User engagement in an open collaboration community after the insertion of a game design element: An online field experiment

Completed Research

Ana Paula O. Bertholdo
Department of Computer Science – University of São Paulo
ana@ime.usp.br

Claudia de O. Melo
Department of Computer Science - University of Brasília
claudiam@unb.br

Artur S. Rozestraten
Faculty of Architecture and Urbanism - University of São Paulo
artur.rozestraten@usp.br

Marco Aurélio Gerosa
University of São Paulo, Northern Arizona University
Marco.Gerosa@nau.edu

Heather L. O'Brien
School of Library, Archival and Information Studies - University of British Columbia
h.obrien@ubc.ca

Abstract

Gamification has been proposed as a possible solution to low user engagement in open collaboration communities. However, most studies do not present statistical analyses and few studies analyze the criterion validity between behavioral and self-reported engagement measures. This study seeks to understand whether gamification contributed to greater behavioral and self-reported engagement in an open collaboration community. We conducted an online field experiment to analyze user engagement in two versions of a new feature, with or without a game design element (Progress bar), with 36 and 37 users, respectively. A subset of the participants (18 users) answered an online questionnaire about their engagement with the system. We found that the group of users with the highest self-reported engagement scores performed the most actions, and users who accessed the Progress bar performed the highest number of actions. More studies are needed to better understand the relationship between each action and the engagement.

Keywords

Engagement, UES, gamification, GLAMs, open collaboration communities, progress bar.

Introduction

A major challenge in open collaboration systems is fostering a sense of community around artifacts (O'Brien and Toms, 2008). The community also needs to attract a sufficient number of participants, i.e., a critical mass (Burke et al., 2009). Galleries, Libraries, Archives, and Museums (or GLAMs) around the world are leveraging the potential of outsourcing specific activities to the community, i.e., crowdsourcing (Oomen and Aroyo, 2011). In this context, participation can be fostered by inviting users to assist in the selection, cataloging, contextualization, and curation of collections (Lankes et al., 2007). These active ways of interacting with collections can lead to a higher level of engagement (Huvila, 2008), which is key to community sustainability (Oomen and Aroyo, 2011). In creating online communities, designers face the critical mass problem: the system does not yet have content capable of attracting new members and, at the same time, there are few participants to create content that attracts other users. In small communities,

the impact of members actions is limited, reducing motivation to contribute (Kraut and Resnick, 2011). Consequently, communities focused on open collaboration (Forte and Lampe, 2013) need to engage participants. Gamification, which describes the use of game design elements in non-game contexts (Deterding et al., 2011), has been employed to promote engagement in a variety of contexts (Hamari, 2017; Seaborn and Fels, 2015). Adopting gamification aims to encourage participation and engage people (Deterding et al., 2011). The literature has shown positive effects for gamification, but few studies compare gamified and non-gamified system versions. There is also a need to explore how gamification can be implemented in specific domains and real settings (Seaborn and Fels, 2015). Studies about gamification lack statistical treatment of empirical data, meaning that standard measures of effect size are unavailable (Seaborn and Fels, 2015). This study seeks to fill this gap, aiming to understand whether a specific game design element (the progress bar) contributed to fostering user engagement in an open collaboration online community about architecture and urbanism in the context of a GLAM. We conducted an online field experiment comparing two versions of the Arquigrafia system (http://arquigrafia.org.br): with and without gamification. We measured behavioral metrics for users in both versions. A subset of the participants reported their engagement with the system answering an online questionnaire based on User Engagement Scale (UES) (O'Brien et al., 2018). We finally analyzed whether relationships exist among behavioral and self-reported user engagement metrics.

Background

Measuring Online User Engagement: Concepts

O'Brien and Toms (2018), define user engagement as "a quality of user experience characterized by the depth of an actor 's investment when interacting with a digital system". Engagement is a process comprised of four stages (O'Brien and Toms, 2008): point of engagement, period of sustained engagement, disengagement, and reengagement. The process is characterized by attributes of engagement that pertain to user, system, and user-system interaction. Overall, engagement is a quality of user experience characterized by attributes of challenge, positive affect, endurability, aesthetic and sensory appeal, attention, feedback, variety/novelty, interactivity, and perceived user control. Subjective measurements are useful for capturing users' motivations, beliefs, and attitudes but may be susceptible to self-reporting biases (O'Brien and Lebow, 2013). Objective measures may be behavioral or physiological. Behavioral metrics include number of distinct and returning users, number of visits and page views, and time spent interacting with a website per session and over several days. These metrics have been used as indicators of user engagement, activity, loyalty, and popularity of websites. Physiological metrics can help to understand an individual's experience because they provide "high-resolution, continuous representation" of subjects over time (Lin et al., 2008). Behavioral and physiological metrics explain "what" user's behaviors are, but not "why" users behave in certain ways. Subjective measures, such as self-report questionnaires, address the "why" as they explore users' motivations, preferences, and attitudes. As a counterpoint, users may under- or over-report their experiences (O'Brien and Lebow, 2013).

Besides that, there are two levels of engagement: micro level involvement and macro level involvement (Calleja, 2011). For Calleja, 2011, p. 40, micro level involvement relates to "the moment-by-moment engagement of gameplay", whereas the macro level relates to "longer-term motivations as well as off-line thinking and activities that keep players returning to the game", covering both "postgame and pregame experiences". Iacovides et al. (2015) present that macro level expectations are informed by prior experience, other players and the wider community; and Repeated micro involvement depends on expectations being met, in-game factors, such as rewards, and external factors. In the context of online communities, the participation of users differs when considering access from newcomers users, those who visit the website for the first time, and existing users (Kraut and Resnick, 2011). Studies have distinguished the engagement from existing users and newcomers (Backstrom et al., 2008; Burke et al., 2009). According to Feild and Allan (2009), frustration is a function not only of the current interaction but of the previous state of frustration. This behavior may imply that users who have experienced frustration in a system in previous accesses are prone to having a frustrating experience in new accesses.

Theoretical foundations of Gamification

Common purposes for gamification relate to motivation, behavior change, and engagement. The psychological theory of intrinsic and extrinsic motivation developed by Ryan and Deci (2000) is a consistent choice among authors. Gamification is consistently positioned as a tool for fostering extrinsic and intrinsic motivation to accomplish specific tasks through the selective use of game elements (Seaborn and Fels, 2015). Liu et al. (2017) offer a list of disciplinary perspectives that can be applied to the future gamification studies, including information systems, marketing, organizational behavior, psychology, and social psychology. From the psychology perspective, authors describe Flow Theory and Self-Determination Theory (SDT) for gamification research. The Flow Theory (Csikszentmihalyi and Nakamura, 1979) describes a mental state of focus and immersion in one activity. According to the Flow Theory, the feedback people experience by doing one activity should be ideal. The activity must be neither too difficult nor too easy, so the user is engaged, focused and absorbed, in a state of flow. Therefore, it is important to confront the user with interesting challenges (Groh, 2012). The scaffolding for those challenges should increase the difficulty for reaching the next level. Also it is desirable to vary the difficulty inside the flow region, where people are neither underchallenged nor overchallenged. The SDT proposes that if one of the three psychological needs of intrinsic motivation - autonomy, competence or relationship - is not supported within a social context, the well-being of those involved will be negatively impacted (Ryan and Deci, 2000). Based on the tenets of SDT, the mechanism by which either of these systems motivates effort and enjoyment may depend on whether relatedness and connectedness are relevant factors within the context. We examine our research questions in light of SDT, Flow Theory, and aspects of micro and macro player involvement.

Related work

We highlight in this section studies performed with the game design element Progress Bar, on which the analysis will be developed in the following sections. According to Kim (2012), the progress dynamic is often responsible for making people go through all steps of the signup process for an online service because people like to move a progress bar to 100 percent. The progress bar makes people feel goal oriented and it is a positive feedback. Progress bars are used to track and display the overall goal progression. In an educational game, progress bars are used as a display mechanism to motivate people who are close to achieving their educational goal or sub-goals. Besides that, progress bar was defined as a enhancing service that increases the perceived value of filling in all details by making use of progress-related psychological tendencies (Huotari and Hamari, 2017).

Pedro et al. (2015) conducted a case study with two groups of students to investigate their behavior during their interaction with a system with and without gamification. The gamified version presented points, progress bar, badges and ranking of each student logged in. The results indicate that the gamification implemented contributed to improve student performance in the case of boys. Yet, improvement was not observed in the case of girls. Codish and Ravid (2014) performed two quasi-experiments in an academic course with students. Game elements used were the immediate feedback game mechanics such as points, rewards, and badges, and comparative feedback mechanics or presentation mechanics such as leaderboards and progress bars. Progress bars were related to the goal setting theory (Locke and Latham, 2002) since they are a form of “summary feedback that reveals progress in relation to their goals” (Locke and Latham, 2002, p. 708). Kim (2012) presented recommendations for how to, and how not to gamify the library experience, and one of the recommendations was: ”Show the progress bar in library catalog”. We analyzed progress bar in the domain of a GLAM. We aim to analyze behavioral and self-reported engagement metrics to understand how self-reported user experience relates to user behavior in the context of a small open collaboration community about architecture and urbanism.

Method

The online field experiment was designed to collect behavioral engagement metrics. In addition, an online questionnaire was designed to collect self-reported engagement metrics. The overall goal of our online field experiment is to investigate whether gamification increases engagement in an open collaboration online community. We thus aim to answer the following research questions: **RQI)** Does gamification contribute to increased behavioral engagement with the content of an open collaboration online

community?; **RQII)** Does gamification contribute to increased self-reported engagement with the content of an open collaboration online community?; and **RQIII)** How does self-reported user experience relate to user behavior? We planned this experiment in the context of the Arquigrafia online community. Arquigrafia is a public, nonprofit digital collaborative community dedicated to disseminating architectural images (www.arquigrafia.org.br). Since Arquigrafia is still small, it needs to foster a community around images and information. The system makes available images from institutional collections as well as user images. However, many images from institutions are old and lack some of the data about the architecture presented in them. A Data Review feature was developed to mitigate this problem; it fosters collaboration among system members by reviewing and editing data on the architectures portrayed in images posted by users and institutions. This is the target feature of the online field experiment. Originally, the experiment was designed to evaluate three game elements: the Progress Bar, Points and Enduring play, according to definition from (Deterding et al., 2011). However, only one person accessed the Points element (after entering the data review interface) and none accessed the interface which Enduring play had been applied (after the completion of a data review). Points and Enduring Play were pieces of Data review feature, and there were no users who reviewed data from an image during the experiment period. However, the Progress Bar is presented in the page that displays information about a selected architecture image by a user. Therefore, we focused our analysis on the Progress Bar. The Progress Bar in the context of the Arquigrafia system presents how far the architecture image is from reaching the goal of 100% of completed and reviewed data. Two versions of the system were analyzed:

Gamified version (Treatment) Data Review Feature with Progress bar.

Non-gamified version (Control) Data Review Feature without Progress bar.

Data Collection

The online field experiment was performed over 20 days (from November 16, 2017 to December 5, 2017). The metrics were calculated based on a mean for the number of days of the experiment. According to Zichermann and Cunningham (2011), engagement can be best analyzed through a series of interrelated metrics that are combined to form a whole. We adopted two types of metrics: self-reported and behavioral as can be noted in Table 1.

Type	Metric	Description	Calculation
UES	Focused attention (FA)	Refers to feeling absorbed in the interaction and losing track of time	Sum(score of items in the questionnaire related to FA)/number of items
UES	Perceived usability (PUS)	Refers to negative affect experienced as a result of the interaction and the degree of control and effort expended	Sum(Code reverse of score of items in the questionnaire related to PUS)/number of items
UES	Aesthetic appeal (AE)	Refers to the attractiveness and visual appeal of the interface	Sum(score of items in the questionnaire related to AE)/number of items
UES	Reward (RW)	Refers to endurability (or the overall success of the interaction), novelty, and felt involvement	Sum(score of items in the questionnaire related to RW)/number of items
UES	UES total score	Overall self-reported engagement score	Sum(score of items from FA, PUS, AE, and RW)/number of items
Behavioral	Frequency	Number of user accesses to the system	Sum(user sessions per day)/20
Behavioral	Recency	Time between visits	Last day access - Penultimate day access in 20 days of each user
Behavioral	Duration	How long users spend time in each connection	Sum(user sessions in seconds per day)/20
Behavioral	Virality	How many other users are influenced by a certain user to engage with the object	Sum(User Accesses from Facebook+Google+Notifications per day)/20
Behavioral	Ratings	User evaluation in terms of quality, quantity, or some combination of both	Sum(User Comments+Likes per day)/20
Behavioral	Actions	User actions during the considered period	Sum ((Photo evaluations; Uploads; Downloads; selections; editions;searches), User selections, chats, Leaderboard views, and Accesses to the *Data Review* feature per day)/20

Table 1. User engagement metrics used in this study.

The self-reported engagement metrics used in this experiment were drawn from the User Engagement Scale (UES) (O'Brien et al., 2018). UES can be analyzed concerning its sub-scales or dimensions — focused attention (FA), perceived usability (PUS), aesthetic appeal (AE), and reward (RW) — or aggregated as an overall engagement score. For self-reported metrics, an online questionnaire was sent to users who used the system in the experiment period. This questionnaire collected user profile data and the UES. The behavioral engagement metrics — *frequency, recency, duration, virality, ratings* (Zichermann and Cunningham, 2011), and *actions* (Hamari, 2017; Iacovides et al., 2015) — can be aggregated as a pragmatic score. For behavioral metrics, event logs were inserted in the system to collect real usage data. We developed an algorithm (in the Java programming language) to convert event logs into behavioral engagement metrics. According to the metric "Average session duration", collected from Google Analytics, 90% of the Arquigrafia user sessions are up to 600 seconds (or 10 minutes). For this reason, we define a single user session as 600 seconds, and it was used to calculate *frequency* and *duration* metrics.

Data Preparation and Analysis

In our experiment, **Group 1** comprises users who logged into the system during the experiment period (n = 73). **Group 2** is a sub-group of Group 1 users that answered the online questionnaire with the UES (n = 18). Group 1 (73 users) has 36 users that accessed the gamified version and 37 users that accessed the non-gamified version. Behavioral metrics were analyzed for Group 1 and behavioral and self-reported metrics were analyzed for Group 2. Only 18 out of 73 participants answered to the self-reported user engagement questionnaire due to low response rate. However, several profiles of typical users of different age groups are represented in this sample: professional architects, architecture and urbanism students, architecture and urbanism professors, librarians, library science students, and professional photographers; with ages between 20-68 years old; and 59.3% male.

In this study, **existing users** (n = 16 for gamified version; and n = 17 for non-gamified version) are users who visited the system before the experiment period (from November 16, 2017 to December 5, 2017) and after the insertion of actions logs in the system (June 14, 2015). **Newcomers** (n=20 in each version) are users who visited the system for the first time during the experiment period or users that accessed for the first time since actions logs were inserted in the system, i.e., users who have not accessed the system for 2 years and 5 months, which is a period that the system has been completely reformulated. We also divided users according to the overall score for the UES: high-UES-score users (n = 3 for gamified version; and n = 2 for non-gamified version) had scores above 4.02; medium-UES-score users (n = 4 for gamified version; and n = 3 for non-gamified version) had scores between 3.27 and 4.02; and low-UES-score users had scores below 3.27 (n = 4 for gamified version; and n = 2 for non-gamified version). We calculated Cronbach's alpha (∂) for the self-reported metrics and examined the internal consistency of subscales based on guidelines from DeVellis (2003): 0.7–0.9 is optimal. The mean for FA subscale was 2.7 (sd 1.2) and ∂ 0.95; the mean for PUS was 3.7 (sd 0.89) and ∂ = 0.86; the mean for AE was 3.6 (sd 0.61) and ∂ = 0.84; the mean for RW was 4 (sd 0.83) and ∂ = 0.91. For overall engagement, the mean was 3.5 (sd 0.74) and ∂ = 0.92. Therefore, the UES was highly reliable. Originally, 12 items were considered in the UES. After the analysis of Cronbach's alpha (∂), one item was dismissed from AE (AE3) to improve the value of Cronbach's alpha for the AE (from 0.65 to 0.84) subscale. According to the Shapiro-Wilk normality test, the data for the self-reported engagement were fairly normally distributed.

Results

Does gamification increase engagement? RQI and RQII results

Although both groups accessed the *Data Review feature*, no user performed a data review over the 20-day experiment period. The total number of actions in the system during the Experiment Period was 553 actions from 37 users who accessed the non-gamified version and 783 actions from 36 users in the gamified version. The group who accessed the gamified version was the one that performed the higher number of actions (230 more actions) when compared to the non-gamified group. Of the 36 users who accessed the gamified version, 16 were in the group of existing users and 20 in the group of newcomers Users. Of 37 users who accessed the non-gamified version, 17 were in the group of existing users and 20 in the group of newcomer users. Existing users (Table 2) performed 308 and 389 actions in the gamified and

non-gamified versions, respectively. Newcomers performed 475 actions in the gamified version and 164 actions in the non-gamified version. The analysis (Table 2) was performed with Wilcoxon rank sum test, and it presents scenarios between gamified and non-gamified versions from newcomer and existing users.

Behavioral Metric	User	W	Gamified Version (Mean/SD)	Non-Gamified Version (Mean/SD)	p-value	r
Duration	Existing Users	96.5	12.42 (sd 17.35)	25.36 (sd 27.42)	0.1597	-0.27
	Newcomers	228.5	18.93 (sd 23.77)	10.0 (sd 9.91)	0.4486	0.23
Frequency	Existing Users	97	0.06 (sd 0.04)	0.11 (sd 0.11)	0.09765	-0.28
	Newcomers	201	0.08 (sd 0.05)	0.07 (sd 0.04)	0.9867	0.10
Recency	Existing Users	127.5	0.31 (sd 0.87)	0.76 (sd 1.85)	0.6441	-0.15
	Newcomers	209.5	0.75 (sd 2.51)	0.75 (sd 3.12)	0.6718	0
Virality	Existing Users	144.5	0.01 (sd 0.03)	0.008 (sd 0.03)	0.5633	0.03
	Newcomers	224.5	0.03 (sd 0.06)	0.007 (sd 0.01)	0.3519	0.25
Ratings	Existing Users	136	0 (sd 0)	0 (sd 0)	NA	NA
	Newcomers	200	0 (sd 0)	0 (sd 0)	NA	NA
Actions	Existing Users	129.5	0.96 (n 308, sd 2.46)	1.14 (n 389, sd 1.76)	0.8283	-0.04
	Newcomers	209	1.18 (n 475, sd 2.60)	0.41 (n 164, sd 0.55)	0.817	0.20

Table 2. Behavioral metrics from Group 1 (73 users).

Scenario	t	df	Mean of Gamified Users	Mean of Non-Gamified Users	p-value	Cohen's d
Existing Users	-0.973	5.9744	3.24 (sd 0.97)	3.75 (sd 0.64)	0.3683	-0.62
Newcomers	0.2275	5.9009	3.70 (sd 0.53)	3.60 (sd 0.61)	0.8277	0.17

Table 3. Overall UES scores from Group 2 (18 users).

Cohen (1988) reported the following guideline to interpret the effect size r: 0.1 for a small effect; 0.3 for an intermediate effect; 0.5 and higher for a strong effect. There were small effect sizes for the comparison between gamified and non-gamified versions of existing users: *duration* (-0.27), *frequency* (-0.28), and *recency* (-0.15). There were small effect sizes for the comparison between gamified and non-gamified versions of newcomers: *duration* (0.23), *frequency* (0.10), *virality* (0.25), and *actions* (0.20). The comparison of overall UES scores (Table 3) presented a mean for gamified/existing users (3.24) lower than the mean of non-gamified/existing users (3.75). For newcomers, the comparison for UES total scores presented a mean for gamified/newcomer users (3.70) higher than the mean of non-gamified/newcomer users (3.60). This behavior is in accordance with Table 2, where higher behavioral metrics were collected for non-gamified/existing users and gamified/newcomer users. Cohen (1988) reported the following guideline to interpret the effect size d (Cohen's d): 0.25: small effect; 0.5: medium effect; 0.8 and higher: large effect (Keppel and Wickens, 2004). A medium effect size was identified for the comparison between gamified and non-gamified versions for UES total scores from existing users (-0.62).

How does self-reported user experience relate to user behavior? RQIII Results

Table 4 presents results from the Kruskal Wallis Test for behavioral metrics classified according to high, medium, and low UES total scores. The group high-UES-score users was related to higher *frequency*, lower *recency*, higher *virality*, and higher *actions*. The *duration* was lower for the group high-UES-score users. However, the standard deviation is high for the mean of *duration*. We used guidelines to interpret Epsilon-squared (ε^2) effect size (Keppel and Wickens, 2004) presented in Table 4: 0.01 for a small effect; 0.06 for a medium effect; 0.15 and higher for a large effect. *Frequency* (0.06) and *recency* (0.14) show

medium effect sizes. *duration* (0.03) and *actions* (0.02) present small effect sizes. Only *virality* present large effect size (0.18). However, the Kruskal Test has not presented statistically significant results, which

	High, mean (SD)	Medium, mean (SD)	Low, mean (SD)	Kruskal Wallis chi-squared	df	p-value	ε^2
Frequency	0.11 (sd 0.06)	0.07 (sd 0.05)	0.08 (sd 0.06)	1.17	2	0.556	0.06
Recency	0 (sd 0)	0.42 (sd 1.13)	2.83 (sd 4.66)	2.39	2	0.302	0.14
Duration	17.7 (sd 21.76)	27.18 (sd 37.93)	17.18 (sd 29.15)	0.51	2	0.7748	0.03
Virality	0.06 (sd 0.08)	0	0.02 (sd 0.06)	3.17	2	0.2046	0.18
Ratings	0	0	0	NA	2	NA	NA
Actions	2.45 (n 245, sd 3.78)	1.33 (n 187, sd 2.44)	0.41 (n 50, sd 0.59)	0.50	2	0.7754	0.02

Table 4: Behavioral metrics according to high, medium, and low overall UES scores.

indicates the need to analyze a higher sample size because there are indications of medium and large effects sizes between behavioral metrics and high, medium, and low scores from overall UES. For the metric of *actions* from high-UES-score users in Group 2, 245 actions were performed, among which 182 actions are from users that accessed the gamified version. For medium-UES-score users, 187 actions were performed, among which 52 actions are from gamified version. This behavior can explain the highest duration from medium-UES-score users. For low-UES-score users, 50 actions were performed and 17 which belonged to the gamified version. For medium and low-UES-score users, users who performed more actions (135 for medium-UES-score users, and 33 for low-UES-score users) accessed the non-gamified version. Table 5 presents the correlation between UES subscales and behavioral metrics through Spearman's rank correlation rho (ρ).

Metric	Score	FA (ρ (p-value))	PUS (ρ (p-value))	AE (ρ (p-value))	RW (ρ (p-value))	UES (ρ (p-value))
Frequency	High	0.66 (0.2189)	-0.5 (0.391)	0 (1)	0.54 (0.3431)	0.86 (0.06134)
	Medium	-0.16 (0.721)	-0.28 (0.5301)	**-0.98 (4.41e-05)**	0.35 (0.4414)	-0.40 (0.368)
	Low	0.49 (0.3205)	0.37 (p-value 0.4679)	0.10 (0.8439)	0.68 (0.1324)	0.46 (0.3551)
Recency	High	NA (NA)	NA (NA)	NA (NA)	NA (NA)	NA (NA)
	Medium	**-0.76 (0.04566)**	0 (1)	-0.64 (0.1174)	0 (1)	-0.61 (0.1392)
	Low	0.60 (0.2056)	0.43 (0.3832)	0.31 (0.5454)	0.54 (0.2594)	0.56 (0.2417)
Duration	High	-0.1 (0.95)	-0.66 (0.2189)	0.70 (0.1817)	0 (1)	0.22 (0.7177)
	Medium	-0.13 (0.7752)	-0.07 (0.8694)	**-0.79 (0.03432)**	0.11 (0.8106)	-0.34 (0.4523)
	Low	-0.12 (0.8158)	0.31 (0.5639)	-0.08 (0.8679)	0.55 (0.2574)	0.23 (0.6584)
Virality	High	0.57 (0.3081)	0.44 (0.4535)	-0.61 (0.2722)	0.40 (0.495)	0.64 (0.2394)
	Medium	NA(NA)	NA(NA)	NA(NA)	NA(NA)	NA(NA)
	Low	0.14 (0.7893)	0.13 (0.8047)	-0.26 (0.6053)	0.66 (0.1502)	0.13 (0.8019)
Ratings	High	NA(NA)	NA(NA)	NA(NA)	NA(NA)	NA(NA)
	Medium	NA(NA)	NA(NA)	NA(NA)	NA(NA)	NA(NA)
	Low	NA(NA)	NA(NA)	NA(NA)	NA(NA)	NA(NA)
Actions	High	0.3 (0.6833)	-0.66 (0.2189)	0.35 (0.5594)	0 (1)	0.44 (0.4502)
	Medium	0.13 (0.7752)	-0.30 (0.5007)	**- 0.79 (0.03432)**	0.43 (0.3351)	-0.10 (0.8175)
	Low	-0.23 (0.6542)	0.23 (0.6584)	-0.17 (0.7342)	0.51 (0.2961)	0.13 (0.8026)

Table 5: Correlations between UES and behavioral metrics (high, medium, and low scores).

There are statistically significant correlations for medium-UES-Scores Users: (1) the larger AE, the smaller *frequency* (-0.98, p-value 4.41e-05); (2) the larger FA, the smaller *recency* (-0.76, p-value 0.04566); (3) the larger AE, the larger *duration* (0.79, p-value 0.03432); and (4) the larger AE, the smaller *actions* (- 0.79, p-value 0.03432). For UES total and behavioral metrics, there are no statistically

significant correlations. However, it is interesting to note that, for high-UES-Scores Users, the larger UES, the larger *frequency, duration, virality, and actions*. For medium-UES-Scores Users, the larger UES, the smaller *frequency, recency, duration,* and *actions*. High-UES-Scores Users and low-UES-Scores users present the same behavior, except for the metric *recency*, for which a positive correlation with UES was measured for low-UES-Scores Users. This may occur because recency is calculated only for the experiment period. Therefore, users who had one access during the period had *recency* equal to 0 (zero). For high-UES-Scores Users, *recency* is not applicable (NA) because all users in this group accessed the system only once in the experiment period. For low-UES-Scores Users, users with higher *recency* accessed the system more than once during the experiment period, and had higher UES scores.

Discussion and Implications

Only users who were already intrinsically motivated (Self Determination Theory) to view complete and correct information about the architecture could be influenced by the Progress bar, and users engage in activities that they have sufficient knowledge to perform (Flow theory). In this case, the Data Review was impaired by relying on previous users' knowledge to be willing to contribute. For aspects of micro and macro player involvement, macro level can be related to self-reported engagement metrics, and micro level can be related to behavioral engagement metrics. In the context of this study, Progress Bar relates to the progress feedback on a goal: to complete the data review of an architecture image. Intrinsic motivation is the self-desire to do something because it is inherently interesting, which it is distinct from extrinsic motivation, which describes taking an action because it leads to a separable outcome (Ryan and Deci, 2000). Our progress bar seeks to engage users in reviewing image data that are incomplete or incorrect.

Intrinsic motivation relates to data reviews because they can be seen as inherently interesting, and perceived as valuable for the community. In accordance with Siemens et al. (2015), Progress bars may be perceived as goal-oriented motivators, in that they may motivate users to exert effort toward the goal. If intrinsic motivation is increased, this would, in turn, lead to more positive attitudes associated with the system. However, based on the Self-Determination Theory, the mechanism by which either of these systems motivates effort and enjoyment may depend on whether relatedness and connectedness are relevant factors within the context. For example, users motivated by the Progress Bar may seek to fill in the image data because incomplete and incorrect information personally bothers them. Neither version has led users to perform data reviews in the experiment period. This can be explained because the Data Review feature depends strongly on the users' knowledge about the architecture represented in the image. This assumption makes it more difficult to act on data review, since accessing it depends on knowledge of architecture information and verification that the information presented in the system is either incorrect or incomplete.

According to flow theory (Csikszentmihalyi and Nakamura, 1979), the feedback people experience from completing an activity should be ideal. The activity must be neither too difficult nor too easy, so that the user is engaged, focused and absorbed - in a state of flow. If users who accessed the system over the experiment period had no knowledge about architecture for images they accessed, they only explored the system, looking for other actions that they could collaborate on or other images that they knew information about. However, users in the gamified version performed more *actions* in the system. This behavior could indicate that the Progress Bar played the role of motivator, leading users to search other possible goals. This explanation can be reinforced when comparing existing users and newcomers.

The gamified version contributes to increased objective engagement for newcomers, although with small effect sizes. Newcomers do not know the system and the behavior of searching to achieve other possible goals makes sense for this group. For existing users, the non-gamified version presented increased objective engagement. Nevertheless, the gamified version has not reduced user engagement. This can be observed from the fact that among the high-UES-score users are existing users that accessed the gamified version. The medium effect size identified for the comparison between gamified and non-gamified versions for UES scores from existing users (-0.62) indicates that Progress bar have not contributed to increased engagement in the context of this study. However, users with high UES scores performed more actions in the system (Group 2), and users who accessed the gamified version presented higher number of *actions (Group 1)*. Therefore, for RQI and RQII, we have not found statistically significant evidences that the Progress Bar has contributed to increased user engagement. However, the effect sizes founded for the comparison of behavioral metrics between gamified and non-gamified versions can indicate the need to

analyze a higher sample size in future studies. For RQIII, since users from the two versions of the system did not use the Data Review feature, they evaluated the system based on the use of other features and prior experiences, as macro-level expectations. Aspects of both micro and macro user involvement have been investigated (Calleja, 2011; Iacovides et al., 2015), but there has been little consideration of how they relate to each other. The highest UES scores from Group 2 were related to users who performed more actions and the user with more actions accessed the Progress bar. In the Group 1, the highest number of actions was performed by users who accessed the Progress bar (230 more actions). Therefore, we realized that for the users who accessed the Progress Bar there was an increased interest in exploring the system and performing more *actions*, or greater micro-level involvement. This may suggest that this version meets user expectations.

Limitations

This analysis is limited by the size of the sample: 73 users for analysis of behavioral metrics and 18 users for comparison of behavioral and self-report metrics. Therefore, other factors may have impacted the self-reported engagement, which may explain why the UES total score does not differ significantly between the two versions. For example, only 8 of 18 respondents are newcomer users, with 4 users for each version of the system. Therefore, most participants had already formed a perception of the system, and this may have contributed to their experience. This perception, whether positive or negative, can be difficult to alter only by the entry of a new feature, with or without Gamification. Additionally, the study was performed in a naturalistic configuration which is dependent of user accesses, and in a system with engagement problems to evaluate the actual use and not a simulated environment in a laboratory. Besides that, the study was short-term and one-off in nature (Seaborn and Fels, 2015). One-off studies need to be replicated, comparative and longitudinal designs employed, to draw stronger, generalizable conclusions.

Conclusion and Future work

The metric *actions* had an important role in understanding the exploratory behavior of users over the experiment period. However, according to O'Brien et al. (2018), multiple measures of experience are important because it is difficult to infer users' motivations and perceptions and the quality of their experience solely from their actions. In the context of this study,we found that users who accessed the Progress bar performed the highest number of actions (Group1), and the group of users with the highest self-reported engagement scores performed the most actions (Group 2). However, there was a negative correlation between the overall UES scores and actions for Medium UES-Score Users. Actions performed by users are directly related to user frustration or engagement. Future work could investigate each action performed in these three groups to better understand why more actions in some groups lead to lower UES scores, but in other groups lead to higher UES scores. Besides that, different actions (e.g., photo evaluations vs. leaderboard views) might indicate varying amount of effort and thought put into the task. This might result in different categories of action based on the amount of effort, existing knowledge, experience, and the accessed user interaction element (e.g., collaborative or game design elements). The isolation of actions and user interaction elements can improve the analysis of the real effects from an intervention in the system on the user engagement.

REFERENCES

Backstrom, L., Kumar, R., Marlow, C., Novak, J., and Tomkins, A. (2008). "Preferential behavior in online groups." Proceedings of the 2008 International Conference on Web Search and Data Mining. ACM, 117–128.

Burke, M., Marlow, C., and Lento, T. (2009). "Feed me: motivating newcomer contribution in social network sites." Proc. of the SIGCHI conference on human factors in computing systems. ACM, 945–954.

Calleja, G. (2011). In-game: From immersion to incorporation. MIT Press.

Codish, D. and Ravid, G. (2014). "Academic course gamification: The art of perceived playfulness." Interdisciplinary Journal of E-Learning and Learning Objects 10.1, 131–151.

Cohen, J. (1988). Statistical power analysis for the behavioral sciences 2nd edn.

Csikszentmihalyi, M. and Nakamura, J. (1979). "The concept of flow." Play and learning, 257–274.

Deterding, S., Dixon, D., Khaled, R., and Nacke, L. (2011). "From game design elements to gamefulness: defining gamification." Proc. of the 15th international academic MindTrek conference: Envisioning future media environments. ACM, 9–15.

DeVellis, R. F. (2003). Scale development: Theory and applications.

Feild, H. and Allan, J. (2009). "Modeling searcher frustration." Proceedings from HCIR.

Forte, A. and Lampe, C. (2013). "Defining, Understanding, and Supporting Open Collaboration: Lessons From the Literature." AMERICAN BEHAVIORAL SCIENTIST 57.5, SI, 535–547.

Groh, F. (2012). "Gamification: State of the art definition and utilization." Institute of Media Informatics Ulm University 39, 31.

Hamari, J. (2017). "Do badges increase user activity? A field experiment on the effects of gamification." Computers in human behavior 71, 469–478.

Huotari, K. and Hamari, J. (2017). "A definition for gamification: anchoring gamification in the service marketing literature." Electronic Markets 27.1, 21–31.

Huvila, I. (2008). "Participatory archive: towards decentralised curation, radical user orientation, and broader contextualisation of records management." Archival Science 8.1, 15–36.

Iacovides, I., Cox, A. L., McAndrew, P., Aczel, J., and Scanlon, E. (2015). "Game-play breakdowns and breakthroughs: exploring the relationship between action, understanding, and involvement." Human–computer interaction 30.3-4, 202–231.

Keppel, G and Wickens, T. (2004). "Effect size, power, and sample size." Design and Analysis. A Researcher's Handbook, ed 4, 159–801.

Kim, B. (2012). "Harnessing the power of game dynamics1: Why, how to, and how not to gamify the library experience." College & Research Libraries News 73.8, 465–469.

Kraut, R. E. and Resnick, P. (2011). "Encouraging contribution to online communities." Building successful online communities: Evidence-based social design, 21–76.

Lankes, R. D., Silverstein, J., and Nicholson, S. (2007). "Participatory networks: the library as conversation." Information technology and libraries 26.4, 17.

Lin, A., Gregor, S., and Ewing, M. (2008). "Developing a scale to measure the enjoyment of web experiences." Journal of Interactive Marketing 22.4, 40–57.

Liu, D., Santhanam, R., and Webster, J. (2017). "Toward meaningful engagement: a framework for design and research of gamified information systems." MIS quarterly 41.4.

Locke, E. A. and Latham, G. P. (2002). "Building a practically useful theory of goal setting and task motivation: A 35-year odyssey." American psychologist 57.9, 705.

O'Brien, H. L. and Lebow, M. (2013). "Mixed-methods approach to measuring user experience in online news interactions." Journal of the Association for Information Science and Technology 64.8, 1543–1556.

O'Brien, H. L. and Toms, E. G. (2008). "What is user engagement? A conceptual framework for defining user engagement with technology." Journal of the American Society for Information Science and Technology 59.6, 938–955.

Oomen, J. and Aroyo, L. (2011). "Crowdsourcing in the cultural heritage domain: opportunities and challenges." Proceedings of the 5th International Conference on Communities and Technologies. ACM, 138–149.

O'Brien, H., Cairns, P., and Hall, M. (2018). "A Practical Approach to Measuring User Engagement with the Refined User Engagement Scale (UES) and New UES Short Form." International Journal of Human-Computer Studies.

Pedro, L. Z., Lopes, A. M., Prates, B. G., Vassileva, J., and Isotani, S. (2015). "Does gamification work for boys and girls?: An exploratory study with a virtual learning environment." Proceedings of the 30th annual ACM symposium on applied computing. ACM, 214–219.

Ryan, R. M. and Deci, E. L. (2000). "Self-determination theory and the facilitation of intrinsic motivation, social development, and well-being." American psychologist 55.1, 68.

Seaborn, K. and Fels, D. I. (2015). "Gamification in theory and action: A survey." International Journal of human-computer studies 74, 14–31.

Siemens, J. C., Smith, S., Fisher, D., Thyroff, A., and Killian, G. (2015). "Level up! the role of progress feedback type for encouraging intrinsic motivation and positive brand attitudes in public versus private gaming contexts." Journal of interactive marketing 32, 1–12.

Zichermann, G. and Cunningham, C. (2011). Introduction - Gamification by Design: Implementing Game Mechanics in Web and Mobile Apps. Sebastopol, California: OŔeilly Media. p. xiv. ISBN 1449315399, 1st edition.