08%			
Guild Master				0.33%	0.21%	0.13%	0.40%	0.25%	0.15%
Guild Warrior				0.33%	0.21%	0.13%	0.40%	0.25%	0.15%
Proactive				0.42%	0.29%	0.21%			
Talkative				0.33%	0.25%	0.13%	0.50%	0.50%	0.50%
Apprentice							0.50%	0.50%	0.50%
Artist							0.50%	0.50%	0.50%
Blacksmith							0.50%	0.50%	0.50%
Master Builder							0.50%	0.50%	0.50%
Replier Extraordinaire							0.25%	0.25%	0.25%
Squire							0.50%	0.50%	0.50%

APPENDIX B Performance Measurements

Table B.1. Student performance metrics per cluster, in the first year.

Property	Cluster 1	Cluster 2	Cluster 3	All	Significant Differences (p < 0.016)
Quizzes Grade (%)	82.22	77.50	66.00	74.19	(A, C)
Labs Grade (%)	95.96	85.45	81.32	87.28	(A, B), (A, C)
Presentation Grade (%)	91.08	78.88	76.27	81.94	(A, B), (A, C)
Exam Grade (%)	76.62	63.37	60.82	66.82	(A, B), (A, C)
Final Grade (%)	90.16	77.82	71.88	79.51	All
Attendance (%)	98.02	97.02	86.03	92.65	(A, C)
Posts (#)	36.58	22.88	16.67	24.91	(A, C)
First Posts (#)	4.92	1.63	2.00	2.91	None
Reply Posts (#)	31.67	21.25	14.67	22.00	(A, C)
Challenge Posts (#)	18.08	12.00	9.47	13.00	(A, C)
XP from Challenges (%)	97.37	89.47	72.46	84.89	(A, C)
Theoretical Challenge Posts (#)	12.42	6.75	6.27	8.49	(A, B), (A, C)
XP from Theoretical Challenges (%)	98.15	87.50	75.93	86.19	None
Lab Challenge Posts (#)	5.67	5.25	3.20	4.51	None
XP from Lab Challenges (%)	96.67	91.25	69.33	83.71	(A, C)
Quest Posts (#)	5.25	1.50	1.33	2.71	(A, B), (A, C)
XP from Quests (%)	98.00	24.50	36.53	54.86	(A, B), (A, C)
Badges (#)	38.17	29.63	21.67	29.14	(A, C)
XP from Achievements (%)	78.24	60.32	49.63	61.88	(A, B), (A, C)
Completed Achievements (#)	11.75	8.50	5.20	8.20	(A, C)
Explored Achievements (#)	17.67	14.00	12.07	14.43	(A, B), (A, C)
[A] Postmaster	2	1.25	0.666667	1.2571	(A, C)
[A] Bookworm	3	3	2.733333	2.8857	None
[A] Proficient Tool User	2.833333	2.625	1.866667	2.3714	(A, C)
[A] Rise to the Challenge	2.666667	2.25	1.866667	2.2286	None
[A] Attentive Student	2.583333	2.25	1.2	1.9143	(A, C)
[A] Class Annotator	2.25	1.875	0.6	1.4571	(A, C)
[A] Challenger of the Unknown	2.5	0.625	0.733333	1.3143	(A, B), (A, C)
[A] Lab Master	1.333333	0.375	0.666667	0.8286	(A, B)
[A] Wild Imagination	1	0.875	0.8	0.8857	None
[A] Right on Time	2.833333	2.75	1.933333	2.4286	(A, C)
[A] Amphitheatre Lover	2.833333	2.75	2.2	2.5429	(A, C)
[A] Lab Lover	2.75	2.375	1.933333	2.3143	None
[A] Popular Choice Award	1.583333	1.375	0.4	1.0286	(A, C)
[A] Hollywood Wannabe	2.166667	1.75	1.4	1.7429	None
[A] Presentation Zen Master	1	1	1	1	None
[A] Golden Star	1.333333	0.875	0.8	1	None
[A] Lab King	0.416667	0	0.266667	0.2571	None
[A] Presentation King	0.25	0.125	0.133333	0.1714	None
[A] Exam King	0.166667	0	0	0.0571	None
[A] Course Emperor	0.083333	0	0	0.0286	None
[A] Quiz King	0.166667	0.125	0	0.0857	None
[A] Quizmaster	0.333333	0.125	0	0.1429	None
[A] Good Host	2.083333	1.25	0.466667	1.2	(A, C)

Table B.2. Student performance metrics per cluster, in the second year.

Property	Cluster 1	Cluster 2	Cluster 3	Cluster 4	All	Significant Differences (p < 0.008)
Quizzes Grade (%)	93.67	89.78	81.61	77.95	85.37	(A, C), (A, D), (B, D)
Labs Grade (%)	98.90	96.10	95.54	91.65	95.38	(A, C), (A, D)
Presentation Grade (%)	91.10	89.57	84.82	85.54	87.58	None
Exam Grade (%)	80.75	74.52	70.09	65.65	72.31	None
Final Grade (%)	98.74	91.31	86.00	79.22	88.29	All
Attendance (%)	98.18	85.15	84.42	79.72	86.10	(A, B), (A, D)
Posts (#)	73.60	50.40	33.36	15.38	41.52	(A, C), (A, D), (B, D), (C, D)
First Posts (#)	4.70	2.20	1.14	0.77	2.04	(A, C), (A, D)
Reply Posts (#)	68.90	48.20	32.21	14.62	39.48	(A, C), (A, D), (B, D), (C, D)
Challenge Posts (#)	46.40	37.33	27.93	12.38	30.31	(A, C), (A, D), (B, D), (C, D)
XP from Challenges (%)	100.00	94.67	95.51	72.75	90.44	(A, D), (B, D), (C, D)
Theoretical Challenge Posts (#)	29.50	21.93	20.14	7.38	19.27	(A, D), (B, D), (C, D)
XP from Theoretical Challenges (%)	100.00	92.73	97.40	69.93	89.69	(A, D), (B, D), (C, D)
Lab Challenge Posts (#)	16.90	15.40	7.79	5.00	11.04	(A, C), (A, D), (B, C), (B, D)
XP from Lab Challenges (%)	100.00	97.95	92.31	77.51	91.72	(A, D), (B, D)
Skill Tree Posts (#)	6.00	0.47	0.43	0.31	1.48	(A, B), (A, C), (A, D)
XP from Skill Tree (%)	30.00	0.00	2.38	0.00	6.41	(A, B), (A, C), (A, D)
Explored Skill Tree Nodes (#)	4.00	0.27	0.36	0.00	0.94	(A, B), (A, C), (A, D)
Quest Posts (#)	3.50	4.33	2.93	1.77	3.15	(B, D)
XP from Quests (%)	100.00	94.44	78.57	35.90	76.60	(A, D), (B, D)
Badges (#)	47.30	37.53	32.71	26.00	35.23	All
XP from Achievements (%)	97.07	85.68	77.56	60.37	79.36	All
Completed Achievements (#)	14.90	9.60	8.57	6.08	9.46	(A, B), (A, C), (A, D), (B, D), (C, D)
Explored Achievements (#)	20.20	17.20	15.43	13.08	16.27	All
[A] Postmaster	3	2.466667	1.642857	0.461538	1.8462	(A, C), (A, D), (B, C), (B, D), (C, D)
[A] Talkative	1.6	0.933333	0.214286	0.153846	0.6731	(A, C), (A, D)
[A] Bookworm	3	2.866667	3	2.923077	2.9423	None
[A] Proficient Tool User	3	2.866667	2.5	1.846154	2.5385	(A, D), (B, D)
[A] Rise to the Challenge	3	2.6	2.857143	1.692308	2.5192	(A, D), (B, D), (C, D)
[A] Attentive Student	2.2	1.2	0.428571	0.076923	0.9038	(A, C), (A, D), (B, D)
[A] Class Annotator	2.6	1.333333	0.571429	0.307692	1.1154	(A, B), (A, C), (A, D), (B, D)
[A] Challenger of the Unknown	3	2.666667	2.142857	0.923077	2.1538	(A, D), (B, D)
[A] Lab Master	2.9	2.8	2.785714	2.461538	2.7308	None
[A] Wild Imagination	1	1	1	1	1	None
[A] Right on Time	2.8	1.933333	1.785714	1.538462	1.9615	(A, B), (A, D)
[A] Amphitheatre Lover	2.9	2.133333	2.142857	1.923077	2.2308	(A, B), (A, D)
[A] Lab Lover	3	2.8	2.857143	2.692308	2.8269	None
[A] Popular Choice Award	1.1	0.666667	0.642857	0.692308	0.75	None
[A] Hollywood Wannabe	2.9	2.733333	2.571429	2.692308	2.7115	None
[A] Presentation Zen Master	1	1	1	1	1	None
[A] Guild Warrior	2.2	2.333333	2.428571	2.615385	2.4038	None
[A] Guild Master	1.6	1.333333	1	0.769231	1.1538	None
[A] Golden Star	0.8	0	0	0	0.1538	None
[A] Lab King	0.6	0.133333	0.071429	0	0.1731	(A, C), (A, D)
[A] Presentation King	0.1	0.133333	0	0	0.0577	None
[A] Exam King	0.1	0	0	0	0.0192	None
[A] Course Emperor	0.1	0	0	0	0.0192	None
[A] Quiz King	0.2	0.066667	0	0	0.0577	None
[A] Quizmaster	1.5	1.066667	0.357143	0	0.6923	(A, C), (A, D), (B, D)
[A] Archivist	0	0	0.214286	0	0.0577	None
[A] Proactive	1.1	0.466667	0.5	0.230769	0.5385	(A, D)

Table B.3. Student performance metrics per cluster, in the third year.

Property	Cluster 1	Cluster 2	Cluster 3	Cluster 4	All	Significant Differences (p < 0.008)
Labs Grade (%)	96.24	94.28	89.23	73.36	89.06	(B, D)
Presentation Grade (%)	78.50	83.33	82.23	64.18	78.54	None
Exam Grade (%)	63.67	65.50	56.02	55.67	60.98	(B, C)
Final Grade (%)	91.79	84.28	68.09	54.94	75.38	(A, C), (A, D), (B, C), (B, D), (C, D)
Attendance (%)	82.31	87.78	50.55	33.77	67.11	(A, D), (B, C), (B, D)
Posts (#)	76.71	48.09	21.77	10.55	37.81	(A, B), (A, C), (A, D), (B, C), (B, D)
First Posts (#)	3.86	2.39	0.46	0.55	1.74	(A, C), (A, D), (B, C)
Reply Posts (#)	72.86	45.70	21.31	10.00	36.07	(A, B), (A, C), (A, D), (B, C), (B, D)
Rated Posts (#)	44.57	28.22	14.08	7.55	22.72	(A, B), (A, C), (A, D), (B, C), (B, D)
Mean Rate	3.46	3.35	3.33	3.01	3.29	None
Challenge Posts (#)	16.86	16.26	9.69	5.27	12.52	(A, C), (A, D), (B, C), (B, D)
XP from Challenges (%)	100.00	95.65	71.79	40.26	79.19	(A, C), (A, D), (B, C), (B, D)
Theoretical Challenge Posts (#)	8.71	8.65	5.54	2.82	6.72	(A, D), (B, C), (B, D)
XP from Theoretical Challenges (%)	100.00	95.65	71.79	36.36	78.40	(A, D), (B, C), (B, D)
Lab Challenge Posts (#)	8.14	7.61	4.15	2.45	5.80	(A, C), (A, D), (B, C), (B, D)
XP from Lab Challenges (%)	100.00	95.65	71.79	45.45	80.25	(A, D), (B, C), (B, D)
Skill Tree Posts (#)	20.86	12.43	4.62	3.18	9.76	(A, B), (A, C), (A, D), (B, C), (B, D)
XP from Skill Tree (%)	100.00	64.57	20.77	17.95	49.12	(A, B), (A, C), (A, D), (B, C), (B, D)
Explored Skill Tree Nodes (#)	11.71	7.91	2.69	2.27	6.00	(A, B), (A, C), (A, D), (B, C), (B, D)
MCP Quest Posts (#)	2.29	2.30	1.31	0.18	1.63	(A, D), (B, D)
XP from MCP Quest (%)	100.00	95.65	69.23	18.18	74.07	(A, D), (B, D)
AvatarWorld Posts (#)	17.57	5.17	3.62	0.09	5.37	(A, B), (A, C), (A, D), (B, D)
XP from AvatarWorld (%)	97.62	25.36	10.26	1.52	26.23	(A, B), (A, C), (A, D)
AvatarWorld Submissions (#)	21.85714	6.26087	2.461538	0.181818	1.5926	(A, B), (A, C), (A, D)
Badges (#)	50.43	40.70	28.77	20.09	34.89	All
XP from Achievements (%)	100.00	93.88	69.97	44.36	78.83	(A, C), (A, D), (B, C), (B, D), (C, D)
Completed Achievements (#)	14.57	11.52	6.92	4.55	9.39	All
Explored Achievements (#)	22.71	18.91	15.69	11.09	17.04	All
[A] Postmaster	3	2.130435	0.692308	0.272727	1.5185	(A, B), (A, C), (A, D), (B, C), (B, D)
[A] Talkative	0.714286	1.173913	0.307692	0	0.6667	(A, D), (B, D)
[A] Bookworm	3	2.956522	2.692308	2.636364	2.8333	None
[A] Proficient Tool User	3	2.869565	2.153846	1.363636	2.4074	(A, D), (B, C), (B, D)
[A] Rise to the Challenge	3	2.869565	2.153846	1.090909	2.3519	(A, D), (B, C), (B, D)
[A] Apprentice	0.571429	0.130435	0	0	0.1296	(A, C), (A, D)
[A] Attentive Student	0.857143	0.173913	0	0	0.1852	(A, C), (A, D)
[A] Class Annotator	2.571429	1.173913	0.153846	0	0.8704	(A, C), (A, D), (B, D)
[A] Blacksmith	3	0.956522	0.384615	0.090909	0.9074	(A, B), (A, C), (A, D)
[A] Master Builder	2.857143	0.565217	0.230769	0	0.6667	(A, B), (A, C), (A, D)
[A] Squire	1.571429	0.217391	0.076923	0	0.3148	(A, B), (A, C), (A, D)
[A] Challenger of the Unknown	1	0.956522	0.692308	0.181818	0.7407	(A, D), (B, D)
[A] Lab Master	2.714286	2.913043	2.615385	2.181818	2.6667	(B, D)
[A] Replier Extraordinaire	3	2.956522	2.769231	2.181818	2.7593	(B, D)
[A] Wild Imagination	1	1	1	0.818182	0.963	None
[A] Artist	3	2.565217	1.461538	0.818182	2	(A, C), (A, D), (B, C), (B, D)
[A] Right on Time	1.857143	2	0.615385	0.363636	1.3148	(A, D), (B, C), (B, D)
[A] Amphitheatre Lover	2	2.26087	0.923077	0.545455	1.5556	(A, D), (B, C), (B, D)
[A] Lab Lover	2.857143	2.652174	2.307692	2	2.463	None
[A] Popular Choice Award	0.714286	0.304348	1.153846	0.363636	0.5741	(B, C)
[A] Hollywood Wannabe	2.142857	2.304348	1.307692	1.545455	1.8889	(B, C)
[A] Presentation Zen Master	1	1	1	0.818182	0.963	None
[A] Guild Warrior	2	2	2	1.818182	1.963	None
[A] Guild Master	1.857143	1.217391	1	0.545455	1.1111	(A, D)
[A] Golden Star	1	1.130435	0.846154	0.454545	0.9074	(B, D)
[A] Lab King	0.142857	0.086957	0	0	0.0556	None
[A] Presentation King	0	0	0.230769	0	0.0556	None
[A] Exam King	0	0.086957	0	0	0.037	None
[A] Course Emperor	0	0.043478	0	0	0.0185	None

APPENDIX C Satisfaction Questionnaire

Table C.1. Satisfaction questionnaire item description and usage history.

Question	Year 1	Year 2	Year 3	Code
1. How would you classify, in general, the gamification experiment performed this year in the PCM course? (1-terrible; 5 - excellent)	x	x	x	Overall
2. How would you classify the new gamified course (1-Much less; 5 - Much more), when compared to regular courses, in terms of:	x	x	x	
2.1. Motivation	x	x	x	Motivation
2.2. Interest	x	x	x	Interest
2.3. Amount of Work	x	x	x	Work
2.4. Difficulty	x	x	x	Difficul
2.5. Easiness in learning course materials	x	x	x	Easiness
3. How would you classify the amount of study you performed for this course when compared with regular courses (1 - Far Less; 5 - Far More), in terms of:	x	x	x	
3.1. Quantity	x	x	x	Quantity
3.2. Continuity	x	x	x	Continuity
4. How much did you feel you were playing a game, instead of just attending a regular course? (1 - Not at all; 5 - A lot)	x	x	x	Game Feeling
5. Do you think the achievements should account for a higher parte of the grade (instead of the exam, etc.) (1-definitely not; 5 - definitely yes)	x	x		Achievement Grade
6. When faced with non-mandatory tasks that would earn you an achievement, did you perform them solely for the grade's sake, or also to get a better score in the game (new achievement, better position in the leaderboard,etc) (1-grade only; 5 - game only)	x	x	x	Non-mandatory
7. Currently, different levels of the same achievement have diminishing returns (Level 1 will earn you more XP than Level 2, and that more than Level 3). Would you prefer: (1 = "as it is", 2 = "the same number of XP per level (less on the first to have more on the latter)", 3 = "less XP on the first levels. more on the last")	x	x		Level Distribution
8. Do you think achievements that required extra actions, such as "Class Annotator", "Quests" and "Theoretical Challenges" contributed to your learning experience? (1-Not at all; 5 - definetely)	x	x	x	Extra Actions
9. Would you think it is a good idea to extend gamification to other courses? (1-definitely not; 5 - definitely yes)	x	x	x	Extend
10. How would you classify (1-Uneven; 5-Even) the distribution of the amount of work asked of you throughout the semester?		x	x	Workload
11. How would you classify (1-Uneven; 5-Even) the distribution of the amount of work you put into this course throughout the semester (accessing moodle, reading the slides, participating in the challenges, posting, etc)?		x	x	Workload MCP
12. Was the game collaborative (1) or competitive (5)?		x	x	Competitive
13. How would you rate the skill tree (1-Terrible; 5-Great)?		x		Rate Skill Tree
14. Rate from 1-Totally Disagree to 5-Totally Agree, the following statements:		x	x	
14.1. The course allowed me to get rewards for things I like to do		x	x	Things I Like
14.2. The course allowed me to be creative		x	x	Creative
14.3. The course taught me useful skills for my future		x	x	Useful

Continuation of Table C.1

15. Do you think you would have learned more if this would have been a "traditional" course: an exam and a large(ish) programming project? (1 = Yes, 2 = No, 3 = The Same)			x	Learn Traditional
16. Do you think you would have gotten a better grade if this would have been a "traditional" course: an exam and a large(ish) programming project? (1 = Yes, 2 = No, 3 = The Same)			x	Grade Traditional
17. What would you prefer: a final exam or several small quizes ("minitestes"), for instance, five questions each, every other lecture on the subjects? (1 - Exam; 5 - Quizes)			x	Exam vs. Quizzes
18. The Quest. How did you find it? (1-Uninteresting; 5-Interesting)			x	The Quest
19. In your opinion, how do the following course features make you feel engaged with the course? Please rate each of them on the following scale: 1 = not engaged 2 = slightly engaged 3 = somewhat engaged 4 = moderately engaged 5 = extremely engaged			x	
19.1. AvatarWorld			x	Engaged AW
19.2. Badges			x	Engaged Badges
19.3. Challenges			x	Engaged Challenges
19.4. Experience Points			x	Engaged XP
19.5. Lab classes			x	Engaged Labs
19.6. Leaderboard			x	Engaged Leaderboard
19.7. Levels			x	Engaged Levels
19.8. MCP Quest			x	Engaged Quest
19.9. Skill Tree			x	Engaged Skill Tree
19.10. Theoretical classes			x	Engaged Lectures

APPENDIX D Adapted Student Course Engagement Questionnaire

To what extent do the following behaviors, thoughts, and feelings describe you, in this course. Please rate each of them on the following scale:

1 = not at all characteristic of me
2 = not really characteristic of me
3 = moderately characteristic of me
4 = characteristic of me
5 = very characteristic of me

1. Making sure to study on a regular basis
2. Working hard
3. Doing all the homework problems
4. Reading the lecture slides regularly
5. Looking over class notes between classes to make sure I understand the material
6. Being organized
7. Taking good notes in class
8. Listening carefully in class
9. Coming to every class
10. Finding ways to make the course material relevant to my life
11. Applying course material to my life
12. Finding ways to make the course interesting to me
13. Thinking about the course between class meetings
14. Really desiring to learn the material
15. Raising my hand in class
16. Asking questions when I don't understand the professor
17. Having fun in class
18. Participating actively in small-group discussions
19. Using moodle forums or talking to the professors outside of class to review assignments or ask questions
20. Helping fellow students
21. Getting a good grade
22. Doing well on the different class assignments
23. Being confident that I can learn and do well in the class

Table D.1. Adaptations made to the Student Course Engagement Questionnaire.

Question	Change (=>)	Reason
2	Putting forth effort => Working hard	Easier to understand by non-native english speakers.
4	Staying up on the readings => Reading the lecture slides regularly	Easier to understand by non-native english speakers.
9	Coming to class every day => Coming to every class	We did not have classes every day. Easier to understand by non-native english speakers.
16	Asking questions when I don't understand the instructor => Asking questions when I don't understand the professor	Easier to understand by non-native english speakers.
19	Going to the professor's office hours to review assignments or tests or to ask questions => Using moodle forums or talking to the professors outside of class to review assignments or ask questions	Office hours ocurred on Moodle forums, in a dedicated chatroom.
22	Doing well on the tests => Doing well on the different class assignments	We had several assignments and evaluation checkpoints other then tests.