Effect of personalized gameful design on student engagement

Alberto Mora
Estudis d'Informàtica, Multimedia i Telecomunicació
Universitat Oberta de Catalunya
Barcelona, Spain
amoraca@uoc.edu

Gustavo F. Tondello
HCI Games Group, Games Institute
University of Waterloo
Waterloo ON, Canada
gustavo@tondello.com

Lennart E. Nacke
HCI Games Group, Games Institute
University of Waterloo
Waterloo ON, Canada
lennart.nacke@acm.org

Joan Arnedo-Moreno
Estudis d'Informàtica, Multimedia i Telecomunicació
Universitat Oberta de Catalunya
Barcelona, Spain
jarnedo@uoc.edu

***Abstract*—Many recent studies of gamification applied to higher education have demonstrated a wide range of positive results. However, most of them fail to consider any personalization factor for the student experience, despite recent studies having shown that gameful systems may be more engaging when they are personalized to each user. Therefore, the goal of this work is to investigate if gameful learning experiences can better motivate and engage students if they are personalized. In this way, we present the design and analysis of a personalized gameful learning experience within a Computer Network Design course. The general purpose of this study is to determine whether a personalized gameful learning experience affects both the students' behavioral and emotional engagement. The results of a descriptive analysis reveal that personalization works better than generic approaches in all items regarding the behavioral and emotional engagement of the students, being a promising standpoint to further investigate in subsequent studies.**

Keywords—gameful design; gamification; personalization; learning; preferences; Hexad.

I. INTRODUCTION

The lack of interest and motivation of the students in the learning process is a concern among the educational community and higher education studies are not an exception. Given the motivational properties of game elements, gamification—commonly known as the use of game-design principles and elements in non-game environments [1]—has emerged as a promising and powerful technique for shaping behaviour in learning environments since six-seven years ago. The goal of gameful design is to create motivating experiences (in this scope, learning experiences), to promote the same feelings and the same commitment that people experience when playing a game, even when the main objective is not pure entertainment. Diverse studies on the use of elements and principles of game design in educational contexts have been developed [2] revealing promising outcomes to motivate students. Existing literature also shows a great interest in the topic reporting many practises with a diverse range of success in our context education [3], being almost, if not all, one-size-fits-all approaches.

Thus, gamification of learning in higher education seems to be a key to increase student motivation and commitment; accordingly, the interest of the researches and practitioners is reflected in the variety of proposed design frameworks in education [4]. However, the effect described in the amount of study cases available in the literature seems to be always limited because they have not been designed considering the characteristics of each student. Diverse psychological viewpoints agree that people are not equal, therefore, they cannot be motivated effectively in the same way. Personalization could be an approach on gameful design to motivate students more effectively, but it has been minimally explored yet and less applied [5]. In recent times, the term "Gamification Persona" [6] has emerged as the way of gathering and keeping visible the objectives of the appropriate users, considering issues of the personality and its motivations. Therefore, acquiring knowledge about the students with the purpose of building personalized experiences is a relevant task.

Consequently, early studies on personalized interactive systems seem to be more effective than one-size-fits-all approaches [5], requiring an adaptation of the gameful experience to the user's preferences. These approaches are usually system-tailored contents and services that fit different users' characteristics [7]. In this way, diverse authors [8, 9] studied the relationship between player types and personality traits in gameful systems, aiming to identify potential relationships with game design elements. The purpose was to obtain some findings for designing gameful systems, specially targeting users' intrinsic motivation, based on their knowledge and experience.

Therefore, regarding the relationship between users and their specific preferences when interacting with gameful environments, there is an open research niche to be covered. Accordingly, an exploratory study previously investigated the different interaction types with gameful digital applications based on user types and preferences for different game design elements revealing relationships between gender, age, and culture among and between user types and game design elements [10]. Considering the evidence that users experience gameful systems differently, according to their preferences, we set the following research questions in the context of a gameful personalized learning experience:

- **RQ1:** Does an online personalized gameful learning experience have a greater impact on students´ engagement than a generic gameful learning experience?
- **RQ2:** Does the same gameful learning experience have a greater impact on students' engagement when they are assigned to different tailored versions based on their user type compared to students that are assigned to different versions randomly?
- **RQ3:** Do the different personalized gameful learning experiences by user types engage students by the same factor?

Thus, this paper is structured as follows: first, in order to better understand our approach, we describe the design of the personalized gameful experience in Section II. Then, we describe the principles and methodology used for the analysis process in Section III. After (Section IV), we present the results obtained in the study case. In the discussion (Section V), we analyse and try to answer the proposed research questions. Finally, we conclude this work and provide some insights and future work.

II. DESIGN

The design process was based on the SPARC (Sense, Purpose, Autonomy, Relatedness and Competence) model [11], which has been successfully used in other gameful experiences directed to adult learners. This model structures the design process according to three dimensions that must be defined: metaphor, rules, and tool. It also provides the basic guidelines to later evaluate of the emotional engagement of the experience.

A. Metaphor

The metaphor contextualizes the activity in the learning context, giving sense to the whole activity, as well as introducing the narrative. On that regard, students were presented with the following introductory text, which summarizes such narrative:

"We are in the year 2025. A consortium of high technology companies decides to act as benefactors of the Cisco Dharma Initiative (CDI), a project to stimulate research, innovation and development (R & D & I) in all types of areas. In its initial phase, the program has deployed four underwater stations where it is expected to develop leading projects in a high tech environment."

Students become recruits who are assigned to four different underwater stations, according to their aptitudes. An overseer will monitor the recruits' actions during their tour of duty, which requires completing several challenges and solving dangerous situations that will unexpectedly arise (as well as uncovering some mysteries that will appear at a later stage). The metaphor mostly follows the mythology laid out by the TV series "Lost" [12], but also takes a bit from videogames such as SOMA [13] and the Fallout and System Shock series [14, 15].

B. Rules

The rules describe the basics of the activity, which interactions can be carried out by the players and which are the expected results.

Onboarding

Recruits must complete the "station assignment test" in order to start their adventure, which proposes several situations while asking how they would act in each one of them, or asks recruits to self-evaluate their preferences when interacting in a gameful environment. This is actually, unbeknown to the recruits, through use of the Hexad User Types scale [9]. The results of this test determine to which station they well be assigned. Each station has a motto that lets recruits guess what they are about.

- Alpha: "Progress relies on competition"
- Beta: "Small circles in harmony"
- Delta: "Free spirits"
- Gamma: "One for all, all for one"

Midgame

There are two separate scoreboards for each station: Research and Development. The "Overseer" periodically publishes challenges (exercises, such as lab practices or tests) assigned to one of them. The station scores points in Research or Development whenever a challenge is successfully completed by their recruits. Rewards are unlocked when some combination of scores is reached in both scoreboards. For instance, when they achieve 10 Research points and 15 Development points, they get bonus questions or extra days for a class deliverable.

However, the specific rules about who gets the rewards or how scoring is achieved change depending on the station. That's the personalization focus. For instance, station Alpha is competitive (not everybody gets the reward) and recruits work individually, whereas station Beta is mostly collaborative and they must work in subgroups. At a later stage, a new scoreboard appears unexpectedly: Crisis. Here, additional challenges are sometimes proposed, each assigned to a narrative. For instance, a reactor meltdown, a hull breach, or some enemy that has to be defeated, such a polar bear or the black mist (both from the lost mythology). In case the recruits are unable to solve the challenge before a deadline expires, it is Game Over.

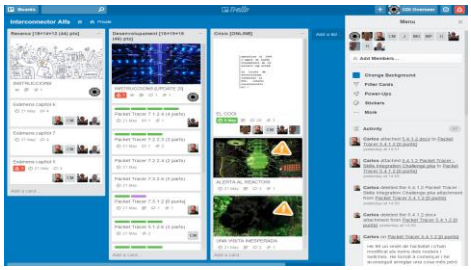
Endgame

The gameful experience is considered finished at the end of the course. Recruits are considered to "win" if they survived their tour of duty (i.e. averted every Crisis) and uncovered the mystery. There is no special reward for this, apart from the satisfaction of having been able to succeed. The outcome is purely in from a narrative standpoint.

C. Tool

At the end, everything that is defined in the previous dimensions must be somehow implemented and deployed in the online learning environment. Therefore, it is relevant to take into account from the very beginning which tools exist, or must be created, that will be capable of this. Two tools were mainly used during the experience. On one hand, a micro-blogging widget embedded in the virtual classroom. The "Overseer" would use this tool to frequently send messages about the underwater stations' status (e.g. new challenges, scores, goading or congratulating players about their actions, etc.). On the other hand, all challenges were published and managed using Trello boards, one for each station (see Figure I). A deck of cards was used for each scoreboard type: Research, Development and Crisis.

FIGURE I. ALPHA'S STATION IN TRELLO BOARD



III. METHODOLOGY

The analytical process described in the following subsections was conducted using the Minitab statistical tool (version 17) for Windows platform [16], a software environment for statistical computing.

A. Participants

This study was held at Universitat Oberta de Catalunya[1] (UOC), a fully on-line university in Barcelona, Spain. All courses at UOC are usually conducted in virtual environments. In this study, students belonged to one of the two groups of the Computer Network Design course (CND), Catalan (CAT) and Spanish (CAS), during the academic season 16/17. It is an optional subject of the Computer Engineering degree program. Students' submissions in the course are done in their native language (Catalan or Spanish), although the contents of the subject are presented in the same language: English. In total, 81 students enrolled in both groups: 60 of them in CAT (74.07%) and 21 to CAS (25.92%).

Regarding age, students in CAT present an average age of 36.15 (SD=7.55), whereas students in CAS had an average age of 39.29 (SD=6.30) (see descriptive summary in Table I). Thus, in order to know the normality of the samples, we run the Anderson-Darling normality test. CAT and CAS present a p-value of 0.151 and 0.634 respectively. Both locations failed to reject the null hypothesis at α=0.05 significance level. Therefore, there is not enough evidence to conclude the data do not come from normally distributed populations.

To compare the variability of the two samples, we use confidence intervals to contrast the standard deviation ratio for the two samples. The confidence interval for normally distributed data contains 1; therefore, it fails to reject the null hypothesis that the ratio equals 1. The p-values for both tests are well above α=0.05, so we fail to reject the null hypothesis that the ratio of the standard deviations is one. The results suggest there is no difference in the standard deviations of the students' age of two groups and their normality distribution.

TABLE I. DEMOGRAPHIC SUMMARY

Group	Descriptive analysis					
	N	***Mean***	***StDev***	***Min***	***Median***	***Max***
CAT	*60*	*36.15*	*7.55*	*20.00*	*35.00*	*56.00*
CAS	*21*	*39.29*	*6.55*	*28.00*	*39.00*	*54.00*

Regarding gender, 54 of students were male (90.00%) and 6 were female (10.00%) in CAT meanwhile 20 of them were male (95.00%) and 1 was female (5.00%) in CAS. Pearson's Chi-Square test allows us to test how likely it is that the distribution of males and females in each group. The null hypothesis would be that gender and the group are independent of one another. It returns a value of $\chi^2 = 0.541$; df = 1; p = 0.462. Thus, the null hypothesis cannot be rejected. We therefore operate under the assumption that there is not an influence about the group to which the student belongs by gender.

Regarding personal situation, 12 of students in CAT already had a university degree (20.00%), 42 worked full-time (70.00%), and 17 had children (28.33%). Regarding CAS, 6 of them already had a university degree (28.57%), 17 of them worked full-time (80.96%), and 6 had children (28.57%). The Chi-Square test returned a value of $\chi^2 = 0.260$; df = 2; p = 0.878. Thus, the null hypothesis cannot be rejected. We therefore operate under the assumption that there is no influence in the group to which the students belong and their personal situation.

B. Data collection

Student engagement was measured regarding the proposed objective (to motivate grade-level students to solve non-evaluative activities) through the collection of both quantitative and qualitative data. Quantitative data were collected from the users' logs generated by Trello and qualitative data came from an anonymous survey at the end of the course. Therefore, two viewpoints have been considered respectively by means of these kind of analysis: behavioral and emotional engagement [17]. On the one hand, behavioral engagement concerns involvement in learning and academic tasks and includes behaviors such as effort and persistence, among others [18]. The development of non-evaluative tasks (self-assessment and practices) is measured within student's behavioral engagement and individually by means of the number of active students (those who have participated in at least one non-evaluative activity) and the total of developed tasks by each group. On the other hand, emotional engagement refers to students' affective reactions in the classroom, including interest, boredom, happiness, sadness, and anxiety [18]. They revealed how motivation leads not only to behavioral engagement, but also to emotional engagement.

[1] www.uoc.edu

C. Realibility

In order to measure the reliability of the measurement scale of items presented in the final survey, we use Cronbach's alpha for the dataset of the 12 items presented in a five-level Likert scale. Cronbach's alpha assumes that the questions are only measuring one latent variable or dimension. In our case, we are only measuring emotional student engagement during the gameful learning experience. The alpha coefficient for the 12 items presented in the CAT survey is 0.915, while 0.943 was obtained in CAS. This suggests that the items have high internal consistency (note that a reliability coefficient of 0.70 or higher is considered "acceptable" in most social science research [19], although some authors suggest higher values of 0.90 to 0.95 and the obtained value is located within that interval).

D. Procedure

The CND course comprises a total of twenty school weeks within the second quarter of the term, and started on February 2017. The gameful experience was encompassed in fourteen weeks. Students belonged to the CAS or CAT group according to the native language recorded in their academic profile. Once the students were assigned to each group, a survey was proposed to assign them to each sub-group (called Alfa, Beta, Delta, Gamma) in the CAT group (CAS does not have subgroups; it presents the same one-size-fits-all gameful experience to all students). A survey was enabled for two weeks and published in the native language of students. The survey was adapted to this context from the Gamification User Types Hexad Scale [9] in order to apply the metaphor from the current gameful learning experience but remaining the original statements of the scale.

Once all students in CAT completed the survey (an online test), they were assigned to one of the sub-groups through the following algorithm (see Scheme I) where some variables are used: 'S' refers to the different groups (stations) meanwhile 'G' (generic) to the specific user types described in the Gamification User Types Hexad Framework [20]. Moreover, 'H' refers to the combination of user types (hybrid user types, a fine-grained combination of user types) with more presence in an exploratory study previously conducted [10] and the pair (V1,V2) are the highest values of each user type resulting from the test.

Thus, the purpose of the assignment function is linking each student to the most adequate gameful experience (according to the previous related theory [9, 10]). In this sense, by means of the result of the Hexad User Types Scale, we tried to fit some of the hybrid profiles (participants that score high in more than one Hexad User Type) with the groups closer to their highest user type scores. When this was not the case, we only considered the participant's highest user type score to make the assignment. In the case of the primary user type was Disruptor (the less frequent by far), we only considered the second highest user type. This assignment is a procedure of its own and based on the findings in a wide-range exploratory study [10]. Within, the authors investigated different perspectives of measuring user types, from a coarse-grained (generic), to a fine-grained considering combination of them (hybrid user types) and how motivation is affected by the game design elements tailoring to particular user types. However, the process described in this work can be replicated by any researcher in further studies.

SCHEME I. ASSIGMENT PROCEDURE

```
S = [Delta, Gamma, Beta, Alfa];
G = [Free Spirit, Philanthropist, Socializer, Player, Achiever, Disruptor];
H = [(Free Spirit, Achiever),(Philanthropist, Free Spirit),(Philanthropist,
Achiever),(Philanthropist, Socializer),(Player, Free Spirit),(Player,
Achiever)];

Function Assignment (Student) {
  For each Student in CAT {
    Profile=Hexad User Type Test (Student)
      For each (V1,V2) in Profile; {
        if (V1,V2) or (V2,V1) == H[1], return (Student, S[1]));
        if (V1,V2) or (V2,V1) == H[2], return (Student, S[2]));
        if (V1,V2) or (V2,V1) == H[3], return (Student, S[2]));
        if (V1,V2) or (V2,V1) == H[4], return (Student, S[3]));
        if (V1,V2) or (V2,V1) == H[5], return (Student, S[4]));
        if (V1,V2) or (V2,V1) == H[6], return (Student. S[4]));
        if V1 == Free Spirit, return (Student, S[1]));
        if V1 == Philanthropist, return (Student, S[2]));
        if V1 == Socializer, return (Student, S[3]));
        if V1 == Player, return (Student, S[4]));
        if V1 == Achiever, return (Student, S[4]));
        if V1 == Disruptor {
            if V2 == Free Spirit, return (Student, S[1]));
            if V2 == Philanthropist, return (Student, S[2]));
             if V2 == Socializer, return (Student, S[3]));
            if V2 == Player, return (Student, S[4]));
             if V2 == Achiever, return (Student, S[4]));
       }
      }
    }
}
```

Accordingly, within the 48 hours after the completion of the survey, students were individually informed to sign into the specific Trello dashboard (station). In the end, previous the publication of marks (on the 20th week of the course), students from both groups (CAT and CAS) were asked to run a voluntary post-survey in their native language. A total of 49 of students in CAT (81.66%) and 17 in CAS (80.95%) completed the post-survey.

IV. RESULTS

A. Impact of personalization for the students' engagement

In order to answer the proposed **RQ1**, the alignment with data collection methods and data analyses are described in Table II, as follows:

TABLE II. ANALYSIS SUMMARY

RQ1	Does an online personalized gameful learning experience have a greater impact on students´ engagement than a generic gameful learning experience?	
Process	A comparison of CAT/CAS as whole from two student engagement perspectives: behavioral and emotional	
Method	Descriptive analysis	Mann Whitney U test
Input	Trello log	Survey

First, we start the analysis of the student behavioral engagement through the development of non-evaluative activities (self-assessment and practices) being measured by the number of developed tasks by each group and the number of active students (those who have participated in at least one non-evaluative activity), thanks to Trello's log.

Actives students (who developed at least one non-evaluative tasks) were 39 in CAT (65.00% out of the 60 total), detailed by types of tasks as follows: 28 students (46.67%) participated of the simulation tasks, meanwhile 34 students (56.67%) participated of the self-assessment exams. Deeping into each CAT subgroup (personalized gameful experiences), the Alpha subgroup presented a student participation of 54.55% of the simulation tasks and 81.82% of the self-assessment exams; Beta revealed a participation of 47.05% and 41.17%; Delta, 35.29% and 41.17%; and finally, Gamma showed 53.34% and 73.34%, respectively.

Regarding CAS, 11 students were identified as active (51.00% out of the total 21), with 8 of them having completed simulation tasks (38.09%), and 8 (38.09%) completed self-assessment exams. In the following Table (Table III), we summarize the descriptive analysis of the two types of activities as a whole (number activities that the student participated) of both groups. Note that the ratio student/task is the same regardless of the size of each group and subgroup in this study.

TABLE III. STUDENT TASK DEVELOPMENT COMPARISON

Group	**Descriptive analysis**					
	N	***Mean***	***StDev***	***Min***	***Median***	***Max***
CAT	*60*	*6.42*	*8.85*	*0.00*	*2.00*	*34.00*
CAS	*21*	*3.90*	*5.30*	*0.00*	*2.00*	*16.00*

Next, we run the U-Mann-Whitney Test, commonly used to check the heterogeneity of two ordinal samples. The starting assumption is that the observations of both groups are independent. Therefore, under the null hypothesis, the starting distribution of both groups is the same, whereas the alternative hypothesis reflects that the values of one of the samples tend to exceed those of the other (personalized vs one-size-fits-all). We can be 95.10% confident that the difference between the population medians is between -0.001 and 3.002. The null hypothesis states that the difference in the median of participated task in a group is 0. Because the p-value of 0.0913 is higher than confidence level of 0.05, we cannot reject the null hypothesis, and cannot conclude that there is a statistical significance between the groups.

Now, if we only consider the two types of activities with active students, we proceed to run the following descriptive analysis (see Table IV). U-Mann-Whitney test is not significant at p = 0.7781 with a confidence level of 0.05; therefore, we cannot reject the null hypothesis, and we cannot conclude that there is a statistical significance between the groups.

TABLE IV. ACTIVE STUDENT TASK DEVELOPMENT COMPARISON

Group	**Descriptive analysis**					
	N	***Mean***	***StDev***	***Min***	***Median***	***Max***
CAT	*39*	*9.87*	*9.31*	*1.00*	*7.00*	*34.00*
CAS	*11*	*7.45*	*5.20*	*2.00*	*7.00*	*16.00*

Regarding emotional engagement of students, we analyzed the emotion-related intrinsic principles described in SPARC model [11] (Sense, Purpose, Autonomy, Relatedness and Competence) through the feedback from survey. Students where asked in a five-level Likert scale about the level of emotional perception regarding the gameful learning experience which they were involved on (from value 1 "very irrelevant" to 5 "very relevant"). Thus, the following Table (Table V) shows a comparison between CAT/CAS for each SPARC item. Additionally, students valued the perception of the experience as 3.84 and 3.47 respectively.

Additionally, another element to be considered regarding the emotional engagement is the interest in having similar experiences in other subjects in the future. Students were asked about it and results reveal that 78.43% of students in CAT agree, a much higher value than 52.63% in CAS. Moreover, we run the U-Mann-Whitney Test to analyze the statistical significance of the different value of each item between the groups, all being non-significant at p = 0.9142, 0.2664, 0.4200, 0.2491, and 0.8653 respectively, with a confidence level of 0.05. Based on these results, we cannot reject the null hypothesis and cannot conclude that student emotional perception is different between the two groups.

TABLE V. EMOTIONAL PERCEPTION COMPARISON (AVERGAGE)

Group	**Intrinsic motivation item**				
	S	***P***	***A***	***R***	***C***
CAT (N=49)	*3.82*	*4.10*	*3.82*	*3.28*	*4.00*
CAS (N=17)	*3.76*	*3.82*	*3.74*	*2.94*	*3.88*

B. Impact of condition assignment based on the student's user type in comparison to random assignment

Next, in order to answer the proposed **RQ2**, the alignment with data collection methods and data analyses is described in Table VI, as follows:

TABLE VI. ANALYSIS SUMMARY

RQ2	Does the same gameful learning experience have a greater impact on students' engagement when they are assigned to different tailored versions based on their user type compared to students that are assigned to different versions randomly?	
Process	A comparison of Gamma1 (CAT) / Gamma2 (CAS) from two student engagement perspectives: behavioral and emotional	
Method	Descriptive analysis	Mann Whitney U test
Input	Trello log	Survey

As we described before, we start the analysis of the student behavioral engagement through the development of non-evaluative activities (self-assessment and practices). Therefore, we analyze and compare subgroups Gamma1 (CAT) and Gamma2 (CAS), where the gameful learning experiences designed are the same. Students were assigned to subgroup Gamma1 based on the result of their initial user type test, meanwhile Gamma2 is formed by all kinds of students independent of their preferences because students in the CAT group have not performed the initial test. Thus, Gamma1 is comprised by 15 students and Gamma2, 21 students. Active students where 11 in Gamma1 (73.34% of the total), detailed by types of tasks as follows: 8 students (53.34%) completed the simulation tasks, while 11 students (73.34%) completed self-

assessment exams. Regarding Gamma2 in CAS, 11 students were identified as active (52.38% of total), with 8 (38.09%) of them having completed simulation tasks and also 8 (38.09%) having completed self-assessment exams. As follows, we summarize the descriptive analysis of two types of activities as a whole (number activities participated by a student) of both subgroups on Table VII:

TABLE VII. STUDENT TASK DEVELOPMENT COMPARISON

Group	Descriptive analysis					
	N	*Mean*	*StDev*	*Min*	*Median*	*Max*
Gamma 1(CAT)	*15*	*7.33*	*6.88*	*0.00*	*5.00*	*20.00*
Gamma2 (CAS)	*21*	*3.90*	*5.30*	*0.00*	*2.00*	*16.00*

Next, we run the U-Mann-Whitney test. Under the null hypothesis, the starting distribution of both groups is the same, whereas the alternative hypothesis reflects that the values of one of the samples tend to exceed those of the other (student assigned by test vs random assignment). The point estimate of the population median for the difference in the number of task completed by students in the two groups is 3.00. We can be 95.30% confident that the difference between the population medians is between 0.001-7.000. Because the p-value is 0.0913, which is more than the significance level of 0.05, we cannot reject the null hypothesis and cannot conclude that student performance in the two subgroups is different.

If we only consider the two types of activities with active students, we run the following descriptive analysis (see Table VIII). U-Mann-Whitney test is not significant at p = 0.1467 with a confidence level of 0.05; therefore, we cannot reject the null hypothesis, and we cannot conclude that there is a statistical significance between the groups.

TABLE VIII. ACTIVE STUDENT TASK DEVELOPMENT COMPARISON

Group	Descriptive analysis					
	N	*Mean*	*StDev*	*Min*	*Median*	*Max*
Gamma 1(CAT)	*11*	*10.00*	*6.08*	*3.00*	*10.00*	*20.00*
Gamma2 (CAS)	*11*	*7.45*	*5.20*	*2.00*	*7.00*	*16.00*

Regarding emotional engagement of the students, we compare the emotion-related intrinsic principles described in the SPARC model. Thus, the following Table (Table IX) shows a comparison between Gamma1/Gamma2 for each item. Additionally, students value the perception of the experience as 3.93 and 3.47 respectively. Another element to be considered regarding the emotional engagement is the interest of students to join similar experiences in other subjects in the future: results reveal that 86.67% of students in Gamma1 agree with this affirmative, a much higher value than 52.63% in Gamma2.

TABLE IX. EMOTIONAL PERCEPTION COMPARISON (AVERGAGE)

Group	Intrinsic motivation item				
	S	*P*	*A*	*R*	*C*
Gamma1 (N=15)	*3.80*	*4.13*	*4.20*	*3.53*	*4.00*
Gamma2 (N=17)	*3.76*	*3.82*	*3.74*	*2.94*	*3.88*

Moreover, we run the U-Mann-Whitney test to analyze the statistical significance of the difference on the responses to each item between the subgroups, with all being non-significant at p = 0.9053, 0.2530, 0.6733, 0.1286, and 0.9366 respectively with a confidence level of 0.05. Based on these results, we cannot reject the null hypothesis and cannot conclude that student emotional perception in the two subgroups is different.

C. Comparison of the engagement impact factor between different personalized experiences

Finally, in order to answer the proposed **RQ3**, the alignment with data collection methods and data analyses is described in Table X, as follows:

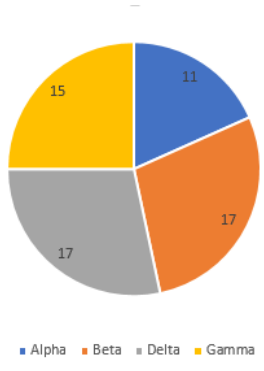
TABLE X. ANALYSIS SUMMARY

RQ3	Do the different personalized gameful learning experiences by user types engage students by the same factor?	
Process	A comparison of Alpha, Beta, Delta, Gamma (CAT) from two student engagement perspectives: behavioral and emotional	
Method	Descriptive analysis	ANOVA
Input	Trello log	Survey

Similar to the previous steps, we start the analysis of the student behavioral engagement through the development of non-evaluative activities (self-assessment and practices). In this case, we analyze and compare subgroups Alpha, Beta, Delta and Gamma (from CAT), where students were assigned by means of the previous test and the gameful learning experiences are personalized. Thus, the total of 60 students enrolled in CAT were distributed in the subgroups as showed in Figure II. It seems the result of the assignment process distributed students uniformly in each subgroup, none of them being much decompensated.

Thus, Alpha is composed of 11 students; Beta, 17; Delta, 17; and Gamma, 15. Thus, students revealed they generally agreed (34.69%) or agreed very much (32.65%) with the station they were assigned after running the initial test and considering their preferences (Alpha: “Every progress is based on competition,” Beta: “Little circles in harmony,” Delta: “Free spirits,” Gamma: “One for all and all for one”). Only four students (8.16%) did not agree with their assignment at the end of the experience, two of them in Alpha, one in Beta, and the other in Gamma. In contrast, all the students in Delta agreed with the assignment.

FIGURE II. STUDENT DISTRIBUTION



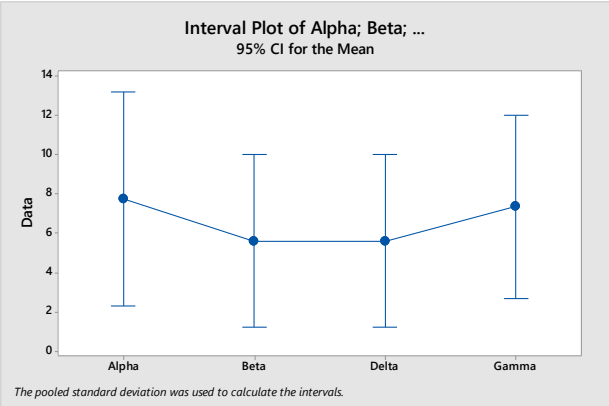
Deeping into each subgroup, we summarize the descriptive analysis of the two types of activities in Table XI. Alpha's log reveals a total of 54.55% of students participated of the simulation tasks, while 81.82% of the self-assessment exams.; Beta, 47.06% and 41.18%; Delta, 35.29% and 41.18%; and finally, 53.34% and 73.34% in Gamma respectively.

TABLE XI. STUDENT TASK DEVELOPMENT COMPARISON

Group	Descriptive analysis					
	N	*Mean*	*StDev*	*Min*	*Median*	*Max*
Alpha	*11*	*7.73*	*10.37*	*0.00*	*3.00*	*34.00*
Beta	*17*	*5.59*	*9.16*	*0.00*	*1.00*	*34.00*
Delta	*17*	*5.59*	*9.64*	*0.00*	*1.00*	*30.00*
Gamma	*15*	*7.33*	*6.88*	*0.00*	*5.00*	*20.00*

As follows, we run the one-way analysis of variance (ANOVA) with four factors. In these results, the null hypothesis states that the average hardness values of the four different groups are the same. Since the p-value (0.879) is more than the significance level of 0.05, we cannot reject the null hypothesis and cannot conclude that some of the groups have different means. The interval graph is show in the following graph (see Figure III), where Beta and Delta have the lowest average in comparison to Alpha, the highest.

FIGURE III STUDENT TASK DEVELOPMENT INTERVAL GRAPH



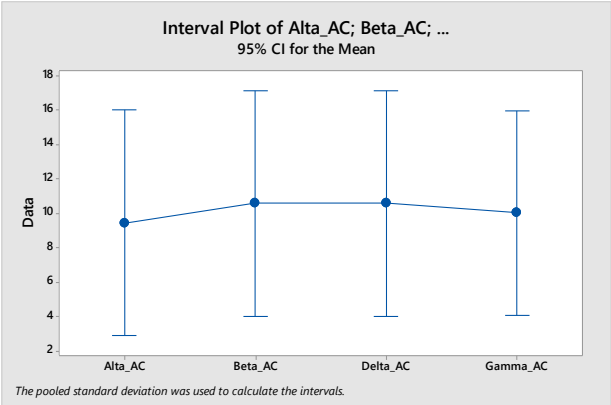
Furthermore, the distribution of active students was 9 in Alpha (81.82% of total), 9 in Beta (52.94%), 10 in Delta (58.82%), and 11 in Gamma (64.70%). Therefore, if we only consider the two types of activities with active students, we run the following descriptive analysis (see Table XII):

TABLE XII. ACTIVE STUDENT TASK DEVELOPMENT COMPARISON

Group	Descriptive analysis					
	N	*Mean*	*StDev*	*Min*	*Median*	*Max*
Alpha	*9*	*9.44*	*10.78*	*1.00*	*4.00*	*34.00*
Beta	*9*	*10.56*	*10.36*	*1.00*	*9.00*	*34.00*
Delta	*9*	*10.56*	*11.26*	*1.00*	*7.00*	*30.00*
Gamma	*11*	*10.00*	*6.08*	*3.00*	*10.00*	*20.00*

Next, we run the one-way ANOVA again with four factors. In these results, the null hypothesis states that the average hardness values of the four different groups are the same. Since the p-value (0.994) is more than the significance level of 0.05, we cannot reject the null hypothesis and cannot conclude that some of the groups have different means. The interval graph is shown in see Figure IV, where Beta and Delta are now the highest in comparison to Alpha, the lowest.

FIGURE IV ACTIVE STUDENT TASK DEVELOPMENT INTERVAL GRAPH



Regarding emotional engagement of students, we compare the emotion-related intrinsic principles described in SPARC model. Thus, the following Table (Table XIII) shows a comparison between the Alpha, Beta, Delta, and Gamma subgroups for each SPARC item. Additionally, students valued the perception of the experience as 3.67, 3.92, 3.75 and 3.93 respectively in a five-level Likert scale (from value 1 "very irrelevant" to 5 "very relevant"). Another element to be considered regarding the emotional engagement is the interest to have similar experiences in other subjects in the future: results reveal that 88.89% of students in Alpha agree with this affirmative, 61.54% in Beta, 91.67% in Delta, and 86.67% in Gamma. However, the one-way ANOVA revels that the f-values and p-values of each element in "S", "P", "A", "R", "C" were respectively 0.01 and 0.999, 0.23 and 0.877, 1.33 and 0.277, 1.98 and 0.131, and 0.87 and 0.462. Thus, we cannot reject the null hypothesis and cannot conclude that some of the groups have different means.

TABLE XIII. EMOTIONAL PERCEPTION COMPARISON (AVERGAGE)

Group	Intrinsic motivation item				
	S	*P*	*A*	*R*	*C*
Alpha (N=9)	*3.78*	*3.89*	*3.67*	*3.11*	*3.56*
Beta (N=13)	*3.85*	*4.08*	*3.62*	*3.62*	*4.15*
Delta (N=12)	*3.83*	*4.25*	*3.67*	*2.75*	*4.17*
Gamma (N=15)	*3.80*	*4.13*	*4.20*	*3.53*	*4.00*

V. DISCUSSION AND LIMITATIONS

Upon completion of this study, the descriptive statistics suggest that personalization of gameful design for student engagement in the learning process seems to work better than generic approaches, since the metrics related to behavioral and emotional engagement were higher for the personalized condition in average. However, the standard deviation for all metrics was very high in relation to the means, in general. This issue, as well as the great difference in sample sizes, prevent us from reporting any significant results, even with the resulting differences in means. Moreover, the high standard deviations also show that the difference in the means could reflect a scenario in which some students were very active students (more than the expected student/performance ratio) as opposed to many students who did act as expected. Next, we aim to answer the research questions proposed at the beginning of the study, based on the outcomes of the descriptive statistics presented in the previous section.

First, does an online personalized gameful learning experience have a greater impact on students´ engagement than a generic gameful learning experience? (RQ1). Regarding behavioral engagement, the outcomes reveal that 65.00% of all the students who joined in a personalized gameful learning experiences were active, meaning that they participated of at least one task. Additionally, the participation of the self-evaluation type of tasks was slightly more common. Nevertheless, a total of 51.00% of the students from the non-personalized gameful course were active. Thus, there is a considerable difference of 14 percentage points between the active students in the personalized condition in comparison to the non-personalized condition. While the ratio student/task is the same regardless of the size of each group and subgroup in this study, students participated of an average of 6.42 tasks per student (being 9.87 if we only consider the active ones) in the personalized condition; in contrast, the one-size-fits-all condition showed a total of 3.90 tasks per student (only considering active students, 7.45). The difference is higher than 1.5 times between conditions in case of all students.

Regarding emotional engagement, students valued the emotional perception of the experience as 3.84 and 3.47 respectively in a five-level Likert scale with values from 1 to 5. Moreover, the interest to join in a similar experience was mentioned by 78.43% of the students who experienced personalization, a much higher value than 52.63% in the one-size-fits-all condition. Deeping into the emotion-related principles of intrinsic motivation, all items were highly valued in the personalized experience. In consequence, these findings lead to the conclusion that personalized gameful learning experience might have a greater impact on students' engagement than a one-size-fits-all gameful learning experience. In this way, a future work could be the application of other mechanisms of personalization and what would be the adequate ratio of design effort and impact of personalization on the students' experiences.

Second, does the same gameful learning experience have a greater impact on students' engagement when they are assigned to different tailored versions based on their user type compared to students that are assigned to different versions randomly? (RQ2). Regarding behavioral engagement, the outcomes reveal that 73.34% of all the students who joined the gameful learning experience by means of their user type were active, meaning that they participated of at least one task. Additionally, the participation of the self-evaluation tasks was slightly more common. Nevertheless, a total of 51.00% of the students randomly assigned to the gameful course were active. Thus, there is a considerable difference of 22 percentage points between the active students in the subgroups assigned through the participants' user types in comparison to the subgroups with random assignment. Since the ratio student/task is the same regardless of the size of each group and subgroup in this study, students completed an average of 7.33 tasks per student (being 10.00 if we only consider the active ones) in the user type-assigned subgroups; in contrast, the randomly assigned subgroups showed a total of 3.90 tasks per student (only considering active students, 7.45). The difference is higher than 1.8 times between the two conditions.

Regarding emotional engagement, students valued the emotional perception of the experience as 3.93 and 3.47 respectively in a five-level Likert scale with values from 1 to 5. Furthermore, the interest to join in a similar experience was mentioned by 86.67% of the students who were assigned to groups based on their user type, a much higher value than 52.63% in the groups formed by diverse user types. In consequence, these findings lead to the conclusion that a gameful learning experience would have a greater impact on students' engagement when personalized based on their user types, compared to students assigned randomly to groups. In this way, a future work could investigate the assignment of participants to personalized groups using the Hexad user types scale with a larger sample in order to validate the findings described here.

Third, do the different personalized gameful learning experiences by user types engage students by the same factor? (RQ3). Regarding behavioral engagement, the outcomes from the study reveal a little difference between the averages of tasks completed between the four personalized gameful experiences. Thus, the Alpha and Gamma subgroups show a participation rate of 7.73 tasks per student, meanwhile subgroups Beta and Delta, 5.59. The difference is smaller when only considering the active students (Alpha: 9.44, Beta: 10.56, Delta: 10.56, and Gamma: 10.00). Moreover, the number of students assigned to each experience was balanced; there was no station (personalized gameful experience) with fewer students, which could have influenced the student engagement (e.g. feeling of loneliness).

Regarding emotional engagement, students valued the emotional perception of each experience as 3.67, 3.92, 3.75 and 3.93 respectively in a five-level Likert scale with values from 1 to 5. Regarding the interest to join in a similar experience, 88.89% of students in the Alpha subgroup agree with the affirmative, 61.54% in Beta, 91.67% in Delta, and 86.67% in Gamma. Only the Beta subgroup presents a lower rate than the other. Regarding emotional-related items, the differences were low and above neutral (3), except for Relatedness in Delta. This can be explained because this group was designed for Free Spirits; however, student interaction was not promoted by the gameful design elements employed in the course. In consequence, these findings lead to the conclusion that the

different personalized gameful learning experiences by user types engage students roughly by the same factor from the behavioral viewpoint, but this cannot be assumed from the emotional point of view. Hence, a future work could be the expansion of the survey items to gather additional, more detailed data to further investigate the students' emotional engagement with the personalized gameful learning experience.

Finally, as a characteristic of the present study, it should be highlighted that the student profile at our online-based University, which has the main purpose of providing access to higher education to non-traditional degree students in face-to-face studies, is somewhat different if compared to traditional Universities. Note that the average student at UOC is approximately more than 30 years old (64% are 30+, and 27% are 40+), most work full-time (95%) and have children (58%) (data collected from an internal source). Therefore, the use of the approach described in this paper poses an interesting research challenge in some educational contexts, such as with adult learners.

VI. CONCLUSION

In this paper, we presented the design and analysis of a personalized gameful learning experience within a Computer Network Design course. The general purpose of this study was to determine how a personalized gameful learning experience affects the students' engagement in comparison with a non-personalized one. A total of 81 students voluntarily joined the experience in two groups: 60 of them in the experimental group (CAT) and 21 in the control group (CAS). The experience was assessed at the end of the course from behavioral and emotional engagement viewpoints. Both quantitative and qualitative standpoints were considered through the logs of the tool and a survey. A descriptive analytic process was conducted as well as the development of different non-parametric tests.

Results reveal that personalization seems to work better regarding the behavioral and emotional engagement of the students with the course by considering the descriptive analysis conducted. However, the characteristics of the sample did not lead to any statistically significant result, which would have allowed us to provide a more definitive answer to the proposed research questions. As future work, we intend to develop a new A/B test with larger and balanced samples of students, and at the same time increasing the level of personalization of the experience, to verify if a greater degree of personalization could lead to a statistically significant difference between personalized and one-size-fits-all experiences.

ACKNOWLEDGMENT

This work was partly funded by Agència de Gestió d'Ajuts Universitaris i de Recerca (Generalitat de Catalunya) through the Industrial Doctorate programme 2014-DI-006 and the Spanish Government by means of the project TIN2013-45303-P "ICT-FLAG". It also has been carried out in collaboration with Grupo ICA Barcelona. Author Tondello thanks the CNPq, Brazil for his funding. Author Nacke thanks NSERC (RGPIN-418622-2012), SSHRC (895-2011-1014, IMMERSe), CFI (35819), and Mitacs (IT07255) for funding his research team.

REFERENCES

[1] S. Deterding, D. Dixon, R. Khaled, and L. Nacke. From game design elements to gamefulness: Defning "gamification". *Proceedings of the 15th International Academic MindTrek Conference: Envisioning Future Media Environments, MindTrek '11*, ACM, 2011, pp. 9-15.

[2] F.F.H. Nah, Q. Zeng, V.R. Telaprolu, A.P. Ayyappa, and B. Eschenbrenner. Gamification of education: a review of literature. *In International Conference on HCI in Business.* Springer, Cham, 2014, pp. 401-409.

[3] J. Vargas-Enríquez, L. García-Mundo, M. Genero, and M. Piattini. Análisis de uso de la gamificación en la enseñanza de la informática. *Actas de las XXI Jornadas de la Enseñanza Universitaria de la Informática*, Andorra La Vella, Andorra. JENUI, 2015, pp. 105–112.

[4] A. Mora, D. Riera, C. Gonzalez, J. Arnedo-Moreno, Gamication: a systematic review of design frameworks, *Journal of Computing in Higher Education*. Springer International, 2017, 29(3), pp. 516-548.

[5] M. Busch, E. Mattheiss, R. Orji, A. Marczewski, W. Hochleitner, M. Lankes, L. E. Nacke, M. Tscheligi. Personalization in serious and persuasive games and gamied interactions. *Proceedings of the 2015 Annual. Symposium on Computer-Human Interaction in Play, CHI PLAY '15, ACM*, 2015, pp. 811-816.

[6] D. M. Popa, *Gamification of productivity software: A user-centered design approach*, Master thesis, University of Copenhagen, Denmark 2013.

[7] R. Orji, J. Vassileva, R. L. Mandryk, Modeling the efficacy of persuasive strategies for different gamer types in serious games for health, *User Modeling and User-Adapted Interaction* 24 (5), 2014, pp 453-498.

[8] L. S. Ferro, S. P. Walz, S. Greuter, Towards personalised, gamified systems: An investigation into game design, personality and player typologies. *Proceedings of The 9th Australasian Conference on Interactive Entertainment: Matters of Life and Death, IE '13*, ACM, 2013, pp. 7:1-7:6.

[9] G. F. Tondello, R. R. Wehbe, L. Diamond, M. Busch, A. Marczewski, and L. E. Nacke. The Gamification User Types Hexad Scale. In *Proceedings of the 2016 Annual Symposium on Computer-Human Interaction in Play - CHI PLAY '16*. ACM, 2016, pp. 229–243

[10] A. Mora, G.F. Tondello, L. Calvet, C. González, J. Arnedo-Moreno, and L.E. Nacke. Towards personalized gamification: exploring user types' preferences on game design elements. *International Journal of Human-Computer Studies*. Elsevier. (under review).

[11] A.Mora, E. Planas, and J. Arnedo-Moreno. Designing game-like activities to engage adult learners in higher education. *In Proceedings of the Fourth International Conference on Technological Ecosystems for Enhancing Multiculturality*. 2016. (pp. 755-762). ACM.

[12] ABC Studios. Lost. 2004-2010.

[13] Frictional Games. SOMA. 2017

[14] Interplay Productions. Fallout. 1997

[15] Looking Glass Studios. System Shock II. 1999

[16] Inc Minitab. *MINITAB release 17: statistical software for windows*. Minitab Inc, USA. 2017.

[17] J.A. Fredricks, P.C. Blumenfeld, and A.H. Paris. School engagement: Potential of the concept, state of the evidence. *Review of Educational Research*, *74*(1), 2004, pp 59-109.

[18] E.A. Skinner, & M. J. Belmont. Motivation in the classroom: Reciprocal effects of teacher behavior and student engagement across the school year. *Journal of Educational Psychology*, *85*(4), 1993, p 571.

[19] J.A. Gliem, & R.R. Gliem. Calculating, interpreting, and reporting Cronbach's alpha reliability coefficient for Likert-type scales. *Midwest Research-to-Practice Conference in Adult, Continuing, and Community Education*. 2003.

[20] A. Marczewski. User Types. In *Even Ninja Monkeys Like to Play: Gamification, Game Thinking & Motivational Design*. CreateSpace Independent Publishing Platform, 2015, pp 69–84.