SAARLAND UNIVERSITY

Faculty of Mathematics and Computer Science
Department of Computer Science
Master Thesis

Measuring the Effect of User Designed Gamification in an Image Annotation Setting

Dominic Buchheit

July 2018

Advisors
Pascal Lessel,
German Research Center for Artificial Intelligence (DFKI),
Saarbrücken, Germany

Maximilian Altmeyer,
German Research Center for Artificial Intelligence (DFKI),
Saarbrücken, Germany

Supervisor
Prof. Dr. Antonio Krüger,
German Research Center for Artificial Intelligence (DFKI),
Saarbrücken, Germany

Reviewers
Prof. Dr. Antonio Krüger,
German Research Center for Artificial Intelligence (DFKI),
Saarbrücken, Germany

Prof. Dr. Dietrich Klakow,
Lehrstuhl für Sprachsignalverarbeitung (LSV),
Saarbrücken, Germany

Submitted
July 30, 2018

Saarland University
Faculty of Mathematics and Computer Science
Department of Computer Science
Campus - Building E1.1
66123 Saarbrücken
Germany

Statement in Lieu of an Oath:

I hereby confirm that I have written this thesis on my own and that I have not used any other media or materials than the ones referred to in this thesis.

Saarbrücken, 30th of July, 2018

Declaration of Consent:

I agree to make both versions of my thesis (with a passing grade) accessible to the public by having them added to the library of the Computer Science Department.

Saarbrücken, 30th of July, 2018

Acknowledgments

At this point, I would like to thank the persons helping me with the realization of this thesis.

I sincerely thank my advisors Pascal Lessel and Maximilian Altmeyer for their incredible high commitment, invaluable feedback, vivid discussions and their continuous support. Furthermore, I thank Prof. Dr. Antonio Krüger and Prof. Dr. Dietrich Klakow for their feedback, advice, and suggestions, as well as reviewing this thesis.

Moreover, I thank my girlfriend Julia Grub and my dear friends Daniel Bruch and Peter Wita for helping me through the ups and downs of writing this thesis. Furthermore, I thank them and Monika Schutz for their indispensable help in proofreading. Many thanks go to Lea Schmeer for helping me in rating thousands of tags. To my friends and family I want to say thank you for supporting me through the entire process.

Lastly, I thank all study participants. Without their help, this thesis would not be possible.

Abstract

It has been shown that allowing users to gamify systems on their own during runtime is appreciated and has an impact on their behavior. Those bottom-up approaches allow users to become designers of their own game experience. We conducted an online experiment that systematically examined how self-designed games influence performance (tag quality and quantity) in an image annotation task without restricting users to a set of game elements. To achieve this, we implemented an appropriate image annotation framework allowing both users to textually describe game concepts as well as developers to quickly implement them. Compared to a control condition without gamification, users with self-designed games produced significantly more tags, while their tag quality decreases significantly as well. Those results provide further insight into bottom-up approaches. Furthermore, self-reports provide evidences that game elements did not significantly affect users' intrinsic motivation. Our data suggests that users are able to design game concepts motivating them in producing more tags.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Gamification describes the usage of game-design elements in non-game contexts [16]. Since its definition in 2011 by Deterding et al., it has been a trending topic (see Figure 1.1). According to Hamari et al. [25], gamification has been "*... subject to much hype as a means of supporting user engagement and enhancing positive patterns in service use, such as increasing user activity, social interaction, or quality and productivity of actions*" ([26], page 3025).

It was shown that approaches where each user experiences the same gamified elements ("one-size-fits-all") may have drawbacks [36]. As an example, employees of Disneyland perceived a gamified system introduced by management as "electronic whip" [1]. Subsequent research has tried to answer questions about whether gamification works at all (e.g., [26]) and whether there are differences in gamification effectiveness based on individual users' needs (e.g., [52]). Such needs are caused, for example, by people's different personalities and player types [17]. If users' individual needs, such as their autonomy (e.g., being able to change system components), are not considered properly, gamified experiences may lead to frustration and pressure [29]. Furthermore, users may feel forced to use the system, which may result in "mandatory fun" [48]. Systems which consider such individual needs and adjusting accordingly are called *personalized*.

In the proceedings of this thesis, we gave people the possibility to design gamified experiences on their own. Research has shown that such a bottom-up approach is perceived positively [41] and has positive behavioral effects [42]. The question what these positive effects can be attributed to, is still an open issue. Reasons could be, for example, that users design game concepts actually fitting their personal needs [27] or that they identify with their work [20]. A further explanation could be the choice given to create their own gamified setup, as shown in other contexts [39], or a combination of those aspects.

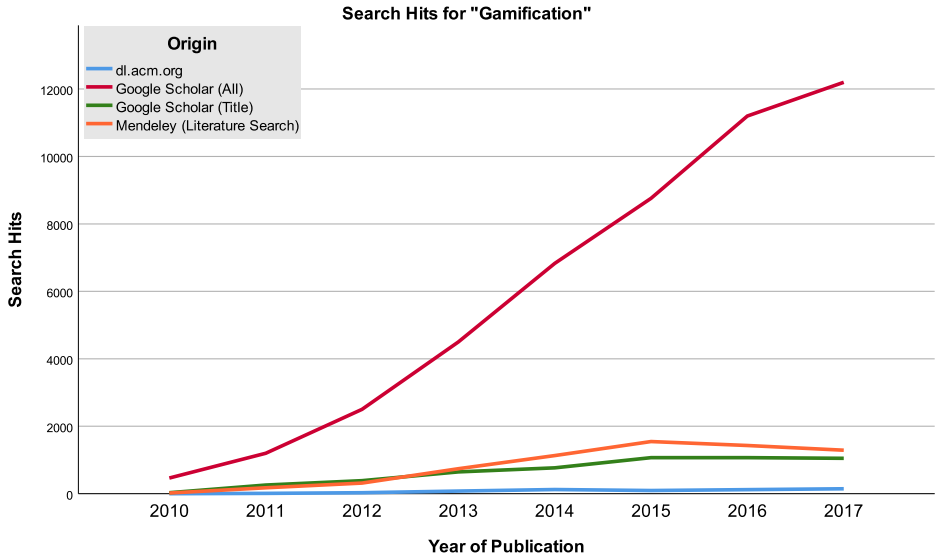

Figure 1.1: Search hits for "Gamification" on different platforms.

With personalization we would identify each user's individual needs. In contrast, by using this bottom-up [41] approach, we wanted to find out whether people are able to design game elements on their own motivating them in doing a certain task better. This is in accordance with participatory design [49]. Whereas in user-centered approaches a professional team focuses on user needs, this approach involves the user deeply and enables an active contribution to the design process.

The resulting game concepts were then implemented in a way that users can actually test their individual designs. Finally, we tried to find hints about whether such an approach has positive effects (i.e., *do people with self-designed games perform better*[1]*?*).

The following sections give more detailed information about the thesis' problem statement and goals.

1.1 Problem Statement

As mentioned above, gamified approaches can work [26]. But it was shown that "one-size-fits-all" approaches actually do not suit for every individual user. *Personalized gamification* tries to improve users' experiences by taking personal needs into account, for example, personality traits and player types. In contrast, *customized gamification* aims for the user's autonomy by

[1]Here, "better" depends on the actual game design setting.

providing control over the gamified interaction (e.g., giving the user the possibility to change game elements during runtime).

While using elements adjusted to the user's needs, personalized gamification provides the user with a fixed set of game elements that they need to rate or use (e.g., [17, 32, 63]). It was shown that systems need a broad range of different game elements to offer a choice for every participant [41]. In contrast, bottom-up approaches [41, 42] tried to overcome this need by allowing users to set up their own gamification setup. From a usability point of view, users still were restricted, since those gamification setups rely on pre-defined – but customizable – elements. In this work, we lifted those restrictions by letting our participants textually describe how their gamification would look like.

Lifting those restrictions, again, aims for the user's autonomy. Autonomy is one of the *three basic psychological needs* as identified by the Self Determination Theory (*SDT*) [59], a theory which is concerned "*... with supporting our natural or intrinsic tendencies to behave in effective and healthy ways*"[2]:

1. **Competence**: Seek to control the outcome and experience mastery.
2. **Relatedness**: Will to interact, be connected to, and experience caring for others.
3. **Autonomy**: Desire to be causal agents of one's own life and act in harmony with one's integrated self.

According to the SDT [59], those needs must be satisfied in order to provide optimal function for a (gamified) system by increasing the chance to develop *intrinsic motivation*. In contrast to *extrinsic motivation*, where the motivation raises from external stimuli as monetary rewards, intrinsic motivation arises from the *inside*. Hence, it is concerned with the motivation behind people's choices without external influence and interference. In this thesis we wanted to give people the autonomy to design their own game concepts. Previous work showed that people with no design or IT background may overlook important details as technological limits [21]. It will be interesting to see whether our participants were able to design game concepts which motivate them to perform better.

To measure whether participants actually perform better, we had to provide a setting that is easy to understand and use. Otherwise, participants may struggle with the study's framing instead of solving the actual task. We wanted to avoid such a situation since it could potentially bias the study's results. Furthermore, it should provide sufficient control to measure whether

[2] `https://selfdeterminationtheory.org`, last accessed on 29.07.2018

Figure 1.2: User interface of the ESP game [64]: Players try to "agree" on as many images as they can in 2.5 minutes. The thermometer at the bottom measures how many images partners have agreed on.

those gamified elements improve the respective user's overall performance. In general, each setting which provides us those options would be sufficient to test our approach. Studies of Mekler et al. [46, 47] show that image annotation (or *tagging*) is an appropriate setting where top-down gamification[3] was used with success. This allowed us to test our bottom-up approach. While tagging, the subject sees an image for a certain – possibly arbitrary large – amount of time and assigns appropriate words to it, where "appropriate" depends on the actual image annotation task. As an example, stock photograph platforms as shutterstock[4] and Adobe Stock[5] use tagged images allowing customers to easily find keyword-related photographs (e.g., "holiday" or "hiking"). Furthermore, the field of deep learning uses such images to train computer vision algorithms ([23], pages 95 – 119 and 440 – 445). The *ESP game* [64] (see Figure 1.2), or *Game With A Purpose* (GWAP), is a popular example where users have the task to annotate images in a gamified manner. The idea is that annotating images is an easy task for human intelligence, but still quite difficult for even the most advanced systems in image processing (e.g., [50]).

[3] E.g., managers, developers, or researchers decide about gamified elements.
[4] `https://www.shutterstock.com`, last accessed on 29.07.2018
[5] `https://stock.adobe.com`, last accessed on 29.07.2018

1.2 Thesis Goals

In contrast to personalized gamification, in this thesis, we let our participants describe their individual gamified experience in a bottom-up manner at design time. To the best of our knowledge, we are the first to follow such an approach where users can textually design and test their proposed game concept after later implementation. To achieve this goal, we needed to do both implement a flexible platform and conduct the actual study. Consequently, this thesis splits into two parts.

The first part (*building the environment*) was to build a platform which allows participants to annotate images while being assigned to different conditions (i.e., use pre-defined gamified approaches or design a game concept). Having tag quality and quantity allowed us to compare those conditions. To answer the question whether people are able to design game concepts which motivate themselves, the platform had to provide a dedicated condition that introduced the image annotation task and provided functions to textual describe such a concept.

In the second part (*conducting the study*), participants were asked to use the platform. Depending on their randomly assigned condition, they had either to annotate images or to create a game concept. Once the feasibility of each game design has been assessed (i.e., *is the game realizable in the image annotation setting?*), the feasible ones had to be implemented before each participant was allowed to try out the game they designed by themselves. A non-realizable concept in this context would, for example, change the setting (e.g., annotate songs instead of images), making it impossible to compare with the other concepts.

Implementing the platform and using it to carry out the study results in a data set where we could compare the mentioned conditions to each other. Besides "raw" performance measurements (tag quantity and quality), self-reported questionnaires were used to measure, for example, the user's individual intrinsic motivation and the platform's usability score.

In summary, the following goals had to be achieved:

Goal1 expects that image annotation is an adequate setting to measure people's performance increases (or decreases). This was shown in prior works as the studies by Mekler et al. [46, 47]. Therefore, we can assume that the setting is indeed adequate for us to test our approach.

By the help of the two control groups of **Goal2a**, we fulfill all preconditions to reproduce the results of Mekler et al. [47] and, hence, show the platforms validity. That people in general are able to design games (e.g., [21]) motivates to check the same in a gamified image annotation task where we are able to compare the realized self-designed game (**Goal2b**) to other (pre-defined)

Goal1	Build a platform providing functionality to *tag images* and *textually describe* game concepts.
Goal2a	Carry out a study with *no* and *pre-defined* gamified elements (*control groups*).
Goal2b	Carry out a study where participants are allowed to *create their own game concept* (*test group*).
Goal3	Implement the created concepts in the setting of the image annotation platform and let the respective participants use them to tag images.
Goal4	Evaluate both tag quality (*are the tags fitting?*) and tag quantity (*how many tags created?*) for each group.
Goal5	Carry out and evaluate questionnaires to receive additional measurement parameters.

conditions (**Goal2a**).

Implementing the user created game concepts (**Goal3**) allows us to compare the individual user performances over several conditions, i.e., compare **Goal2a** and **Goal2b**. Previous works show that people with no design or IT background may overlook important details (e.g., technological limits) [21]. This motivates us to check each game concept for feasibility, which, in addition, represents the basis for the respective concept's implementation.

In their previous works [46, 47], Mekler et al. used tag quality and quantity to measure the participants' individual performance. They showed that those are indeed appropriate values to measure, since they allow to see differences between the conditions, if exist. This motivates to evaluate and compare the tags of all groups as required by **Goal4**.

Assuming that participants in the self-designed game condition perform better, we still would not exactly know why this is the case. This motivates to carry out and evaluate appropriate questionnaires (**Goal5**) to get additional measurement parameters. Such parameters could be, for example, the participants' developed intrinsic motivation as well as their personality traits.

1.3 Outline

The thesis is structured as follows: First, we give an overview about related work in the gamification and image annotation task domain (Chapter 2). After this, we will introduce the image annotation framework (Chapter 3) we developed to conduct the study. Then, we move to the actual study (Chapter 4) where we show the used method and participants acquired. After presenting the study's results, we summarize and discuss them. Finally, we provide a conclusion and recommendations for future work (Chapter 5).

Chapter 2

Related Work

This chapter provides an overview of conceptually related works. Multiple studies and systems which consider the effectiveness of gamification as well as works which survey participatory designs are presented. To impose structure, these works are ordered by their main types: personalized and customized gamification, participatory design, and gamification in image annotation domain. At the end of this chapter, a comparison between the presented approaches and ours is made.

2.1 Gamification

Gamification, the usage of game design elements in non-gaming contexts, was introduced and defined 2011 by Deterding et al. [15, 16]. It was shown that "one-size-fits-all" approaches, which provide the same gamification elements to all users, have drawbacks [7, 52]. For example, employees of Disneyland perceived an introduced system using leaderboards as "electronic whip" [1] instead of motivation to work faster. Those drawbacks can arise because of different user-dependent factors as culture, age, and personality traits [36, 62] and manifest in decreased positive effects as dropped motivation or pressure to perform. Mollick and Rothbard [48] denote this phenomena as "Mandatory Fun".

In order to address those problems, several types of gamification emerged from the original definition. While a gamified system that is adjusted to the user's needs is called *personalized*, a system that can be adjusted by the user itself is called *customizable*.

2.1.1 Personalized Gamification

Studies have shown the benefits of personalized over generic approaches in domains like user interface design [3], persuasive technologies [34], and games [5, 53]. Extending this personalized approach to the field of gamification, design decisions and mechanics are still done by researches, managers, or developers, with the difference that systems now try to figure out the individual user's needs and adapt accordingly. Tondello et al. [62] observed that personalized gamification is still in its infancy and existing studies are mostly of theoretic nature. As an example, Oyibo et al. [56] found relationships between age and gender for the game elements reward, social competition, social learning and social comparison (e.g., males perceive competition better than females). Other studies revealed relationships between player types and game elements [22], between the Big Five personality traits [58] and game elements [32, 54], and between player types / personality and game elements [63]. Those relationships help to design systems which fit the target audience's preferences and to build recommender systems[6] [61]. Subsequent works as the studies of Orji et al. [52, 55] showed that tailoring is beneficial, and personalization is perceived positively.

Recently, Tondello and Nacke [61] published a work-in-progress paper where they aim at demonstrating the viability of designing gamified interactive systems according to user preferences. The goal is to find out whether, if participants are allowed to choose between gamified elements, they make choices corresponding to the theoretical relationships with user types, personality, gender and age, as reported in the studies above. Actually, this goes beyond personalization, introducing the concept of *customized gamification* that will be considered in the next section.

2.1.2 Customized Gamification

Customized gamification allows users to adjust the system for their needs by themselves. Similar to the personalized case, Orji et al. [52] found out that customized gamification is perceived positively. They also report that their participants see issues in using customized gamification, i.e., that it might be difficult, time consuming, or distracting. They claim that, while personalized systems increase factors as usefulness, ease of use, and relevance, customized systems improve trust, feeling of control, and freedom.

Lessel et al. [41, 42] measured the effect of bottom-up gamification, where users are allowed to decide at runtime *whether and how* they want to use gamification. Therefore, they used a micro-task setting, enhancing the process of optical character recognition (OCR). Figure 2.1 exemplary shows

[6] e.g., Spotify recommending songs you may like.

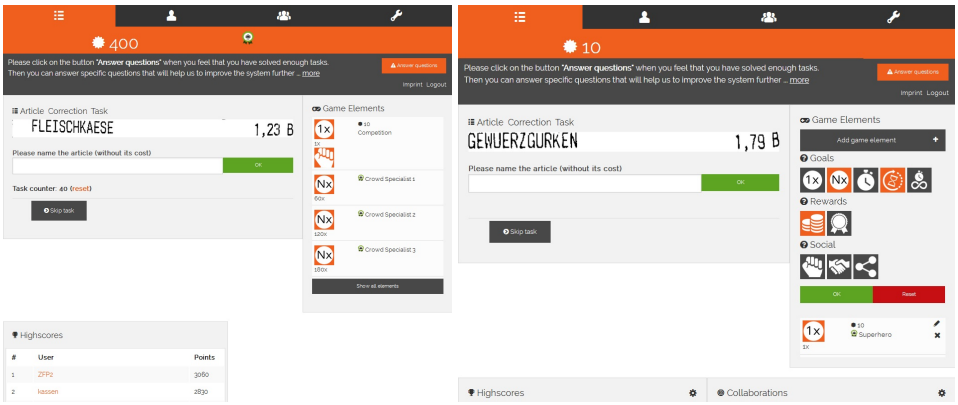

Figure 2.1: Expense Control user interface comparison. **Left:** Top-down interface showing the actual task on the left side and the pre-defined game elements on the right side and on the bottom. **Right:** Bottom-up interface. The task on the left side stays the same but the game elements on the right side are customizable [41].

the used top-down[7] and bottom-up user interfaces. In contrast to top-down, the bottom-up version allows users to define their own game configurations. Besides those two types, they also considered a *selective* version of both, where the user was allowed to select certain game configurations, and, in the case of selective bottom-up, even to edit them (without creating new elements/configurations). The authors report positive results in terms of significantly more solved micro-tasks when the participants had a choice and used it ($Avg = 82.2$, $SD = 87.0$, $Mdn = 50$), compared to those who had no choice ($Avg = 49.4$, $SD = 66.1$, $Mdn = 25$) or no gamification ($Avg = 37.2$, $SD = 40$, $Mdn = 25$) at all, even though more thought and effort needs to be invested by the users. But they also stated that further research is needed, especially to find out whether the actual choice or the self-selection of a fitting game configuration was the motivating factor for the participant.

Similar to the work of Lessel et al., we wanted to investigate the effect of bottom-up gamification. But instead of allowing the participants to select and edit game configurations during the runtime, we wanted our's to design game elements completely by themselves, with only being restricted to the setting but, i.e., not being restricted in adapting or using pre-defined components. It will be interesting to see whether this freedom of design also had an impact on the number and quality of tags the participants create.

In their study, Lee et al. [40] state that it is hard for experts (e.g., software designers and engineers) to account every user's unique goals and, therefore, also suggest to allow users to customize systems on their own. Such bottom-up approaches are in accordance with Nicholson [51], who states

[7]Users can not decide whether and how they want to use gamification.

that providing users with the option to customize games to their needs is important to achieve "meaningful" gamification setups. The Self Determination Theory [59] (SDT) gives additional support. According to it, the chance to develop intrinsic motivation is increased by the user's autonomy as, for example, being able to change components of the system. We wanted to achieve this by giving the participants the autonomy to design game elements on their own.

2.2 Participatory Design

A design approach that actively includes all stakeholders (e.g., users, customers, and developers) is called *Participatory Design* [49]. We wanted to include our study's participants as much as possible by considering them as designers and giving them the maximum freedom while designing their individual game elements. Therefore, works that cover those topics were part of our literature research. In the following, we start with works that consider users as designers, move then beyond user participation, and finally emphasize the role of stimuli in the process of game idea generation.

2.2.1 Users as Designers

Treating users as designers is relevant for participatory designs. As an example, Arroyo et al. [4] built a technology-based paradigm with the goal to support educational games by the help of embodied smart devices. As part of the study, students could create games as finite state machines (FSMs), resulting in a large variety of games. Such variety demonstrates the importance of user autonomy and led to our expectation in receiving a large variety of games as well. This also may be enforced due to the fact that we do not restrict our participants in using FSMs.

Gaye and Tanaka [20] described a case study where groups of teenagers ($15 - 18$ years old) changed place with designers, conceptualizing and implementing an interactive educational information pack. After the implementation of the prototypes, the research team conducted group-based interviews, with the result that the prototypes were received very positive by the teenagers. According to Gaye and Tanaka, one reason for this positive feedback is that they felt a strong sense of ownership over the project's outcome. It will be interesting to see whether our participants also felt a sense of ownership and, assuming such, if it had an impact on tag quantity and quality.

A further example where end-users took the place of designers is the workshop of Kanstrup [33]. In this study, 17 families with at least one member

having diabetes were asked to design IT services to support everyday living with the chronic illness. Kanstrup concluded that end-users (the families) are indeed able to derive ideas, even without an IT background, which is also important for us, since it suggests that users are generally able to derive game ideas.

An approach with end-users and game design experts was carried out by Gerling et al. [21]. Here, both groups designed wheelchair-based games and the authors report that both were able to design approaches providing valuable insights into the design of such games. But they also state that the non-experts miss important details as technical constraints or relationships between game mechanics. With regard to our approach, this study provided some insight into the results of such a non-expert design task and it will be interesting to see whether our participants were able to design game elements that are realizable, detached from the motivational aspect.

The main difference to the approaches above is that those were group efforts, while, in our case, individual participants had the opportunity to design their own game elements. This could again increase the effect of ownership and finally the overall effectiveness of the gamified elements. A further difference to the works of Kanstrup, Gaye and Tanaka is, that, while the design decisions still entirely remain on the end-user's side, we implemented the designs for them, allowing us to objectively measure their individual performance.

2.2.2 Moving Beyond User Participation

In their work, Wagner and Piccoli [66] considered different levels of participation in order to create successful systems. They claim that the difficulty in creating interface systems is the translation between the software designers and those who use the software after development. In our study, the participants had the opportunity to design their game elements on their own, without forcing them through development processes as creating requirements catalogs or participating in developer meetings. This is in accordance to the authors' suggestion: "*Rather than fighting human nature by trying to force the participation of user groups throughout a project, we should broaden our thinking about development and implementation methodologies to reflect what happens in practice*" ([66], page 51).

Overall, Wagner and Piccoli suggest that it is not enough to involve the user, but instead, timing and focus of user participation should be carefully considered. We hoped to address those points by letting the participants design their own game elements and move beyond "usual" user participation. As carried out in the last section, Gerling et al. reported that non-expert designers tend to miss important parts when designing games. At this point,

it will be interesting to see whether our "non-forced" approach suggested by Wagner and Piccoli resulted in realizable (i.e., are the designs feasible in the image annotation setting) and successful (i.e., does the implementation have an effect on the user's tag quantity and quality?) designs.

2.2.3 Stimuli in Game Idea Generation

Participants who design their own game elements were a crucial part of our study. To know what might restrict the participants' creativity, works about game idea generation were also part of our literature research. Kultima [38] presents in her paper findings of an ideation experiment considering the role of stimuli in game idea generation. In the study, participants had the task to brainstorm ideas as a pair, in order to find a connection between provided word stimuli and generated ideas. Only two of a total of 100 ideas were almost identical, even though some stimuli were adapted literally. On the other side, some atomistically ideas[8] had strong connections to the stimuli. Kultima concluded that, even if the final ideas diverged, the core ideas had a strong connection to the stimuli provided to the pair.

This work shows that stimuli can play a role when designing games or, as in our case, gamified elements. With this knowledge, we needed to make sure that we did not provide such stimuli on purpose or by accident (e.g., providing too detailed examples or leading questions), which could bias the study.

2.3 Image Annotation

To validate the effect of gamification, we needed a unit that is easy and objective to measure. For example, in an educational domain, such a unit could be the grade of an exam before and after the gamified intervention (e.g., [4]), while in a physical activity domain, steps taken (e.g., [43]) or stairs taken (e.g., [19, 35]) could be a good unit to measure.

The use of classification micro-tasks was already reported on previous studies of gamification in general [47], as well as customizable gamification [2, 42]. In fact, we used the same image annotation setting as described in the studies of Mekler et al. [46, 47], allowing us to compare our results with theirs. In the image annotation task, participants see an image and have the task to describe what they see, or in the case of Mekler et al. and our study, which moods cross their minds. This process is also known as *image tagging*. Figure 2.2 exemplary shows parts of the image annotation interface Mekler

[8] Game element ideas directly coming from the stimuli.

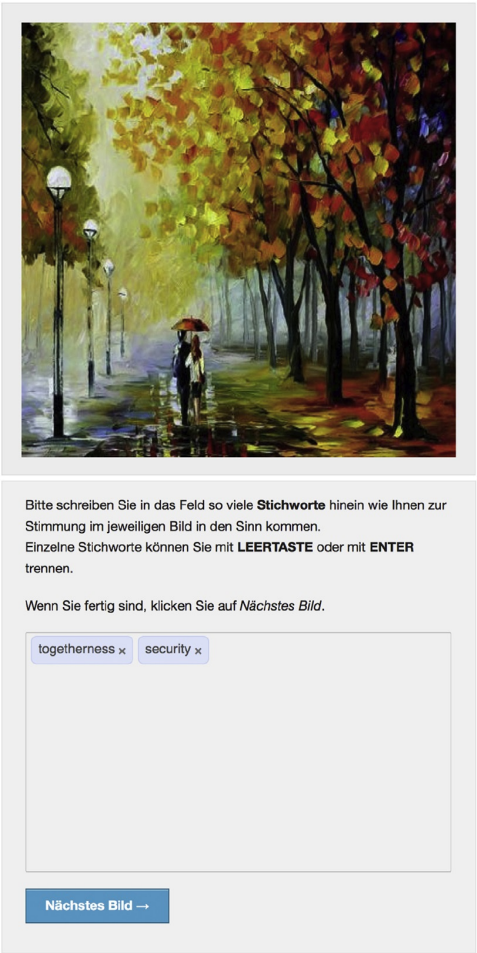

Figure 2.2: Image annotation view as used in [47]. **Left**: The image to describe. **Right**: A input where participants can enter the moods crossing their minds when seeing the image.

et al. used in their studies. Units to measure in such a setting could be the quantity and quality of tags created by the participants.

Mekler et al. conducted an online experiment that systematically examined how different gamified elements, as well as participants' goal causality orientation, influence intrinsic motivation, competence, and performance (*tag quality* and *quantity*). Participants sequentially saw 15 images and, while the image was visible, each participant had the task to think about moods crossing the mind when seeing the image. After 5 seconds, the image disappeared and the participant had to write down the moods. The gamified elements were visible during the image annotation task and depended on the condition participant's randomly assigned condition: points, leaderboard, levels, and plain (no gamified elements). Figure 2.3 exemplary shows the

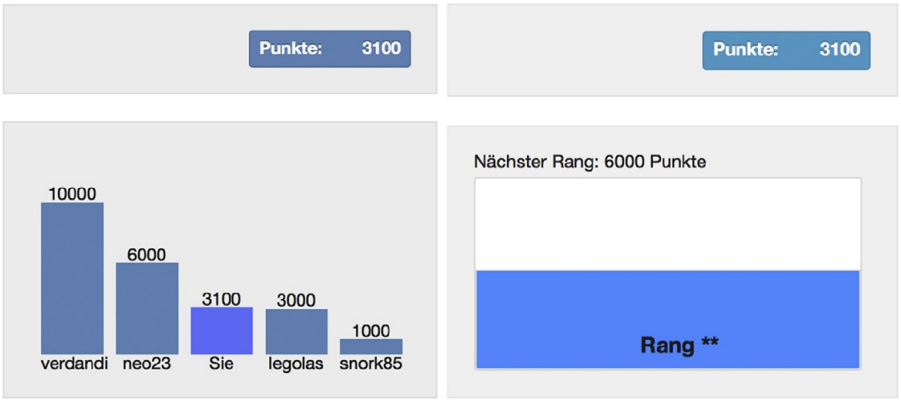

Figure 2.3: Left: The points and leaderboard condition. **Right**: The condition showing several levels the participant could reach. The respective levels correspond to the leaderboard entries in the leaderboard condition [47].

leaderboard and levels condition as used by Mekler et al. [46, 47].

The authors report that participants in the game element conditions generated significantly more tags than participants in the control condition ($F(3, 265) = 10.09$, $p < .001$, $\eta_p^2 = .10$). Furthermore, participants in the points condition significantly outperformed participants in the plain condition ($t(141) = 2.61$, $p = .01$, $d = .44$) and were in turn significantly outperformed by participants in the leaderboard ($t(138) = 2.3$,$p = .02$, $d = .39$) and levels ($t(138) = 2.03$, $p = .04$, $d = .35$) condition. Between leaderboard and level, they found no difference in mean. For us, this means that one of those two conditions (leaderboards and levels) suffices to compare with the results of Mekler et al. and we can expect a significant difference between a non gamified condition and one of those two conditions. The authors report that points, levels and leaderboards by themselves neither make nor break users' intrinsic motivation in non-game contexts. On the contrary, they assume they act as "progress indicator", guiding through the progress and improving user performance. In our study, it will be interesting to see whether the open design approach affected our participants' intrinsic motivation.

As mentioned above, besides participatory performance, there are further measurements one can take into account. *Intrinsic motivation* describes the motivation which arises from the inner (e.g., participating a contest because you find it enjoyable). In contrast, *extrinsic motivation* arises from external factors (e.g., participating a contest to win awards). Studies provide evidence that certain forms of rewards, feedback, and other external events can have detrimental effects on intrinsic motivation [12]. Furthermore, a recent study by Hanus and Fox [28] suggests that this may be also true in gamified environments. But in contrast to this study, multiple studies argue

that – providing a non-controlling setting – implementations of gamified elements may indeed satisfy people's innate psychological needs for autonomy, competence, and relatedness, and, hence, improve intrinsic motivation (e.g., [14, 18]). Mekler et al. measured the intrinsic motivation of their participants, using the Intrinsic Motivation Inventory (IMI, [31]), but found no significant main effect of game elements on intrinsic motivation. At this point, it will be interesting to see whether in our approach a significant effect could be found when participants design their own concepts.

2.4 Comparison to our Approach

We have seen that gamification as introduced and defined by Deterding et al. [15, 16] is indeed a wide research domain. In extension to the original definition, we presented studies about personalized and customized gamification, which overcome some of its drawbacks [7, 52] and allow a better gamified user experience [51, 52, 55]. Knowing that customization can work [41, 42], we used a bottom-up approach where the participants were able to choose which gamified elements they want, and even *if* they want such elements at all. This is in accordance to the works regarding participatory design we considered, showing that users treated as designers are able to design their own game elements, even without having such background knowledge (e.g., [33]). This approach goes beyond customization and we had to keep in mind that there are drawbacks (i.e., non-expert users missing important parts in development) [21]. It will be interesting to see whether our participants were able to create feasible game concepts that could motivate them in creating more tags and if they perceived the interaction as "mandatory fun" [48].

Wagner and Piccoli [66] suggest not trying to force the participation of users throughout a project, possibly leading them to be overwhelmed. While our "open" approach allowed the participant to freely decide which game elements to use, overpowering can still be an issue if. As an example, participants who never designed a game (or anything else) could be overwhelmed by the task. In line with this, and as suggested by Kultima [38], we had to be aware that certain stimuli can influence a participant's process of creativity. Taking this together, we had to design the study in a way that it is absolutely clear to the participant what the task is, without providing stimuli that possibly influence the concept creation. This can be done, for example, by properly introducing our participants to the image annotation task and avoiding influencing phrases (e.g. "you have" or "you should").

Finally, we presented some work which considered the task of image annotation. Mekler et al. reported that gamified interventions in such a setting are indeed beneficial, i.e., using game elements as points and leaderboards

can increase the number of tags the participants create during the image annotation (e.g., [46, 47]). A further advantage of such an approach is that user performance can readily be measured through the quantity and quality of tags. Intrinsic motivation is measurable by using (self-reported) questionnaires as the mentioned IMI. Those findings indicate that image annotation is a valid setting for us to test whether users are able to design game concepts which motivate them in creating more tags.

Chapter 3

Image Annotation Framework

In order to conduct the study, we needed an image annotation platform that allowed us to do both tagging images and designing gamified elements. This chapter describes the usage of such a platform in the context of our study and the requirements we had while implementing it. Furthermore, it gives details about the platform's structure.

3.1 Usage within the Study

Our framework is conceptually based on the platform used by Mekler et al. [46, 47], and similar to them we wanted to compare several conditions using different kinds of gamification. As seen in Section 2.3, Mekler et al. used four different conditions: *no gamification*, *points*, *points and leaderboard*, and *points and levels* (see also Figure 2.3). They showed that *points and leaderboard* and *points and levels* performed similary and both outperformed the *point* condition. While tagging, each participant's view was split into two parts as illustrated in Figure 3.1. One part shows the actual image annotation task while the other contains game elements as defined in the respective condition. Based on this, we opted the conditions *no gamification* and *points and leaderbord* as baseline for our test condition and to be able to compare the results of Mekler et al. with ours. By means of a test condition we wanted to answer this thesis' main question, i.e., whether people are able to design game concepts which motivate them to create more tags in an image annotation setting. In the following, each mode the platform provides will be explained in more detail.

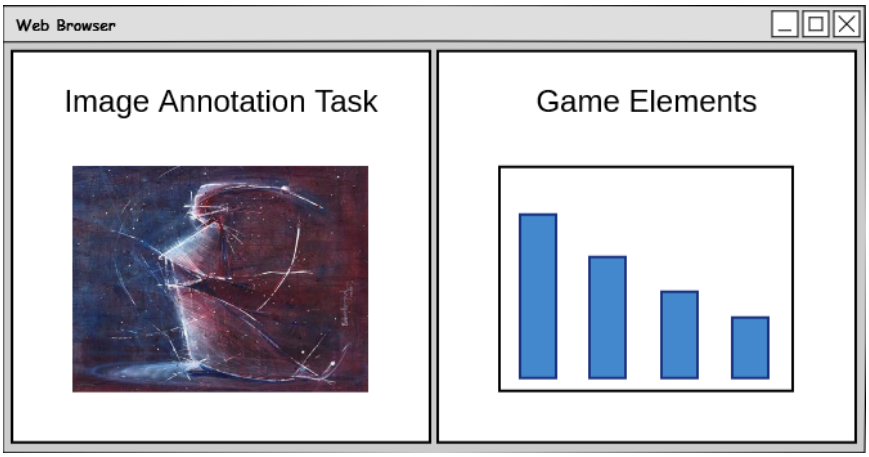

Figure 3.1: Schematic view of the image annotation platform while tagging. The left-hand side stays the same for each condition, i.e., it shows the image to tag and provides input fields for this purpose. The right-hand side contains the game elements depending on the participant's condition and stays empty in the *no gamification* condition.

3.1.1 No Gamification

In the *no gamification* condition, we did not provide any gamified elements to the participants. In other words, the right-hand side of Figure 3.1 was left empty and participants received no feedback about the tags entered.

3.1.2 Top-Down Gamification

In contrast to *no gamification*, the *top-down gamification* condition provided the game elements *points* and *leaderboard* in the right-hand side of Figure 3.1. Specifically, for each tag, participants received 100 points that were accounted to a total score. This score, again, decides on the participant's position in a leaderboard. Figure 3.2 shows the platform as seen by a participant in this condition. To reach the lowest position on the leaderboard, a participant had to create at least 10 tags, or, in other words, earn 1000 points. Subsequent positions were reached after a score of 2900, 6000, and 10200. These step sizes were adapted from Mekler et al. [47] and chosen to allow participants to reach a reasonably high position on the leaderboard while, at the same time, it was expected to be still reasonably challenging to create more than 100 tags.

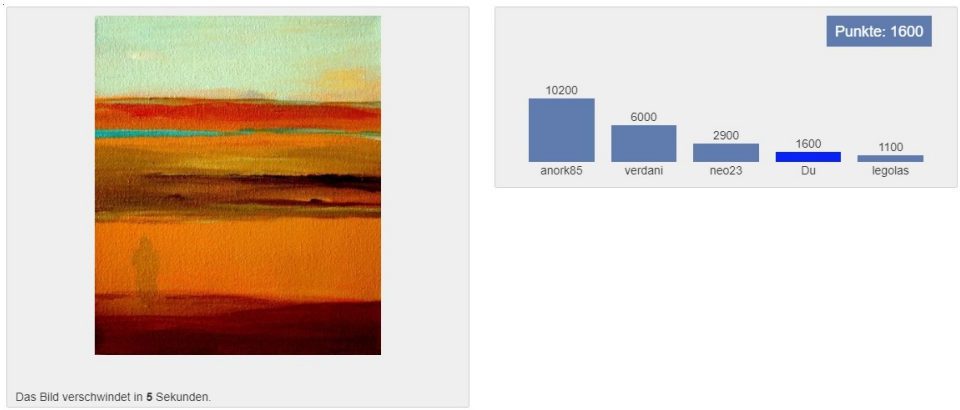

Figure 3.2: The image annotation platform in the top-down condition. The image on the left side is visible for five seconds and is then replaced by a field the participant can use to enter tags. The right side shows the participant's current points (100 points per tag) and the position in a leaderboard.

3.1.3 Design Task

In the *design task* condition, participants were asked to design their own gamified elements in the image annotation setting. These designs then had to be implemented so that the right-hand side of the view (see Figure 3.1) contains the participant-designed game elements. After this, the participants were asked to annotate images using their designed games. Hence, the design task required to interact with the platform twice; first when designing their game, and second when actually tagging images.

3.2 Requirements

It was shown that the platform developed and used by Mekler et al. [46, 47] provides functionality to compare different gamified approaches. In addition to their platform, we needed a further condition, representing our design task group. Furthermore, implementing concepts submitted by participants of this group required more flexibility in creating, adding, and adapting game elements than their platform provides. Hence, we decided to build such a platform from scratch on our own, extended by a framework which provides us with such flexibility. From now on, "platform" describes the actual image annotation platform, while "framework" denotes a scaffold around the platform that allows us to quickly add and adapt game elements. The following sections explain the framework's requirements we addressed during its implementation.

Figure 3.3: Tutorial overlay introducing the game elements *points* and *leaderboards* to the control groups' *top-down* condition.

3.2.1 Basic Functionality

In Section 3.1, we identified our study's conditions. Therefore, at a bare minimum, the platform provides functionality to *view* and *annotate* images, as well as to *display gamified elements* and *submit* participant-designed ones. Similar to [46, 47], the image annotation task consists of a specified number of images that are visible for a certain time[9]. After disappearing, an image did not become visible anymore and reveals an input field for entering tags. This is supposed to prevent participants from viewing the image for more than the specified time, which could possibly interfere the *spontaneous* moods the image caused. After a certain amount of time, this restriction was removed, so that participants are allowed to see the current image again and can continue with tagging images.

To introduce each participant to the process of image tagging and their game elements (if present), they have to complete a *tutorial* (see Figure 3.3) before tagging images or designing a concept. Each tutorial consists of three example images and explains both the image annotation platform and the participant's game elements. For the design task, the same tutorial as for participants with no gamification is used. After implementing their concepts and right before the actual image annotation task, a second tutorial which consisted of a single image introduces their self-designed game elements to them. For each tutorial image, the platform suggests three example moods which fit to that image to help participants in finding a feeling of what is meant by "mood".

[9] In the case of [46, 47], there were 15 images for 5 seconds each.

3.2.2 Game Concept Implementation

The framework's most important feature is its ability to provide methods to *implement game concepts* in a fast manner. It is designed in such a way that, when implementing user game concepts, we just have to focus on the game elements: While the framework provides all needed basic functions (i.e., displaying and tagging images), it only takes to implement the game elements (i.e., the right-hand side of Figure 3.1) in order to realize a user's concept. Section 3.4 describes this process in more detail on a technical level.

In order to quickly implement game concepts, it would be beneficial to re-use game elements appearing in multiple concepts. This has the advantage that a participant designing a view using the game elements *points*, *leaderboard*, and *achievements*, can be assigned to "groups" representing the respective game elements: points group, leaderboard group, and achievement group. Assuming the views for points, leaderboard, and achievements already exist, implementing such a design by re-using existing elements would be less expensive than building it from scratch.

3.2.3 Additional Functionality

In addition to the basic functionality and the ability to quickly implement game concepts, we had further requirements.

First, the framework *prevents multiple participations* to avoid participants being assigned to different conditions after finishing their respective attendance. For example, a participant who already annotated images in the top-down condition and then had to create a game design, might have been biased by the potential stimuli provided from points and leaderboard (see also Section 2.2.3). Hence, the framework is capable of storing and retrieving user sessions to allow participants to continue after having been interrupted.

Second, in addition to the main task where each participant has to tag a certain amount of *basic images*, the platform provides *extra rounds* consisting of additional images to annotate. Those extra rounds behave like the basic round (i.e., provide the same game elements), but are voluntary and may be used as a further variable to measure.

Finally, a built-in *questionnaire environment* allows operators to ask questions directly on the platform, i.e., without the need to forward participants to a dedicated survey platform.

3.3 Structure

The framework is split into two independent parts. The part that is visible to the participants (*front-end*) is web-based and defines the actual image tagging platform. In order to handle and store the participants' input (e.g., tags and questionnaires), the front-end is able to communicate with the *back-end*. This software "knows" which images have already been tagged by the participant and which questionnaires have been answered. Therefore, this part of the framework can be considered as the point where the data collection happens.

Both front-end and back-end software are explained in more detail in the following sections. We published the framework under MIT license[10] on GitHub[11]. Besides the code, this is where you find information about installing and using the software, as well as additional documentation.

3.3.1 Back-End Software

The back-end (or *API*) is entirely written in Java and consists of four parts: client interface, message interpreter, session manager, and data storage. Each of them is explained in the following sections.

3.3.1.1 Client Interface

The *client interface* forms the point of intersection between front-end and back-end, based on WebSockets. The WebSocket protocol[12] is a TCP-based networking protocol which supports bidirectional connections, invented to easily connect web-based applications with WebSocket servers. API administrators can choose whether the connections are encrypted or not. It is worth to mention that some browsers (e.g., Firefox 59 and Internet Explorer 10) may not permit to use encrypted WebSocket connections in some situations. In such cases, we recommend to encrypt the platform using a HTTPS proxy which forwards each message to the (unencrypted) API.

3.3.1.2 Message Interpreter

Each message that arrives at the client interface is forwarded to the *message interpreter*. First, the message interpreter checks whether the incoming message is valid and, if not, discards it or sends an error back to the client.

[10] `https://en.wikipedia.org/wiki/MIT_License`, last accessed on 29.07.2018.
[11] `https://github.com/dextreem/GIAP`, last accessed on 29.07.2018
[12] `https://tools.ietf.org/html/rfc6455`, last accessed on 29.07.2018.

This depends on the actual content of the message. In the version of the framework we used for the study, all messages sent from and to the server are in the well known JSON[13] format. Each message has a type, defining the content and purpose of it and its interpretation. All types are explained on a GitHub wiki page[14].

3.3.1.3 Session Manager

The *session manager* has two main tasks. On the one hand, it ensures that only eligible users can access the platform and, hence, the underlying data storage. On the other hand, it holds and handles user sessions and connection information, which allows the API to send messages to the client. A further important feature is that the session manager handles the case where users try to log in multiple times, preventing them from "cheating" (e.g., to create the same tags for the same image twice).

3.3.1.4 Data Storage

The *data storage* is used to store both image tags and questionnaire answers. Furthermore, administrators can define custom commands, allowing to create new and edit existing procedures without the need of recompiling the entire API. The latter is needed to implement self-designed game concepts without restarting the API, which would log out every participant who is currently actively on the platform. In the current version, the data storage uses a MySQL database to store all data. It is implemented as a Java interface so that it is easily extensible to new storage methods as other databases or even the bare file system.

3.3.1.5 General Design Decisions

In this section, we present general design decisions we made during the back-end's implementation. First, we decided to use *Java* as programming language. We made this decision because it is easy to get it running on multiple operating systems and there is a wide choice of WebSocket implementations we can use for the client interface.

The second design decision was to build the API as a "closed-in-itself" application. This has the big advantage that the API can be executed on every system that can establish a connection to the data store, i.e., it does not rely on the system where the front-end's web server is hosted. In

[13] `https://www.json.org`, last accessed on 29.07.2018.
[14] `https://github.com/dextreem/GIAP/wiki`, last accessed on 29.07.2018.

consequence, running the back-end and the front-end on different systems is supported by construction. This is especially interesting in a use case where the framework operators want the API (and hence the data store) in a controlled and secure environment, whereas the front-end can be hosted basically anywhere.

As already mentioned, we did not want to rely on web servers to provide the API. For the client interface, this had the consequence that we did not use PHP, since this would have implied the need of a web server. Therefore, we decided to use WebSockets, since they are widely implemented in the majority of all programming languages and easier to handle than, for example, bare sockets.

Those design decisions allow the API to be used in a very generic scenario, i.e., it is not restricted to the image annotation task or any web-based platform at all. As an example, it can be used to build mobile applications for image tagging with no changes and for any other purpose where it is necessary to properly distinguish between users (or participants) with little effort compared to developing it from scratch.

3.3.2 Front-End Software

The front-end describes the part of the image annotation platform which is visible for the participants. It uses the back-end to obtain information about the respective participant (e.g., which images are already tagged) and to store input made (e.g., the tags). Besides being the base of this study, this front-end serves as a reference implementation which shows how to make use of the back-end.

3.3.2.1 Approaching the Platform

Upon entering (i.e., when logging in), the platform retrieves the user's group assignment and initializes the respective game elements. This is beneficial in two ways: First, it ensures that the participant only sees game elements as defined in the condition, and, second, the front-end only needs to initialize required game elements (i.e., it does not waste computational resources). As a further consequence, it "forces" the framework operator to build views and game elements in a modular fashion.

3.3.2.2 Distinction of Concept Implementations

It is of high importance that the concept implementations run isolated from each other. In other words, game elements of concept A must not be visible

in the implementation of concept *B*. Each implementation is split into two parts, the *view* and the *logic*. Views consist of at least a *basic view* and *tutorial texts*. The basic view describes how the participants see their game elements (i.e., the right side of Figure 3.1), while tutorial texts are shown during the tutorial and provide information for participants on how to use their implementation.

The logic behind each view is always a single JavaScript. This file provides generic functions as rendering its dedicated view, but also functions to manipulate parts of it (e.g., increasing points or showing achievements). See Section 3.4 for more details on how to implement and use the logic file.

While logging in, each participant is assigned to a unique group identifier that is then used to load only the logic needed for the implementation. This ensures that no other logic is loaded and, in consequence, no unwanted game elements are visible.

3.3.2.3 Survey Engine

The framework has a built-in survey engine, based on the open-source JavaScript library survey.js[15]. The platform is capable of displaying both questionnaires stored on the web server (i.e., in a linked JavaScript file) and questionnaires stored in the API's data store. Furthermore, each questionnaire can only be displayed as long as it is not answered. This ensures that no questionnaire will be answered twice by the same user. Besides the "simple" display of questions, it is easy to configure which questions are required to be answered in order to continue, and which messages are to be shown when a question is not answered. Finished questionnaires are stored in the data storage by being sent to the API.

3.3.2.4 General Design Decisions

As mentioned in the back-end software section (see Section 3.3.1), there would be the option to use *PHP* as an alternative to the API. But since we decided to build and use the API using WebSockets, there is, by design, no need to use PHP in the front-end.

The front-end is web-based and written using HTML and JavaScript with some extensions such as jQuery[16] and Mustache[17]. Using Mustache, it is possible to separate the logic of an HTML document from the actual view, as illustrated in Figure 3.4. We use the features of Mustache when creating the

[15] `https://surveyjs.io`, last accessed on 29.07.2018
[16] `https://jquery.com`, last accessed on 29.07.2018.
[17] `https://mustache.github.io`, last accessed on 29.07.2018.

Listing 3.1: High-level HTML pattern using mustache syntax

```
<h1>{{header}}</h1>
{{#bug}}
{{/bug}}

{{#items}}
  {{#first}}
    <li><strong>{{name}}</strong></li>
  {{/first}}
  {{#link}}
    <li><a href="{{url}}">{{name}}</a></li>
  {{/link}}
{{/items}}

{{#empty}}
  <p>The list is empty.</p>
{{/empty}}
```

Listing 3.2: Data to render into the pattern

```
{
  "header": "Colors",
  "items": [
    {"name": "red", "first": true, "url": "#Red"},
    {"name": "green", "link": true, "url": "#Green"},
    {"name": "blue", "link": true, "url": "#Blue"}
  ],
  "empty": false
}
```

Listing 3.3: Rendered output, combining pattern and data

```
<h1>Colors</h1>
<li><strong>red</strong></li>
<li><a href="#Green">green</a></li>
<li><a href="#Blue">blue</a></li>
```

Figure 3.4: Mustache allows to render data into a high-level defined pattern. In this example *header* is a simple variable and is replaced in the pattern. Since *bug* is not present in the data and *empty* is set to *false*, both are not visible in the rendered output. The *items* array is automatically displayed as a list without the need of manually iterating over all items (Example adapted from `https://mustache.github.io/#demo`, last accessed on 29.07.2018.)

participants' individual game elements, but also in the basic versions (no gamification and top-down gamification), when, for example, displaying information as the participant's score.

Each game concept's implementation is encapsulated in a JavaScript object (see Figure 3.5). Using this approach, it is possible to retrieve a *class-instance*-like object containing all needed elements, wrapped in a single JavaScript object. As an alternative, we could use JavaScript classes which provide the same functionality but, since they are rather new, even browsers such as the latest versions of *Internet Explorer* and *Opera*, do not support classes at the this thesis is being written[18]. For that reason we decided to use the approach as described above.

Angular.js[19] is an open-source front-end web application framework based on JavaScript and developed by Google. It provides features to automatically update the view without the need of manual DOM manipulation. In Angular.js, view and logic are designed to be isolated from each other, making it suitable for our purposes. Nevertheless, we renounced to use it, since the features of Mustache are sufficient for our needs and using Angular.js would add a complexity level that is not necessary for our purposes.

Instead of JavaScript we could have make use of *TypeScript*[20], a strict syntactical superset of JavaScript developed and maintained by Microsoft. However, we decided not to use TypeScript for the same reason we did not use Angular.js: We only needed a small part of the features and TypeScript would add a further complexity level.

3.4 Adding a Condition

Adding a condition to, for example, *implement a user-designed concept*, is straight forward and boils down to five steps:

1. Register the condition in the API.
2. Build the view.
3. Build the logic script.
4. Build the tutorial dialogs.
5. Assemble all prior steps.

The following sections explain each of those steps (i.e., describe the steps we had to take for each concept during implementation) by implementing a simple example concept:

[18] `https://goo.gl/Dj6XgJ`, last accessed on 29.07.2018.
[19] `https://angularjs.org`, last accessed on 29.07.2018.
[20] `https://www.typescriptlang.org`, last accessed on 29.075.2018.

Listing 3.4: Logic File in JavaScript

```
LOGIC = function () {
  var POINTS = 0, // A field holding the participants points

  setView = function () {
    // Render the view
  },

  addNewTag = function(tag){
    // Handle a new tag, e.g. check validity
  },

  updatePoints = function(new_points){
    // Update points in the object and on the view
  },

  privateMethod = function(){
    // This method is not accessible using the object
    // but only in the object's scope.
  };

  // Return all the method that are accessible
  return{
    setView: setView,
    addNewTag: addNewTag,
    updatePoints: updatePoints
  }
}

// Example usage
var logic = LOGIC();
logic.setView();
logic.addNewTag('example');
logic.updatePoints(123);

// Create a second (independent) object
var logic2 = LOGIC();
```

Figure 3.5: A minimal exemplary logic file providing functions to show the view, handle a new tag, and update the participant's points. Basically, all functions are wrapped in a JavaScript function that is returned, allowing "public" and "private" functions.

> My concept looks like this: For every tag I enter, I'd like to receive 100 points to my score. My point score should always be visible and update automatically after entering a new tag. To compare with other players I want to see a leaderboard that shows how I perform compared to others. Therefore, there must be a feature allowing me to enter my nick name others can see in the leaderboard.

This concept describes the top-down condition. In the following sections, bold phrases describe what to do to achieve the respective goal.

3.4.1 Condition Registering

Registering a new condition takes to **creating a new user group** and **assigning the appropriate participant's user** to this group. Since the client receives group information while logging in, this can be used to load the respective views. Within the scope of our implementations, we defined a generic group that allowed the user to query the database with pre-defined commands (e.g., query tags of other users) and a unique group for each concept. The latter is solely used to identify the view and, hence, allow the front-end to load the respective view and logic.

3.4.2 Building the View

The view describes the right side of the image annotation window (see Figure 3.2). In the scope of the framework, each view is a **Mustache enhanced HTML file**. It is loaded freshly for each image. In the example, the view has to contain two elements; one to display points the participant has gathered, and one to display the leaderboard.

3.4.3 Building the Logic

The logic has to take care of how to **display** and **manipulate the view**. Basically, a logic file is a JavaScript file looking as shown in Figure 3.5. It has to provide methods to show the tutorial view, the tagging view, and the end view, i.e., it is expected to provide these three functions *setTutorialView()*, *setView()*, and *setEnd()*. In most cases, all of those three functions refer to an internal function that renders the view with different content, but in general, they are independent of each other. Besides those basic functions, the logic has to provide functions for updating parts of the view.

In the example, we can make use of some of the three basic functions. While *setEnd()* is not needed (since the end view is identical to the tagging view), *setTutorialView()* could be used to show some sample content (i.e., points and leaderboard) that is not needed in *setView()*. Furthermore, we need to

write functions which update the participant's points and the leaderboard as soon as a tag was entered or removed.

3.4.4 Building Tutorial Dialogs

After implementing both the view and the logic it is important to build dialogs which **explain the elements** to the participant. They are Mustache enhanced HTML files. The example concept is quite simple and requires a single tutorial dialog explaining the score system and the leaderboard (e.g., see Figure 3.3).

3.4.5 Assembling

Finally, all prior steps have to be assembled by **registering the logic's functions** to the so-called *game element handler*. While approaching the platform, the respective handlers are loaded, depending on the participant's individual user groups. Those handlers are triggered at several points in the platform. After the user has seen an image for five seconds, the image disappears and actions as, for example, starting a timer, can be executed. A list of important events follows along with when they are triggered and for what they are used:

- **TaggingEnvironment**: Triggered when the platform renders the tagging environment, used to render each participant's game elements.
- **AfterGetImage**: Triggered after the front-end successfully received an image from the API, used to query additional image information as, for example, frequently used tags of other participants.
- **AfterFlipImage**: Triggered after an image got flipped, used to, for example, make certain elements visible only after the image disappeared or to start timers.
- **AfterTagAdded**: Triggered after the participant added a tag to the tagging field, used to further process the respective word.
- **AfterTagRemoved**: Triggered after the participant removed a tag from the tagging field, used to further process the respective word.

In the example above, we have to define a **TaggingEnvironment** handler which renders the game elements (points and leaderboard) to the platform's view. Furthermore, **AfterTagAdded** and **AfterTagRemoved** have to be defined in order to increase or decrease the participants score when they enter or remove tags. No further handlers are needed in this case. When these steps are concluded, the new condition is ready to be tested.

3.5 Feasibility Test

After having implemented the framework and before we started the actual study, we performed a feasibility test to check the integrity, usability and functionality of the design task mode. We also wanted to gain insights on what type of ideas and designs potential participants may have submitted and how fast we would be able to implement such designs by using the framework.

We casted four participants (gender: 3x female, 1x male; age: 2x $18-24$, 1x $25-31$, 1x $46-52$) with German nationality from family and friends and manually assigned them to the design task. Each of them was introduced to the image annotation task using a single image from the same set as the other images and no gamified elements (i.e., the right side of the view remains empty). Subsequently, they had to textually describe a game concept using at least 700 characters. After two days, each participant had submitted an individual concept which allowed us to start with the implementation. Since our main goal was to test the framework and not the participants' performance, we implemented all of the concepts as far as possible, inspite of potential lacks in applicability. Before sending the games to the participants, we reviewed each game for quality assurance reasons. Overall, it took two weeks from the time the participants submitted their designs until they received their games. We carried out semi-structured interviews with each participant after they designed their game concepts and after they tagged images using their individual implementations. In the interviews, we asked for technical issues (e.g., "*Have there been any crashes while designing or tagging?*"), as well as issues with the design task itself (e.g., "*Did you know what to do at all time?*").

While tagging, the participants created 95.00 tags on average ($SD = 20.77$, $Min = 69$, $Max = 118$). Besides typos on the platform, we extracted the following key findings: A single tutorial image was perceived as insufficient by all participants. This led us to use three tutorial images for each condition. Furthermore, three participants reported that it was not obvious when the study was finished, i.e., when they could close the browser window. This was fixed by introducing a final page containing a small textual debriefing. Finally, every participant perceived the 700 character limit as too much. Keeping that in mind, after discussion, we did not reduce the number of characters in order to receive properly described game designs. No further changes were needed.

Chapter 4

Study Evaluation

We conducted a study to investigate the effects of bottom-up gamification compared to top-down gamification. We used the same image annotation setting as described and used in [47]. Mekler et al. already analyzed the effects of top-down gamification on people's performance (i.e., tag quantity and quality) and showed that the number of created tags is higher in gamified settings than in a baseline without gamification. We defined four conditions: one condition was confronted with no gamified elements (see Section 3.1.1, in the following abbreviated with **none**), one with top-down gamification (see Section 3.1.2, in the following abbreviated with **TD**), one could create their own concept and received its implementation (see Section 3.1.3, in the following abbreviated with **own**), and the last condition could create their own concept but received a different implementation (in the following abbreviated with **other**). For concept creation, participants were asked to textually describe game elements that would motivate them to create more tags. **None** and **TD** were taken from [47] and formed our control conditions, **own** and **other** were our test conditions.

The following sections describe the hypotheses, the study's method, as well as its results, a discussion, and limitations.

4.1 Hypotheses

The following describes our hypotheses we wanted to investigate within the image annotation setting.

$H_{Quant,1}$	**TD**, **own**, and **other** (i.e., with gamification) will create significantly more tags than **none** (i.e., without gamification).
$H_{Quant,2}$	**Own** will create significantly more tags than **TD** and **other**.
$H_{Quant,3}$	**TD** will create significantly more tags than **other**.

The first batch aimed for *tag quantity*, which was our main measurement. Points and leaderboards are two commonly used game elements providing users with performance feedback (e.g., [57, 68]). Previous research has shown their effectiveness for promoting certain behavior by forming a clear connection between user action and performance (e.g., [9, 25]), as well as their positive effect on user performance in an image annotation setting [46, 47]. Hence, our first hypothesis ($H_{Quant,1}$) stated that participants with gamified interactions create significantly more tags than those with no gamification. $H_{Quant,2}$ stated that participants with the autonomy (see Section 2.1.2 and Section 2.2.1) to design and receive their own game elements create significantly more tags than those with pre-defined gamification or those who receive a game different from what they designed. Reasons could be, for example, that users design game concepts that fit their personal needs or that they associate with their work [20]. $H_{Quant,3}$ stated that receiving a game different from what was designed has a reduced positive effect: Participants with pre-defined gamification will create significantly more tags than those who designed a game but received a different implementation. Although there was no design task, Mollick and Rothbard [48] reported such an effect.

$H_{Quali,1}$	Tag quality of **none** will be significantly higher than of **TD** and **other**.
$H_{Quali,2}$	Tag quality of **own** will be significantly higher than of **none**, **TD**, and **other**.

In accordance with the results of Mekler et al. [47], we assumed that there would not be any difference in tag quality between participants with gamification and those with no gamification. $H_{Quali,1}$ followed from the assumption that participants with no gamification would enter only tags which they consider as "fitting", while participants in other conditions may enter non-fitting tags to, for example, receive more points. According to the SDT [59], the increased autonomy of participants designing their own games

would lead to an increased intrinsic motivation and, hence, to an improved tag quality, leading to $\mathbf{H_{Quali,2}}$.

$\mathbf{H_{Auton}}$	**Own** will significantly increase autonomy satisfaction, compared to **none**, **TD**, and **other**.

Based on the results of Mekler et al. [47], we assumed that our conditions would have no significant effect on competence need satisfaction and interest / enjoyment, respectively. While they did not find any correlation between autonomy need satisfaction and their conditions, we assumed that – according to the SDT [59] – giving participants the autonomy to design a game on their own would have significant positive effects, leading to $\mathbf{H_{Auton}}$.

4.2 Method

We created a German online platform using the image annotation framework (see Chapter 3) to compare four conditions by means of a between-subject study design. While people in the control conditions (i.e., **none** and **TD**) were able to finish the study in a single session, while people in the test conditions (i.e., **own** and **other**) had to participate twice: First, to design individual gamified elements, and second to actual tag images by using their respective games (after a 6- to 44-day-long implementation delay). While tagging, each participant had the task to create as many tags as possible describing "moods" crossing their minds when viewing a certain image.

On the platform's start page we briefed participants with the same cover story as used in [47]: Participants annotate images which helps us in improving affective image classification. As participants were unpaid, we emphasized that the study was set in the context of a master thesis to provide an additional incentive. This was supposed to lead to a general positive framing effect as shown in [45]. The platform's link was distributed via Facebook survey groups[21], a subreddit[22], SurveyCircle[23], a student mailing list (consisting of computer science and psychology students), and the authors' social networks and working places. Along with the link we sent a short explanation that participation will help the master thesis and improve affective image classification; no hint towards a motivational study was given. Participation was not rewarded to rule out potential extrinsic motivation as confounding variable. In the following sections we first describe the process

[21] E.g., `https://goo.gl/nbCCVe`, last accessed on 29.07.2018.
[22] `https://www.reddit.com/r/SampleSize`, last accessed on 29.07.2018.
[23] `https://www.surveycircle.com`, last accessed on 29.07.2018.

each participant had to go through (see Section 4.2.1) and we then move to additional steps that the participants in the design task had to undertake (see Section 4.2.2). In addition, Figure 4.1 illustrates the method as explained in the following sections by showing a flow chart.

4.2.1 Image Tagging Process

After the participants confirmed that they had read the information on their data's privacy, the framework was used to distribute them to one of our four conditions. Since we had assumed that we would receive concepts that could not be implemented within the scope of this study, we set the probability of being assigned to the design task (i.e., to **own** and **other**) at twice as high as to be assigned to **none** or **TD**. Demographic questions about the participant's age, gender and nationality and the Big Five Inventory followed. After that, a tutorial (same three images for everybody) introduced the image annotation task and the game elements, if available. Additionally, we provided three exemplary tags per image and "forced" each participant to create at least one tag per image, to make sure they had tried out the process at least once. The tutorial showed the procedure as it was used later while tagging study-relevant images.

The actual image annotation task was the same for each participant and differed only as to which game elements were visible. Participants were shown 15 abstract paintings in random order, taken from Machajdik and Hanbury's study [44] on affective image classification and used by Mekler et al. [46, 47] as well. Each image was visible for five seconds before it disappeared and revealed the area where participants were to enter their tags. After 15 images were tagged, an *end screen* was shown to every participant. On this screen, we thanked them for tagging images and asked them to answer a final questionnaire needed to finish their participation. While the game elements were still visible, participants of **TD**, for example, were allowed to enter a nickname on the leaderboard. In addition to previous studies [46, 47], we provided every participant with the option to annotate *extra* images in the form of extra rounds consisting of five images each. This allowed them to create more tags and, in consequence, continue to play (i.e., earn more points or unlock more elements) and was intended to emphasize their "free-choice", which can have positive effects on intrinsic motivation [13, 39]. For us, the number of absolved *extra* rounds served as an additional value to measure their motivation in solving annotation tasks. In total, we provided images to absolve 20 *extra* rounds (i.e., 100 images), taken from the same set as the 15 *basic* images.

In the final questionnaire, we asked about the platform's usability using

a German version[24] of the System Usability Scale (*SUS*) [8] and the participants' intrinsic motivation using a German version [67] of the Intrinsic Motivation Inventory (*IMI*) [12]. Finally, we thanked the participants for their help and displayed a small textual debriefing explaining the study's goals. Tags produced by the participants during the interaction were stored and analyzed quantitatively and qualitatively (see Section 4.4).

4.2.2 Design Task Specific Steps

In contrast to our base conditions (i.e., **none** and **TD**), after the tutorial, design task participants had the task to design their own game in the image annotation context. After implementing their concepts, the same participants were asked to tag images by using their self-designed games as described in Section 4.2.1.

We required them to textually describe the game elements by using at least 700 characters, followed by two additional free-text questions ("*Why do you think that your game concept will motivate you?*" and "*Which element of your concept is most important for you?*"). Six statements on a 5-point scale with the labels *Disagree*, *Somewhat disagree*, *Neither agree nor disagree*, *Somewhat agree*, *Agree* had the goal to assess the participants' satisfaction with their self-designed concepts and whether they would estimate their realization as motivating for them or others to create more tags. Finally, three drop-down statements asked for participants' prior experiences with game design and image tagging, as well as how difficult they would rate their concept's implementation. The entire questionnaire is available in Section B.2. Before submitting their concepts, participants were asked to enter an e-mail address for receiving their implemented game. We analyzed the submitted concepts to learn which gamified elements were designed. The complete concept analysis is available in Section 4.3. Based on that we classified each concept as either *feasible* (fully implementable), *infeasible* (contains elements we could not implement, e.g., monetary rewards or setting changes), or *not sufficient* (not a game concept). We implemented all *feasible* concepts, forming the **own** condition. Participants with *infeasible* concepts received an implementation differing from their's (i.e., at most two game elements coincided; zero coincidences: 52%; one: 22%; two: 26%), forming **other**. As it turned out, a feasible concept sufficiently differed from the others (see the aquarium concept in Section A.1.5) so that we considered it as suitable for this purpose. All implemented concepts are available in Section A.1, a selection of infeasible ones as examples in Section A.2. The image annotation task started after the participants' game were implemented, reviewed and sent to them (in the case of **others**, a *different game*). They also had to go through

[24] `https://goo.gl/shC7qE`, last accessed on 29.07.2018

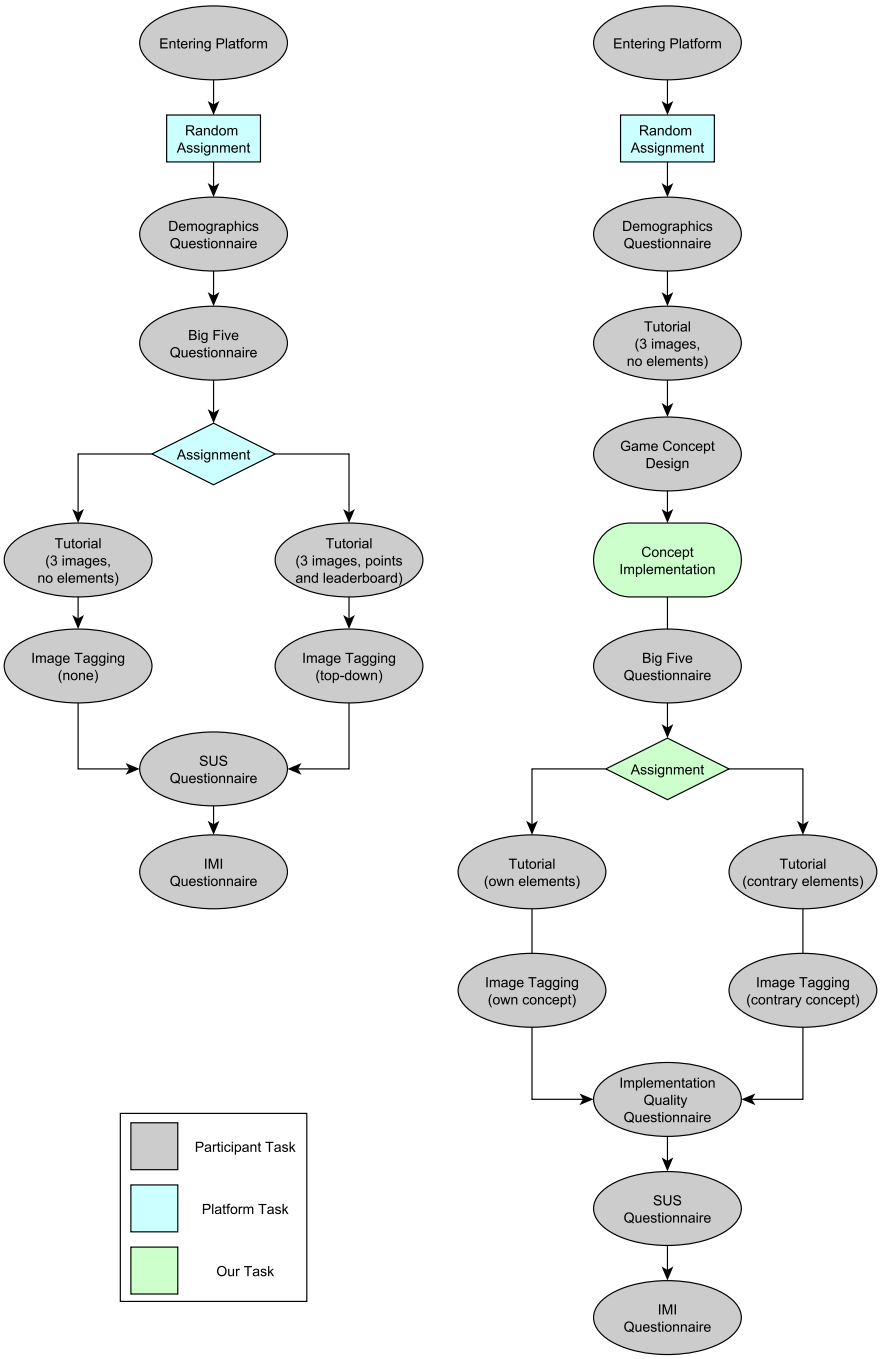

Figure 4.1: The study's method flow chart. **Left**: for **none** and **TD**. **Right**: for **own** and **other**.

a second tutorial consisting of a single image to introduce the implemented game elements. In the final questionnaire, participants in **own** and **other** additionally had to answer statements about their concept's implementation (e.g., "*The game visible while tagging corresponded with the concept I submitted.*").

4.3 Coding Process of Concepts

Goal of the coding process was to identify the user-designed game elements by conducting a content analysis [30] and creating a code book. For the code book development we utilized an existing one at the chair which we adapted for our needs. Two independent coders inspected all submitted game concepts separately and made adjustments to the code book, if necessary. The full code book containing all 45 codes along with small explanations and how often the respective codes were coded is available in Appendix C.

Certain codes had relationships, e.g., if *Real Challenge* was coded *Goals*, was coded, as well. After coding and implementing the code book adjustments, the two coders discussed the codings and solved deviations. A design was then considered as *feasible* if the concept itself did not contain any logic flaws (e.g., unreachable goals) and if it was implementable within the frame of this study (i.e., in the provided setting and the provided time). Concepts changing the setting (e.g., "*I want to see the image for 10 seconds.*") or taking too much time to implement (e.g., concepts using complex 3D graphics) and those containing logic flaws were considered as *infeasible*. *Insufficient* concepts (see Section 4.2) were not coded. When looking at all submitted valid concepts, 41.67% of the participants who created feasible concepts ($N = 12$) reported prior design experience, while only 10.71% of participants who created infeasible concepts ($N = 27$) reported prior design experience.

To check the validity of the coding, a third coder coded all concepts we rated as *feasible*. The inter-rater agreement (Cohens Kappa [11]) for every code was calculated, which was on average $\kappa = .79$ ($Min = .57$, $Max = 1.00$). This can be considered as "substantial" [60]. We created a total of 235 codings, resulting in a mean of 6.03 game elements ($SD = 2.33$, $Md = 7$, $Min = 1$, $Max = 11$) per concept. Most coded elements were "elements for multiple players" (see Appendix C for more amounts): *Multiplayer* (used in 72% of all concepts), *Social Competition* (62%), and *Social Comparison* (54%). Besides that, *Points* were used in about every second concept (49%), usually serving as a basis for other game elements (e.g., competition or progression).

4.4 Tag Quality Rating

Two independent raters rated the quality of each tag, allowing us to make a point on tag quality. Before the rating, we used the open source tool OpenRefine[25] to fix typos and to receive a more consistent tag set. To reduce work further, tags appearing multiple times per image were pooled and rated together. After those refinements, all tags were rated. Since we did not specify the kind of "mood" while tagging, we received tags of several perspectives as, for example, psychology or art. This complicated the rating task, since some of the moods may not be sensible at first sight (e.g., "colorful" - art, or "threatening" - psychology) and heavily depend on individual image cognition.

The raters' task was to assign each tag the value 1 when a tag *does not fit*, 2 when a tag *reasonably fits*, and 3 when a tag *fits perfectly*. Initially, we derived basic rules that helped in rating the tags (e.g., assign 2 if a tag fits the image but is not a "mood"). Those rules were refined after a first iteration, since we experienced a large deviation between the two raters. While refining, the two raters discussed 100 controversial tags (see Appendix D) and, based on this, derived advanced rules for rating tags. The refined rules were used to rate tags whose ratings deviated in the first iteration for a second time. For validating reasons, an inter-rater agreement (Choen's Kappa [11]) for all tags and images of $\kappa = .84$ was calculated. This can be considered as "substantial" to "almost perfect" [60]. Based on the two raters' ratings, mean tag quality scores were calculated for each image-tag pair and, in addition, for each participant.

4.5 Participants

The image annotation platform was used until we received 20 valid responses for each of our control conditions (**none** and **TD**) and more than 40 game concepts in the test condition. In total, we received 42 game concepts that were all coded (see Section 4.3) and evaluated. 12 of them were considered as *feasible*, 27 as *infeasible*, and three as *insufficient*. We implemented all *feasible* concepts (**own**) and assigned all participants with *infeasible* concepts to **other**. Some of **other** claimed that the game they saw during tagging corresponded to what they submitted. Those were removed from further evaluation, since we were not sure about this claim's origin (e.g., social desirability [24]).

After those adjustments, we had a total of $N = 60$ valid responses (**none**: 20; **TD**: 20; **own**: 10; **other**: 10). Most of the participants are between 18

[25] `http://openrefine.org`, last accessed on 29.07.2018

and 24 (43.3%) or 25 and 31 (38.3%) years old ($32-38$: 3.3%; $39-45$: 1.7%; $53-59$: 8.3%; 60 and older: 3.3%). Their gender distributes evenly over all conditions (female: 50%; male: 48.3%; not specified: 1.7%). Most of them (93.3%) have the German nationality. All participants answered with at least a neutral response to our gaming related questions (e.g., "*I characterize myself as game affine*", see also B.1) and, hence, can be considered as open to games in general.

4.6 Results

In the following, we first present results about the conditions' comparability to each other, then move to tag quantity and quality, and, finally, present results to the participants' intrinsic motivation. For the tag analysis, we split the images into two *image types*: *basic* describes the 15 images each participant had to annotate and *extra* the (optional) additional images. For each result, we conducted a Kruskal-Wallis ANOVA in order to find significant differences between the conditions. As post-hoc test for the Kruskal-Wallis tests, we used Dunn-Bonferroni tests with corrected p-values to neutralize alpha inflation errors and to find which conditions differed significantly. As of now, $H(x)$ denotes the *test statistics* of a Kruskal-Wallis ANOVA with the *degrees of freedom* x. *Significance values* are represented by p, Dunn-Bonferroni *test statistics* by z, and *effect sizes* by r.

4.6.1 Condition Comparability

To rule out the participant's game affinity and personality traits, as well as the system's usability score as potential confounding variables, we evaluated the respective questionnaires and tested them for differences.

First, for each participant, we calculated a mean game affinity (**none**: Mean = 3.36, SD = 1.09; **TD**: Mean = 2.97, SD = 1.00; **own**: Mean = 3.28, SD = .73; **other**: Mean = 3.56, SD = .92) based on their average answers to the game affinity questions (see Section B.1). A Kruskal-Wallis ANOVA provided no evidence that the game affinity significantly differs over the conditions ($H(3) = 2.60$, $p = .46$).

A Kruskal-Wallis ANOVA over the evaluated Big Five Inventory provided no evidence that extraversion ($H(3) = 2.28$, $p = .52$), neuroticism ($H(3) = 2.29$, $p = .52$), openness ($H(3) = 3.90$, $p = .27$), and agreeableness ($H(3) = 1.06$, $p = .79$) significantly differs over the conditions, but for conscientiousness it does ($H(3) = 9.65$, $p = .02$). The post-hoc test found a significant difference between **other** and **none** ($z = 2.85$, $p = .03$, $r = .52$). No further condition pair showed a significance level larger than .05 (**other-own**:

	N	Mean	SD	Min.	Max.	Median
None	20	81.38	13.99	50.00	97.50	85.00
TD	20	83.13	12.16	47.50	95.00	86.25
Own	10	86.00	10.01	67.50	97.50	86.25
Other	10	81.00	15.10	45.00	97.50	85.00

Table 4.1: System Usability Score for each condition. The means do not significantly differ.

$z = .65$, $p = 1.00$; **other-TD**: $z = 1.76$, $p = .47$; **own-TD**: $z = 1.00$, $p = 1.00$; **own-none**: $z = 2.10$, $p = .22$; **TD-none**: $z = 1.34$, $p = 1.00$).

Third, in average, each condition's *system usability score* (see Table 4.1) was between 81 and 86 which, according to Bangor et al. [6], is a *good* to *excellent* score. We conducted a Kruskal-Wallis ANOVA and were not able to find a significant difference ($H(3) = .60$, $p = .90$).

Summarizing, we found no evidence that the participants' game affinity or the system's usability score significantly differ between our conditions, ruling out those as potential confounding variables. But we found a significant difference between **other** and **none** when considering the participants' personality trait *conscientiousness*, which may be an additional limitation to those described in Section 4.8.

4.6.2 Tag Quantity

For the quantitative analysis, all tags ($N = 6868$) were considered, i.e., no tag was excluded, despite its quality (see Section 4.4). This is valid, since the participants' task was to create as much tags as possible and no requirement for quality was suggested; tag quality is considered in the next section. Table 4.2 shows the number of tags created per condition for image types *basic*, *extra*, and *basic + extra*, respectively.

Looking at descriptive statistics, one can see that – as assumed – participants in **own** created the most tags, followed by **other** and **TD**, and finally **none** (illustrated in Figure 4.2). Against our assumptions, on average, **other** and **TD** created the same amount of tags. A Kruskal-Wallis ANOVA showed significant differences between conditions and tag quantity for image types *basic* ($H(3) = 15.32$, $p = .00$) and *basic + extra* ($H(3) = 18.11$, $p = .00$).

For *basic* images, post-hoc tests found a significant difference between **none** and **other** ($z = -2.74$, $p = .04$, $r = .50$), as well as between **none** and **own** ($z = -3.53$, $p = .00$, $r = .64$). All other pairs produced non-significant values (see Table 4.3). For *basic + extra* images we found a significant difference between **none** and **TD** ($z = -3.08$, $p = .01$, $r = .49$), **none** and **other**

		Basic	Extra	Basic + Extra
None	**N**	20	1	20
	Mean	46.10	16.00	46.90
	SD	18.71	0.00	19.13
	Median	46	16	46
	Min	23	16	23
	Max	102	16	102
TD	**N**	20	12	20
	Mean	66.45	20.17	78.55
	SD	30.00	12.64	32.88
	Median	54	17	74
	Min	33	6	38
	Max	140	54	144
Own	**N**	10	4	10
	Mean	119.00	113.25	164.30
	SD	100.30	132.02	158.99
	Median	89	58	111
	Min 32	27	32	
	Max	381	310	479
Other	**N**	10	2	10
	Mean	73.80	23.00	78.40
	SD	27.85	1.41	30.77
	Median	70	23	70
	Min	26	22	26
	Max	126	24	126

Table 4.2: Number of tags created per condition for each image types *basic*, *extra*, and *basic + extra*.

($z = -2.67$, $p = .05$, $r = .49$), and **none** and **own** ($z = -3.78, p = .005$, $r = .69$). All other pairs produced non-significant values (see Table 4.3).

Looking at all tags created (i.e., *basic + extra* images), those values suggest that participants with gamified interactions tend to create more tags than participants with no gamification, confirming the observations of Mekler et al. [47]. This provides evidences to $\mathbf{H_{Quant,1}}$. While the descriptive values suggest that participants with self-designed games created more tags than those in **TD** or **other**, we found no significant effect and, hence, no evidence to $\mathbf{H_{Quant,2}}$. Lastly, against our assumption, we found no evidence for $\mathbf{H_{Quant,3}}$, i.e., participants in **TD** did not create more tags than those in **other**.

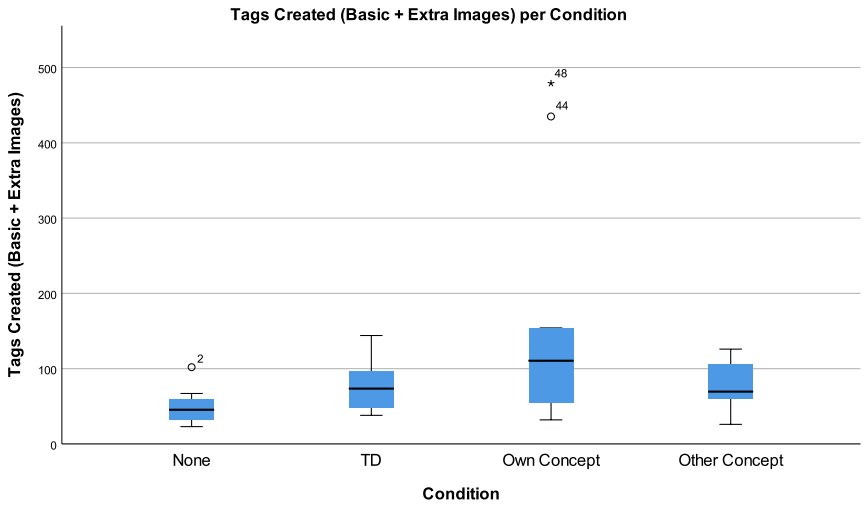

Figure 4.2: Boxplot of the participants' tag quantity for *basic + extra* images. Horizontal lines denote the respective medians. The blue box around the median illustrates the range within which the middle 50% of scores fall. Whiskers (the lines at the top and the bottom of the blue boxes) extend to the most and least extreme scores respectively. Furthermore, outliers can be found in **none** and in **own**.

Image Type	Compaired Conditions	Std. Test Statistic (z)	Adj. Sig (p)
Basic	**None - TD**	-2.21	.17
	None - Other	-2.74	.04
	None - Own	-3.53	.00
	TD - Other	$-.94$	1.00
	TD - Own	-1.73	.50
	Other - Own	.69	1.00
Basic + Extra	**None - TD**	-3.08	.01
	None - Other	-2.67	.05
	None - Own	-3.78	.00
	TD - Other	$-.15$	1.00
	TD - Own	-1.26	1.00
	Other - Own	.96	1.00

Table 4.3: Dunn-Bonferroni tests with adjusted significance values for tag quantity and *basic* and *basic + extra* images. Highlighted values denote significant differences ($p < .05$). $N = 60$ for both, *basic* and *basic + extra*.

		Basic	Extra	Basic + Extra
None	**N**	20	1	20
	Mean	2.66	2.66	2.65
	SD	.25	.00	.25
	Median	2.8	2.7	2.8
	Min	2.15	2.66	2.15
	Max	2.96	2.66	2.96
TD	**N**	20	12	20
	Mean	2.64	2.60	2.63
	SD	.23	.23	.22
	Median	2.7	2.6	2.7
	Min	1.92	2.17	2.0
	Max	2.94	2.90	2.92
Own	**N**	10	4	10
	Mean	2.55	2.54	2.55
	SD	.22	.23	.21
	Median	2.5	2.5	2.5
	Min	2.24	2.29	2.25
	Max	2.87	2.80	2.85
Other	**N**	10	2	10
	Mean	2.57	2.66	2.57
	SD	.25	.19	.24
	Median	2.6	2.7	2.6
	Min	2.15	2.52	2.15
	Max	2.90	2.79	2.90

Table 4.4: Tag quality, based on the participants' mean tag ratings, per condition for each image type.

4.6.3 Tag Quality

For each participant, we calculated a mean tag quality using the mean of the average rating per tag (illustrated in Figure 4.3). Table 4.4 shows the mean tag quality per condition for image types *basic*, *extra*, and *basic + extra*, respectively. We conducted a Kruskal-Wallis ANOVA, but were not able to find a significant difference between the conditions and tag quality for none of the image types (*basic*: $H(3) = 2.82$, $p = .42$; *basic + extra*: $H(3) = 2.70$, $p = .44$). For **none** and **TD**, this is in accordance to the results of Mekler et al. [47].

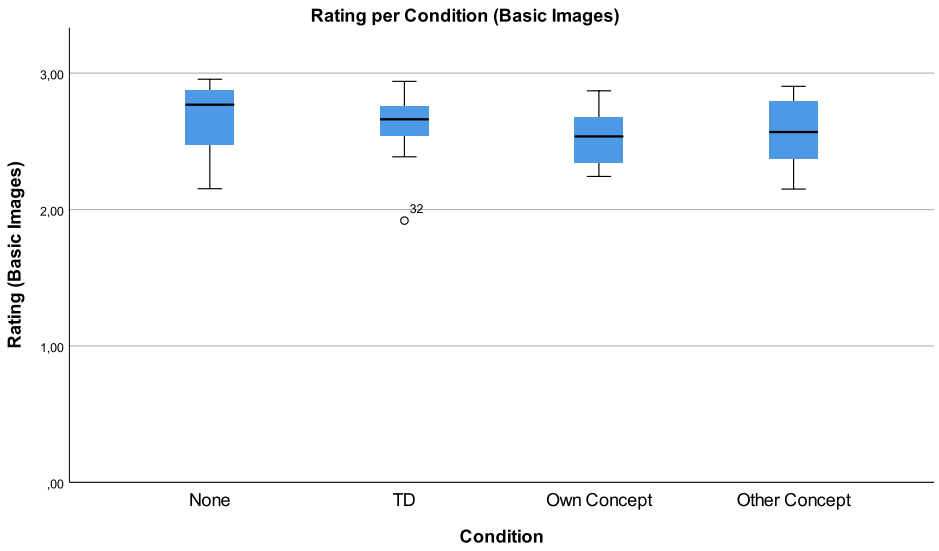

Figure 4.3: Boxplot of the participants' tag quality rating per condition for the 15 basic images (based on participants' mean rating).

Image Type	Compaired Conditions	Std. Test Statistic (z)	Adj. Sig (p)
Basic	**None - TD**	$-.27$	1.000
	None - Other	7.17	.00
	None - Own	10.59	.00
	TD - Other	7.38	.00
	TD - Own	11.14	.00
	Other - Own	-2.52	.070
Basic + Extra	**None - TD**	.02	1.00
	None - Other	7.51	.00
	None - Own	10.80	.00
	TD - Other	7.42	.00
	TD - Own	10.65	.00
	Other - Own	-1.74	.49

Table 4.5: Dunn-Bonferroni tests with adjusted significance values for tag quality and *basic* and *basic + extra* images. Highlighted values denote significant differences ($p < .05$). $N_{basic} = 5057$, $N_{basic+extra} = 5057 + 826 = 5883$.

To verify our hypotheses $\mathbf{H_{Quali,1}}$ and $\mathbf{H_{Quali,2}}$ we decided to analyze the tag quality using all created tags over all conditions. In other words, we used the rated quality of each tag ($N = 6848$) to perform a Kruskal-Wallis ANOVA over all conditions, instead of using participant means where we would loose information (e.g., the participants' tag quality standard deviation). Here, we found significant differences between the conditions when looking at *basic* images ($H(3) = 179.83$, $p = .00$) and *basic + extra* images ($H(3) = 178.45$, $p = .00$). For *basic* images, post-hoc tests (see Table 4.5) found significant differences between **other** and **none** ($z = 7.38$, $p = .00$, $r = 1.35$), **own** and **none** ($z = 7.17$, $p = .00$, $r = 1.31$), **other** and **TD** ($z = 11.14$, $p = .00$, $r = 2.03$), and **own** and **TD** ($z = 10.59$, $p = .00$, $r = 1.93$).

For *basic + extra* images we found significant differences (see Table 4.5) between **other** and **none** ($z = 7.51$, $p = .00$, $r = 1.37$), **own** and **none** ($z = 7.42$, $p = .00$, $r = 1.35$), **other** and **TD** ($z = 10.80$, $p = .00$, $r = 1.97$), and **own** and **TD** ($z = 10.65$, $p = .00$, $r = 1.94$).

Those values suggest that participants without gamification tend to create tags of higher quality than the other groups (except for **TD**). Hence, we found partial evidence for $\mathbf{H_{Quali,1}}$. Against our assumptions, giving participants the autonomy to design their game elements on their own provided no evidence to increase the quality of the created tags, compared to the other conditions. Hence, no evidence for $\mathbf{H_{Quali,2}}$ is provided.

4.6.4 Intrinsic Motivation

Table 4.6 shows the participants' means for *interest / enjoyment*, *perceived competence*, *perceived choice*, and *pressure / tension* as measured with the Intrinsic Motivation Inventory. We performed a Kruskal-Wallis ANOVA, but found no significant differences between intrinsic motivation and any of the conditions (*interest / enjoyment*: $H(3) = 1.34$, $p = .72$; *perceived competence*: $H(3) = 3.73$, $p = .29$; *perceived choice*: $H(3) = 1.62$, $p = .65$; *pressure / tension*: $H(3) = 1.54$, $p = .67$). For *interest / enjoyment* and *perceived competence*, this result is as expected and reported by Mekler et al. [47]. On the other side, while the descriptive results of *perceived choice* are slightly higher in the **own** condition, we did not find a significant difference between the conditions. Hence, we can not provide evidence to a significant increase in autonomy satisfaction as posited in $\mathbf{H_{IMI}}$.

		Interest / Enjoyment			Perceived Competence			Perceived Choice			Pressure / Tension		
	N	**Mean**	**SD**	**Mdn.**	**Mean**	**SD**	**Mdn.**	**Mean**	**SD**	**Mdn.**	**Mean**	**SD**	**Mdn.**
None	20	6.95	3.50	8.0	5.95	2.61	6.0	7.00	3.01	7.0	5.80	3.09	6.0
TD	20	7.05	3.68	8.0	5.85	2.76	6.0	7.55	2.40	8.0	5.65	3.13	5.0
Own	10	7.60	4.14	8.0	6.30	2.79	7.0	8.10	3.54	9.5	4.50	2.76	4.0
Other	10	5.80	3.77	5.5	4.40	2.17	4.0	7.70	2.58	7.5	6.00	3.09	6.5

Table 4.6: *Interest / Enjoyment*, *Perceived Competence*, *Perceived Choice*, and *Pressure / Tension*, as measured with the Intrinsic Motivation Inventory. While the *perceived choice* of **own** is slightly increased, compared to the other conditions, *pressure / tension* is slightly decreased. *Interest / enjoyment* and *perceived competence* of **other** is slightly decreased compared to all other conditions.

Compaired Conditions	**Std. Test Statistic (z)**	**Adj. Sig (p)**
None - Other	$-.76$	1.000
None - Own	-2.13	.20
None - TD	-3.62	.00
Other - Own	1.19	1.00
Other - TD	2.20	.17
Own - TD	.83	1.00

Table 4.7: Dunn-Bonferroni tests with adjusted significance values for the number of extra rounds. The highlighted value denotes significant differences ($p < .05$) between **none** and **TD**. No further significant difference was found.

In addition to the IMI and as a further value to measure the participants' intrinsic motivation, we provided each participant the option to perform extra rounds before answering the final questionnaire. The results (see *extra* column of Table 4.2 or Table 4.4) show that 60% of **TD**, 40% of **own**, 20% of other, and only 5% of **none** absolved at least one extra round. We performed a Kruskal-Wallis ANOVA and found significant differences between the number of absolved extra rounds and conditions ($H(3) = 14.54$, $p = .00$). The post-hoc tests showed a significant difference between **none** and **TD** ($z = -3.62$, $p = .00$, $r = .57$). All other pairs produced non-significant values (see Table 4.7). In our case, this might be a hint that points and leaderboards (**TD**) may have actually functioned as extrinsic incentives [10] and the extra rounds were taken in order to, for example, earn additional points and climb up the leaderboard.

4.7 Discussion

In this study, goal was to test a bottom-up gamification approach and to investigate whether it can compete with or exceeds top-down gamification approaches. At the end of this section, Table 4.8 summarizes which hypotheses we were able to provide evidence for. In the following, each paragraph represents a finding to discuss.

Gamified conditions will create more tags than non-gamified conditions

When looking at all images after the tutorial (i.e., *basic* + *extra* images) and as hypothesized, we were able to show that participants with gamified elements produced significantly more tags than the condition with no gamification, providing evidence to $\mathbf{H_{Quanti,1}}$. This is in accordance to prior research [47], even though other types of gamification were used in their work, i.e., only pre-defined game elements.

Participants using self-designed games create more tags than those using pre-defined games

We showed that participants with self-designed games create more tags on average than all other conditions. At a descriptive point of view, this suggests that bottom-up gamification can have a positive effect on the number of created tags. But we were not able to find significant differences, which may be because of the low number of participants in the respective condition. Hence, no significant evidence for $\mathbf{H_{Quanti,2}}$ was found.

Participants using pre-defined games create more tags than those receiving games differing from what they designed

We showed that participants who designed their own game concepts but received a different one performed similar to a pre-defined top-down gamification approach. We could not find a significant difference between those groups and, hence, can make no statement about what the motivating part was; the game creation process or the actual game? Hence, no evidence for $\mathbf{H_{Quanti,3}}$ was found. Further research to this direction is needed, for example, by deploying an additional top-down group tagging images without designing their own game first.

Participants with no gamification create tags of higher quality

When looking at the participant's mean tag quality, we found no significant differences between the respective conditions. Again, this is in accordance to prior research [47] which showed the same with top-down gamification, although our mean tag quality ($N = 60$, $Mean = 2.61$, $SD = .23$) differs from theirs ($N = 273$, $Mean = 1.42$, $SD = .22$) and, hence, more conventions are required.

Looking at the ratings of all images (instead of calculating a mean tag quality for every participant), we found evidence that participants in the no gamification condition created tags with a significantly higher quality than those designing their own game elements. Here, it did not matter whether they received their self-designed game or a different one. We found no further evidence that the quality significantly differs between our control conditions, i.e., no and top-down gamification. Hence, we found partial evidences for $\mathbf{H_{Quali,1}}$. The assumed increased autonomy of participants using their self-designed games provided no evidence that they produce tags of higher quality, compared to all other conditions. Hence, no evidence to $\mathbf{H_{Quali,2}}$ was provided.

Self-designed games did not increase autonomy satisfaction

As expected, we found no evidence that self-designed games significantly affect participants' *interest / enjoyment*, *perceived competence*, or *pressure / tension*. Additional to that, we also found no evidence for significantly affecting their *perceived choice*, although the descriptive results show a slightly increased mean, compared to the other conditions. Having more participants in both test groups may support those observations by finding significant differences, but, at this point, the results do not allow to make any general statement about the influence of bottom-up gamification on intrinsic motivation and, hence, no evidence to $\mathbf{H_{IMI}}$ is provided.

Prior design experience

An important question is whether the level of prior design-experience influenced the study's outcome. As assessed with our questionnaires, participants with designs we implemented (i.e., *feasible* concepts) had more experience when it comes to game design, compared to those with designs we did not implement (i.e., *infeasible* concepts). Hence, the performance increase of participants with self-designed games may be because of a "better" designed game instead of the actual game.

Image annotation setting

Finally, the results of this study may be not as much due to the game elements or the self-designed games, respectively, but rather due to the image annotation task setting. Participants who received games which differed from their submission and those in the top-down condition, respectively, were only scored for tag quantity and did not receive any feedback on whether a tag was fitting an image or not. On the other side, participants who received their self-designed games had the chance to design more "responsive" games that, for example, consider tag quality (e.g., Section A.1.6) or provide more sophisticated social competition / comparison (e.g., Section A.1.8) than the leaderboard of the top-down condition or even both (e.g., Section A.1.2). Previous research as, for example, the studies of von Ahn and Dabbish [64, 65], turn such tag quality into a game element by pairing two players having the task to guess words their respective opponent creates. Points, leaderboards and unlockable elements might need such an additional quality level in order to be perceived as "meaningful" or "informational". Additionally, finding a setting which is more relevant to the participants might emphasize the effects of bottom-up gamification.

Hypothesis	Description	Evidence provided?
$H_{Quant,1}$	**TD**, **own**, and **other**, will create significantly more tags than **none**.	yes
$H_{Quant,2}$	**Own** will create significantly more tags than **TD** and **other**.	no
$H_{Quant,3}$	**TD** will create significantly more tags than **other**.	no
$H_{Quali,1}$	Tag quality of **none** will be significantly higher than of **TD** and **other**.	partially
$H_{Quali,2}$	Tag quality of **own** will be significantly higher than of **none**, **TD**, and **other**.	no
H_{IMI}	**Own** will significantly increase autonomy satisfaction, compared to **none**, **TD**, and **other**.	no

Table 4.8: Summary about which hypotheses we found evidences for.

4.8 Limitations

Our study had limitations. First, the *delay* between a design's submission and the participant actually tagging images was on average 29.45 days ($SD = 9.02$, $Min = 6$, $Max = 44$), which might be a limiting factor. It is worth to mention that this delay occurred mainly due to the fact that we tried to use synergies between the design and not because of the platform's capabilities. Furthermore, this would be a possible explanation why some participant in the **other** condition identified the game as theirs (see Section 4.5).

Second, the *number of participants*, especially in **own** and **other**, can be seen as a limiting factor. Having more than a single concept developer would have allowed to implement more concepts in the same time and, hence, to increase the test groups' sizes. Platforms as, for example, Amazon Mechanical Turk could be used to acquire more participants, although, paying them for participating (extrinsic motivation) may harm their intrinsic motivation [12].

Third, the *image annotation task* itself may have influenced the outcome of the study. In contrast to other image annotation tasks [65] we asked our participants to describe a "mood" when viewing an abstract painting, instead of, for example, describing the actual image's content. This rather free-form image annotation task may have made it difficult for participants to come up with suitable tags. As already suggested by Mekler et al., it might be necessary to "*... also study tasks with relatively clearly defined quality metrics, which make it easier to assess and provide feedback on performance quality.*" ([47], page 533). In order to retain the comparability with the results of Mekler et al., we intentionally decided to follow the same "open" tag approach as they did.

Finally, our participants interacted with the platform for a *relatively short time*. Studies [26, 37] reported that gamification may improve user behavior for a short time. Hence, follow-up studies might increase the study duration by introducing additional game elements as, for example, badges (e.g. "Tag images five days in a row!"). Furthermore, while the study in general is adaptable to other applications, our results are specific to the image annotation context, i.e., no other gamified applications were tested.

Chapter 5

Conclusion and Future Work

In this thesis, we presented both the implementation of an image annotation framework allowing to quickly deploy gamified conditions, and a study making use of such. The goal was to investigate the effectiveness of bottom-up gamification. After recapturing the thesis goals and summarizing the study results, a small outlook for possible future work is given.

5.1 Thesis Goals

We start with elaborating which thesis goals (see Section 1.2) we reached. First of all, we built a platform providing functionality to tag images and textually describe game concepts (**Goal1**). The platform was extended by a framework which allows operators to quickly implement these game concepts. Using this framework, we carried out a study with no gamification and top-down gamification (**Goal2a**) with participants who were allowed to create their own game concepts (**Goal2b**). Participants were then asked to use their games in association with the image annotation task (**Goal3**). We evaluated tag quantity, tag quality (**Goal4**), and self-reported questionnaires (**Goal5**). In consequence, each thesis goal as described in Section 1.2 has been met.

5.2 Conclusion

We reproduced previous research of Mekler et al. showing that participants with gamified elements produce significantly more tags than without. While participants who created their own gamification created significantly more tags than participants without gamification, we found no evidence for a significant difference between this bottom-up approach and our top-down conditions. Letting people design their own games and serving them with different ones still has significant positive effects compared to no gamification, but no evidence for a significant difference between this approach and our gamified top-down conditions was found.

The tag quality was relatively high compared to the previous study of Mekler et al. [47]. In contrast to them, we found that – when looking at the quality of each tag – participants with no gamification produced tags of significant higher quality than participants who were subject to one of the bottom-up conditions.

In accordance with previous results [47], we found no significant differences between our conditions and the participants' intrinsic motivation and their perceived autonomy in particular.

To summarize, we can affirm the question whether people are able to design games motivating them in creating more tags compared to using no game, providing further insight to such bottom-up approaches. While we are not able to provide evidence on what the motivator was (i.e., the actual game or the design process), we showed that using self-designed games can lead to a positive performance change. These findings provide evidence that treating users as designers may also be beneficial in settings other than the image annotation task. While this setting is rather restricted, it would be interesting to see whether people could design games (or software in general) in a setting that is more relevant to them. People may, for example, design games motivating them to exercise more often or save more energy at home. We also found no significant evidence that participants with self-designed games create more tags than people with pre-defined games. One reason for this may be the low number of participants in our test groups. Nevertheless, extending the test groups and finding a significant performance improvement would emphasize the role of bottom-up approaches besides top-down gamification. In addition, we found no significant influence of our gamified interventions on the participants' intrinsic motivation.

5.3 Future Work

Possible future work may aim to eliminate our study's limitations. First, the delay between participants designing their game and actually tagging images could be decreased, as well as the number of implemented concepts increased. This can be achieved by, for example, having more than one developer implementing the game designs right after they were submitted. A further approach would be to define a "building set" containing as many game elements as possible in the image annotation task's context. Implementing a game concept would then take only the merging of those pre-defined game elements. Considering the wide range of submitted game elements in our study, the latter might be a challenging approach. In addition to this, we experienced that two thirds of the submitted concepts were infeasible in the context of the image annotation platform. Hence, it might be advisable to increase the number of participants assigned to the design task.

Future studies could carry out the game of the second test group as a top-down condition, i.e., without the need to design a game concept first. This would allow a comparison of this pseudo-bottom-up game to a true top-down condition and also the deduction of statements about the motivational part (i.e., the design task or the actual game). In fact, while writing this thesis, such a condition has been set up and deployed, it is waiting for additional participants.

Even if in general the study is applicable to a wide field, our results are restricted to the image annotation setting. It remains to be seen whether people in other settings perform better when using a bottom-up gamification approach.

Appendix A

Concepts

In this chapter, we present all concepts we implemented (i.e., *feasible* concepts) and three of those we did not (i.e., *infeasible* concepts) as examples. Each concept was translated to English for readability reasons.

A.1 Concept Implementations

For each *implemented* concept we present the participant's individual description. Besides that, we describe our implementations and provide some screenshots.

In the following, we describe the implementations and decisions we made on a rather high level. Due to space reasons, we renounce to go in too much detail. The titles of the following sections are derived from the platform's intern unique identifier. This is helpful when you want to inspect the respective code in the platform. For example: *Beat Friend 51 (CID 1)* refers to the concept with ID *1* of the user with ID *51* and the internal codename *id_beat_friend_51* in the platform. Here, *code/frontend/js/ge/id_id_beat_friend_51.js* and *code/frontend/views/id_beat_friend_51.js* would be a good starting point for inspecting the GitHub repository.

A.1.1 Beat Friend 51 (CID 1)

Concept:
Title: Beat your Friend by Tags
Goal: Find more tags for this image as your friend. There has to be an average of "moods" below each image. Thus, one can see how many moods there are in average to a certain image. But, at this point, one does not know what the highest amount of tags is. This would lead to try to fiend at least as many tags as the average. After

Figure A.1: Tagging screen of Beat *Friend 51*. While tagging, the user only sees the average tags for this image (first line). After confirming the tags, the ranking and the statistics below are getting visible. In the end screen, only the ranking is visible.

submitting the tags, a small ranking shows how many tags I and friends reached. How many participants felt the same mood and similar. Additionally, it should be possible to "like" other tags (which then would count). With this, one would prevent that someone enters random values without connection to the image.

Why motivating?
Leads to competition. One can compete with or compare to other users.

Most important element:
Competition

The participant explained that the concept would motivate to create "at least the average number of tags". We defined the average number of tags by providing a random value between 6 and 8, resulting in an average of 7 tags per image. This is in accordance to our target of about 100 tags over 15 images. The ranking is identical to the top-down condition. The statistics below the ranking are real data from other participants, as well as the three tags to rate. Rating words has no effect for any participants. Figure A.1 shows the tagging screen, after confirming the tags. The participant is able to enter a nickname for a highscore (similar to the top-down condition) after tagging all images.

A.1.2 Complex 115 (CID 8)

Concept:
I imagine a game in which each image represents a level. To get to the next level, at least 3 out of 5 of sample keywords that are invisible to the image should fall. These keywords must be those that are often associated with the particular image (also by experts, or other participants in the study). The number of matching keywords is displayed (for example 2/5), each hit gives a point. As soon as there are at least three hits, the display turns green (before it is red). You are welcome to enter more terms before moving to the next level, the counter keeps counting. Thus one can reach by the hits maximally 5 points per picture. If you go with less than 3 hits to the next picture, then for each missing hit in each case 2 points are subtracted from the total score. The total score is displayed constantly, and is red if you average less than 3 points per image, and green if you're above or equal. At the end of the 15 images all entered terms should be compared again with an additional sample keyword list of 10 keywords (instead of 5). For every term that matches this list, you get 1/4 bonus point. The bonus points are added up and then calculated on the total score. The score should be constantly listed (with no bonus) with other participants, and at the very end with a bonus, so that you can compare with them.

Why motivating?
I have a guideline that I can and must orient myself to, so that I do not lose too many points. It compels me to compose as many meaningful keywords as possible in the "showdown" with other participants.

Most important element:
Constant transparency, which is given to me by the keyword control and comparison score and motivates me to write many meaningful keywords to lead the score

We implemented most of the concept's parts as described by the participant. As a difference, we increased the *sample word list* from 5 to 15 to make it easier to hit the maximal points of 5. We did this, because, in our tests, 5 of 5 words was almost impossible to hit for a single person. Furthermore, we correlated the *bonus points* with the participant's amount of tags, again targeting to our tag goal of 100. Figure A.2 shows the game elements while tagging. The participant is able to enter a nickname for a highscore (similar to the top-down condition) after tagging all images.

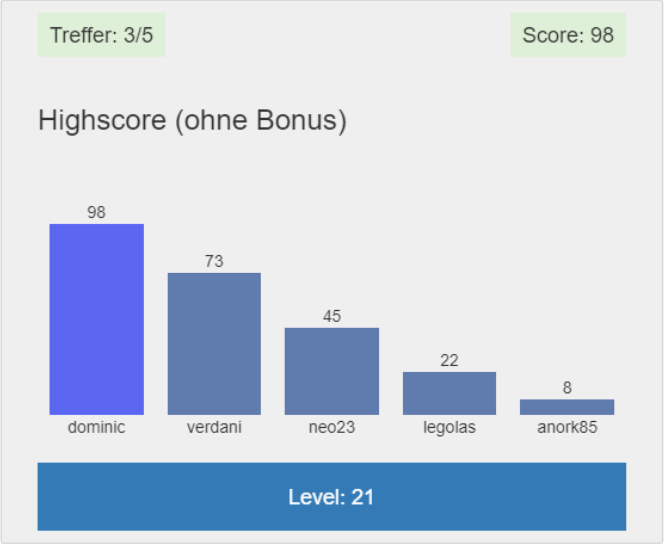

Figure A.2: Tagging screen of *Complex 151*. The *Hits* and *Score* (first line) are implemented as defined by the participant: At the beginning, their background is red and changes to green when certain conditions are met. The highscore is adapted from the top-down condition to match our 100 tag goal. The level describes the number of image that currently get tagged. After tagging all images, only the highscore is visible, but this time with bonus points.

A.1.3 Points 96 (CID 5)

Concept:
A game would be for 1-2 people. You give the moods that you have to the picture. The specified moods are compared with those of other people and based on that, depending on whether the word occurs frequently or less frequently you get points for it. At the end you have a "high score" or point comparison with the other players. For example, the first picture has 3 impressions, these impressions are stored anonymously. The impressions are then compared with other stored impressions. For example, if you have the impression that you are "quiet" and "quiet" was taken most often, you get 100 points, then scale down in percent, starting from the maximum value with the number of mentions.

Why motivating?
Just the fact that there are points ensures that the reward center is activated in the brain, which makes me or others more motivated.

Most important element:
The most important thing is that there are points and so the psychological effect can be used.

For each word, the participant gets randomly 50 to 100 point. Again, this adaption was needed to aim for the 100 tag goal and the competition aspect as described in the concept. Figure A.3 shows the game elements as they are visible in the end screen. The participant is able to enter a nickname for

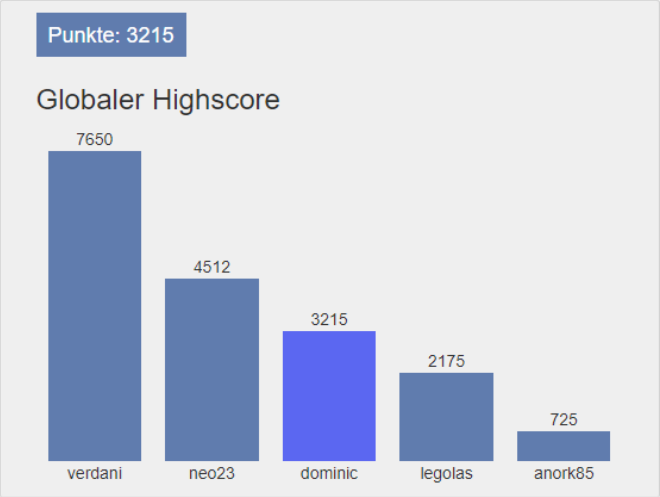

Figure A.3: End screen of *Points 96*. The upper line displays the participant's current points. The highscore below is the same (but scaled) as in the top-down condition and is only visible in the end screen and the tutorial.

a highscore (similar to the top-down condition) after tagging all images.

A.1.4 Points 151 (CID 17)

Concept:
My idea would be to compete against a global high score. Each player should get points for his keywords. The points are summed up and give the score of the player. Example Usecase: The player chooses his nickname, then he gets the first picture for 5 sec and gets the chance to enter his tags. After some time he can confirm his input and receives the points for his keywords. Then he will be forwarded to the next picture. Similar to the previous example. After all 15 images have been processed, the points are added up and the nickname of the player is entered along with the points in the leaderboard. The keywords for keywords could vary. For example, frequently-used keywords might score a few points, while poorly used or unused words may give more points.

Why motivating?
It motivates because you compete against a global high score and you have the incentive to be better than other players.

Most important element:
The concept to give points for tags.

The implementation is related to *Points 96* with the difference that here, instead of the total score, the participant sees the points achieved for the last image. As in *Points 96*, the participant gets points for every word, but this time with a bit of randomness to mimic the variance of points for different tags. Figure A.4 shows the game elements as they are visible on the end

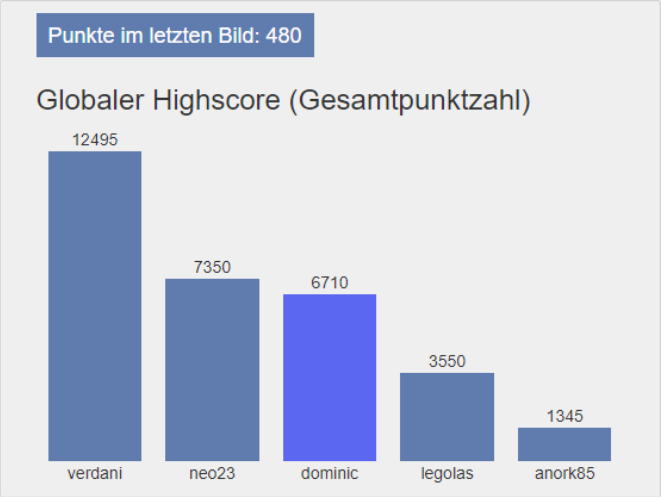

Figure A.4: End screen of *Points 151*. Besides the points of the last image, the participant sees his success compared to other participant. The highscore is the same as in the top-down condition, scaled to fit the 100 tag target, and only visible in the end screen and tutorial.

screen. The participant is able to enter a nickname for a highscore (similar to the top-down condition) after tagging all images.

A.1.5 Points Aquarium 339 (CID 21)

Concept:
In order to motivate myself, I would like to achieve something that I can look forward to with every word entered or after a certain number of words. For this reason, I imagine an aquarium, which is initially drab and empty. After each e.g. I wish the third mentioned mood that other plants and fish fill the aquarium and thus create a colorful bustle. To make this also not too monotonous, I wish for special fish, which I have been promised in advance as a surprise (surprise 1 in X tags available, surprise 2 ..., etc.). It should therefore always be displayed to me how many keywords are missing until the next surprise, but without revealing WHAT this surprise will be. Here would be e.g. my idea that every 30th tag adds a special fish / plant in the pelvis (pufferfish, rays, sharks, etc., I am very happy to be surprised :)!). After tagging, if I have exceeded a certain number of tags, I would like to get an extra surprise in the form of another special fish.

Why motivating?
Due to the fact that there is always something new waiting for someone, I can easily imagine that this concept motivates me. One is curious about what is to come and of course wants to tag as many pictures as possible in order to unlock all secrets. I have chosen the aquarium, because I, but also many other people like to observe aquariums. That way I can imagine that it could motivate others too.

Most important element:

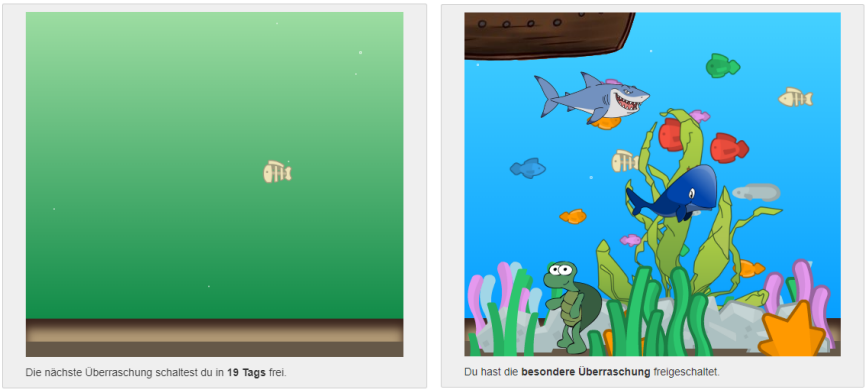

Figure A.5: Tagging and end screen of *Aquarium 339*. **Left:** Small amount of tags, only a few elements visible. **Right:** All elements unlocked.

> The construction of an aquarium with many interesting fish; Not necessarily the surprise ad; Please do not spoil the surprise in advance (for example, by outlines)

We developed elements so that the participant has to enter 100 tags to unlock all of them. While every third tag reveals a new element, big surprises are unlocked after 30, 60, and 90 tags, respectively. There is a note under the aquarium, giving hints on how many tags are needed to unlock the next big surprise. After tagging and with 100 (or more) tags entered, the participant unlocks the special surprise (shark, see figure A.5). As wished by the participant, we renounced in spoiling the experience by showing hints on what's the next big surprise. Figure A.5 shows the aquarium when the participant entered both a low and a high amount of tags.

A.1.6 Points Review 125 (CID 9)

> **Concept:**
> Maybe as a virtual card game with others realize, but rather on "honest cooperation" builds. A picture is revealed as a card and each player has some time (previously defined) to think of as many keywords as possible. The keywords of all players are collected.
> Now each player will see the keywords of the OTHER players (which is not his own), but without revealing who they are from. Each chooses the keywords that he / she finds appropriate for the image seen. If at least X players agree to the proposal (e.g., half), then each player who made that proposal will receive one point. If the proposal was made by a single player, he will receive three points. The goal is to collect as many points as possible.
>
> **Why motivating?**
> The incentive to win the game
>
> **Most important element:**

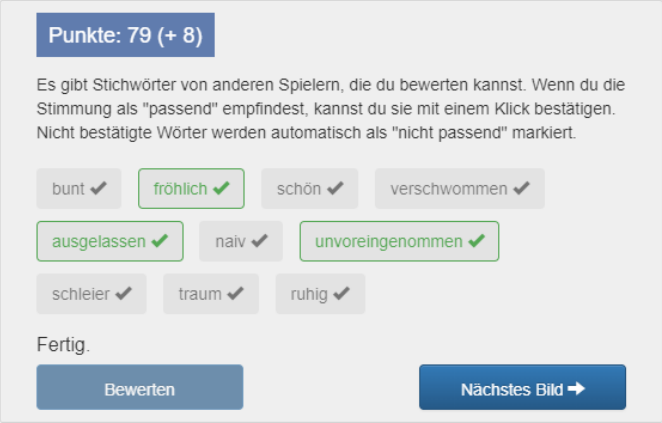

Figure A.6: Tagging screen of *Points Review 125*. The first line shows the participant's total points and – if tags were confirmed – the amount of points achieved in the current image in brackets. The elements below are only visible if the participant confirmed the tags and allow to rate tags of other participants.

The honest cooperation of the players, so that no keywords for tactical reasons by "weak" players are rejected

We gave the user the illusion to play with two other participant, while, in fact, he played alone against the tags of all participants' tags. This is needed, since it is infeasible to find other participants playing at the same time as the participant. Furthermore, with this approach, we were able to aim for the 100 tag goal. In consequence, the participant's tags are compared to other participants' tags and achieve points related to the respective frequency of tags. As tags for the other participants, we used the 25 most used tags for the respective image. Moreover, the buttons to rate the tags as "fitting", have no function except getting highlighted when clicked. The end screen shows a small information about which place the participant reached, compared to the other "participants", again supporting the 100 tag goal. Figure A.6 shows the game elements as they are visible while tagging.

A.1.7 Points Review 210 (CID 15)

Concept:
Each "player" sees the respective picture for 5 seconds at the same time. After that, each player notes down his own keywords. Once everyone has made his notes, they are compared with each other. Now each player gets points for points that match those of the other players: for a match 1 point, for 2 pieces 2 points and so on. If a player A has noted down a point that no other player has, there is still a chance of getting points, as other players consider them suitable. For each player who thinks this is appropriate, Player A gets 2 points (His idea was thus more original than the

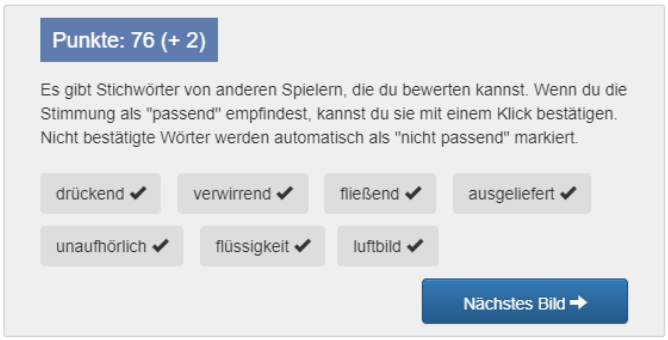

Figure A.7: Tagging screen of *Points Review 210*. The points (first line) are visible all the time. While tagging and after confirming the tags, the points achieved are visible in brackets behind the point score. Below that, the participant can rate other participant's tags.

other, because no one else has come to it) In the end now wins the player with the most points.

Why motivating?
There is a competitive idea.

Most important element:
That players do not play each other out by failing to pinpoint the unique bullet points, if applicable.

Calculation and displaying (achieved) points, as well as rating other participants' tags, is identical to *Points Review 210*. The main difference is that the participant does not have to rate and explicitly confirm other tags. Some of the tags are displayed delayed in time in order to support a better feeling of opponent participants' presence. As tags for the other participants, we used the 25 most used tags for the respective image. The end screen shows a small information about which place the participant reached, compared to the other "participants", again supporting the 100 tag goal. Figure A.7 shows the game elements as they are visible while tagging.

A.1.8 Points Review 438 (CID 31)

Concept:
You could generate an internet game that is built much like City Land River. Each participant must write down as many keywords as possible for each image that come to mind. All players have to do this in parallel and after 1 minute the processing time is over. When all players have finished editing, a screen will open showing all submitted suggestions. The more keywords match, the more points each player gets. Scoring can be staggered if 2 players use the same word, then there are 2 points, if 3 players use the same word, there are 3, etc. So you have to try to make the most accurate statements about the picture and therefore give more Keywords, as there

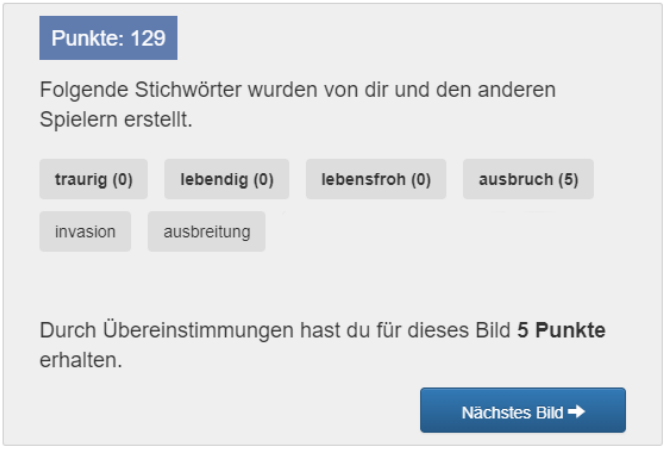

Figure A.8: Tagging screen of *Points Review 438*. The first line indicates the points the participant achieved over all previous images. Below are the participant's entered tags (bold), as well as tags of other participants (non-bold). The numbers in brackets indicate the number of correlations with other participants and, hence, the number of points the participant achieved in this image.

are more points for it

Why motivating?
So you have to try to make the most accurate statements about the picture and therefore enter more keywords, as there are more points for it

Most important element:
Earn points

As one can see in figure A.8, the view is related to *Points Review 210* and *Points Review 125*. As tags for the other participants, we used the 25 most used tags for the respective image. Since there is no direct competition, there is no scaling in tag amount, and, in consequence, this concept does not explicitly aim for the 100 tag goal. Furthermore, here, clicking the tag buttons has no effect, since there is no rating-system in this concept. Since the participant explicitly mentioned that gathering points is the most important element and there is no ranking or similar in the concept, the end screen only shows the amount of points achieved and no ranking, highscore, or similar.

A.1.9 Points Similar 61 (CID 4)

Concept:
After I have entered the tags for a picture, a small line appears over the next tag window in which is how many tags I have given and how many were. Below a line in which is how many tags the previous average has submitted to users, for example,

Figure A.9: Tagging screen of *Points Similar 61*. The first line indicates the number of words the participant achieved over all previous images. Below are statistics about the participant's number of tags for this image, the average number of tags, as well as the most used tags for this image.

if there were 3, I get the 3 most frequently displayed, 7 tags the 7 most common. In the end, there could be a high score list with nicknames (which of course need to be replaced by a dirty word filter). (I find it weird to have to use at least 700 words, good concepts do not necessarily need a big explanation, MauMau does not have a manual of> 700 words and it's still one of the most popular games.)

Why motivating?
You always want to be better than others

Most important element:
Only afterwards to find out how many tags the average had

The average tag quantity is defined as a random value between 4 and 8 to roughly hit the 100 tag goal. The most used tags are taken from the 10 most used tags for the respective image. While the participant described that he wants to see the statistics "over the next tagging window", we decided to move it to the right side. We did this for two reasons: Firstly, it would not be clear to see the image to tag, since the statistics would move it down. And secondly, having the game elements on the right side makes the implementation better comparable to the other implementations in terms of visibility. The participant is able to enter a nickname for a highscore (similar to the top-down condition) after tagging all images. Figure A.9 shows the game elements after confirming the tags for an image.

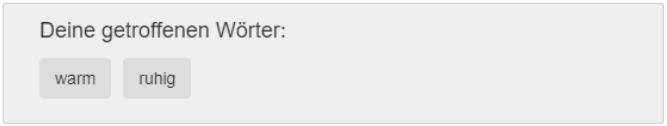

Figure A.10: Tagging screen of *Single Teams 407*. The windows shows the words the participant already hit. They are dynamically faded in when the participant enters the respective word to the tagging field.

A.1.10 Single Teams 407 (CID 27)

Concept:
The concept would be played either with fellow players or with a programmed automatism. For several players it's about finding the right words at the same time. It is recognized by the program when a certain number of words were called by both players. One can continue this idea, so that a competition between several teams of players, each consisting of two people arises. Thus, a team of players can win against another team. My concept is a kind of bingo that you can play too far against several. As a single player, the program detects when a certain number of preprogrammed words have been named. Thus, single players can play against each other. One could push the concept further by replacing an AI with the second team player who is configured for a human combination gift.

Why motivating?
By the attraction of competition and speed

Most important element:
Varied images to challenge the combination of the players again and again.

This implementation is split in two modes: In the *single-player mode*, the goal is to hit 4 pre-defined words faster than an opponent. The *multi-player mode* is quite similar, with the difference that there are two teams where each has to enter the same 4 tags for the image. The implementation for both is identical: The user has to hit 4 of the 25 most used tags to win against the opponent or opposing team, respectively. Depending on the mode, the wording is different (e.g., "You won this round." versus "Your team won this round."). If the participant was not able to win the round after 60 to 80 seconds, an invisible timer stops the round and decides for the winner. Here, the participant has a 50% chance to win (the opponent gave up) and a 50% chance to loose (the opponent hit the target tags). In the tutorial, we show the participant the single player-mode and explain that there is a multi-player mode. After that, we give the illusion to search for other players and give the option to choose between the two modes. Figure A.10 shows the tagging screen while already hit two tags.

A.1.11 Teams Simple 104 (CID 6)

Concept:
So I have not really understood how a game concept should be created now, even though it is about my own motivation? One possible game concept could be to make it a multiplayer game, where there are two teams and each team has some time to look at the picture and then have some time to find words. Who has found the most "logical" words wins the round and it goes to the next picture. Here, the concept of the images shown is implemented. It could be difficult that if the game is played more often, you already know some pictures and therefore has already made associations or words of the "opponent" or the one has noticed can write down.

Why motivating?
Actually, creating associations with pictures is fun for me, and that's why I do not really need this game concept.

Most important element:
The competition, as this always encourages to be better than the other, so to write more verbal than the opposite.

Similar to *Single Teams 407*, the user has to enter 4 of the 25 most used image tags to win the round. Each team has 90 seconds to enter and confirm tags. If the timer expires before confirming the tags, they are confirmed automatically and the winner gets decided. This is needed to support the illusion of multiple players and should be sufficient large when looking at the participants typing speed (about 55 words per minute). The teams with the most wins over all images win the game and are displayed on the end screen. Figure A.11 shows the tagging screen after the participant confirmed his tags.

A.1.12 Levels 498 (CID 41)

Concept:
Competition concept with multiple game levels Prerequisite would be a game in which it is possible to reach several levels. At the beginning of the game, the player asks any other player of the same level to play against him. Both are shown the same pictures. The aim of the game is to create more matching keywords per picture than the other players. A picture corresponds to a game round. At the end the won rounds are summed up. The player with the most won rounds has won. This then automatically rises to a higher level and can then play against other players of the same level. If a player loses on a level, he automatically drops one level. In order for players to "stay on the ball" and share their winnings with others, you could create a kind of leaderboard.

Why motivating?
Comparison with other players encourages faster and more targeted thinking

Most important element:
Play together with other players

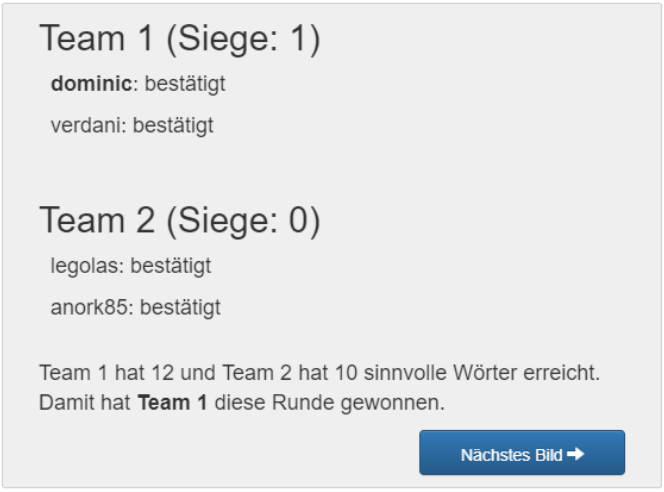

Figure A.11: Tagging screen of *Teams Simple 104* after confirming the tags. The teams can see their number of wins and which player is typing or already confirmed his tags. Below, one can see the number of "meaningful" tags and which team won the round.

Similar to *Single Teams 407*, the user has to enter 4 of the 25 most used image tags to win the round. We gave the illusion to do both waiting for a challenger and challenging a (unknown) player. Both functions are implemented as simple random timeouts showing the player a challenger or a challengee was found, respectively. The end screen is shown in figure A.12.

A.1.13 Teams Points 369 (CID 23)

Concept:
I would play this game with the team. Each team secretly records which keywords the other team will truly say. Each guess word gives points to the team. If a team has reached the maximum score of 10 points, all teams are resolved and re-rolled. If a team has not guessed any keywords, the team gets minus points. If in a round no team has guessed a single keyword, the same pictures must be reused, only one must not repeat the keywords now. If there is still no match, the teams are dissolved and new diced. The points are credited to the people who had the highest score when rolling the dice.

Why motivating?
It makes a fun competition out of it

Most important element:
That it is fun and not too long

As demanded, there are two teams having the goal to guess tags the respective opponent team probably enters. For each correlation, the respective

Figure A.12: End screen of *Levels 498*. The current level is always visible. The highscore is only visible a after tagging all 15 images. Under the highscore, there is a notification about whether the player won/lose this round (a single image) or the entire game (15/5 images).

team gets a point. As soon as one of those two teams hit 10 points, the teams are shuffled. It is programmatic excluded that a team hits no opposing tags. This is needed, since "re-using the same images" would make this implementation incomparable to the other concepts. For each image, the participant has 90 seconds to enter tags. This is needed to support the illusion of multiple players and should be sufficient large when looking at the participant's typing speed (about 20 words per minute). There is a logical flaw in this concept: The teams have to hit words that the opponent team probably enters. This would mean that they always have the same amount of (negative) points so that no team can win over another. That's why we **excluded this implementation** from the second part of the study.

Figure A.13: Tagging screen of *Teams Points 369* after confirming the tags. The teams can see whether they gain (green) or lost (red) points and the amount of points they gathered.

A.2 Not Implemented Concepts

This section shows an exemplary selection of concepts we considered as *infeasible*, along with a small explanation why we renounced to implement them. All of them are available here[26] (uncorrected and untranslated). The following sections' titles are derived from the project intern concept identification (e.g., *UID 123 – CID 456* refers to the concept ID *456* of the participant with user ID *123*). This is helpful when you want to inspect the respective concept in the raw data (e.g., the database).

A.2.1 UID 55 – CID 2

Concept:
- The game is only for me. Only when I'm done, I get the opportunity to look at answers from other players.
- I do NOT want to see examples of moods among the pictures because the examples affect my perception.
- If I have not written anything after a few seconds, tips might appear.
- I want to play the "" game "" on the PC.
- You could distribute points for each word.

I can not say more about that. I can not say more about that. I can not say more about that. I can not say more about that. I can not say more about that. I can not say more about that. I can not say more about that. I can not say more about that. I can not say more about that.

[26] `http://dx.doi.org/10.13140/RG.2.2.26801.68960`, last accessed on 29.07.2018.

Why motivating?
Due to the competitive and comparative thoughts (points, what others have chosen)

Most important element:
It's pretty simple. :)

Why not implemented: Filled up with words to reach 700 characters and content too vague. Hence, it's not a concept, but more tips to build a concept.

A.2.2 UID 183 – CID 13

Concept:
It would have to be a reward system. There is a star or similar for each keyword given. and a ranking list with names where you could climb up if you have more keywords and get a lot of stars. Much like Facebook does with its requests for site changes. This could then be evaluated according to positive, negative or neutral aspects. This could be special points. For example: with 100 positive keywords there is a happysmile or other ...
Alternatively, you may be able to work on the rail with real rewards like PAYBACK. So a shop from which I can choose something for my points something that I get sent then.

Why motivating?
Rewards are always on. ;)

Most important element:
Fun

Why not implemented: Real prizes as the mentioned "PAYBACK points" are not feasible for us.

A.2.3 UID 398 – CID 26

Concept:
Several people get together to analyze the meaningfulness of the image together and to facilitate a group dynamic within a discussion. Various people let the picture initially take a little more than 5 seconds to work on it. The effect is then discussed in detail in the group and sometimes questioned why individual observations have occurred only in certain individuals. The experience that different types of positive as well as negative effects are caused in the viewer, may help another person to better understand or understand his way of thinking. The game goal or a kind of points distribution would have to be considered, otherwise the game sense might be questioned.

Why motivating?
Because it is discussed in the group and you will spend so much time and more time on it..

Most important element:
group dynamics

Why not implemented: Not a game concept but an approach to tag images in groups.

Appendix B

Questionnaires

The following sections contain questionnaires we used in addition to the established ones (e.g., SUS and IMI). Unless otherwise stated, all questionnaire statements had to be answered using a 5-point Likert scale (*Disagree*, *Somewhat disagree*, *Neither agree nor disagree*, *Somewhat agree*, *Agree*). Each questionnaire was translated to English for readability reasons.

B.1 Game Affinity

1. I characterize myself as game affine.
2. I often play video games.
3. I have a passion for video games.
4. I often play board games.
5. I have a passion for board games

This was followed by an optional free-form text asking for how frequently the participant plays board or video games.

B.2 Design Task

It follows the design task's description:

> "At the previous page you have seen how to enter mood tags. In the following, you will have the opportunity to build a concept that should motivate yourself in creating more tags per image. For that, you can assume that you will see a sequence of **15 images** for **5 seconds** each.
>
> The concept should be a **game** or **game-like** in **the context of this website**. **Not** in the context of this website is, for example, a smartphone application, a board game, or an application without relation to the previous shown tagging creation. You can freely decide how your game concept should look and, for example, if it should be a game *just for you* or a game with others.
>
> After your submission, we will review and implement your concept. You will *receive an e-mail* containing further information about how you can test it."

This was followed by three free-text questions:

1. How would this game look?
2. Why do you think that your game concept will motivate you?
3. Which element of your concept is most important for you?

The following statements had to be answered:

1. I am satisfied with my concept.
2. I perceived it as easy to create a game concept in the image annotation setting.
3. I think my game concept would motivate others in tagging more images.
4. I think my game concept would motivate me in tagging more images.
5. Image tagging is relevant to me.
6. Image tagging already motivates me – I do not need a game to motivate myself.

Three drop-down statements were asked to answer:

1. How hard do you think is it to implement your concept? (*Easy*, *Hard*, *Not implementable*, or *Not assessable*, optional free-form text when answering with anything else than *easy*.)
2. I already designed a game or was part of such a process.
3. The principle of image tagging was known to me a priori.

Appendix C

Coding Book

Table C.1 shows all codes according to our code book, along with an explanation and their respective frequency. All received concepts, except of those we estimated as *insufficient*, were considered in this table. Table C.2 shows all assignments as coded by the two coders for the concepts presented in this document (i.e. Section A.1 and Section A.2).

Game Element	Explanation – The participant mentions ...	Total
Multiplayer	... a multiplayer component	28 (72%)
Social Comparison	... that it is possible to compare one's own performance to others	24 (62%)
Social Competition	... competition between human players where one will be the winner	21 (54%)
Points	... an entity that can be accumulated in the game	19 (49%)
Goals	... specific (sub-) goals that need to/can be achieved	11 (28%)
Progression	... progression in the game overall or specific game attributes	8 (21%)
Bonus	... that it is possible to achieve a bonus (e.g., 100 bonus points)	8 (21%)
Peer pressure	... that other people can see whether or how I (fail to) do something	7 (18%)
Fairness	... concepts that make the game fair (e.g., only fitting tags count)	7 (18%)
System assistance	... an assistance function that eases something for the player	7 (18%)
Real challenge	... a particularly challenging aspect or beating one's own scores	6 (15%)
Time pressure	... actions that need to be done within a certain time	6 (15%)
Surveillance	... a control instance that monitors progress in the game	6 (15%)
Visible Progress	... an indication that shows progress or distance to the next goal	5 (13%)
Single Player	... that the concept is (also) usable for a single player	5 (13%)
Teams	... that players can be grouped into teams or guilds	5 (13%)
Social Collaboration	... that players cooperate to reach a goal	5 (13%)
Punishment	... penalties (e.g., subtracting points)	5 (13%)
Skill improvement	... that the game (or players) conveys knowledge (to others)	5 (13%)
Appearance	... something that belongs to the look/appearance of the game	5 (13%)
Prizes	... physical (e.g., cinema tickets) or personal reward (e.g., watch TV)	4 (10%)
Analogy	... an existing game or concept that he/she adapts to his/her concept	4 (10%)
Customization	... that the player can adjust components to personal needs	3 (8%)
Surprise	... unexpected elements that surprise the player in the game	3 (8%)
Periodicity	... something that is or should be done regularly (e.g., weekly)	2 (5%)
Social Recognition	... that progress made is visible to others (to show it to them)	2 (5%)
Friends	... that the game can be played with friends/family	2 (5%)
Security/Privacy	... security or privacy aspects	2 (5%)
Anti-Cheating	... something that prevents cheating in the game	2 (5%)
Communic. Tools	... that it is possible to communicate with others in the game	2 (5%)
Unsp. reward	... rewards, but the kind of the reward is not further specified	2 (5%)
Unsp. competition	... a competition against other players or AI-controlled entities	1 (3%)
Socialization	... that the game helps to get to know other people	1 (3%)
Achievements	... achievements (e.g., badges, ranks) can be gained	1 (3%)
Unlockables	... that new game content or features can be unlocked	1 (3%)
Collecting	... collecting a specific entity explicitly as a motivational factor	1 (3%)
Encouragement	... that the game or other players encourage one to reach goals	1 (3%)
Personalization	... that the systems adapts to the player's needs itself	1 (3%)
Virtual Character	... character(s) that do not represent the player but other entities	1 (3%)
Care Taking	... that the player needs to care about others (e.g., virtual pets)	1 (3%)
Lottery/Gambeling	... that players can bet in the game or have a raffle	1 (3%)
Mini Games	... mini games that can be played within the game	1 (3%)
Story	... story components	1 (3%)
Computer Game	... a computer component of the game	1 (3%)
Mobile Game	... a mobile component of the game	1 (3%)

Table C.1: The game elements, an explanation and their frequency.

	1	2	4	5	6	8	9	13	15	17	21	23	26	27	31	41
Goals	1					1		1			1	1				1
Real challenge																
Visible Progress						1					1					
Progression						1		1			1					1
Time pressure					1		1								1	
Periodicity																
Multiplayer	1		1	1	1	1	1	1	1	1		1	1	1	1	1
Single Player		1									1			1		
Teams					1							1		1		
Social Competition	1		1	1	1	1	1	1	1	1		1		1		1
Social Collaboration												1		1		
Social Comparison	1	1	1	1	1	1	1	1	1	1		1		1		1
Social Recognition										1						
Peer pressure / Accountability	1		1				1		1						1	
Unspecific competition																
Friends	1															
Socialization																
Fairness					1					1					1	1
Security / Privacy				1			1									
Anti-Cheating	1															
Surveillance	1					1	1		1							
System assistance		1	1													
Communication Tools																
Unspecific reward																
Points		1		1		1	1	1	1	1		1			1	
Bonus				1		1	1	1	1							
Punishment						1						1				1
Achievements								1								
Prizes																
Unlockables											1					
Collecting/Collectables															1	
Encouragement																
Personalization																
Customization																
Virtual Character (other)											1					
Care Taking											1					
Knowledge / Skill improvement																
Surprise											1	1				
Lottery/Gambeling																
Appearance	1					1										
Mini Games																
Story																
Computer Game		1														
Mobile Game																
Analogy														1	1	

Table C.2: The codings as coded by our coders. Headers denote the project intern concept ID. You can use this ID to identify the actual concept in either Section A.1 or Section A.2. Rows denote the codings as defined in the code book (Table C.1). Codings that were coded for a specific concept are denoted with 1's.

Appendix D

Controversial Tags

This chapter shows all controversial tags we discussed (untranslated). The list numbers correspond to image names, i.e., "123" would refer to the image with the name "abstract_123.jpg" in the repository. Full size images can be found on GitHub[27].

0001 fantasievoll, feindselig, heiter, schmerzlich, sommer, uninteressant

0002 abends, stimmungsvoll, verträumt

0003 konzentriert, quatsch, winterlich

0010 frisch, mediativ, ruhend, tiefsinnig, verliebt

0056 konfus, nebelig, senkrecht, unbedeutsam, warm

0061 düster, leben, skurril, verzerrt

0075 adam, behütend, herbstlich, lebensfroh, verliebtes

0106 bedrückend, blätter, blau, fenster, stimmungsvoll, urwald, zerrissen

0119 aufregend, behutsam, deteiert, emotional, verbunden, zuneigung

0120 einig, stimmungsvoll

0134 entspannend, glücklich, urlaubsstimmung, verlassen

0168 interesse, stimmungsvoll, unantastbar, unzufrieden, verwirrt

0197 abend, allein, frei, munter, vergänglichkeit

0202 aufwühlend, ausgeglichen, desinteressiert, eintönig, fade, melancholisch, öde, stadt, standardisiert, weiß

0207 herbst, malerisch, nächtlich, verlassen

0239 gereizt, tiefe, tödlich, ungutes, unklar, wild

[27] `https://github.com/dextreem/GIAP`, last accessed on 29.07.2018.

0256 abschied, lila, sommerlich, weltvergessen, wohlfühlen

0278 geerdet, herbst, regnersich, undeutlich, verloren

0004 gegensätzlich, trauer

0006 cool

0007 einsam

0027 unergrundbar

0030 rot

0047 exotisch, gemeinsam, lebendig, spannend

0048 lieblos

0050 dunkel

0069 abstrakt, bewegt, natürlich, tödlich

0091 feinfühlig, gemeinsam

0092 schlecht

Bibliography

[1] Frederick E. Allen. 2011. Last accessed on 29.07.2018. *Disneyland Uses 'Electronic Whip' on Employes*. https://www.forbes.com/sites/frederickallen/2011/10/21/disneyland-uses-electronic-whip-on-employees

[2] Maximilian Altmeyer, Pascal Lessel, and Antonio Krüger. 2016. Expense Control: A Gamified, Semi-Automated, Crowd-Based Approach for Receipt Capturing. *Proceedings of the 2016 International Conference on Intelligent User Interfaces. International Conference on Intelligent User Interfaces, IUI '16* (2016), 31 – 42. https://doi.org/10.1145/2856767.2856790

[3] Ofer Arazy, Oded Nov, and Nanda Kumar. 2015. Personalityzation: UI Personalization, Theoretical Grounding in HCI and Design Research. *AIS Transactions on Human-Computer Interaction* (2015). https://doi.org/10.17705/1thci.00065

[4] Ivon Arroyo, Matthew Micciollo, Jonathan Casano, Erin Ottmar, Taylyn Hulse, and Mercedes Rodrigo. 2017. Wearable Learning. *Proceedings of the Annual Symposium on Computer-Human Interaction in Play - CHI PLAY '17* (2017), 205 – 216. https://doi.org/10.1145/3116595.3116637

[5] Sander Bakkes, Chek Tien Tan, and Yusuf Pisan. 2012. Personalised Gaming: A Motivation and Overview of Literature. *Proceedings of The 2012 Australasian Conference on Interactive Entertainment: Playing the System - IE '12* (2012). https://doi.org/10.1145/2336727.2336731

[6] Aaron Bangor, Philip Kortum, and James Miller. 2009. Determining what Individual SUS Scores Mean: Adding an Adjective Rating Scale. *Journal of Usability Studies* (2009). https://doi.org/66.39.39.113

[7] Martin Böckle, Jasminko Novak, and Markus Bick. 2017. Towards Adaptive Gamification: A Synthesis of Current Developments. *Proceed-*

ings of the 2017 European Conference on Information Systems - ECIS '17 (2017).

[8] John Brooke. 1996. SUS - A Quick and Dirty Usability Scale. *Usability Evaluation in Industry* (1996). https://doi.org/10.1002/hbm.20701

[9] Jared Cechanowicz, Carl Gutwin, Briana Brownell, and Larry Goodfellow. 2013. Effects of Gamification on Participation and Data Quality in a Real-World Market Research Domain. In *Proceedings of the 2013 International Conference on Gameful Design, Research, and Applications - Gamification '13*. https://doi.org/10.1145/2583008.2583016

[10] Christopher P. Cerasoli, Jessica M. Nicklin, and Michael T. Ford. 2014. Intrinsic Motivation and Extrinsic Incentives Jointly Predict Performance: A 40-year Meta-Analysis. *Psychological Bulletin* (2014). https://doi.org/10.1037/a0035661

[11] Jacob Cohen. 1960. A Coefficient of Agreement for Nominal Scales. *Educational and Psychological Measurement* (1960). https://doi.org/10.1177/001316446002000104

[12] Edward L. Deci, Richard Koestner, and Richard M. Ryan. 1999. A Meta-Analytic Review of Experiments Examining the Effects Of Extrinsic Rewards on Intrinsic Motivation. (1999). http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.588.5821&rep=rep1&type=pdf

[13] Edward L. Deci, Richard Koestner, and Richard M. Ryan. 2001. Extrinsic Rewards and Intrinsic Motivation in Education: Reconsidered Once Again. *Review of Educational Research* (2001). https://doi.org/10.3102/00346543071001001

[14] Sebastian Deterding. 2014. Eudaimonic Design, or: Six Invitations to Rethink Gamification. In *Rethinking Gamification*.

[15] Sebastian Deterding, Dan Dixon, Rilla Khaled, and Lennart Nacke. 2011. From Game Design Elements to Gamefulness. *Proceedings of the 2011 International Academic MindTrek Conference on Envisioning Future Media Environments - MindTrek '11* (2011), 9. https://doi.org/10.1145/2181037.2181040

[16] Sebastian Deterding, Miguel Sicart, Lennart Nacke, Kenton O'Hara, and Dan Dixon. 2011. Gamification. Using Game-Design Elements in Non-Gaming Contexts. In *Proceedings of the 2011 Annual Conference Extended Abstracts on Human Factors in Computing Systems - CHI EA '11*. https://doi.org/10.1145/1979742.1979575

[17] Lauren S. Ferro, Steffen P. Walz, and Stefan Greuter. 2013. Towards personalised, gamified systems: an investigation into game design, personality and player typologies. *Proceedings of the 2013 Australasian Conference on Interactive Entertainment: Matters of Life and Death - IE '13* (2013). https://doi.org/10.1145/2513002.2513024

[18] Andrés Francisco Aparicio, Francisco Luis Gutiérrez Vela, José Luis González Sánchez, and José Luis Isla Montes. 2013. Gamification: Analysis and Application of Gamification. *New Trends in Interaction, Virtual Reality and Modeling* (2013). https://doi.org/10.1007/978-1-4471-5445-79

[19] Yuichi Fujiki, Konstantinos Kazakos, Colin Puri, Pradeep Buddharaju, and Ioannis Pavlidis. 2008. NEAT-o-Games: Blending Physical Activity and Fun in the Daily Routine. (2008), 1 – 22. https://doi.org/10.1145/1371216.1371224

[20] Lalya Gaye and Atau Tanaka. 2011. Beyond Participation: Empowerment, Control and Ownership in Youth-Led Collaborative Design. *Proceedings of the 2011 ACM Conference on Creativity and Cognition - C&C '11* (2011), 335 – 336. https://doi.org/10.1145/2069618.2069684

[21] Kathrin M. Gerling, Conor Linehan, Ben Kirman, Michael R. Kalyn, Adam B. Evans, and Kieran C. Hicks. 2015. Creating Wheelchair-Controlled Video Games: Challenges and Opportunities when Involving Young People with Mobility Impairments and Game Design Experts. *International Journal of Human Computer Studies* (2015), 64 – 73. https://doi.org/10.1016/j.ijhcs.2015.08.009

[22] Borja Gil, Iván Cantador, and Andrzej Marczewski. 2015. Validating Gamification Mechanics and Player Types in an E-Learning Environment. In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*. https://doi.org/10.1007/978-3-319-24258-3_61

[23] Ian Goodfellow, Yoshua Bengio, and Aaron Courville. 2016. Deep Learning. *MIT Press* (2016). https://doi.org/10.1142/S1793351X16500045

[24] Pamela Grimm. 2010. Social Desirability Bias. In *Wiley International Encyclopedia of Marketing*. https://doi.org/10.1002/9781444316568.wiem02057

[25] Juho Hamari. 2013. Transforming Homo Economicus into Homo Ludens: A Field Experiment on Gamification in a Utilitarian Peer-to-Peer

Trading Service. *Electronic Commerce Research and Applications* (2013). https://doi.org/10.1016/j.elerap.2013.01.004

[26] Juho Hamari, Jonna Koivisto, and Harri Sarsa. 2014. Does Gamification Work? - A Literature Review of Empirical Studies on Gamification. *Proceedings of the 2014 Annual Hawaii International Conference on System Sciences - HICSS '14* (2014), 3025 – 3034. https://doi.org/10.1109/HICSS.2014.377

[27] Juho Hamari and Janne Tuunanen. 2014. Player Types : A Meta-Synthesis. In *Selected Articles from the DiGRA Nording 2012 Conference: Local and Global-Games in Culture and Society*. https://doi.org/10.1111/j.1083-6101.2006.00301.x

[28] Michael D. Hanus and Jesse Fox. 2015. Assessing the Effects of Gamification in the Classroom: A Longitudinal Study on Intrinsic Motivation, Social Comparison, Satisfaction, Effort, and Academic Performance. *Computers and Education* (2015). https://doi.org/10.1016/j.compedu.2014.08.019

[29] Carrie Heeter, Yu-Hao Lee, Brian Magerko, and Ben Medler. 2011. Impacts of Forced Serious Game Play on Vulnerable Subgroups. *International Journal of Gaming and Computer-Mediated Simulations* (2011). https://doi.org/10.4018/jgcms.2011070103

[30] Hsiu-Fang Hsieh and Sarah E Shannon. 2005. Three Approaches to Qualitative Content Analysis. *Qualitative Health Research* (2005). https://doi.org/10.1177/1049732305276687

[31] Intrinsic Motivation Inventory. 1994. Intrinsic Motivation Inventory (IMI). *The Intrinsic Motivation Inventory, Scale Description* (1994). https://doi.org/www.selfdeterminationtheory.org

[32] Yuan Jia, Bin Xu, Yamini Karanam, and Stephen Voida. 2016. Personality-Targeted Gamification. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems - CHI '16*. https://doi.org/10.1145/2858036.2858515

[33] Anne Marie Kanstrup. 2012. A Small Matter of Design. *Proceedings of the 2012 Participatory Design Conference on Research Papers: Volume 1 - PDC '12* (2012), 109. https://doi.org/10.1145/2347635.2347651

[34] Maurits Kaptein, Boris De Ruyter, Panos Markopoulos, and Emile Aarts. 2012. Adaptive Persuasive Systems. *ACM Transactions on Interactive Intelligent Systems* (2012). https://doi.org/10.1145/2209310.2209313

[35] Konstantinos Kazakos, Thirimachos Bourlai, Yuichi Fujiki, James Levine, and Ioannis Pavlidis. 2008. NEAT-o-Games. In *Proceedings of the 2008 International Conference on Human Computer Interaction with Mobile Devices and Services - MobileHCI '08*. https://doi.org/10.1145/1409240.1409333

[36] Ana C. T. Klock, Isabela Gasparini, Marcelo S. Pimenta, and Jose Palazzo M. de Oliveira. 2015. "Everybody is Playing the Game, but Nobody's Rules are the Same": Towards Adaptation of Gamification Based on Users' Characteristics. *Bulletin of the Technical Committee on Learning Technology* (2015).

[37] Jonna Koivisto and Juho Hamari. 2014. Demographic Differences in Perceived Benefits from Gamification. *Computers in Human Behavior* (2014). https://doi.org/10.1016/j.chb.2014.03.007

[38] Annakaisa Kultima. 2017. The Role of Stimuli in Game Idea Generation. *Proceedings of the 2017 International Academic Mindtrek Conference - AcademicMindtrek ' 17* (2017), 26 – 34. https://doi.org/10.1145/3131085.3131127

[39] Ellen J. Langer and Judith Rodin. 1976. The Effects of Choice and Enhanced Personal Responsibility for the Aged: A Field Experiment in an Institutional Setting. *Journal of Personality and Social Psychology* (1976). https://doi.org/10.1037/0022-3514.34.2.191

[40] Jisoo Lee, Erin Walker, Winslow Burleson, Matthew Kay, Matthew Buman, and Eric B. Hekler. 2017. Self-Experimentation for Behavior Change: Design and Formative Evaluation of Two Approaches. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems - CHI '17*. https://doi.org/10.1145/3025453.3026038

[41] Pascal Lessel, Maximilian Altmeyer, Marc Müller, Christian Wolff, and Antonio Krüger. 2016. "Don't Whip Me With Your Games". *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems - CHI '16* (2016), 2026 – 2037. https://doi.org/10.1145/2858036.2858463

[42] Pascal Lessel, Maximilian Altmeyer, Marc Müller, Christian Wolff, and Antonio Krüger. 2017. Measuring the Effect of "Bottom-Up" Gamification in a Microtask Setting. *Proceedings of the 2017 International Academic Mindtrek Conference - AcademicMindtrek '17* (2017), 63 – 72. https://doi.org/10.1145/3131085.3131086

[43] James J. Lin, Lena Mamykina, Silvia Lindtner, Gregory Delajoux, and Henry B. Strub. 2006. Fish'n'Steps: Encouraging Physical Activity with

an Interactive Computer Game. (2006), 261 – 278. https://doi.org/10.1007/11853565_16

[44] Jana Machajdik and Allan Hanbury. 2010. Affective Image Classification Using Features Inspired by Psychology and Art Theory. In *Proceedings of the 2010 International Conference on Multimedia - MM '10*. https://doi.org/10.1145/1873951.1873965

[45] Elisa D. Mekler, Florian Brühlmann, Klaus Opwis, and Alexandre N. Tuch. 2013. Disassembling Gamification: The Effects of Points and Meaning on User Motivation and Performance. *CHI '13 Extended Abstracts on Human Factors in Computing Systems* (2013), 1137 – 1142. https://doi.org/10.1145/2468356.2468559

[46] Elisa D. Mekler, Florian Brühlmann, Klaus Opwis, and Alexandre N. Tuch. 2013. Do Points, Levels and Leaderboards Harm Intrinsic Motivation? *Proceedings of the 2013 International Conference on Gameful Design, Research, and Applications - Gamification '13* (2013), 66 – 73. https://doi.org/10.1145/2583008.2583017

[47] 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* (2017), 525 – 534. https://doi.org/10.1016/j.chb.2015.08.048

[48] Ethan R. Mollick and Nancy Rothbard. 2013. Mandatory Fun: Gamification and the Impact of Games at Work. *SSRN Electronic Journal* (2013), 1 – 54. https://doi.org/10.2139/ssrn.2277103

[49] MJ Michael J. Muller and Sarah Kuhn. 1993. Participatory Design. *Communications of the ACM* (1993), 24 – 28. https://doi.org/10.1145/153571.255960

[50] Venkatesh N. Murthy, Subhransu Maji, and R Manmatha. 2015. Automatic Image Annotation Using Deep Learning Representations. In *International Conference on Multimedia Retrieval - ICMR '15*. https://doi.org/10.1145/2671188.2749391

[51] Scott Nicholson. 2012. A User-Centered Theoretical Framework for Meaningful Gamification. *Games + Learning + Society* (2012). https://doi.org/10.1007/978-3-319-10208-5_1

[52] Rita Orji, Regan L. Mandryk, Julita Vassileva, and Kathrin M. Gerling. 2013. Tailoring Persuasive Health Games to Gamer Type. *Proceedings of the 2013 SIGCHI Conference on Human Factors in Computing Systems - CHI '13* (2013), 2467. https://doi.org/10.1145/2470654.2481341

[53] Rita Orji and Karyn Moffatt. 2016. Persuasive Technology for Health and Wellness: State-of-the-Art and Emerging Trends. *Health Informatics Journal* (2016). https://doi.org/10.1177/1460458216650979

[54] Rita Orji, Lennart E. Nacke, and Chrysanne Di Marco. 2017. 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. https://doi.org/10.1145/3025453.3025577

[55] Rita Orji, Kiemute Oyibo, and Gustavo F. Tondello. 2017. A Comparison of System-Controlled and User-Controlled Personalization Approaches. *Adjunct Publication of the 2017 Conference on User Modeling, Adaptation and Personalization - UMAP '17* (2017), 413 – 418. https://doi.org/10.1145/3099023.3099116

[56] 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. In *Adjunct Publication of the 2017 Conference on User Modeling, Adaptation and Personalization - UMAP '17*. https://doi.org/10.1145/3099023.3099071

[57] Andrew K. Przybylski, C. Scott Rigby, and Richard M. Ryan. 2010. A Motivational Model of Video Game Engagement. *Review of General Psychology* (2010). https://doi.org/10.1037/a0019440

[58] 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* (2007). https://doi.org/10.1016/j.jrp.2006.02.001

[59] Richard M. Ryan and Edward L. Deci. 2000. Intrinsic and Extrinsic Motivations: Classic Definitions and New Directions. *Contemporary Educational Psychology* (2000). https://doi.org/10.1006/ceps.1999.1020

[60] Steve Stemler. 2001. An Overview of Content Analysis. (2001), 137 – 146.

[61] Gustavo F. Tondello and Lennart E. Nacke. 2018. Towards Customizing Gameful Systems by Gameful Design Elements. *CEUR Workshop Proceedings* (2018), 102 – 110.

[62] Gustavo F. Tondello, Rita Orji, and Lennart E. Nacke. 2017. Recommender Systems for Personalized Gamification. *Adjunct Publication of the 2017 Conference on User Modeling, Adaptation and Personalization - UMAP '17* (2017), 425 – 430. https://doi.org/10.1145/3099023.3099114

[63] Gustavo F. Tondello, Rina R. Wehbe, Lisa Diamond, Marc Busch, Andrzej Marczewski, and Lennart E. Nacke. 2016. The Gamification User Types Hexad Scale. *Proceedings of the 2016 Annual Symposium on Computer-Human Interaction in Play - CHI PLAY '16* (2016), 229 – 243. https://doi.org/10.1145/2967934.2968082

[64] Luis von Ahn and Laura Dabbish. 2004. Labeling Images with a Computer Game. In *Proceedings of the 2004 Conference on Human Factors in Computing Systems - CHI '04*. https://doi.org/10.1145/985692.985733

[65] Luis von Ahn and Laura Dabbish. 2008. Designing Games With a Purpose. *Communications of the ACM* (2008). https://doi.org/10.1145/1378704.1378719

[66] Erica L. Wagner and Gabriele Piccoli. 2007. Moving Beyond User Participation to Achieve Successful IS Design. *Communications of the ACM* (2007). https://doi.org/10.1145/1323688.1323694

[67] Matthias Wilde, Katrin Bätz, Anastassiya Kovaleva, and Detlef Urhahne. 2009. Überprüfung einer Kurzskala intrinsischer Motivation (KIM). *Zeitschrift für Didaktik der Naturwissenschaften* (2009).

[68] José P. Zagal, Michael Mateas, Clara Fernández-Vara, Brian Hochhalter, and Nolan Lichti. 2007. Towards an Ontological Language for Game Analysis. *Worlds in Play: International Perspectives on Digital Games Research* (2007).