Investigating User-Created Gamification in an Image Tagging Task

Marc Schubhan[1], **Maximilian Altmeyer**[1], **Dominic Buchheit**[2], **Pascal Lessel**[1]
[1]German Research Center for Artificial Intelligence (DFKI), [2]Saarland University
Saarland Informatics Campus, Saarbrücken, Germany
[1]firstname.lastname@dfki.de, [2]dominic.buchheit@gmail.com

ABSTRACT
Commonly, gamification is designed by developers and not by end-users. In this paper we investigate an approach where users take control of this process. Firstly, users were asked to describe their own gamification concepts which would motivate them to put more effort into an image tagging task. We selected this task as gamification has already been shown to be effective here in previous work. Based on these descriptions, an implementation was made for each concept and given to the creator. In a between-subjects study (n=71), our approach was compared to a no-gamification condition and two conditions with fixed gamification settings. We found that providing participants with an implementation of their own concept significantly increased the amount of generated tags compared to the other conditions. Although the quality of tags was lower, the number of usable tags remained significantly higher in comparison, suggesting the usefulness of this approach.

Author Keywords
Customization; user-driven game design; bottom-up; motivation; replication

CCS Concepts
•**Human-centered computing** → **Empirical studies in HCI;**

INTRODUCTION
"One-size-fits-all" gamification approaches, i.e., every user receives the same game elements in a system, have been shown not to be ideal [6, 22, 34]. A reason for this is that users have different preferences and expectations based on individual aspects like their player type [34] or personality traits [7].

Research on tailoring gamification tries to mitigate the drawbacks of generic solutions by providing individual gamification settings. Two main approaches are currently being investigated [21]: personalization, i.e., the system creates a user model and adapts the gamification towards the user automatically, and customization, i.e., users can adapt game elements to their needs. With this paper we add to the latter by investigating a particularly strong form of customization: allowing users to first describe which gamification concept they want to have in a system, and in a second step, to receive an implementation of their described concept. In contrast to "bottom-up" approaches in which users can change the gamification at the runtime of the system but only with a predefined set of game elements [11, 12], in our approach users are only limited by their imagination and their own ideas. As such a flexibility for users entails a high effort for developers, it is necessary to investigate whether and which beneficial effects it can produce. As customization approaches have been shown to result in positive effects (e.g. [8, 12, 29]), even when the available customization options are limited [13], it could be assumed that describing a gamification configuration and having it implemented for a task might have even stronger effects.

A first step in this direction has been taken by Lessel et al. [10]. Here it was investigated which kinds of gamification approaches users suggested that would motivate them for different tasks (e.g., cleaning the kitchen) and it was found that the concepts were quite diverse (again, highlighting the need for tailoring gamification). Participants' self-reports indicated that these concepts might be motivational for them, but the authors did not implement the concepts and thus could not determine whether an implementation would indeed lead to positive effects as intended by the participants. We continue this line of work by also requesting participants to describe a concept that would motivate them for an image tagging task, similar to [10]. Unlike the other work, we implemented the suggested concepts for the creators. The area of image tagging was chosen as gamification has been shown to be effective here [15, 17] and it was already used in customization studies [13]. Importantly, it makes measuring the impact of the intervention easily measurable based on tag quality and quantity.

We conducted a between-subjects user study in this context (n=71) and could replicate that gamification is beneficial in general (i.e., participants who tagged in a gamification condition produced a higher number of tags compared to those that had no gamification). In addition, we found that those who had the chance to describe their own gamification concept and receive its implementation later produced significantly more tags than those who only received a generic gamification approach. Although the tag quality dropped, the amount of usable tags was still higher compared to other gamification conditions, i.e. tags sufficiently related to the image shown.

CHI'20, April 25–30, 2020, Honolulu, HI, USA

ISBN 978-1-4503-6708-0/20/04…$15.00
DOI: https://doi.org/10.1145/3313831.3376360

This paper contributes, to our knowledge, the first investigation of a truly user-driven gamification scenario, in which users suggest a form of gamification they would like to have in a given scenario and receive its implementation. Although previous research has shown that users can suggest such gamification concepts and that it might lead to positive effects when implemented for the corresponding users [10], the latter has not yet been shown. The paper demonstrates that providing users with their self-created gamification concepts leads to positive effects compared to no gamification or fixed gamification. Thus, with this paper we provide a foundation for continuing this line of research, especially in light of the increased effort that such an approach would mean for developers.

RELATED WORK

As described above, driven by the need for individualization to account for interpersonal differences in the perception of gamified interventions, two main approaches, personalization and customization, have been studied. We will situate our work by presenting related research in both fields.

Personalization of Gamified Systems

Many factors have been shown to play a role to select suitable gamification elements. For instance, Jia et al. [7] investigated the role of personality traits to adapt gamified systems. They found that personality influences how people perceive certain gamification elements, showing the potential of a personality-based adaptation of gamified systems. This is supported by findings from Orji et al. [20] who studied the relationship between personality traits and persuasive strategies within systems for health. In line with the results by Jia et al. [7], they found correlations between personality traits and the perceived persuasiveness of the presented strategies. Besides personality, demographic differences have also been investigated. Birk et al. [3] analyzed the impact of age on several game-related factors and found that preferences and play motives change with increasing age from focusing on performance towards focusing on completion and enjoyment. Also, gender has been identified to have an impact on the perception of motivational elements. Orji [19] found that the perception of persuasive strategies differs between male and female participants in the domain of healthy eating, which is supported by Oyibo et al. [23] who found that competition and rewards are more relevant for male users. Moreover, user type models have been developed to inform decisions about the selection of suitable gamification elements. For instance, the Hexad user types model by Tondello et al. [34] distinguishes between six different user types, provides a validated questionnaire to derive a user's type [33] and has been shown to have an impact on the perception of gamification elements [1, 34]. The dynamics of behavioral intentions and their impact on the perception of gamification elements has been researched by Altmeyer et al. [1]. They replicated the correlations between gamification elements and Hexad user types which have been found by Tondello [34] and showed that behavioral intentions should be considered and combined with the Hexad user types to select gamification elements for gamified fitness systems.

While the aforementioned studies show promising results of using certain factors to tailor gamified systems automatically, a "perfect" user model accounting for every user preference has not yet been found, as far as we know. The huge amount of potential relevant factors and the problem that not all personal preferences can yet be formalized by respective dependent variables makes finding such a perfect user model unlikely. Therefore, in this paper, we investigate whether giving users full control of describing their own gamification concept without any technical or game-element-related restrictions has positive effects on their enjoyment and performance. Besides mitigating the aforementioned limitations of personalization, this approach might lead to increased feelings of autonomy, which has been shown to have various positive effects on enjoyment and motivation. This will be discussed in the following section.

Customizable Gamification

Autonomy is, besides relatedness and competence, one of three main drivers of intrinsic motivation, according to the Self-Determination Theory [5]. Allowing users to alter the gamification setup and giving them the choice of selecting gamification elements according to their preferences may support this need. Consequently, customizing gamified systems not only has potential for selecting suitable gamification elements and mitigating the problem of interpersonal differences, but may also increase the feeling of autonomy, which affects intrinsic motivation positively [5]. The offered choice is also likely to have beneficial effects on motivation (e.g, in [32] it was shown that providing a choice reduces anxiety and in [37] that it increases intrinsic motivation). In addition, users may experience higher levels of ownership when having an influence on their gamification setup [24].

The aforementioned benefits of autonomy and having a choice provide solid reasons for investigating customizable gamification in addition to the ongoing personalization efforts. For instance, Nicholson recommends to put users in the loop and states that gameful systems should allow users to "*create their own tools to track different aspects of the non-game activity, to create their own leveling systems and achievements, to develop their own game-based methods of engaging with the activity and to be able to share that content with other users*" (p. 4, [18]). Following these recommendations, different degrees of customization have been investigated in games and gamification research. A very basic approach to customization has been considered in [13]: allowing users to enable or disable gamification. The authors found that even such a simple choice has positive effects on the enjoyment and task performance within an image tagging system. On the aesthetic and functional level, Kim et al. [8] have shown that allowing players to change the visual appearance and characteristics of in-game entities positively affects their experience. Additionally, letting users select the type of reward within a game (e.g., receiving points to compete on a leaderboard, receiving a currency for buying virtual items, unlocking a short story or making progress) has been investigated by Siu and Riedl [29]. In a comparative study, the authors were able to show positive effects when participants

were able to select their type of reward themselves, compared to participants that had no such choice. That users actually want to gamify certain tasks on their own to make them more engaging has also been recognized by Donald Roy [26] and recently, in a study about how to motivate safe driving, Steinberger et al. [30] reported that some of their participants tended to self-gamify and want to establish *their* ideas.

Concerning the amount of offered options, Lessel et al. [11, 12] went one step further than the aforementioned approaches by investigating what they call "bottom-up" gamification, i.e. allowing users of gamified systems to adapt the gamification elements that are being offered during the runtime of the system. This meant that users were able to select which gamification elements they want to have, combine them as they see fit and adjust any parameter (for example to adjust the amount of points being rewarded for an activity) according to their preferences. In the context of a task management application [11], the authors found that participants appreciated this freedom of choice and subjectively reported positive behavioral effects. In a follow-up study [12], the authors also found that participants who made use of the "bottom-up" gamification solved more microtasks within an image-labeling context, indicating that such an approach indeed positively affects task performance. However, the set of gamification elements and the way how they could be combined was still restricted to the game elements that were made available by the developers of the system. Therefore, users still did not have full control over their gamification setup and could only define configurations with these given elements.

One question is whether users, who are typically not game designers or experts in gamification, are able to set up motivating game or gamification concepts on their own. This has been investigated in [10]. Here, participants were asked to come up with a gamification concept to motivate themselves within several given contexts (such as saving energy or performing a physical activity) and describe their concept textually. Through a qualitative analysis, it was found that participants created a broad range of game concepts without much overlap (i.e. no gamification concept was described twice), that they perceived their game concepts to be motivating, that it was easy for them to develop these and that they were not overwhelmed by having the full choice over their concepts. These findings indicate that participants are able to create gamification concepts without much guidance and provide strong motivation to investigate what happens when actually implementing the textually described gamification concepts: a gap that we fill with this paper.

Similar results have also been found through the study of "Crowdjump", where players could post ideas and vote for them to adapt and advance a platform game [9]. Results show that players (even those who do not like platform games in particular) enjoyed Crowdjump. The fact that the community came up with an enjoyable game on their own (even as non-designers) supports the idea of letting users design their own gamification approach even further.

Summary

Related work has demonstrated that several approaches to account for the interpersonal differences in the perception of gamified systems exist. One approach is personalization, where several factors like personality [7, 20], age [3], gender [19, 23], player types [34] or behavioral intentions [1] have been shown to have an impact on the perception of gamification elements. However, limitations to this approach are the fact that such measures explain user preferences only to a certain extent, and that not all of these preferences can be formalized by respective measures yet. In contrast, letting users customize gamification elements has been shown to lead to a broad range of positive outcomes: the increased autonomy potentially leads to higher intrinsic motivation [5], which is measurable even when allowing only a very basic form of customization [13]. Also, increased feelings of ownership [24] and the offered choice were shown to lead to various positive outcomes [32, 37]. Driven by these benefits, existing gamification research about customization has demonstrated that users appreciate having customization options [11], users who actually make use of such options have an increased task performance [12] and users are generally able to come up with appealing gamification concepts to motivate themselves [10]. However, it remains unclear whether each user's self-created gamification concept actually has an impact on their performance and enjoyment when it is realized. In this paper, we address this open question by allowing users to describe their own gamification concepts without restricting them (besides focusing on one context in which the concept should be applied; in our case, an image tagging context). Since each proposed and textually-described gamification concept has been implemented specifically for each user, we are able to provide insights about the open question of whether such user-created gamification has measurable effects on their behavior.

APPARATUS

We follow the approach of [13], [16] and [17] and use an image tagging platform. Users are presented 15 paintings (see left part of Figure 1) and are asked to provide tags about "moods" that they associate with them. Since the platform is in German, the screenshots provided in this paper are translated to English[1]. The paintings were taken from Machajdik and Hanbury's study about affective image classification [14] and were the same as used in [13, 16, 17]. In [13, 17] it has been shown that a top-down gamification approach using points and leaderboards has positive effects on the amount of tags generated. When re-implementing the platform, we designed it in a way that allowed us to easily implement new gamification conditions. With this, our aim was to reduce the amount of time it took between implementing and deploying gamification concepts. For each gamification element, the platform loaded the respective HTML template and the corresponding logic written into a single JavaScript file. For the JavaScript file we used an interface where the following events got triggered during the tagging process:

[1] See supplementary material section A for the original German version of the screenshots.

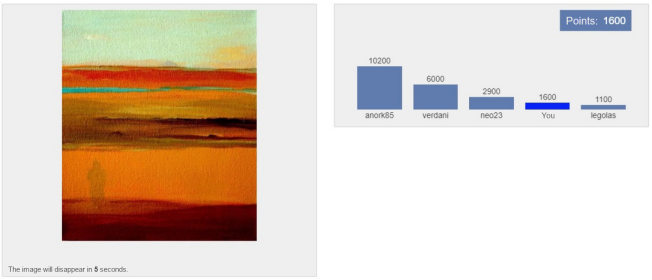

Figure 1. Screenshot of the *Top-Down Leaderboard* condition.

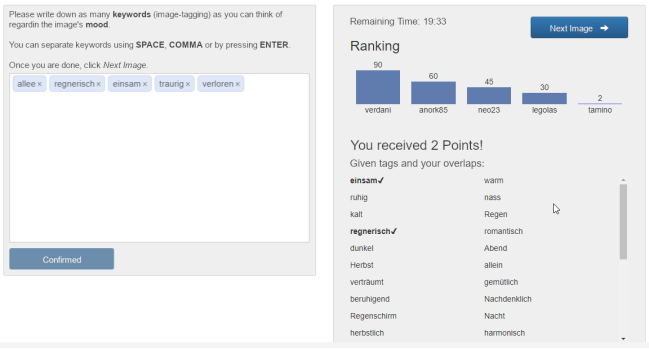

Figure 2. Screenshot of concept 16 from the user-created concepts.

- ***AfterGetImage***: Fired when an image was loaded by the platform. It is used, for example, to get additional information concerning the image, like the tags of other users, if a gamification concept incorporated such information.
- ***AfterFlipImage***: Fired when an image got hidden after being presented to the user for five seconds. It is used, for example, to start timers or trigger visibility of certain elements at the same moment the user is allowed to start tagging.
- ***AfterTagAdded***: Fired when a user has entered a tag. It is used, for example, to increase points.
- ***AfterTagRemoved***: Fired when a user has removed a previously entered tag. It is used, for example, to decrease points.

These events allowed us to adapt the platform to each new game element visually and logically and minimize the effort it took to implement new concepts directly into the platform.

Regarding the overall functionality, we implemented a welcome page, describing the study and providing users with information on data privacy, a questionnaire view, a tutorial, the tagging view (see Figure 2 left) and an end screen, where we thanked participants for their participation. The questionnaire view, the tutorial and the tagging view were implemented following the approach of [13].

Three decisions were made in advance for every gamification concept which might be suggested: First, if a gamification concept incorporated multiplayer components, we chose to "fake" opponents such that every user of the platform would experience the same pre-defined procedure, i.e. they faced the same "opponents". For example, the leaderboard and points element always contained the same leaderboard for every user, similar to [17]. The opponents had 1100, 2900, 6000 and 10200 points, i.e. users needed 103 tags to get to first place as each tag got rewarded with 100 points. Since our participants could only participate once, they were not able to notice the same behavior of the "opponents" across multiple sessions. This way, we were able to preserve comparability within multiplayer concepts, while the participants presumably were not aware of playing against non-human players. Second, if users incorporated tags of other users in their gamification concept, we used the tags that had already been stored in our database for the images to make, for example, statistics like "most used tags for this image" believable. Third, to maintain comparability between the user-created concepts, it was important to set a "standard" for the amount of tags a concept aimed for. Comparability would, for example, be threatened if one concept aims at 200 tags for one "playthrough", while another only expects 50. Based on the leaderboard element, we decided to adopt the goal of approximately 100 tags for all concepts.

STUDY

Our study investigates the following hypotheses:

H1 Participants who use gamification for image tagging perform better than those using no gamification.

H2 Participants who use their user-created gamification concepts perform better than those who use fixed top-down gamification.

H3 The quality of tags in user-created gamification concepts is worse than the quality of tags in top-down gamification or no gamification.

H4 The amount of usable tags will be higher for participants who use their user-created gamification concepts than for those using fixed top-down gamification or no gamification.

H1 is derived from [13] and [17] where such results have already been found, thus, we expect to replicate this here as well. **H2** is based on the results found and suggestions made by [10] and the self-determination theory [28]. **H3** is based on a negative correlation between tag quantity and quality in [17]. As **H2** suggests higher tag quantity with the user-created gamification, we consequently expect a worse overall quality in the user-created gamification condition. With **H2** and **H3**, we expect that the total number of tags of high quality generated through user-created gamification concepts is significantly higher compared to the other conditions (**H4**), i.e. while we expect lower overall tag quality, we expect this to be compensated for by the higher amount of tags generated,thus leading to an overall higher tag quantity of qualitative tags.

Based on the hypotheses, we prepared three conditions: the *Baseline* where each participant tagged the presented images without being accompanied by gamification at all; the *Top-Down Leaderboard* condition, which was based on Mekler et al.'s [17] top-down condition, where participants also tagged

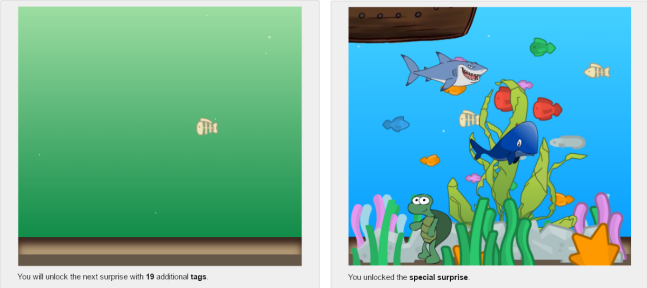

Figure 3. Screenshot of the *Top-Down Aquarium* condition, which was also one of the gamification concepts in *Own*. By providing tags participants could unlock changes to the fish tank. The image on the left shows an empty fish tank, when no tags were entered, the image on the right shows the fish tank after 100 tags were provided.

the presented images, receiving points for each tag they entered and climbing up a leaderboard, which was shown to them next to the images (see Figure 1).

And finally, participants in the*Own* condition were asked to write down their own idea of a gamification concept to go along with image tagging and later on did the image tagging task with their user-created gamification.

After we received and realized the first gamification concepts in *Own*, we decided to include one of the user-created concepts as an additional top-down condition. This way, we wanted to see how both top-down conditions compare to each other, since the *Own* concept offered a different set of game elements than just points and leaderboards. In addition, this allowed us to compare performance in both top-down conditions to *Own*. To further investigate the latter, we wanted to see how the individual who provided the concept in *Own* compared to the participants who received this concept as a top-down variant. We call this additional top-down condition *Top-Down Aquarium*, as participants could gradually fill a fish tank, the more tags they provided (see Figure 3).

Our main dependent variable was tag quantity, i.e. how many tags participants submitted for the pictures they were shown. Similar to [13] and [17], we additionally used tag quality as a dependent variable in order to gain insight on potentially increasing or decreasing quality.

Based on our hypotheses, we expected the following effects:

- If **H1** is true, *Top-Down Leaderboard*, *Top-Down Aquarium* and *Own* should generate more tags than *Baseline*.
- If **H2** is true, *Own* should generate more tags than *Top-Down Leaderboard* and *Top-Down Aquarium*.
- If **H3** is true, *Own* should generate tags of lower quality than *Baseline*, *Top-Down Leaderboard* and *Top-Down Aquarium*.
- If **H4** is true, *Own* should quantitatively generate more tags of reasonable quality than *Baseline*, *Top-Down Leaderboard* and *Top-Down Aquarium*.

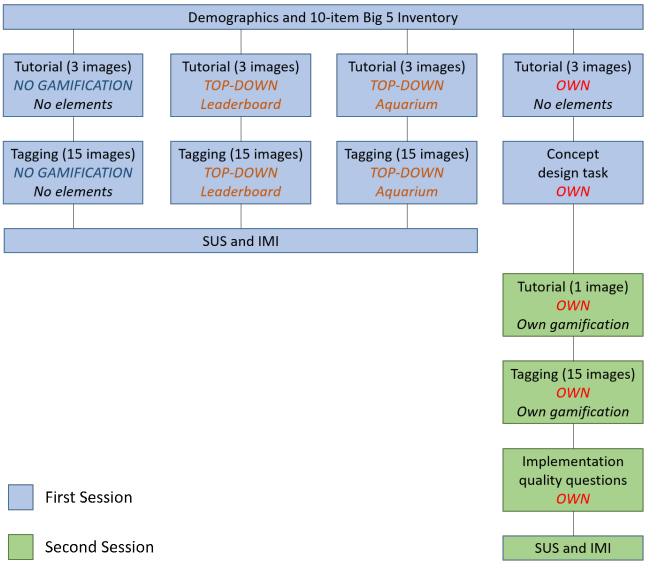

Figure 4. Sequence of the different conditions.

Method

The study was designed as a between-subjects study. It was designed to have one session if a participant was in *Baseline*, *Top-Down Leaderboard* or *Top-Down Aquarium* and two sessions if the participant was assigned to the *Own* condition. Participants were automatically assigned to one of the four conditions upon entering the platform, but we decided to double the chances in favor of *Own*, as we expected some of the concepts to be faulty or unusable, and tried to compensate for this. Consequently, the distribution was 1:1:1:2. Figure 4 gives an overview of the sequence for the different conditions, which is explained in the following.

First Session

Participants received a link leading to the image tagging platform, where they could read a short introduction text stating that the image annotation would help us to see which moods an image can convey. This way, at no point was a hint towards a motivational study provided, so as not to introduce a bias. After following the link and accepting the platform's privacy conditions, participants had to complete a demographic questionnaire about age, gender and nationality, followed by the 10-item Big Five Inventory [25]. A tutorial followed, where the image tagging process was explained and participants were asked to tag three tutorial pictures. In *Baseline* and *Own*, no gamification was active, while in *Top-Down Leaderboard* and *Top-Down Aquarium* gamification was already active. The tutorial was used to explain the conditions participants were in: Participants in *Baseline* and *Own* only received the information that they would see an image for five seconds and they should think about moods they would assign to that image; participants in *Top-Down Leaderboard* additionally received the explanation that they gain points per submitted tag and that they climb the leaderboard that way. Participants in *Top-Down Aquarium* received the same explanation as *Baseline* and *Own* with the addition that they unlock fish or plants for their fish tank every third tag provided. This way, we helped participants understand the system and, if available, the gamification.

Participants who were not in *Own* now continued with the main run of tagging with gamification (*Top-Down Leaderboard*, *Top-Down Aquarium*) or without (*Baseline*). The 15 abstract paintings were shown in randomized order, each for five seconds after which tags for it could be provided (with no time pressure). Once all paintings were tagged a closing questionnaire was provided. It contained questions to evaluate the platform's usability with the help of the German version [27] of the *System Usability Scale* (SUS) [4], as well as the German version of the *Intrinsic Motivation Inventory* (IMI) [36].

Participants in *Own* were asked after the tutorial to describe a gamification concept that would motivate them to do the previously experienced task. Here, we followed the approach of [10] in which participants also needed to develop a gamification concept textually using at least 700 characters[2]. Afterwards they answered four more questions about why they think their gamification concept would motivate them ("*Why do you think that your game concept will motivate you?*"), which element of their concept is most important for them ("*Which element of your concept is most important for you?*"), and whether they had any prior experience with designing games ("*I already designed a game or was part of such a process.*") or image tagging ("*The principle of image tagging was known to me a priori.*"). Finally, participants were asked to enter their e-mail address so we could contact them again once we implemented their concept and were done with the first part of condition *Own*.

Second Session

Only participants in *Own* had a second session. After receiving a concept, we analyzed it and classified it as either *feasible*, *infeasible* or *not sufficient* (see below). All feasible concepts were implemented and the respective participants were again invited to the image tagging platform with the request to complete the second part of our study. On the tagging platform they followed a similar procedure as the other conditions in the first session, starting with a single tutorial run, where the image tagging and their user-created gamification concept were explained. Although they had already gone through the tutorial in the first session, we decided to add another tutorial here to make this run comparable to the other conditions' first session and to give them a proper explanation of the implemented features of their gamification concept. Afterwards they started tagging the 15 paintings, similar to the other conditions, but with their own gamification configuration active. The closing questionnaire also contained the SUS and IMI as above, but also statements on their perception of the implemented gamification concept. These statements were about their satisfaction with their textual concept ("*I'm satisfied with my game concept*" (this will be abbreviated with Concept subsequently)), completeness of our implementation ("*All aspects of my game concept were realized*" (Completeness)), visual realization ("*My game concept was visually satisfying*" (Visuals)), general satisfaction with our implementation ("*The game during image tagging matches my game concept*" (General)), the motivational impact of the design process ("*Creating my own game concept motivated me to tag images*" (Design Process)) and the motivational impact of the game itself ("*The game during image tagging motivated me to tag images*" (Implementation)). All statements had to be rated on a scale from 1 (Disagree) to 5 (Agree).

[2]See supplementary PDF in section B for the provided task description.

	Rating				
Aspect	*M*	*Mdn*	*SD*	*Min*	*Max*
Concept	4.2	4.5	1.0	2	5
Visuals	4.6	5.0	0.7	3	5
Completeness	4.7	5.0	0.6	3	5
General	4.9	5.0	0.3	4	5
Design Process	3.3	4.0	1.2	1	5
Implementation	4.2	5.0	1.1	1	5
Satisfied with Implementation	4.3	4.4	0.6	3	5

Table 1. Mean (M), median (Mdn) and standard deviation (SD) of the different aspects of the user-created implementation rated on a 5-point scale (n=20). The actual questions can be found in the Method section of this paper.

The study was approved by the Ethical Review Board of the Faculty of Computer Science at Saarland University[3] (No. 18-2-2).

Concept Coding and Feasibility

Every concept we received in *Own* was analyzed for which kinds of game elements were used. For this, we used the codebook as presented in Lessel et al. [10]. Two independent coders coded all the concepts and differences among the coders were discussed and solved[4]. As a quality metric, a third independent coder coded the concepts as well, and the resulting inter-rater agreement was κ=0.85. This can be considered as almost perfect [31]. After one consistent set of codes existed for a concept, we continued to discuss the feasibility of a concept. For this purpose we classified the concepts in three different categories: *Feasible*, *infeasible* and *not sufficient*.
A concept was considered *feasible* if we assessed it as implementable in a short time frame and suitable in the context of the described image tagging task. An *infeasible* concept contained elements that we considered as not realizable in the context of the study in terms of comparability to other participants, or that would take too long to implement. An example of this is a gamification concept containing elements that would not be realizable such as granting monetary rewards for every provided tag or changing the nature of the task (such as allowing skipping of images). Lastly, concepts were *not sufficient* if they did not describe a gamification concept. Overall we received 45 concepts, of which 20 were feasible[5], 22 infeasible in the study setting and 3 not sufficient. Of the 20 feasible concepts, no two concepts had the same set of codes, and 34 game elements were suggested. Again, this shows that participants were quite diverse, similar to [10].

[3]https://erb.cs.uni-saarland.de/, last accessed on January 6, 2020

[4]See supplementary PDF in section C for an overview of the codes of our feasible concepts.

[5]See supplementary PDF in section D for the user-created concepts, screenshots and implementation specifics.

		Tag quantity								
		Overall			Tag quality > 1 "Usable"			Tag quality > 2 "Good"		
Condition	*n*	*M*	*Mdn*	*SD*	*M*	*Mdn*	*SD*	*M*	*Mdn*	*SD*
Baseline	19	43.2	45.0	13.7	42.1	45.0	13.2	27.7	25.0	11.3
Top-Down Leaderboard	19	68.0	54.0	30.0	66.5	53.0	29.1	45.9	41.0	23.2
Top-Down Aquarium	15	59.8	62.0	20.1	52.4	56.0	15.7	26.5	25.0	10.3
Own	18	98.3	93.0	41.3	93.1	87.5	36.5	51.1	48.5	19.4

Table 2. Mean (M), median (Mdn) and standard deviation (SD) of the amount of tags generated by participants among the different conditions and different quality levels.

Concept Satisfaction

Participants' level of satisfaction with the realization of their user-created concept is important when it comes to the *Own* Condition. Participants being unsatisfied could mean that the concept was not realized in the way they imagined, and thus, the performance of the participant might confound the data. Table 1 shows how satisfied participants were with our realization of their concept. As can be seen most of the factors, including the "Satisfied with Implementation" score, are within the range of 4 to 5 out of 5. In detail, this means that participants were still satisfied with their concept after playing it and reading it again, were satisfied with the visual realization of their idea, were satisfied with the features we implemented and thought that in general, the presented implementation represents what they imagined for their gamification concept. Nonetheless, the minimum values show that there are outliers as well. Based on the above argument, these need to be excluded so as not to confound the results (see below).

Determining Tag Quality

To determine the tag quality, two independent coders rated participants' tags, similar to [13]. This resulted in three possible values a rating could have. A value of 1 means that the tag was neither a mood nor related to the picture. 2 means that the tag was not a mood but described the picture or vice versa and 3 means that the tag was a fitting mood for the picture. Overall, 5820 tags were coded. In a first run, Cohen's κ was found to be κ=0.49, which implies a moderate agreement between coders [31]. To increase the agreement, 330 controversial tags were discussed to refine the coding rules and to reveal which interpretations were available. After this discussion, all remaining 5490 tags were coded separately again. The κ after this second run was 0.86, which represents an almost perfect agreement [31]. For those tags that still deviated after this process we calculated the mean rating of the coders to account for the remaining differences among the coders.

Participants

Overall, 105 participants took part in our study (*Baseline*: 20, *Top-Down Leaderboard*: 20, *Top-Down Aquarium*: 20, *Own*: 45). Every participant spoke German. They were recruited via the German Facebook survey group "Surveys for study projects", the subreddit "SampleSize", SurveyCircle, a student mailing list (reaching computer science, media informatics and psychology students), as well as the authors' social networks. As previously described, 25 out of the 45 submitted concepts in *Own* were infeasible or not sufficient, resulting in 20 remaining participants in *Own*. We removed outliers

		Tag quality		
Condition	*n*	*M*	*Mdn*	*SD*
Baseline	19	2.7	2.8	0.2
Top-Down Leaderboard	19	2.7	2.7	0.6
Top-Down Aquarium	15	2.4	2.4	0.3
Own	18	2.5	2.6	0.2

Table 3. Mean (M), median (Mdn) and standard deviation (SD) of the tag ratings of participants' tags among the different conditions.

from our dataset using Tukey fences [35], i.e. if they were above 1.5x the interquartile range in regard to tag quantity (5 outliers), underneath an average tag quality rating of 1.5 (3 outliers) and below 1.5x the interquartile range in terms of "Satisfaction with Implementation" (1 outlier). These borders were chosen for outlier detection as we tried to prevent taking data points where the motivational origin might have been something other than the gamification concept (tag quantity), submitted tags were unrelated to the picture (tag quality) or participants were unhappy with our implementation ("Satisfaction with Implementation"). The latter were not representative of *Own* in terms of playing their own gamification concept as described above. Overall, these criteria led to 9 exclusions (1x *Baseline*, 1x *Top-Down Leaderboard*, 5x *Top-Down Aquarium*, 2x *Own*) and a final set of 71 participants across all conditions (gender: 39x female, 30x male, 1x non-binary, 1x no answer; age: 31x 18-24, 29x 25-31, 2x 32-38, 3x 39-45, 4x 53-59, 1x > 60, 1x no answer; nationality: 66x German, 1x Bosnian, 1x Chinese, 1x Austrian, 1x Swiss, 1x Luxembourg).

Results

The following results are mostly based on analyses of variance (ANOVA). In all cases we chose the Kruskal-Wallis ANOVA as the corresponding residuals were not normally distributed. Post-hoc tests were adjusted using the the Benjamini & Hochberg adjustment method [2] in order to prevent type I error accumulation.

R1: Participants in a gamification condition produced significantly more tags than those who tagged without gamification

Table 2 shows the average tag count per condition, separated into an overall tag count and a tag count excluding tags of quality 1 as well as tags of quality 1 and 2. Similar to the related work [13, 17], we first considered only the overall number of tags. An ANOVA revealed that there are significant differences between the conditions ($H(3)=25.2$, $p<.001$). According to the post-hoc tests, all comparisons with

		Enjoyment			Competence			Autonomy			Pressure		
Condition	*n*	*M*	*Mdn*	*SD*	*M*	*Mdn*	*SD*	*M*	*Mdn*	*SD*	*M*	*Mdn*	*SD*
Baseline	19	7.0	8.0	3.6	5.8	6.0	2.6	6.9	7.0	3.1	5.9	6.0	3.1
Top-Down Leaderboard	19	6.9	8.0	3.7	5.8	6.0	2.8	7.4	8.0	2.4	5.7	5.0	3.2
Top-Down Aquarium	15	6.9	6.0	3.3	5.1	5.0	2.3	7.7	8.0	2.1	5.0	6.0	2.4
Own	18	7.1	7.5	3.6	5.4	5.0	2.7	8.2	9.0	3.3	4.7	4.5	2.6

Table 4. Mean (M), median (Mdn) and standard deviation (SD) of the IMI scores in the categories Enjoyment, Competence, Autonomy and Pressure.

Baseline, i.e. with *Top-Down Leaderboard* (p_{adj}=.021; r=.43), with *Top-Down Aquarium* (p_{adj}=.035, r=.37) and with *Own* (p_{adj}=.000, r=.82) were significant. Based on the mean values seen in Table 2 and the tests performed, we can derive that every condition using gamification scored significantly better than *Baseline*. This supports **H1**.

R2: Participants who created their own gamification concept produced significantly more tags compared to those in a top-down condition

The post-hoc tests of the ANOVA also revealed significant differences between *Own* and *Top-Down Leaderboard* (p_{adj}=.027, r=.39) as well as *Own* and *Top-Down Aquarium* (p_{adj}=.022, r=.45). Hence, we found significant differences between our two top-down conditions and *Own*. This supports **H2**.

R3: Top-Down Leaderboard showed no significant difference in tag quality compared to no gamification

Table 3 shows the quality ratings across conditions. *Baseline* and *Top-Down Leaderboard* achieved the highest ratings here. We performed another ANOVA on these values, which revealed that differences between conditions exist (H(3)=14.13, p=.003). Post-hoc tests showed that there is no measurable significant difference between *Baseline* and *Top-Down Leaderboard* (p=.598). This is in line with results from [17].

R4: Top-Down Aquarium and user-created gamification showed significantly lower tag quality compared to no gamification

Said post-hoc tests also show that *Baseline* differs significantly from *Top-Down Aquarium* (p_{adj}=.006, r=-.57) and from *Own* (p_{adj}=.042, r=-.38). Consequently, this supports **H3** in terms of *Baseline* vs. *Own*.

R5: User-created gamification showed no significant difference in tag quality compared to top-down gamification

The post-hoc tests on tag quality additionally revealed that the differences between *Top-Down Leaderboard* and *Own* (p_{adj}=.108) as well as *Top-Down Aquarium* and *Own* (p_{adj}=.338) are not significant. Taking **R4** into consideration as well, **H3** is consequently only partially supported.

R6: Participants who created their own gamification produced a significantly higher amount of usable tags compared to the other conditions

Besides overall tag quantity, Table 2 shows the amount of tags per condition when counting only those with quality level 2 or better ("usable tags") and those with quality level 3 only ("good tags"). Running an ANOVA on these mean values reveals that there are still significant differences when counting usable or good tags only (H(3)=25.02, p<.001; H(3)=22.32, p<.001). After adjusting the p-values, we found significant differences regarding usable tags between *Baseline* and *Own* (p_{adj}<.001, r=-.82), *Top-Down Aquarium* and *Own* (p_{adj}=.006, r=.55) as well as *Top-Down Leaderboard* and *Own* (p_{adj}<.036, r=.37). Regarding good tags only, we found differences between *Baseline* and *Own* (p_{adj}<.001, r=-.65) as well as *Top-Down Aquarium* and *Own* (p_{adj}<.001, r=.63). Although *Own* has the second lowest mean quality score (see Table 3) and its quality is significantly lower than *Baseline*, these results show that *Own* still excels, as the amount of usable or good tags is not exceeded by another condition, and in nearly all cases *Own* produces significantly more tags. This supports **H4**.

Additional Results

The following results support our main results and provide further insights in the context of user-created gamification.

AR1: Participants in Top-Down Aquarium performed worse compared to the participant in Own who created the gamification concept

Running an one-sample t-test on the *Top-Down Aquarium* subset and comparing their tag quantity mean (M=59.8 tags) to the original inventor's amount of tags (Aq_{Cre}=81) revealed that there is a significant difference with a large effect size (t(14)=4.09, p<.001, d=1.06), further supporting **H2**. By inspecting the individual results in *Top-Down Aquarium*, in two cases participants in this condition provided even more tags than the content creator (Aq_{P1}=84, Aq_{P2}=101 tags). Checking the quality scores of the creator ($Aq_{Cre\_quality}$=2.88) with the mean quality score in *Top-Down Aquarium* (M=2.42), we again see a significant difference (t(14)=7.15, p<.001, d=1.84). No other participants in *Top-Down Aquarium* had a higher quality score than Aq_{Cre}. When comparing the quality scores of Aq_{P1} (1.96) and Aq_{P2} (2.09) we also see that they performed worse. In sum, from a motivational standpoint, the participant who created the aquarium gamification concept appeared to profit the most from it.

AR2: No condition produces a significant difference in the IMI subscales

Table 4 shows that *Own* had the highest score for perceived choice. Nonetheless, an ANOVA did not reveal a significant difference across conditions (H(3)=2.75, p=.433). This was also true for the other IMI subscales (Enjoyment: H(3)=.04, p=.998, Competence: H(3)=.58, p=.901, Pressure: H(3)=2.24, p=.524).

AR3: Half of the concept creators would have changed something about their concept after using it for the first time

We did ask participants whether they would change something about their concept after they used it, to which 50% answered yes. Considering that after exclusion participants were satisfied with their concepts, this might be the consequence of noticing flaws or improvements when "testing" the implementation and is likely not related to the implementation itself. This outcome yields the potential that with further iterations of the implementation, results in *Own* might become stronger.

Regarding the Big Five Inventory we found no conclusive results and did not report them here for space reasons.

Discussion

Our results show the benefits of letting users create their own gamification concepts. It has beneficial effects on the amount of tags submitted by a participant (**H1**) and users perform better in that regard compared to using fixed gamifications (**H2**). Although we found worse tag quality when comparing no gamification to user-created gamification (**R3**), *Own* was still able to generate significantly more usable tags in total when compared to the other conditions (**H4**). Looking at the most qualitative tags, i.e., those with a quality rating of 3, *Own* is similar to *Top-Down Leaderboard* and excels the other conditions, but is not significantly worse than any of the other conditions (**R6**). In summary, we can say that allowing users to create their own gamification concept and providing them with it is beneficial for the image tagging task. Although *Top-Down Leaderboard* showed that top-down gamification can be as effective in terms of generating tags of good quality as user-created gamification, *Top-Down Aquarium* showed that this is not always the case. Hence, user-created gamification represents a more reliable approach to generate more tags of good and usable quality. A first insight into a direct comparison between concept creators and the same concept provided as a top-down gamification is given in **AR1**. It indicates that the performance of users who create and receive their own concept, in comparison with users who received the same concept in a top-down fashion, was better. Although this result should not be overestimated (as we would need a top-down comparison group for each of the *Own* concepts), the instance that we investigated indicated that only the creator of the concept showed a high qualitative and quantitative performance.

Even though it is beneficial for the image tagging task, this approach comes with some drawbacks. Giving users their user-created concepts leads to increased effort as concepts need to be implemented first. In our case, it took on average 23 days from concept submission, to coding, implementation, providing it to the participants and them continuing their work, although we prepared the platform for fast realization and integration of new concepts. Implementing individualized gamification concepts requires more effort for developers compared to realizing one gamification configuration and providing every user with it. And even compared to gamification systems that use personalization, it is still more work, as the possible range of personalization options is much larger in our case. In addition, with the work presented in this paper, we did not want to investigate the developers' perspective on user-created gamification yet. Instead, we wanted to first learn whether user-created gamification provides benefits for users in general. From our experiences during the experiment, we reason that in the long run, the set of implemented game elements becomes large enough so that new concepts can be realized faster and with less effort, as developers will be able to re-use many implemented game elements. Given the personalization literature (see related work section), it seems not yet possible to automate this process for all kinds of users (especially in respect to their diversity as shown in this paper, as well as in [10]). Thus, future work should also focus on the developers' perspective and investigate how to lessen the burden for them.

An open question is where our effects originate from in particular. There are several potential explanations for this, for example the user's investment (users were more inclined to provide tags because they has already invested time in writing a gamification concept), good fit (the concept fit the users' expectations well and they therefore provided more tags), ownership (users provided more tags because they played their own gamification) and autonomy (being able to create the gamification motivated the user to provide more tags). As this study was meant to investigate whether a general effect can be found, it was not designed to specifically answer the question where this effect originates from. Nonetheless, **AR2** indicates that autonomy might not be the major factor here, as autonomy scores did not differ significantly across conditions. For the other explanations, our study did not provide more insights, but this is an important direction for future work.

As shown in **AR3**, 50% of our participants in *Own* stated that they would change something about their concept. Since the overall satisfaction with their concepts was rather high, we assume this to reflect possible improvements participants thought of while using their gamification concept for the first time. This could be interpreted as a hint that giving users the opportunity to do multiple iterations of concept creation, and thus further improve the concept to fit what they imagined, might improve our results. On the downside, this approach would further increase effort and time consumed to realize such concepts for developers and users. Nonetheless, future work should explore options that might reduce these efforts or other mechanics. In this sense, approaches such as "bottom-up" gamification [11], that allow users to customize a system at runtime might be used on top of purely user-created gamification.

Our top-down conditions differ significantly in terms of tag quality, although we expected them to perform similarly well. This might result from the different nature of both gamification concepts. As seen in Table 2, *Top-Down Aquarium* is the only condition to lose almost 50% of its submitted tags when looking at usable vs. good tags. Hence, we assume *Top-Down Aquarium* had a larger incentive to unlock everything, compared to reaching first place in *Top-Down Leaderboard*, at the cost of submitting lower quality tags. We conclude that not every gamification might be reasonable for our task. Even though *Own* also loses roughly 45%, we consider this to be less of an issue, given that *Own* significantly produced the highest amount of good tags next to *Top-Down Leaderboard*.

Limitations

There are some limitations to our study. First, as mentioned above, there was an average time gap of approximately 23 days between concept submission and using the realized gamification concept for participants in *Own*. This gap was partly caused by us (time to code the concepts and implement the game) and partly by participants who did not immediately complete the second part of the study. This may have impacted their perception of the implemented gamification concept, as they may have forgotten parts of their concept or may have remembered them differently. Furthermore, **AR1** is a limitation as it is based on a single representation of a concept creator. To get stronger results, the other cases should be investigated as well (i.e., provided in a top-down fashion). Hence, **AR1** should not be overemphasized yet. Results might look different if we had more comparisons between concept creators and concept users.

Another limitation comes from the image tagging platform and the task to tag images with "moods" itself. This limits the possible valid input participants can give for an image and might have had an influence on the motivation of our participants as well as the tag quality rating. For replication and comparison reasons, we chose to keep it that way, as this is what [13] and [17] did. Furthermore, we only tested these results in the image tagging context, hence, results are limited to this context and the question remains whether different task contexts yield similar results. Also, the amount of participants per condition is a limiting factor and could be higher overall.

CONCLUSION

With this paper we showed how giving users the freedom to create their own gamification for a task can increase their performance in comparison with fixed top-down gamification. This builds on and continues the work done by Lessel et al. [10] where users had already created gamification concepts, but were not able to use them. Instead, here, we gave them the opportunity to actually use their user-created gamification and used an image tagging platform already employed in other gamification studies such as [13] or [17]. Our results show that letting users individually create and use their own idea of gamification for an image tagging task can significantly improve their performance in terms of how many tags they provide per image. On the downside, we also found lowered tag quality compared to using no gamification at all. Yet, despite being lower, the tag quality was still reasonably high and when looking at the amount of qualitative tags, user-created gamification still lead to as many or significantly more tags, compared to other conditions. Hence the overall performance of creating and using your own gamification is beneficial in the image tagging task.

While "bottom-up" gamification until now was investigated at run-time, we investigated it at design-time. Future research could investigate it at run- and design time, which makes sense given that half of the participants would have changed something about their gamification after having used it. Such considerations would inform a strong form of participatory gamification and a potential alternative to top-down gamification used today.

In future research (besides the already mentioned aspects), the results of this study should be further investigated in different contexts. As gamification can be used in a broad variety of contexts, image tagging alone can not represent gamification as a whole. Furthermore, as the time gap between concept creation and actual implementation usage could have had an impact on our outcome, it would be reasonable to investigate this as well.

Lastly, it would be important to know the actual reason for the effects we found here. We suggested some reasons and were able to show that autonomy seems not to be one of them, yet there is still a broad range of reasons to be investigated, which should be a next step from here on. For example, time investment could be investigated by letting participants create their own gamifications in a similar way as we did, but provide them with a different top-down gamification afterwards and compare this to participants who did not invest time in the study before and just received the top-down gamification. This way, if results would be similar to our study, one could conclude that investing time into the concept creation itself already is a beneficial factor.

ACKNOWLEDGEMENTS

We thank Lea Schmeer and Tamino Lobert for their help with the tag quality ratings and implementations of the concepts, as well as all the participants and the CHI reviewers.

REFERENCES

[1] Maximilian Altmeyer, Pascal Lessel, Linda Muller, and Antonio Krüger. 2019. Combining Behavior Change Intentions and User Types to Select Suitable Gamification Elements for Persuasive Fitness Systems. In *International Conference on Persuasive Technology*. Springer, 337–349.

[2] Yoav Benjamini and Yosef Hochberg. 1995. Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *Journal of the Royal Statistical Society* 58, 4 (1995), 619–678.

[3] Max V. Birk, Maximilian A. Friehs, and Regan L. Mandryk. 2017. Age-Based Preferences and Player Experience: A Crowdsourced Cross-Sectional Study. *Proceedings of the Annual Symposium on Computer-Human Interaction in Play - CHI PLAY '17* (2017), 157–170.

[4] John Brooke. 1996. SUS - A Quick and Dirty Usability Scale. *Usability Evaluation in Industry* 189, 194 (1996), 4–7.

[5] Edward L. Deci, Richard Koestner, and Richard M. Ryan. 2001. Extrinsic Rewards and Intrinsic Motivation in Education: Reconsidered Once Again. *Review of Educational Research* 71, 1 (2001), 1–27.

[6] Carrie Heeter, Brian Magerko, Ben Medler, and Yu-Hao Lee. 2011. Impacts of Forced Serious Game Play on Vulnerable Subgroups. *International Journal of Gaming and Computer-Mediated Simulations* 3, 3 (2011), 34–53.

[7] Yuan Jia, Bin Xu, Yamini Karanam, and Stephen Voida. 2016. Personality-Targeted Gamification: A Survey Study on Personality Traits and Motivational Affordances. In *Proceedings of the 34th Annual ACM Conference on Human Factors in Computing Systems (CHI '16)*. ACM, 2001–2013.

[8] Keunyeong Kim, Michael G. Schmierbach, Saraswathi Bellur, Mun-Young Chung, Julia D. Fraustino, Frank Dardis, and Lee Ahern. 2015. Is It a Sense of Autonomy, Control, or Attachment? Exploring the Effects of In-Game Customization on Game Enjoyment. *Computers in Human Behavior* 48 (2015), 695–705.

[9] Pascal Lessel, Maximilian Altmeyer, and Nicolas Brauner. 2019. Crowdjump: Investigating a Player-Driven Platform Game. In *Proceedings of the Annual Symposium on Computer-Human Interaction in Play (CHI PLAY '19)*. ACM, 1–11 (to appear).

[10] Pascal Lessel, Maximilian Altmeyer, and Antonio Krüger. 2018. Users As Game Designers: Analyzing Gamification Concepts in a "Bottom-Up" Setting. In *Proceedings of the 22nd International Academic Mindtrek Conference (AcademicMindtrek '18)*. ACM Press, 1–10.

[11] Pascal Lessel, Maximilian Altmeyer, Marc Müller, Christian Wolff, and Antonio Krüger. 2016. "Don't Whip Me With Your Games": Investigating "Bottom-Up" Gamification. In *Proceedings of the 34th Annual ACM Conference on Human Factors in Computing Systems (CHI '16)*. ACM, 2026–2037.

[12] Pascal Lessel, Maximilian Altmeyer, Marc Müller, Chrisitan Wolff, and Antonio Krüger. 2017. Measuring the Effect of "Bottom-Up" Gamification in a Microtask Setting. In *Proceedings of the 21st International Academic Mindtrek Conference (AcademicMindtrek '17)*. ACM, 63–72.

[13] Pascal Lessel, Maximilian Altmeyer, Lea Verena Schmeer, and Antonio Krüger. 2019. "Enable or Disable Gamification?": Analyzing the Impact of Choice in a Gamified Image Tagging Task. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. ACM.

[14] Jana Machajdik and Allan Hanbury. 2010. Affective Image Classification Using Features Inspired by Psychology and Art Theory. In *Proceedings of the 18th ACM International Conference on Multimedia (MM '10)*. ACM, 83–92.

[15] Elisa D. Mekler, Florian Brühlmann, Klaus Opwis, and Alexandre N. Tuch. 2013a. Disassembling Gamification: The Effects of Points and Meaning on User Motivation and Performance. In *Proceedings of the 31st Annual ACM Conference Extended Abstracts on Human Factors in Computing Systems (CHI EA '13)*. ACM, 1137–1142.

[16] Elisa D. Mekler, Florian Brühlmann, Klaus Opwis, and Alexandre N. Tuch. 2013b. Do Points, Levels and Leaderboards Harm Intrinsic Motivation?: An Empirical Analysis of Common Gamification Elements. In *Proceedings of the First International Conference on Gameful Design, Research, and Applications*. ACM, 66–73.

[17] Elisa D. Mekler, Florian Brühlmann, Alexandre N. Tuch, and Klaus Opwis. 2017. Towards Understanding the Effects of Individual Gamification Elements on Intrinsic Motivation and Performance. *Computers in Human Behavior* 71 (2017), 525–534.

[18] Scott Nicholson. 2012. A User-Centered Theoretical Framework for Meaningful Gamification. In *Proceedings of the 8th International Conference on Games + Learning + Society (GLS '12)*. 1–7.

[19] Rita Orji. 2014. Exploring the Persuasiveness of Behavior Change Support Strategies and Possible Gender Differences. *CEUR Workshop Proceedings* 1153, BCSS (2014), 41–57.

[20] Rita Orji, Lennart E. Nacke, and Chrysanne Di Marco. 2017a. Towards Personality-driven Persuasive Health Games and Gamified Systems. *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems - CHI '17* (2017), 1015–1027.

[21] Rita Orji, Kiemute Oyibo, and Gustavo F. Tondello. 2017b. A Comparison of System-Controlled and User-Controlled Personalization Approaches. In *Adjunct Publication of the 25th Conference on User Modeling, Adaptation and Personalization (UMAP '17)*. ACM, 413–418.

[22] Rita Orji, Gustavo F. Tondello, and Lennart E. Nacke. 2018. Personalizing Persuasive Strategies in Gameful Systems to Gamification User Types. In *Proceedings of the 36th Annual ACM Conference on Human Factors in Computing Systems (CHI '18)*. ACM, 435:1–435:14.

[23] Kiemute Oyibo, Rita Orji, and Julita Vassileva. 2017. The Influence of Culture in the Effect of Age and Gender on Social Influence in Persuasive Technology. *Adjunct Publication of the 25th Conference on User Modeling, Adaptation and Personalization - UMAP '17* (2017), 47–52.

[24] Hector Postigo. 2007. Of Mods and Modders: Chasing Down the Value of Fan-Based Digital Game Modifications. *Games and Culture* 2, 4 (2007), 300–313.

[25] Beatrice Rammstedt and Oliver P. John. 2007. Measuring Personality in One Minute or Less: A 10-item Short Version of the Big Five Inventory in English and German. *Journal of Research in Personality* 41, 1 (2007), 203–212.

[26] Donald Roy. 1959. "Banana Time": Job Satisfaction and Informal Interaction. *Human Organization* 18, 4 (1959), 158–168.

[27] Bernard Rummel. 2015. System Usability Scale – jetzt auch auf Deutsch. `https://experience.sap.com/skillup/system-usability-scale-jetzt-auch-auf-deutsch/`. (2015). [Online; accessed 04-December-2019].

[28] Richard M. Ryan and Edward L. Deci. 2000. Self-Determination Theory and the Facilitation of Intrinsic Motivation, Social Development, and Well-Being. *American Psychologist* 55, 1 (2000), 68–78.

[29] Kristin Siu and Mark O. Riedl. 2016. Reward Systems in Human Computation Games. In *Proceedings of the 3rd Annual Symposium on Computer-Human Interaction in Play (CHI PLAY '16)*. ACM, 266–275.

[30] Fabius Steinberger, Ronald Schroeter, Marcus Foth, and Daniel Johnson. 2017. Designing Gamified Applications That Make Safe Driving More Engaging. In *Proceedings of the 35th Annual ACM Conference on Human Factors in Computing Systems (CHI '17)*. ACM, 2826–2839.

[31] Steve Stemler. 2001. An Overview of Content Analysis. *Practical Assessment, Research & Evaluation* 7, 17 (2001), 137–146.

[32] Ezra Stotland and Arthur L. Blumenthal. 1964. The Reduction of Anxiety as a Result of the Expectation of Making a Choice. *Canadian Journal of Psychology* 18, 2 (1964), 139–145.

[33] Gustavo F. Tondello, Alberto Mora, Andrzej Marczewski, and Lennart E. Nacke. 2018. Empirical Validation of the Gamification User Types Hexad Scale in English and Spanish. *International Journal of Human-Computer Studies* (2018).

[34] Gustavo F. Tondello, Rina R. Wehbe, Lisa Diamond, Marc Busch, Andrzej Marczewski, and Lennart E. Nacke. 2016. The Gamification User Types Hexad Scale. In *Proceedings of the 3rd Annual Symposium on Computer-Human Interaction in Play (CHI PLAY '16)*. ACM, 229–243.

[35] John W. Tukey. 1977. *Exploratory Data Analysis*. Addison-Wesley.

[36] M. Wilde, K. Bätz, A. Kovaleva, and D. Urhahne. 2009. Testing a Short Scale of Intrinsic Motivation. *Zeitschrift für Didaktik der Naturwissenschaften* 15 (2009), 31–45.

[37] Miron Zuckerman, Joseph Porac, Drew Lathin, and Edward L. Deci. 1978. On the Importance of Self-Determination for Intrinsically-Motivated Behavior. *Personality and Social Psychology Bulletin* 4, 3 (1978), 443–446.