

**University
of Basel**

Faculty of
Psychology



Master's thesis presented to the Department of Psychology of the University of Basel for the degree of Master of Science in Psychology

Rewards in Gamification: A Conceptual Replication and Extension

Author: Saraceno Sebastian

Immatriculation number: 15-067-937

Correspondence email: sebastian.saraceno@unibas.ch

Examiner: Dr. Brühlmann Florian

Supervisor: Prof. Dr. Opwis Klaus

Center for General Psychology and Methodology, University of Basel

Submission Date: 15.11.2021

**MASTERARBEIT
REGISTRIERT
15.11.2021**

Acknowledgements

The author would like to thank Dr. Florian Brühlmann for his valuable feedback and support in the conception, realization and evaluation of this project and thesis. Further, special thanks go out to Prof. Dr. Klaus Opwis, Dr. Elisa Mekler, and everyone at the department of MMI Basel for making this project possible. Lastly, the author would like to thank Dr. Daniel Lakens, and Dr. Indrajeet Patil for their quick answering to questions concerning their R-packages, and Rowena Waldis for proofreading.

Declaration of scientific integrity

The author hereby declares that they have read and fully adhered the [Code for Good Practice in Research of the University of Basel](#).

Abstract

The aim of this study was to examine the effects of different kinds of rewards in gamification on intrinsic motivation and performance. In a study by Mekler et al. (2013b), the game element points increased performance without the expected negative effect on intrinsic motivation. Therefore, this study was replicated with a few changes to the procedure and material. It was hypothesized that the game element *points* would improve performance and impair intrinsic motivation, as posited by self-determination theory. A new, and more *controlling* tangible reward was expected to amplify these effects further. In this between-subject design study, participants (N = 291) tagged 18 journalistic pictures in an online image tagging task. Neither a statistically significant negative nor positive effect of both gamified conditions *points* ($d < 0.001$, 95%CI $d [-0.27, 0.27]$), and *controlling* ($d = 0.04$, 95%CI $d [-0.25, 0.34]$) on intrinsic motivation were found. Compared to the control group, performance was significantly higher in the *controlling* condition ($r = 0.23$, 95%CI $r [.09, .35]$), whereas the condition *points* did not have a significant effect on performance ($r = 0.13$, 95%CI $r [.008, .25]$). Performance decreased significantly over time in the *controlling* ($r = 0.678$, 95%CI $r [.54, .79]$) compared to the other two conditions. This observation was also supported by a significant interaction effect between condition and time in an ANOVA ($F(4/576) = 9.59$, $p = .006$). Intrinsic motivation decreased significantly over time in all conditions (*plain* ($d = 0.29$, 95%CI $d [0.1, 0.47]$), *points* ($d = 0.24$, 95%CI $d [0.03, 0.47]$)), however the largest decrease was observed in the *controlling* ($d = 0.402$, 95%CI $d [0.21, 0.63]$) condition. The results indicate that tangible and intangible rewards in gamification have varying effects on performance and intrinsic motivation. Whereas points alone had a positive effect on performance without impairing intrinsic motivation, points in combination with a performance-contingent tangible reward led to highly increased initial performance, which quickly decreased over time. Thus, the widely used tangible rewards in gamification might yield negative effects for long-term performance as well as intrinsic motivation and must be further investigated.

Contents

Rewards in Gamification: A Conceptual Replication and Extension.....	4
Motivation and Gamification.....	6
The Original Study.....	6
Aim of This Study.....	11
Methods.....	13
Participants.....	13
Materials.....	14
Measures.....	17
Procedure.....	18
Changes to the Original Study.....	19
Analysis Plan.....	19
Results.....	20
Exploratory Factor Analysis.....	20
Performance.....	21
Cheating Behavior.....	23
Intrinsic Motivation and Need Satisfaction.....	24
Discussion.....	27
Limitations.....	33
Further Research.....	35

Rewards in Gamification: A Conceptual Replication and Extension

As video games are more popular than ever (Nestor, 2021), more and more people are accustomed to game elements. Therefore, using these elements to improve user motivation in non-game contexts is very attractive for areas ranging from online stores to education. This practice can be seen all over the world and is frequently described as gamification.

According to a widely cited definition, gamification is the use of game elements in non-game contexts to improve user motivation and performance (Deterding et al., 2011, p. 1). When looking at contemporary research on google scholar, it stands out that most papers concern gamification in education and health contexts. In the real world though, gamification is also deployed in apps, workplaces, and marketing. For example, gamification was used in marketing campaigns by big companies such as M & M's or Nike (Burmester, 2021). A Swiss example of a gamified online store is digitec-galaxus (*digitec*, 2021), where users can gain points, badges and climb up a level if they participate in reviewing, commenting, or buying products. In turn, the users are rewarded with coupons if they reach a certain level or milestone. Even though gamification seems to be well used and working, there have been many critics of the concept. For example, Bogost (2011) described how gamification primarily consists of simply adding points, badges, and leaderboards to already existing systems in his critique *Gamification is Bullshit*.

Furthermore, this reward-based practice has been critiqued as only externally motivating (Nicholson, 2015), because the users act merely to get rewards, and not for the action itself. These arguments stem from self-determination theory (SDT) by Deci and Ryan (1985). They differentiate between two types of motivation: Intrinsic motivation, which comes from within the person because of interest or enjoyment, and extrinsic motivation, which can be promoted by outside stimuli, mostly rewards. These two types of motivation are not always separately stimulated but intrinsic motivation without external rewards is superior to external motivation in multiple regards (Deci & Ryan, 1985). For instance, intrinsic motivation

is associated with better performance, creativity and overall being a stronger, long-lasting motivation whereas extrinsic motivation is only present if the provided rewards are attractive (Ryan & Deci, 2000b). If the external rewards are taken away, the motivation can thus come to a halt (Ryan & Deci, 2000b). Intrinsic motivation is deeply connected to the three basic psychological needs: autonomy, competence, and relatedness. In the present master thesis, only autonomy and competence will be of interest. Autonomy means the basic need for choice and free will, which external rewards can impair if they are experienced as too controlling. Competence refers to the feeling of being in control of the activity and experiencing the sense of mastery (Ryan & Deci, 2000b).

The self-determination theory and corresponding research deliver reasonable concerns about using reward-based motivation systems, as they might harm intrinsic motivation. In an experiment with college students, Deci (1971) observed how monetary rewards impair intrinsic motivation in a puzzling task. Students who were granted rewards showed significantly less motivation to engage in the task after the reward was taken away when compared to the control group which never received rewards. This effect was called “undermining effect” (Ryan & Deci, 2000b). Numerous experiments have replicated and extended these results as shown in a meta-analysis by Deci et al. (1999). There, it is distinguished between tangible rewards such as money and intangible rewards such as praise. This meta-analysis showed how tangible rewards impair intrinsic motivation, and verbal rewards improve intrinsic motivation with a small to medium effect size. However, in another meta-analysis from the same era (Cameron et al., 2001), results suggested that rewards in general do not harm motivation to perform a task. A negative effect was only observed in very interesting tasks and if the reward was tangible, expected, and loosely tied to performance (Cameron et al., 2001). These inconsistent results of meta-analyses paint an unclear picture of the effects of rewards on intrinsic motivation. So how do these effects transfer to gamification, and can these theories be used to argue against applying

gamification? A closer look on gamification research and its effects on motivation will be taken in the next section.

Motivation and Gamification

The results of recent meta-analyses show a small to medium effect of gamification on overall performance (Huang et al., 2020; Sailer & Homner, 2019). Both meta-analyses measured performance as cognitive learning outcomes (the knowledge of facts, principles, and concepts as well as procedural, strategic and situational knowledge). Sailer and Homner (2019) further investigated if the use of gamification influenced motivational and behavioral learning. Motivational learning refers, among other motivational components, to intrinsic motivation, and behavioral learning encompasses skills, competences, and performance on a task (Sailer & Homner, 2019, p. 85). In an analysis of all included studies, these measures were significantly influenced by gamification in a positive way. However, when only analyzing studies with high methodological rigor, the before found positive effects on motivational and behavioral learning seemed to diminish and were no longer significant. Therefore, the effect of gamification on motivation and behavioral measures is described as less stable than the effect for cognitive learning outcomes. It is further mentioned that the results of the investigated studies are heterogeneous, meaning the observed effects can differ from study to study in size and direction (Sailer & Homner, 2019). Due to these inconclusive results of the effect of gamification on motivation, further investigation on how rewards work in different settings is needed.

The Original Study

Mekler et al. (2017, 2013a, 2013b) conducted a series of experiments that could not find positive or negative effects of gamification on motivation. These experiments were conducted on the image annotation platform *Tag'em* which was first introduced by Mekler et al. (2013a). *Tag'em* is an online platform where the participants have to come up with words describing the emotional mood in pictures of abstract paintings. *Tag'em* was loosely modeled after a platform created by Wang and Yu (2011). Before starting with the task, the

participants were informed that their tags would improve affective image categorization and thus advance science. This framing was based on the idea of a game which was invented by von Ahn and Dabbish (2008). There, people tagged pictures in a game context, and the tags were later used to improve the google picture search machine. Thus, the task had deeper meaning than just being entertaining and engaging in the task had a real impact. The tags in Mekler et al.'s studies did not have a real purpose, however this positive framing of the task was found to improve tag quality in the first experiment (Mekler et al., 2013a).

These studies examined how the game elements points, levels, and leaderboards positively influenced performance in this gamified system. Motivation and the basic psychological needs autonomy and competence, however, were not influenced by the external rewards, which implies that the undermining effect did not take place, contradicting the predictions of SDT (Deci & Ryan, 1985). Because of these results and the present replication crisis as described by Echter and Häußler (2018), the goal of the master thesis at hand was to perform a conceptual replication and extension of one of these studies, of which the second study conducted in 2013 was chosen (Mekler et al., 2013b). This study was selected because, in contrast to the first study (Mekler et al., 2013a), it inquired intrinsic motivation and autonomy, as well as competence need satisfaction, while leaving out the participants goal causality orientation which was surveyed in the third study (Mekler et al., 2017) and would have exceeded the scope of this thesis. Therefore, the selected study (Mekler et al., 2013b) provided a well-rounded opportunity for a replication due to its simplicity and yet elaborated measurement of intrinsic motivation. The following section entails more detailed information about the study design of the chosen study.

Study Design

Mekler et al. (2013b) conducted an online experiment where participants had the task to describe the emotional mood in abstract paintings with one-word tags. Participants were told that the tags helped to improve affective image categorization, giving it a positive frame that should increase the meaning of the task and thus increase intrinsic motivation. There

were three gamified conditions wherein the participants were introduced to a gamification system of either points, points and levels, or points and leaderboards. The control group (*plain*) did not experience any gamification. After tagging the 15 paintings, the participants were asked to inform about intrinsic motivation, experienced autonomy, and competence need satisfaction in a version of the intrinsic motivation inventory (IMI) (Ryan & Deci, 2000a). As in the other studies (Mekler et al., 2017, 2013a), the gamified conditions led to significantly higher performance, measured by counting the tags created. The different conditions did not statistically significantly influence the participants' intrinsic motivation. Participants reported similar intrinsic motivation across all conditions, and there was no effect of gamification on autonomy or competence need found. As said before, these results stand in contrast with self-determination theory and earlier findings (Deci et al., 1999) as no undermining effect was found. However, SDT and its sub-theories provide explanations for these anomalies, which might stem from problems of the study design. The most plausible explanations are expanded on in the following sections.

Interest

A possible explanation for the absence of a negative effect of rewards on intrinsic motivation in the original study might be that the task itself was not intrinsically motivating. Suppose the task was not intrinsically motivated from the beginning. In that case, intrinsic motivation cannot be impaired by gamification, and so it becomes impossible to measure the effects of gamification on intrinsic motivation between the conditions. A low intrinsic motivation might also relate to a generally low level of interest in the task. Intrinsic motivation can be split up into interest and enjoyment (Reeve, 1989). Interest arouses initiation and first exploration of a new activity (Reeve, 1989) and is proven to be connected to intrinsic motivation, but not to extrinsic motivation (Weber, 2003). Enjoyment will ensure persistent and long-term engagement (Reeve, 1989). Since the study was of relatively short duration (15 min), it is assumed that interest played a more prominent role than enjoyment in the rating of intrinsic motivation in the questionnaire. The most apparent flaw which could have

led to low interest is the set of paintings used in the original study. These paintings were taken from a study by Machajdik and Hanbury (2010) concerning affective image classification. In this paper, the paintings are described as consisting only of a combination of color and texture, without any recognizable objects, but the exact origin of the paintings is not indicated (Machajdik & Hanbury, 2010). The images were peer-rated in terms of what emotional category the pictures fit best. Pictures with an inconclusive assignment to an emotional category were omitted. It is not reported how the 15 pictures for the original study were chosen from the 228 pictures of Machajdik and Hanbury (2010). These paintings might be well-fitted for the task of tagging the emotional mood, but it is not known if the paintings were interesting for the participants in the study.

Control

Effects of gamification are often explained in comparison to experiments using monetary, and thus tangible rewards such as in Nicholson (2015). However, tangible rewards are not fully comparable to the elements used in gamification, as these game elements are intangible and do not hold any value in the real world. Therefore, they might act more like verbal rewards instead of a tangible reward. Verbal rewards were found to improve intrinsic motivation in a number of studies (Cameron et al., 2001; Deci et al., 1999). This can be explained using cognitive evaluation theory, a sub-theory of SDT (Deci et al., 1999). Rewards can either be experienced as controllers of the behavior or as an indicator of competence, either improving competence or impairing autonomy. If a person receives verbal rewards or positive feedback, it is much more likely to be experienced as an indicator of competence, which might increase intrinsic motivation. However, these verbal rewards yield much more information than the game elements implemented in the gamification study by Mekler et al. (2013b). Tangible rewards such as money, on the other hand, are expected to be more controlling, which is, in turn, harming intrinsic motivation (Deci et al., 1999). Thus, it remains unclear in which category the elements of gamification really fall, because they

are not tangible but also do not hold as much competence information as a verbal feedback as used in Deci (1971).

Further, it can be differentiated between different types of rewards in terms of how they are distributed. Of particular interest for this thesis are task-contingent and performance-contingent rewards because they will be deployed in this conceptual replication. For example, the points in the gamification conditions of Mekler et al. (2013b) act as a so-called task-contingent reward, i.e., the reward is given for the mere engagement in the task, and does not imply a set amount of tags which have to be reached in order to get the points. One must simply engage in the task and write words; for each word, 100 points are given. Such a reward is likely to be experienced as controlling, because the participants have to work on the task to be rewarded (Deci et al., 1999). At the same time, the points are not a good indicator of competence (Deci et al., 1999), since they are received for each word, no matter how good or bad the word actually fits the picture (Mekler et al., 2013b). Concluding, the points in the gamified conditions can be specified as task-contingent feedback, which is expected to act primarily controlling and thus impair intrinsic motivation.

In the other conditions levels, and leaderboards, additional performance-contingent rewards are implemented. Performance-contingent are rewards that are only given if a specific goal is reached, for example, if a puzzle is successfully solved as in the experiment by Deci (1971). In the experiment by Mekler et al. (2013b), the points were complemented with either levels or a leaderboard where participants could climb up if they performed well. These added mechanics are performance-contingent rewards, as the levels or leaderboard can only be climbed if enough effort is put into the task. Thus, they are more controlling than the points alone as they set a clear performance goal for the participants. Performance-contingent rewards show a strong tendency to undermine intrinsic motivation (Deci et al., 1999). However, they might also be interpreted as positive competence information if the goal is reached (Deci et al., 1999). Concluding, the two conditions levels and leaderboards worked with performance-contingent rewards, and the participants might have reacted to the

game elements in contrasting ways according to their individual interpretation of the rewards, which might also explain the ambiguity of the results of the study (Mekler et al., 2013b).

Aim of This Study

In the present project, a conceptual replication was conducted as a within-subject design experiment with three conditions. The goal of this study was to examine if results differ when two specific changes addressing the above-mentioned problems were applied to the study design of Mekler et al. (2013b). Firstly, a new set of journalistic, and in a pre-study selected pictures was implemented, aiming to increase interest, and thus improving the initial intrinsic motivation of the participants. Secondly, a new experimental condition *controlling* with an expected performance-contingent monetary, and thus tangible, reward loosely tied to performance was created, which is expected to extend the controlling feeling of the task and impair the competence feedback of the platform (Cameron et al., 2001; Deci et al., 1999). The other two conditions were *plain*, which is a replication of the control group of the original study, and *points*, which is a replication of the condition where participants were only introduced to the game element points. It was decided to replicate the gamification condition *points* because this game element was found to be the most common in a recent literature review (Torres-Toukourmidis et al., 2021), and because the lowest score of intrinsic motivation was reported for the points condition (4.54) in the study by Mekler et al. Even though non-significant, this observation might indicate a tendency of the undermining effect predicted by SDT, which made this condition of special interest for this study. Six hypotheses were formulated based on the hypotheses of Mekler et al. (2013b).

Hypotheses

The design, all hypotheses, and analysis procedure were preregistered on the Open Science Framework and is available under <https://osf.io/m8sdj>. For changes to the preregistration see Appendix A.

All hypotheses refer to the measurement at *T2*. Measures of *T1* will only be used for manipulation checks and exploratory analyses. All the hypotheses are directive.

Intrinsic Motivation. Based on SDT theory (Deci & Ryan, 1985), we assume that the gamification acting as external rewards will impair intrinsic motivation. This is close to the hypothesis by Mekler et al. (2013b) which could not be confirmed.

H1: Participants in the *points* conditions report significantly lower interest/enjoyment in the GIMI when compared to the *plain* condition.

Performance And Cheating Behavior. The second set of hypotheses are based on the findings of Mekler et al. (2013b) and examine the effects of the game elements on the quantity and quality of tags generated per condition.

H2a: Participants in the *points* condition generate significantly more tags compared to the *plain* condition.

H2b: Participants in the *points* condition show significantly less cheating behavior (use of nonsensical tags) compared to the *plain* condition.

Controlling. The last set of hypotheses concentrates on the effect of the third condition *controlling*.

H3a: Participants in the *controlling* condition report significantly lower scores in the GIMI subscale interest/enjoyment compared to the *points* condition.

H3b: Participants in the *controlling* condition report significantly higher scores in the GIMI subscale tension/pressure compared to the *points* condition.

H3c: Participants in the *controlling* condition generate significantly more tags compared to the *points* condition.

Methods

To test these hypotheses, an online study with a between-subject design was conducted. Independent variables were the three conditions *plain*, which acted as the control group, *points*, wherein participants experienced gamification with the game element points, and *controlling*, where the game element points was connected to an expected tangible reward loosely tied to good performance. The dependent variables were the performance of the participants, meaning quantity and quality of tags created in the task, and the reported intrinsic motivation as well as need satisfaction from a questionnaire.

Participants

Participants were recruited via e-mail through the university's database, and the study took place on unipark (*Unipark*, 2021) and on the online platform called *Tag'em* (Brühlmann, 2014/2015). Since the effect sizes in the original study (Mekler et al., 2013b) for intrinsic motivation were small ($d < 0.22$) and the ones for performance were medium to large ($d > 0.54$), a power analysis with *G*Power* (Faul et al., 2009) led to the decision that a sample of at least 400 participants was realistic and reasonable for this experiment. Due to an unexpected lack of response to our invitation mails a sample of only 365 participants was reached, of which 291 passed the quality checks (96 male, 189 female, 6 not specified; mean age 39.7 years (SD = 13.1), range 18-82 years). In line with the original study (Mekler et al., 2013b), participants could partake in a raffle to win one of four vouchers for a Swiss online store. This form of reward is called task-noncontingent and is not expected to influence motivation within the task since it is in no way connected with the engagement in the study task (Deci et al., 1999). Due to the third condition, *controlling*, all participants were wrongly informed that the voucher is only 50 CHF instead of the actual 100 CHF. This enabled to present the chance to double the prize in the raffle in the *controlling* condition without disadvantaging participants from the other conditions. All winners, independent of the condition, received 100 CHF vouchers.

Materials


Participants were asked to describe the emotional mood in pictures on the image tagging platform *Tag'em*, which was the same platform as used by Mekler et al. (2013b) with some minor updates. In the *Tag'em* version used here, participants could see a picture under which an input area was shown. A small text invoked participants to describe the emotional mood in the picture with one-word tags. Tags could be separated with either *space* or *enter*. In contrast to the *Tag'em* version in the original study (Mekler et al., 2013b), the pictures were visible until the participant decided to move on to the next picture, instead of flipping over after 5 seconds. Depending on the condition, *Tag'em* was extended with a set of game elements and textboxes (Figure 1).

In the *plain* condition, participants experienced a plain environment, and neither game elements nor the textbox were present. In the conditions *points* and *controlling*, participants were introduced to the game element points as used in Mekler et al.'s study. Thus, they were informed that 100 points are received for each tag, and the total was shown on the right-hand side. Additionally, the participants in the *controlling* condition were informed that the top 10 participants with the most points would get the chance to double their winnings in the raffle. Thus, an increase of the vouchers in the raffle to 100 CHF instead of the 50 CHF earlier stated in the invitation letter was contingent on very good performance. Intentionally, no clear goal was set to ensure that the gratification of success is absent and to create a lack of competence feedback. Additionally, a text box saying that they must perform very well to win double the amount was displayed while tagging (Figure 1; under the points). This resulted in an expected performance-contingent tangible reward loosely tied to performance, which was the only type of reward with detrimental effects on intrinsic motivation in the meta-analysis by Cameron et al. (2001). Since the interpersonal style of administering performance-contingent rewards play a role in how they affect motivation (Deci et al., 1999), special care was laid on the usage of controlling wording such as "you must" to further assure the controlling factor of this condition.

Figure 1

Example of Tagem in the Condition Controlling

Bild



Ihre Punkte

400

Verdoppeln Sie Ihren Gewinn in der Verlosung von 50 CHF auf 100 CHF!
Dafür MÜSSEN Sie möglichst viele Punkte sammeln.

Ihre Stichworte

Here

are

your

tags.

Welche Stimmung wird durch das Bild vermittelt? Bitte schreiben Sie in das Feld so viele **Stichworte** hinein wie Ihnen dazu in den Sinn kommen.

Einzelne Stichworte können Sie mit **LEERTASTE** oder mit **ENTER** trennen.

Nächstes Bild →

Note. For demonstration purposes, four tags were created (“Here | are | your | tags.”), resulting in 400 points.

Like already mentioned, the paintings of the original study (Mekler et al., 2013b) were replaced with a new set of pictures, with the aim of making the task more interesting and thus more intrinsically motivating. These pictures were compared to the original paintings in a pre-study.

Pre-Study

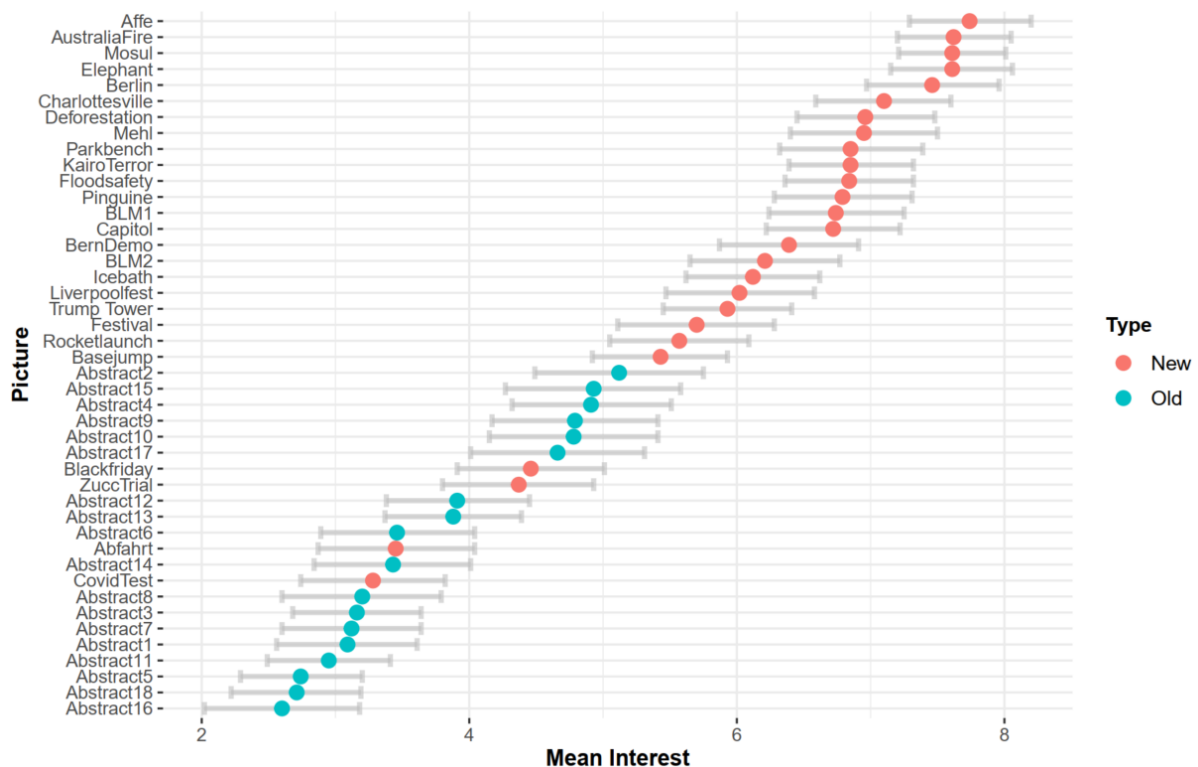
In preparation for the pre-study, new potential pictures were searched for on the world wide web. It was decided that current topics as well as journalistic pictures will be a good fit as they were assumed to be of high interest to a wide range of participants. Specific requirements were defined for a picture to be fitting for this study. It had to obtain a certain level of complexity, leave space for interpretation, and be likely to emit emotions in the

viewer while not being too cruel or shocking due to ethical considerations. The result was a collection of 26 Pictures found on web pages like *World Press Photo* (2021) and *The New York Times* (2021) which mainly were journalistic pictures with fluctuating actuality, such as Berlin after the second world war or the trial of Mark Zuckerberg. Since only 18 pictures were needed for the study and to find out if they were indeed more interesting than the original paintings, a pre-study was conducted.

In this within-subject online study, the new pictures and the paintings used by Mekler et al. (2013b) were shown to 83 students at the university of Basel in randomized order. The students rated their level of interest for each picture on a 10-point-scale ranging from “not at all interesting” to “very interesting”.

Figure 2

Ratings of New and Old Pictures With 95% Confidence Interval



Note. Most of the new pictures were rated as more interesting than the old paintings.

When looking at the individual pictures, the mean rating of the new pictures was higher in almost all accounts, except for the pictures “Blackfriday”, “ZuccTrial”, “Abfahrt” and “CovidTest” (Figure 2). The main study was conducted with the 18 highest-rated new pictures, and all other pictures and old paintings were not further used.

Measures

Intrinsic motivation, autonomy, competence, as well as the pressure and tension as experienced by the participants was measured with a questionnaire. The questionnaire consisted of the intrinsic motivation inventory (IMI) by Deci and Ryan (1985), translated to German by the author, complemented by a German short version which will be referred to as “KIM” (Kurzsкала intrinsischer Motivation) (Wilde et al., 2009). This procedure was chosen due to the unobtainability of the questionnaire used in the study by Mekler et al. (2013b) and a lack of a full German version of the IMI. Further, it allowed analysis of both the KIM items as well as the newly translated items of the IMI in an exploratory factor analysis.

All items of the IMI were translated to German with the help of *DeepL* (2021) and the translated versions were then compared to the items of the KIM. If two items were alike, the item of KIM was preferred. Both items were included if the items did not match, leading to a surplus of items in the subscale choice. In this subscale, the IMI items focused on the freedom of choice of partaking in the task, whereas the items of KIM focused on the freedom of choice within the task. Thus, the meaning of this subscale was different between the two original questionnaires. The final questionnaire consisted of 25 items of four subscales. The subscales were *interest/enjoyment*, measuring intrinsic motivation, *competence*, being the measure of competence need satisfaction, *choice*, measuring the autonomy need satisfaction and *pressure/tension*, which was of interest for the *controlling* condition because of the increased control and external pressure put onto the participants. This new questionnaire will be referred to as German IMI, short “GIMI”. All subscales and questions can be found in Appendix B.

The performance of the participants was also of interest. Thus, performance was measured by tracking the number of tags created per participant in *Tag'em*. To measure performance over time, the chronology of pictures was tracked for each participant. The next section entails more information about the study procedure.

Procedure

As mentioned before, participants were recruited via e-mail, where the study was introduced as an experiment about the perception and classification of pictures. As an incentive to participate in the study, it was mentioned that four vouchers of 50 CHF each will be raffled among all participants. In this mail, they could also find the invitation link to the study. When clicking the invitation link, participants were randomly assigned to one of three conditions: *plain*, *points*, and *controlling*.

First, participants were asked to fill out a brief demographic questionnaire about their age and gender. Participants were then introduced to the image annotation task and informed that their tags would help improve affective image categorization. After a short trial phase of 3 pictures in randomized order, all participants filled out the subscale *interest/enjoyment* of the GIMI questionnaire (T1). The pictures in this trial phase (*BerlinBomb*, *BLM2*, *BernDemo*) were selected at random and were the same for all conditions and participants.

In the second phase, the participants tagged the remaining 15 pictures in randomized order. Before starting the task, the participants in the condition *points* and *controlling* were introduced to the corresponding game elements as described in the materials section. After tagging 15 pictures, all participants returned to the questionnaire (T2) and filled out all subscales *interest/enjoyment*, *competence*, *choice*, and *pressure/tension*. At the end of the study, participants had to inform whether they took the study serious and indicate if they wanted to participate in the raffle. For a visual depiction of the procedure see Appendix C.

Changes to the Original Study

There were a few other differences to the study besides the manipulations of the pictures and the third condition, which are described in the following section. Firstly, the pictures in *Tag'em* were shown for as long as participants wanted to tag them instead of 5 seconds per picture as in the study by Mekler et al. (2013b). It was decided to remove this time limit because it might cause unwanted memory strain and implement an element of stress, which was undesirable for this conceptual replication. Secondly, the questions in the questionnaire (IMI and KIM) were not the same due to the unobtainability of the questionnaire used in Mekler et al.'s study. Lastly, a tweak in the procedure was implemented where participants had to fill out the *interest/enjoyment* subscale of the questionnaire after the trial phase and before the main task, which is called *T1*. This allowed for pre- and post-manipulation analysis and thus enabled observations of the participant's intrinsic motivation over time.

Analysis Plan

As stated in the preregistration, for most hypotheses a t-test was calculated. To test normality, a Shapiro-Wilk Test was calculated. If normality was not given, a Wilcoxon rank-sum test was calculated. If results of the Wilcoxon test are reported, the statistics Z , p value as well as the effect size r and its 95% confidence intervals are reported. If the significance results of the Wilcoxon test were the same as in a t-test, only the results of the t-test are reported with the t statistic, degrees of freedom, p value, as well as the effect size Cohen's d and its 95% confidence intervals. The only exception is hypothesis H2a concerning cheating behavior, where a chi-square test was performed as in the original study (Mekler et al., 2013b). In this case, F and the p value are reported.

Since the effects of gamification on intrinsic motivation were not significant in the study by Mekler et al. (2013b), it was decided to follow the two one-sided tests (TOST) procedure to perform an equivalence test (Lakens et al., 2018) for hypothesis $H1$. This test enables to find out if the difference was merely insignificant or if the results are close enough

to be called equivalent. To conduct such a test, an equivalence bound is needed. If this bound is exceeded, results are not equivalent. To set our equivalence bounds, a smallest effect size of interest (SESOI) had to be set. For replications, it was proposed to set the SESOI to an effect size which the earlier study had 33% power to detect (Simonsohn, 2015). Thus, the equivalence bounds were calculated using *G*power* (Faul et al., 2009) and using the results of Mekler et al. (2013b), a SESOI of $d = 0.25$ was determined to be the equivalence bounds used in this analysis. For all statistical tests, an alpha level of 5% was used. All analyses were calculated in R (R Core Team, 2018) and for Figure 3 and Figure 6, the R package *ggstatsplot* was used (Patil, 2021).

Results

Exploratory Factor Analysis

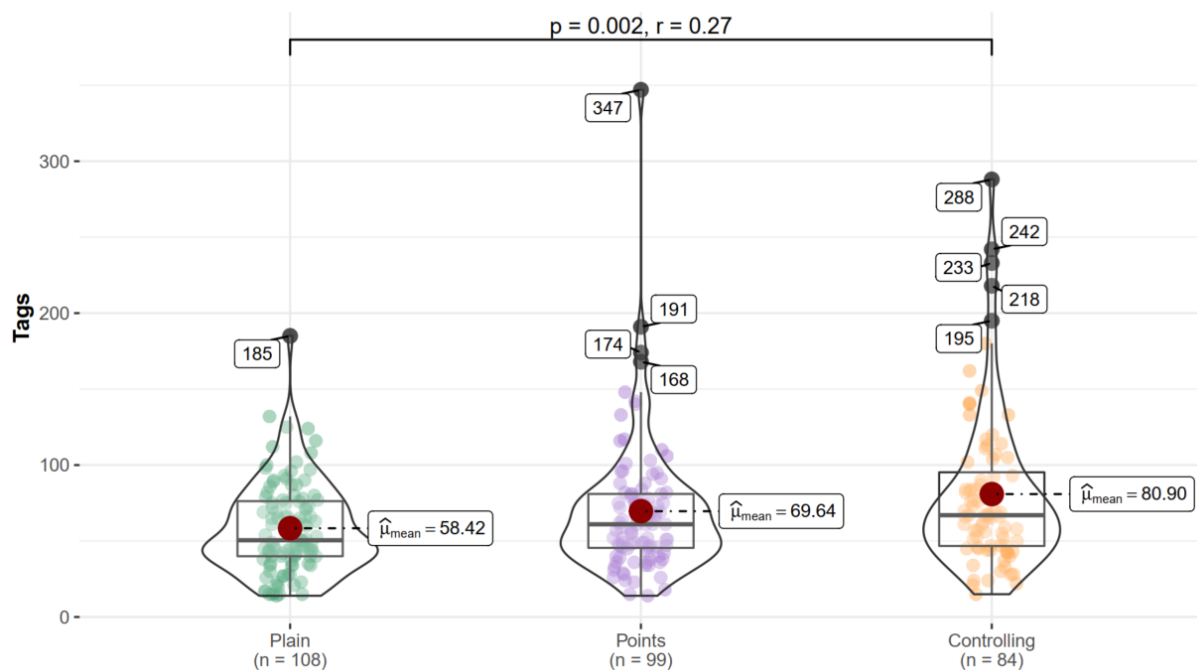
An exploratory factor analysis (EFA) was performed using the *psych* package in R (Revelle, 2021) to evaluate the translated questionnaire (GIMI). It showed that most items performed well with loadings of 0.44 to 0.9 on the main factor except for three items. The items «Das Beschreiben der Bilder hat meine Aufmerksamkeit überhaupt nicht erregt.» from the IMI (interest/enjoyment), «Ich hatte Bedenken, ob ich das Beschreiben der Bilder gut hinbekomme.» from KIM (pressure), and «Ich habe die Bilder beschrieben, weil ich es wollte.» from IMI (choice) were omitted due to cross loadings larger than 0.3. In a second run of the EFA, no items had to be excluded. The items of the subscale *choice* loaded on different factors based on IMI or KIM, but there were no large cross-loadings between the two factors. This might be due to the different focuses of the two questionnaires in this subscale where the IMI items focused on the freedom of choice of partaking in the task, and the KIM items focused on the freedom of choice within the task. This subscale was split up to see if it influences the results, however, no significant difference was found between the two questionnaires ($F(1/288) = 1.86, p = .173$), and thus it was combined for the exploratory analysis (Figure 6). The subscale *choice* was not used for any confirmatory test. All

remaining items were included, and analyses were conducted with the mean values of each subscale. See Appendix D for a detailed table of the EFA results.

Performance

Looking at *H2a* and the performance differences between the conditions *plain* and *points*, a Wilcoxon test revealed that the difference was not statistically significant ($Z = -1.86$, $p = .063$, $r = 0.13$, 95%CI r [.008, .25]). When calculating a t-test however, the difference in performance between the conditions *plain* and *points* was significant ($t(170) = 2.12$, $p = .036$, $d = 0.3$, 95%CI d [0.05, 0.55]), which can be explained with five moderate and two extreme outliers in the *points* condition. In the study by Mekler et al. (2013b, p. 69), all data were square-root transformed to assure homogeneity of variance. Therefore, it can be assumed that the outliers were not excluded from analysis like in the rank sum tests performed here. Concluding, the results of Mekler et al.'s study could not be replicated and the preregistered hypothesis *H2a* could not be confirmed.

For the hypotheses concerning performance differences between the conditions *points* and *controlling*, a t-test did not reveal significant differences between the two groups ($t(164) = 1.56$, $p = 0.12$, $d = -0.23$, 95%CI d [-0.53, 0.05]). The preregistered hypothesis *H3c* was therefore not supported. Concluding, both preregistered hypotheses concerning the effect of rewards in gamification on performance (*H2a*, *H3c*) were rejected. The next section presents the results of the explorative analysis of performance.

Figure 3*Performance Differences Between All Conditions*

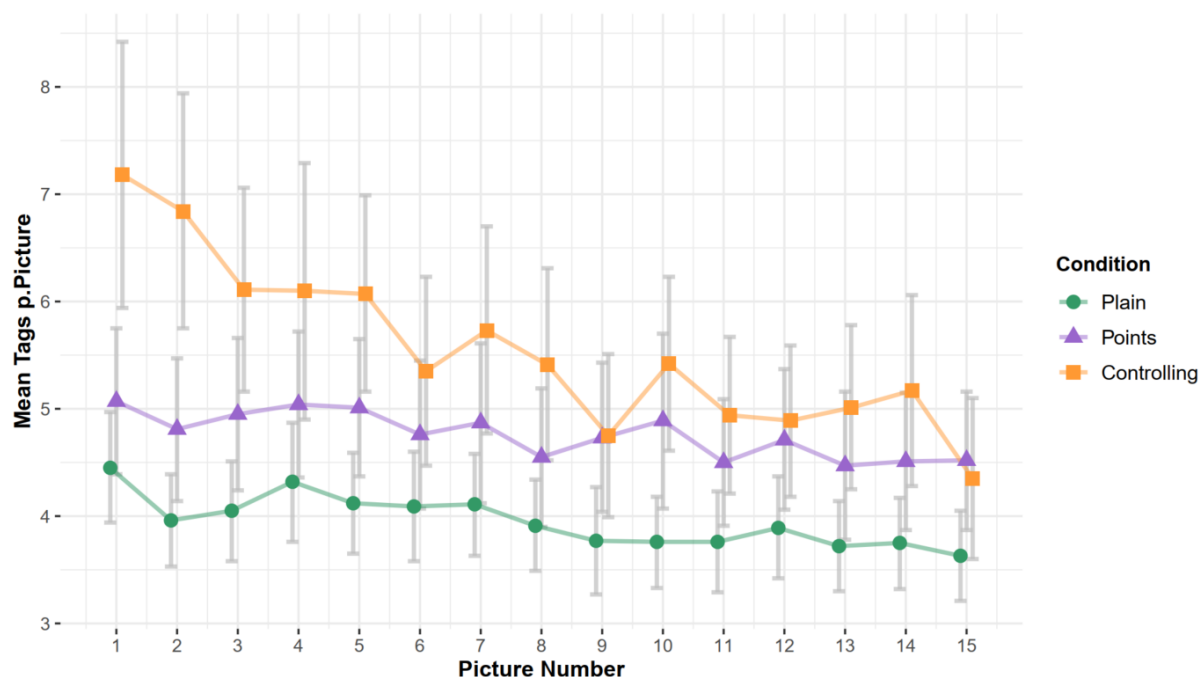
Note. Only significant differences are presented. Reported measures refer to a Wilcoxon test between *plain* and *controlling*.

Explorative Analysis of Performance

Explorative analysis with a Wilcoxon test revealed a significant difference in performance with a small effect between the conditions *plain* and *controlling* ($Z = -3.15$, $p = .002$, $r = 0.23$, 95%CI r [.09, .35]) (Figure 3). Further, an analysis of performance over time was performed with a mixed-effects ANOVA. To measure time, the order of the pictures was split up into three blocks of five pictures each as a within-subject factor. The between-subject factor was condition. A significant effect for condition ($F(2/288) = 7.21$, $p < .001$) as well as a significant effect over time ($F(2/576) = 45.221$, $p < .001$) was found. There was also a significant interaction between condition and time ($F(4/576) = 9.59$, $p < .001$). The ANOVA revealed how performance decreased over time in all conditions like in Mekler et al. (2013b), however, it decreased significantly more in the *controlling* condition, as shown in figure 4.

Figure 4

Performance Over Time With 95% Confidence Intervals



Note. ANOVA revealed a significantly higher decrease in condition *controlling* compared to the conditions *plain* and *points*.

Wilcoxon tests of the difference in performance between the first and last block for each condition supported the findings of the ANOVA, with a significant difference found only in the *controlling* condition ($Z = 3$, $p = .002$, $r = 0.678$, 95%CI r [.54, .79]). The difference in performance between the first and the last block in the other two conditions were not significant (*plain* ($Z = 1.645$, $p = .10$, $r = 0.379$, 95%CI r [.2, .55]), *points* ($Z = 1.34$, $p = .18$, $r = .299$, 95%CI r [.11, .48])). The next section concerns analysis results of cheating behavior.

Cheating Behavior

Similar to Mekler et al. (2013b), all tags were matched with the German and English dictionaries in the package hunspell (Ooms, 2020) to see if people cheated on the task to get more points. Mekler et al. used a similar procedure but worked with another dictionary. When initially looking at cheating behavior, results similar to the outcome of the original study emerged. A chi-square test revealed significant differences between the conditions (χ^2

(2, $N = 291$) = 21.81, $p < .001$), and the percentages for the conditions were comparable to the ones of Mekler et al. (control (8.2%), points (7.4%), leaderboard (6.2%), levels (4.9%)) in the sequence of percentages: *plain* (15.6%), *points* (15%) and *controlling* (13%). Thus, it seemed like cheating behavior decreased when gamification was implemented.

In an additional step, all words marked as FALSE were individually controlled by the author. This investigation revealed that most of the words were merely typos, and the cheating behavior was ill-detected by the dictionaries. A chi-square test of the new and controlled data was still significant (X^2 (2, $N = 291$) = 6.38, $p = .041$), but the percentages were distributed differently to the initially found ones with *controlling* (0.23%) showing the most and *points* (0.07%) the least cheating behavior. The control condition *plain* (0.12%) was the middle ground. To test the preregistered hypothesis *H2b*, a chi-square test between the conditions *plain* and *points* was conducted with the controlled data. The differences in cheating behavior between these groups was not statistically significant (X^2 (1, $N = 207$) = 0.531, $p = 0.47$). Thus, the preregistered hypothesis *H2a* was rejected.

Intrinsic Motivation and Need Satisfaction

Contradicting the first hypothesis *H1*, intrinsic motivation did not significantly differ between the conditions *plain* and *points* at *T2* ($t(204) = -.006$, $p = .995$, $d < 0.001$, 95%CI d [-0.26, 0.28]), and a TOST showed that the observed effect of gamification on intrinsic motivation was statistically equivalent to zero ($t(205) = -1.79$, $p = .037$). Concluding, the results concerning intrinsic motivation are similar to the results of Mekler et al. (2013b) and the preregistered hypothesis *H1* was not confirmed.

Similarly, reported intrinsic motivation did not differ significantly between the groups *points* and *controlling* ($t(177) = 0.3$, $p = 0.77$, $d = 0.04$, 95%CI d [-0.25, 0.33]). A TOST revealed non-equivalence between the two groups ($t(182) = -1.437$, $p = .076$). Thus, *H3a*, stating that participants in the *points* condition report higher intrinsic motivation than the *controlling* condition at *T2*, is also rejected.

When looking at *H3b*, reported pressure did not differ significantly between the two experimental groups *points* and *controlling* ($t(174) = 0.33, p = 0.74, d = 0.05, 95\%CI d [-0.22, 0.34]$). Thus, the preregistered hypothesis *H3b*, expecting that pressure would be greater in the *controlling* condition when compared to the *points* condition, was rejected.

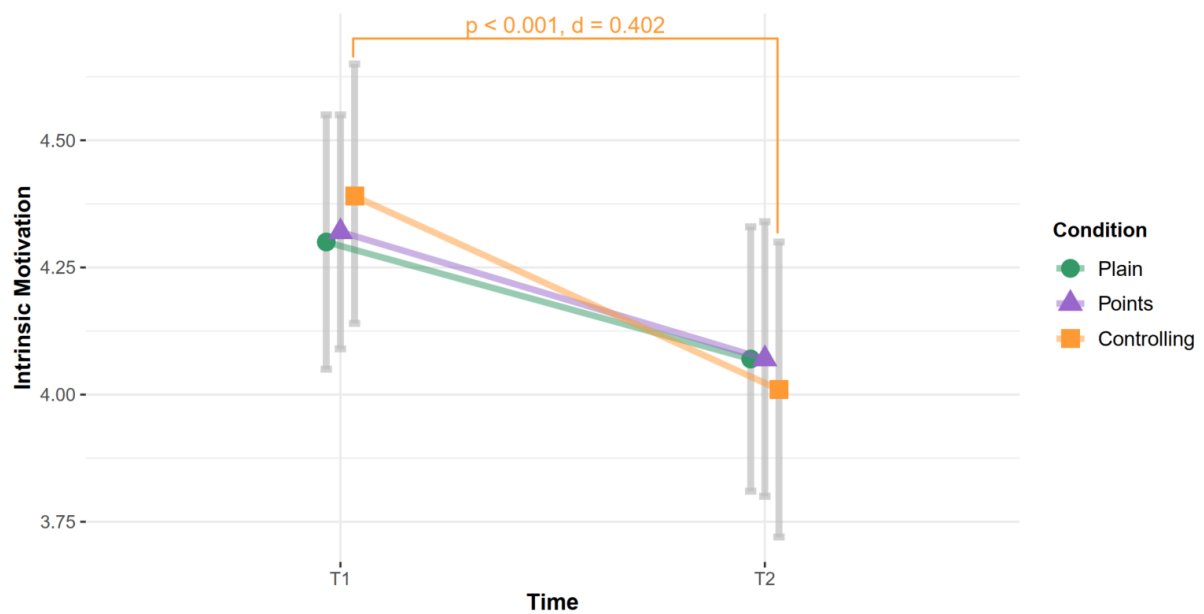
Explorative Analysis of Intrinsic Motivation

In the explorative analysis, the reported intrinsic motivation at *T2* of the *controlling* and *plain* conditions were tested for differences, and the test was not significant ($t(182) = 0.29, p = 0.77, d = 0.04, 95\%CI d [-0.24, 0.31]$). Results of a TOST between the conditions *plain* and *controlling* revealed that the difference, even though not significant, was not equivalent to zero ($t(191) = -1.47, p = .071$), so the *controlling* condition seems to have influenced intrinsic motivation to a small degree which could not be discovered by the t-test.

Explorative analysis considered the differences of intrinsic motivation for each condition between *T1* and *T2*. A mixed-effects ANOVA with time (*T1/T2*) as within-subject factor revealed a significant main effect of time ($F(2/288) = 26.8, p < .001$) but the effect for condition ($F(2/288) = 0.005, p = .994$) and the interaction between time x condition ($F(4/576) = 0.667, p = .51$) were both not significant. In other words, intrinsic motivation decreased significantly in all conditions between the survey after the trial phase *T1*, and the survey after the main phase *T2*. This decrease was similar for all conditions. Yet, individual repeated measures t-tests were conducted for each condition and revealed interesting results. All were significant (*plain* ($t(107) = 2.99, p = .003, d = 0.29, 95\%CI d [0.1, 0.47]$), *points* ($t(98) = 2.34, p = .021, d = 0.24, 95\%CI d [0.03, 0.47]$)), with *controlling* yielding the highest effect size ($t(83) = 3.68, p < .001, d = 0.402, 95\%CI d [0.21, 0.63]$) which can be seen in Figure 5.

Figure 5

Differences of Intrinsic Motivation Between T1 and T2, With 95% Confidence Intervals

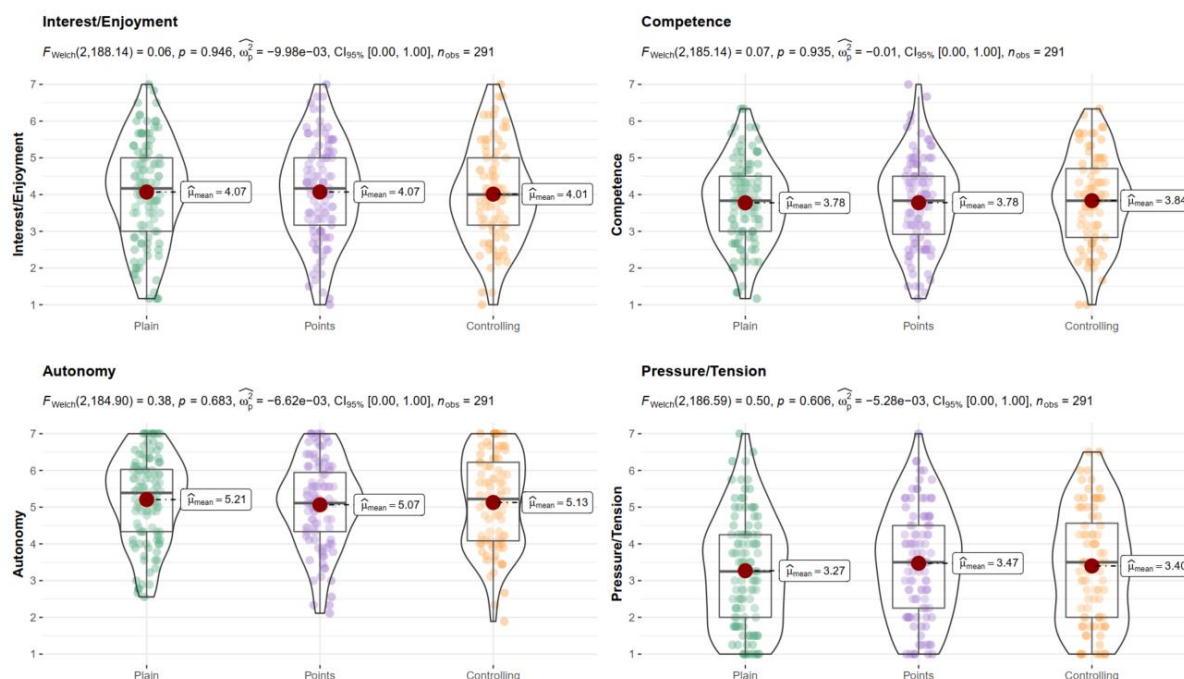


Note. Repeated measures t-tests revealed significant differences in all conditions, with *controlling* showing the most substantial effect (highlighted in orange).

Further, it was investigated if gamification influenced the other subscales of the GIMI. ANOVAs of each subscale did not reveal any significant differences between the conditions. Therefore, the experimental conditions failed to influence the participants feeling of *competence*, *autonomy*, and *pressure/tension* in this task in a statistically significant way as shown in *figure 6*.

Figure 6

All IMI Subscales Between all Conditions at T2



Note. Autonomy is composed of the subscale *choice* of both the IMI and KIM since no meaningful difference between the two questionnaires was found (see table 1).

Discussion

This project had two goals: the first was to determine if the game element points might impact intrinsic motivation to a higher degree when using a more interesting set of pictures in the image annotation task. The second goal concerned the assumption that a more controlling condition would more closely resemble the predictions of SDT about an impairing effect of gamification on intrinsic motivation. This study mainly focused on the effect of rewards in the form of points, which are often implemented in gamification. Since gamified systems can encompass much more than just points (Deterding et al., 2011), this study must be interpreted as an isolated investigation of rewards in gamification. Thus, this study classifies as gamification research just as the study which was replicated and extended upon here. Since research in this field is rather young and diverse, a concise

oversight of the field's challenges (Seaborn & Fels, 2015) and how they were addressed by this study will be given in the next section to enable a more integral understanding of the meaning of this project.

Two central challenges presented by Seaborn and Fels (2015) are the study design, which is often lacking control groups or isolation of the gamification effect, and the missing statistical analysis of the effects in gamification research. The experimental design of this study enabled detailed statistical analyses of the isolated effects of game elements. According to the survey by Seaborn and Fels, there seems to be a disconnect between the theories and the actual application of gamification. Addressing this challenge in the present study, the bottom-up method of creating the condition *controlling* based on SDT and experiments like Deci (1971) enabled new and theory-driven insights. The simplicity of the experiment and its conditions enabled isolated investigation of the game element points and thus addressed the challenge of determining "the usefulness of particular game elements." (Seaborn & Fels, 2015, p. 29). However, some challenges stated by Seaborn and Fels could not be addressed in this project. These are the exploration of new contexts and elements for gamification, as well as the validation of new design approaches to gamification such as the approach by Zichermann (2011) or Nicholson (2015). Nonetheless, this experiment yielded insightful results concerning the effect of the game element points alone, and points combined with an expected tangible reward on intrinsic motivation and performance. The next section entails investigation of each reported effect after a concise overview of the general findings.

Results show how the new set of pictures did not yield different effects of the game element points on intrinsic motivation and performance than the paintings in the original study (Mekler et al., 2013b). In line with previous research by Mekler et al. (2017, 2013a, 2013b), there was no difference in intrinsic motivation measured between the conditions *plain*, *points*, and *controlling* after tagging the pictures (T2). However, a significant negative

effect of the *controlling* condition on performance over time could be observed in the explorative analysis. In the next section, the effect of the new pictures is discussed.

Pictures. When looking at the measures of intrinsic motivation in this study, it becomes evident that the new pictures failed to increase the base intrinsic motivation of the participants. The aim of implementing these pictures was to increase intrinsic motivation in all conditions with the prediction that this will enable the expected negative effects of gamification on intrinsic motivation to show to a stronger degree. However, the mean intrinsic motivation score over all conditions at *T2* (4.05) (Table 1) was lower than the mean score of 4.77 in Mekler et al. (2013b), implying that the newly implemented pictures did not yield higher base intrinsic motivation than the paintings used in the original study. Yet, participants in this conceptual replication created approximately 5 more tags over all conditions compared to the original study, which might indicate a higher informational content of the new pictures. However, this raise in performance could also be due to the picture not flipping over, and the difference is rather small. Concluding, the mere implementation of more interesting pictures was not enough to make the task more intrinsically motivating. Next, the two experimental conditions and the corresponding effects will be discussed.

Points. When looking at performance, participants in the condition *points* generally created more tags when compared to the *plain* condition. However, the difference between the groups was very tightly non-significant, and thus the preregistered hypothesis *H2a* was rejected. Mekler et al. (2013b) found a significant effect of their *points* condition on performance, which was not replicated in this study. Regarding the mean number of tags per group, which was 54.24 for the *plain* condition and 66.01 for the *points* condition in the original study, a similar performance increase was observed between the conditions *plain* (58.4) and *points* (69.6) in this study (Table 1). This leads to our assumption that the condition *points* had a similar impact on the participant's performance in both studies.

Table 1*Mean And Standard Deviation of All Measures for All Conditions*

	Plain (N = 108)		Points (N = 99)		Controlling (N = 85)	
	Mean	SD	Mean	SD	Mean	SD
Tags	58.4 ^c	29.9	69.6	44.3	80.6 ^c	51.9
Competence	3.78	1.47	3.78	1.59	3.84	1.61
Autonomy (IMI) ^a	5.14	1.43	5.03	1.41	5.11	1.44
Autonomy (KIM) ^a	5.33	1.34	5.14	1.38	5.18	1.53
Pressure	3.27	1.81	3.47	1.83	3.40	1.84
IM T1 ^b	4.30 ^d	1.87	4.32 ^d	1.86	4.39 ^d	1.84
IM T2 ^b	4.07	1.83	4.07	1.82	4.02	1.87

Note. All measures other than T1 refer to the questionnaire after the main phase.

^a Autonomy was split into the two questionnaires: Intrinsic Motivation Inventory (IMI) and Kurzskala intrinsischer Motivation (KIM). ^b Refers to intrinsic motivation at the time of measurement. ^c A significant difference in performance was found between the conditions *plain* and *controlling* ($p = 0.002$, $r = 0.23$, 95%CI $r [.09, .35]$). ^d Intrinsic motivation was significantly higher at *T1* in all conditions when compared to *T2* ($p = .014$).

This experiment once again conveyed evidence that points are improving performance without impairing reported intrinsic motivation, nor autonomy or competence in this image tagging task. Therefore, *H1* was rejected. A possible explanation from an SDT standpoint might be that points in gamification do not act as controlling as task-contingent rewards in other contexts. The points might have worked as a competence indicator, and thus motivated the participants to do better and increase intrinsic motivation similar to verbal feedback in Deci (1971). However, if this was the case, the reported intrinsic motivation as well as competence and autonomy need satisfaction in the condition *points* at *T2* should have been higher than in the condition *plain*. In a recent study, achievement-related game

elements such as points were found to be connected with increased competence and autonomy need satisfaction in online communities (Xi & Hamari, 2019). This effect was not found in the present experiment, which could either be explained by the game element points not being a good indicator of competence, or the short timeframe of this study compared to Xi and Hamari (2019). Otherwise, when taking a look at the evidence of this study and the ones by Mekler et al. (2017, 2013a, 2013b), it must be assumed that external rewards can improve performance in short-term tasks without having a substantial effect on the intrinsic motivation of the people engaging in the task. Moreover, these studies show how intangible rewards in gamification such as points present a new type of reward, because they hold less information than verbal feedback but seem to be less controlling than tangible rewards (Deci et al., 1999). The third condition *controlling* investigated the effects of an expected performance-contingent tangible reward loosely tied to performance, and is discussed in the next section.

Controlling. The aim of the controlling condition was to increase the controlling feeling of the task by mimicking the reward of the classic puzzle experiment (Deci, 1971) with an expected performance-contingent tangible reward loosely tied to performance. This type of reward was expected to impair intrinsic motivation to a bigger extent than the points in the *points* condition (Cameron et al., 2001).

Even though participants in the *controlling* condition performed significantly better than participants in the *plain* condition (Figure 3), no significant difference in the reported intrinsic motivation after tagging all pictures (T2) was found. Also, the subscales *choice*, *competence*, and *pressure/tension*, which should have theoretically been negatively influenced by the reward in this condition, did not show any significant difference to the conditions *plain* and *points*. The tangible reward thus had a positive impact on performance but did not impair the participants feeling of autonomy and control and thus did not impact intrinsic motivation and pressure. Therefore, the preregistered hypotheses *H3a* and *H3b* were rejected.

When measuring the total performance in the task, there was no difference between the conditions *points* and *controlling*, which led to the rejection of the last hypothesis *H3c*. However, the results show a significantly larger decrease in performance over time for the *controlling* condition compared to the other conditions (Figure 4). This led to the assumption that the participants in the *controlling* condition were more externally motivated than in the other conditions. Extrinsic motivation is known to be short-lived, and highly dependent on the attractiveness and accessibility of the reward (Ryan & Deci, 2000b). The participants initially performed very well until they thought that the goal of getting into the top 10 was reached. Once their goal was reached, a steep decline of created tags set in, as the performance-contingent reward was no longer motivating. As Zichermann and Cunningham (2011) described, users which are given a reward have to be kept in that reward loop forever to keep up the motivation. Further, the decrease in intrinsic motivation had the largest effect in the *controlling* condition. This indicates that the expected tangible reward had a different and more negative effect on intrinsic motivation than the intangible game element points. Thus, one must be wary of comparing the tangible rewards used in motivation studies such as discussed in the meta-analysis by Deci et al. (1999) to the intangible game elements deployed in gamification as they might have varying effects. In the next section, the measure of cheating behavior will be discussed as it revealed some important inconsistencies compared to the original study's results.

Cheating behavior. Different from the original study (Mekler et al., 2013b), all words found to be false by the spellchecking were individually looked at. Almost all cheat words detected by the spellcheck were revealed to be simply misspelled and not deliberate cheating behavior. The results of the spellchecker resembled the results found by Mekler et al. (2013b). There, participants in the condition *levels* showed significantly less cheating behavior than participants in the other three conditions. Like the spellchecker's results in this study, descriptive analysis of percentages in Mekler et al.'s study showed how cheating behavior was most common in the control condition. After controlling for misspelled words in

this study, however, the amount and the distribution of cheating behavior in the conditions changed drastically and was no longer significant, leading to the rejection of *H2b*. Since the task was replicated to a high degree in this study, it must be assumed that a closer investigation of cheats in the original study would have led to a similar result. In the more recent study by Mekler et al. (2017), a much more sophisticated measure of cheating behavior was deployed, and no significant main effect for game elements was found.

Therefore, it is proposed to treat the results concerning cheating behavior in Mekler et al. (2013b) with caution, as they might not be connected to cheating behavior. This measure might relate more to attention, as less typos might indicate that the participants were more concentrated and attentive. If this was the case, the spellcheck results from the present study indicate that the participants in the *controlling* condition were more concentrated than the participants in the other two conditions *plain* and *points*.

Concluding all results, the predictions about the impairing effect of gamification on intrinsic motivation were partly confirmed with the type of gamification deployed in the *controlling* condition. Intrinsic motivation and performance over time could indeed be negatively influenced if the gamification is connected with an expected tangible reward. In this study, an expected performance-contingent tangible reward loosely tied to performance resulted in higher decrease in performance and intrinsic motivation over time compared to the other conditions. On the other hand, it was shown how the game element points as a task-contingent intangible reward consistently increased performance compared to a control condition without impairing the participant's intrinsic motivation. However, there were several limitations to this study, which are presented in the next section.

Limitations

Firstly, the study's sample ($N = 291$) was smaller than required ($N > 400$) and could have led to a small power to detect effects. This might have been due to the constricted timeframe and timing of the study in the middle of summer.

Secondly, the questionnaire's translation was not precise for all items. Item «Das Beschreiben der Bilder hat meine Aufmerksamkeit überhaupt nicht erregt.» from the IMI (interest/enjoyment) was omitted due to cross loadings and was later detected as different in meaning compared to the English version. The other excluded items also showed cross-loadings, but they were not linked to bad translations in these cases.

Another limitation was the selection of pictures. While they were rated as more interesting than the paintings of the original study (Mekler et al., 2013b), written feedback of the participants in this study revealed that they were partly perceived as burdening and unpleasant due to a prevalent theme of war and misery and thus might have raised negative mood in the participants ($n = 14$). Mood has been shown to influence even the subjective perception of cognitive abilities (Marino et al., 2009), so the negative impact of the pictures might have also led to the decrease of reported intrinsic motivation in all conditions over time (Figure 5). It was also mentioned that the mood in the pictures was redundant, and the same words were used for multiple pictures. For a subsequent study, it might be of interest to collect a more diverse set of pictures.

Feedback from the participants revealed that some had difficulties finding words to describe the mood in the pictures as asked and resorted to describing the content of the picture ($n = 42$). Asking the participants to describe the mood might have also capped the number of possible tags and thus impaired performance. The task to describe the mood in pictures was designed by Mekler et al. (2013a) with the aim of creating a positive frame by informing participants that their work will improve affective image categorization. The same frame and task were used in the present study. Feedback revealed that the task was not straightforward in terms of what in the picture must be described. Future studies might decide to exclude the framing and simply give the task to describe the content of the pictures. This will enable the participants to write more tags and further increase performance differences in the conditions while simplifying the task. Another feedback was that some participants experienced problems with *Tag'em* ($n = 9$). Mostly the lack of a

possibility to go back and correct entered tags was mentioned. Thus, for a next study with *Tag'em*, these usability reports would have to be considered. The next section entails suggestions for further research.

Further Research

The doubt about the image annotation task's ability to promote intrinsic motivation must be considered. Studies might want to further manipulate *Tag'em* or conduct similar experiments on different platforms to inspect the effect of points and other game elements in different contexts as proposed by Seaborn and Fels (2015).

Further research should implement pre- and post-manipulation surveys and maybe even longer study timeframes with multiple or longer sessions. In a longitudinal study over the course of one semester, a decrease in motivation, satisfaction and empowerment was measured in a gamified class compared to a control class (Hanus & Fox, 2015). Students in the gamified condition performed significantly worse in an exam at the end of the semester when compared to students which did not experience gamification. Like the points in the *controlling* condition, the points in this educational study (Hanus & Fox, 2015) were connected to tangible rewards, for example an extension on a paper-deadline. Although it is not reported how often these points were exchanged for rewards, results of the present study indicate that they may have played a central role in the decrease of motivation and performance over time. In another longitudinal study by Koivisto and Hamari (2014), enjoyment of the gamified system decreased over time. The authors suggested that this might be due to a positive novelty effect of gamification, which subsides with time.

Applied to the present experiment, it can be assumed that the performance would have further decreased in the *controlling* condition, together with an ongoing decrease of intrinsic motivation. Of special interest for a longer timeframe would be the *points* condition, as the decrease in performance and intrinsic motivation over time did not differ from the *plain* condition. It is possible that a novelty effect might take place as in the above mentioned study (Koivisto & Hamari, 2014), and thus the performance would diminish over time and

decrease to a level comparable to the performance in the *plain* condition. For a longitudinal extension of this study, a study design similar to the concept of Deci (1971) is suggested, whereas subjects participate in three sessions of the task, with rewards only in the second session. This design would cater to the research agenda of Seaborn (2021) who suggested investigating the effect of removing gamification. With such a design, the effect of tangible and intangible rewards in gamification could be compared in terms of how the performance and intrinsic motivation in the conditions differ over the course of the sessions. Based on the present results and Deci (1971), it could be assumed that the participants receiving a tangible reward in session two would perform worse in the last session and an undermining effect of intrinsic motivation would show, whereas the performance of the participants who received points would only slightly decrease without an impact on intrinsic motivation.

Also, it must be further researched how effects change if gamification is connected to a tangible or monetary reward. Although a survey by Lewis et al. (2016) revealed that many gamification interventions (7/18) implemented tangible rewards, reward-contingency and characteristics are a rarely discussed topic in gamification research. One field study, which investigated tangible rewards in gamification, found that they significantly increased engagement with the online store compared to intangible rewards. No negative long-term effects were found (Meder et al., 2018). Since the present study revealed that performance can decrease significantly faster over time when combining the game element points with a tangible reward, the results seem contradicting and further investigation is needed. Also, tangible rewards are prevalent in web shops such as digitec-galaxus (*digitec*, 2021) and emerged in new videogames in form of cryptocurrencies such as in *Axie Infinity* (2021), and *Decentraland* (2021), indicating that research of these rewards is more relevant than ever.

Concluding, the here presented study delivered new evidence for the specific effects of tangible and intangible rewards in gamification especially over time. Based on the reported results, discussed propositions for further studies might provide deeper insights into

the effect of tangible rewards in gamification, and thus deliver applicable insights for designers of gamified systems.

References

- Axie Infinity*. (2021, November 5). Axie Infinity. <https://axieinfinity.com/>
- Bogost, I. (2011, August). Gamification is Bullshit [Bogost.com]. *Gamification is Bullshit*.
http://bogost.com/writing/blog/gamification_is_bullshit/
- Brühlmann, F. (2015). *Tag'em* [JavaScript]. <https://github.com/psyflo/tagem> (Original work published 2014)
- Burmester, N. (2021). 7 Best Gamification Examples 2021. *Gamify*.
<https://www.gamify.com/gamification-blog/7-best-gamification-examples-2021>
- Cameron, J., Banko, K. M., & Pierce, W. D. (2001). Pervasive negative effects of rewards on intrinsic motivation: The myth continues. *The Behavior Analyst*, 24(1), 1–44.
<https://doi.org/10.1007/BF03392017>
- Decentraland*. (2021, November 5). <https://decentraland.org/>
- Deci, E. L. (1971). Effects of externally mediated rewards on intrinsic motivation. *Journal of Personality and Social Psychology*, 18(1), 105–115.
<https://doi.org/10.1037/h0030644>
- Deci, E. L., Koestner, R., & Ryan, R. M. (1999). A meta-analytic review of experiments examining the effects of extrinsic rewards on intrinsic motivation. *Psychological Bulletin*, 125(6), 627–668. <https://doi.org/10.1037/0033-2909.125.6.627>
- Deci, E. L., & Ryan, R. M. (1985). *Intrinsic Motivation and Self-Determination in Human Behavior*. Springer US. <https://doi.org/10.1007/978-1-4899-2271-7>
- DeepL*. (2021, October 13). <https://www.DeepL.com/translator>
- Deterding, S., Dixon, D., Khaled, R., & Nacke, L. (2011). From game design elements to gamefulness: Defining „gamification“. *MindTrek'11*, 7.
<https://doi.org/10.1145/2181037.2181040>
- Digitec*. (2021, November 5). <https://www.digitec.ch/>

- Echtler, F., & Häußler, M. (2018). Open Source, Open Science, and the Replication Crisis in HCI. *Extended Abstracts of the 2018 CHI Conference on Human Factors in Computing Systems*, 1–8. <https://doi.org/10.1145/3170427.3188395>
- Faul, F., Erdfelder, E., Buchner, A., & Lang, A.-G. (2009). Statistical power analyses using G*Power 3.1: Tests for correlation and regression analyses. *Behavior Research Methods*, 41(4), 1149–1160. <https://doi.org/10.3758/BRM.41.4.1149>
- Hanus, M. D., & Fox, J. (2015). Assessing the effects of gamification in the classroom: A longitudinal study on intrinsic motivation, social comparison, satisfaction, effort, and academic performance. *Computers & Education*, 80, 152–161. <https://doi.org/10.1016/j.compedu.2014.08.019>
- Huang, R., Ritzhaupt, A. D., Sommer, M., Zhu, J., Stephen, A., Valle, N., Hampton, J., & Li, J. (2020). The impact of gamification in educational settings on student learning outcomes: A meta-analysis. *Educational Technology Research and Development*, 68(4), 1875–1901. <https://doi.org/10.1007/s11423-020-09807-z>
- Koivisto, J., & Hamari, J. (2014). Demographic differences in perceived benefits from gamification. *Computers in Human Behavior*, 35, 179–188. <https://doi.org/10.1016/j.chb.2014.03.007>
- Lakens, D., Scheel, A. M., & Isager, P. M. (2018). Equivalence Testing for Psychological Research: A Tutorial. *Advances in Methods and Practices in Psychological Science*, 1(2), 259–269. <https://doi.org/10.1177/2515245918770963>
- Lewis, Z. H., Swartz, M. C., & Lyons, E. J. (2016). What's the Point?: A Review of Reward Systems Implemented in Gamification Interventions. *Games for Health Journal*, 5(2), 93–99. <https://doi.org/10.1089/g4h.2015.0078>
- Machajdik, J., & Hanbury, A. (2010). Affective image classification using features inspired by psychology and art theory. *Proceedings of the International Conference on Multimedia - MM '10*, 83. <https://doi.org/10.1145/1873951.1873965>

- Marino, S. E., Meador, K. J., Loring, D. W., Okun, M. S., Fernandez, H. H., Fessler, A. J., Kustra, R. P., Miller, J. M., Ray, P. G., Roy, A., Schoenberg, M. R., Vahle, V. J., & Werz, M. A. (2009). Subjective perception of cognition is related to mood and not performance. *Epilepsy & Behavior*, *14*(3), 459–464.
<https://doi.org/10.1016/j.yebeh.2008.12.007>
- Meder, M., Plumbaum, T., Raczkowski, A., Jain, B., & Albayrak, S. (2018). Gamification in E-Commerce: Tangible vs. Intangible Rewards. *Proceedings of the 22nd International Academic Mindtrek Conference*, 11–19. <https://doi.org/10.1145/3275116.3275126>
- Mekler, E. D., Brühlmann, F., Opwis, K., & Tuch, A. N. (2013a). Disassembling gamification: The effects of points and meaning on user motivation and performance. *CHI '13 Extended Abstracts on Human Factors in Computing Systems on - CHI EA '13*, 1137. <https://doi.org/10.1145/2468356.2468559>
- Mekler, E. D., Brühlmann, F., Opwis, K., & Tuch, A. N. (2013b). Do points, levels and leaderboards harm intrinsic motivation?: An empirical analysis of common gamification elements. *Proceedings of the First International Conference on Gameful Design, Research, and Applications*, 66–73.
<https://doi.org/10.1145/2583008.2583017>
- Mekler, E. D., Brühlmann, F., Tuch, A. N., & Opwis, K. (2017). Towards understanding the effects of individual gamification elements on intrinsic motivation and performance. *Computers in Human Behavior*, *71*, 525–534.
<https://doi.org/10.1016/j.chb.2015.08.048>
- Nestor, G. (2021). Number of Gamers Worldwide 2021/2022: Demographics, Statistics, and Predictions. *Finances Online*. <https://financesonline.com/number-of-gamers-worldwide/>
- Nicholson, S. (2015). A RECIPE for Meaningful Gamification. In T. Reiners & L. C. Wood (Hrsg.), *Gamification in Education and Business* (S. 1–20). Springer International Publishing. https://doi.org/10.1007/978-3-319-10208-5_1

- Ooms, J. (2020). *hunspell: High-Performance Stemmer, Tokenizer, and Spell Checker* (3.0.1) [R code; R]. <https://docs.ropensci.org/hunspell/>
- Patil, I. (2021). Visualizations with statistical details: The „ggstatsplot“ approach. *Journal of Open Source Software*, 6(61), 3167. <https://doi.org/10.21105/joss.03167>
- R Core Team. (2018). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing. <https://www.R-project.org/>
- Reeve, J. (1989). The interest-enjoyment distinction in intrinsic motivation. *Motivation and Emotion*, 13(2), 83–103. <https://doi.org/10.1007/BF00992956>
- Revelle, W. (2021). *psych: Procedures for Psychological, Psychometric, and Personality Research* (R package version 2.1.9) [R]. Northwestern University. <https://CRAN.R-project.org/package=psych>
- Ryan, R. M., & Deci, E. L. (2000a). Self-determination theory and the facilitation of intrinsic motivation, social development, and well-being. *American Psychologist*, 55(1), 68–78. <https://doi.org/10.1037/0003-066X.55.1.68>
- Ryan, R. M., & Deci, E. L. (2000b). Intrinsic and Extrinsic Motivations: Classic Definitions and New Directions. *Contemporary Educational Psychology*, 25(1), 54–67. <https://doi.org/10.1006/ceps.1999.1020>
- Sailer, M., & Homner, L. (2019). The Gamification of Learning: A Meta-analysis. *Educational Psychology Review*, 32(1), 77–112. <https://doi.org/10.1007/s10648-019-09498-w>
- Seaborn, K. (2021). Removing Gamification: A Research Agenda. *Extended Abstracts of the 2021 CHI Conference on Human Factors in Computing Systems*, 1–7. <https://doi.org/10.1145/3411763.3451695>
- Seaborn, K., & Fels, D. I. (2015). Gamification in theory and action: A survey. *International Journal of Human-Computer Studies*, 74, 14–31. <https://doi.org/10.1016/j.ijhcs.2014.09.006>

- Simonsohn, U. (2015). Small Telescopes: Detectability and the Evaluation of Replication Results. *Psychological Science*, 26(5), 559–569.
<https://doi.org/10.1177/0956797614567341>
- The New York Times*. (2021, Oktober 13). What's Going On in This Picture?
<https://www.nytimes.com/column/learning-whats-going-on-in-this-picture>
- Torres-Toukoumidis, A., Carrera, P., Balcazar, I., & Balcazar, G. (2021). Descriptive Study of Motivation in Gamification Experiences from Higher Education: Systematic Review of Scientific Literature. *Higher Education*, 7.
- Unipark*. (2021, Oktober 13). Unipark. <https://www.unipark.com/>
- von Ahn, L., & Dabbish, L. (2008). Designing games with a purpose. *Communications of the ACM*, 51(8), 58–67. <https://doi.org/10.1145/1378704.1378719>
- Wang, J., & Yu, B. (2011). Labeling Images with Queries: A Recall-based Image Retrieval Game Approach. *Inproceedings*, 8.
- Weber, K. (2003). The relationship of interest to internal and external motivation. *Communication Research Reports*, 20(4), 376–383.
<https://doi.org/10.1080/08824090309388837>
- Wilde, M., Bätz, K., Kovaleva, A., & Urhahne, D. (2009). Überprüfung einer Kurzsкала intrinsischer Motivation (KIM). *Zeitschrift für Didaktik der Naturwissenschaften*, 15, 31–45.
- World Press Photo*. (2021, Oktober 13). <https://www.worldpressphoto.org/>
- Xi, N., & Hamari, J. (2019). Does gamification satisfy needs? A study on the relationship between gamification features and intrinsic need satisfaction. *International Journal of Information Management*, 46, 210–221.
<https://doi.org/10.1016/j.ijinfomgt.2018.12.002>
- Zichermann, G. (2011, Oktober 27). Intrinsic and Extrinsic Motivation in Gamification. *Gamification Co*. <https://www.gamification.co/2011/10/27/intrinsic-and-extrinsic-motivation-in-gamification/>

Zichermann, G., & Cunningham, C. (2011). *Gamification by design: Implementing game mechanics in web and mobile apps* (First edition). O'Reilly.

Appendix A

Changes to the Preregistration

The names for the conditions used in the preregistration were changed to ease understanding: “Gamified” in the preregistration is newly called *points*. “Gamified-Controlling” in the preregistration is newly called *controlling*. The name of the control group remains *plain*.

There were no changes made to the procedure of the experiment. In the preregistration, it was stated that all hypotheses would be tested with an independent two sample t-test. For changes to this procedure, see Section *Analysis Plan*.

Appendix B

These are all questions of the questionnaire. Questions were randomized in each subscale. KIM indicates that these questions are from the KIM, all other questions are from the IMI. All questions marked with "removed" were removed before the analysis due to cross-loadings in the EFA.

Interest/Enjoyment

1. Das Beschreiben der Bilder hat mir Spass gemacht. (KIM)
2. Ich fand das Beschreiben der Bilder sehr interessant. (KIM)
3. Das Beschreiben der Bilder war unterhaltsam. (KIM)
4. Das Beschreiben der Bilder fand ich ziemlich angenehm.
5. Ich fand das Beschreiben der Bilder langweilig.
6. Das Beschreiben der Bilder hat meine Aufmerksamkeit überhaupt nicht erregt.
(Removed)
7. Während ich mich mit dem Beschreiben der Bilder beschäftigte, dachte ich darüber nach, wie sehr ich es genieße.

Competence

1. Mit meiner Leistung im Beschreiben der Bilder bin ich zufrieden. (KIM)
2. Beim Beschreiben der Bilder stellte ich mich geschickt an. (KIM)
3. Ich glaube, ich war ziemlich gut im Beschreiben der Bilder. (KIM)
4. Ich denke, dass ich im Vergleich zu anderen Teilnehmenden ziemlich gut abgeschnitten habe.
5. Nachdem ich mich eine Weile mit dem Beschreiben der Bilder beschäftigt hatte, fühlte ich mich ziemlich kompetent.
6. Das Beschreiben der Bilder war eine Tätigkeit, die ich nicht sehr gut ausführen konnte.

Choice

1. Ich konnte das Beschreiben der Bilder selbst steuern. (KIM)

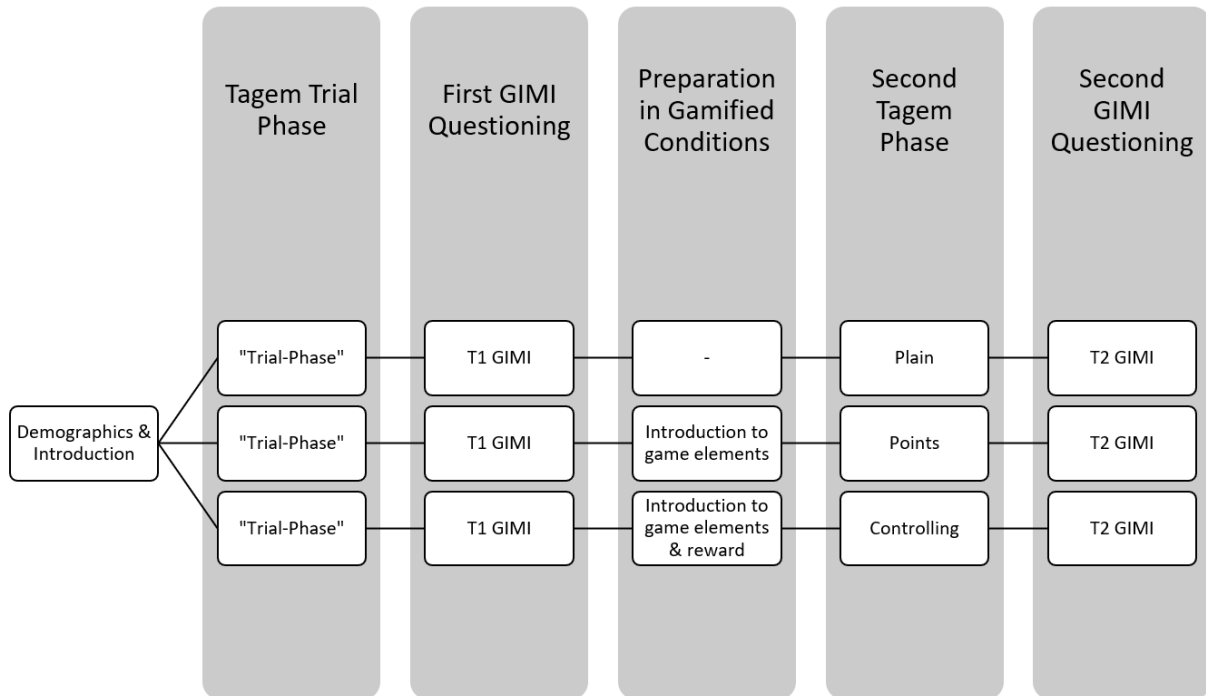
2. Bei dem Beschreiben der Bilder konnte ich wählen, wie ich es mache. (KIM)
3. Bei dem Beschreiben der Bilder konnte ich so vorgehen, wie ich es wollte. (KIM)
4. Ich habe die Bilder beschrieben, weil ich keine andere Wahl hatte.
5. Ich habe die Instruktionen aufmerksam durchgelesen. Um dies zu bestätigen, bitte beantworten Sie diese Frage nicht. (Attention Check)
6. Ich habe die Bilder beschrieben, weil ich es wollte. (Removed)
7. Ich hatte das Gefühl, dass es nicht meine eigene Entscheidung war, die Bilder zu beschreiben.
8. Ich habe die Bilder beschrieben, weil ich es musste.
9. Ich glaube, ich hatte die Wahl, ob ich die Bilder beschreibe oder nicht.
10. Ich hatte nicht wirklich eine Wahl bei dieser Aufgabe.
11. Ich hatte das Gefühl, dass ich das tun musste.

Pressure/Tension

1. Bei dem Beschreiben der Bilder fühlte ich mich unter Druck. (KIM)
2. Bei dem Beschreiben der Bilder fühlte ich mich angespannt. (KIM)
3. Ich hatte Bedenken, ob ich das Beschreiben der Bilder gut hinbekomme. (KIM)
(Removed)
4. Ich habe mich bei dem Beschreiben der Bilder überhaupt nicht nervös gefühlt.
5. Ich war sehr entspannt bei dem Beschreiben der Bilder.

Appendix C

Study Procedure



Appendix D

Table of EFA Results

Item	Number	Factor 1	Factor 3	Factor 2	Factor 4	Factor 5	Origin
t2_enj2	18	0.87	0.01	0.04	0.02	0.01	KIM
t2_enj1	17	0.87	0.03	-0.01	-0.12	0.04	KIM
t2_enj3	19	0.84	0.05	0	-0.01	0.03	KIM
t2_enj5	21	0.75	-0.04	0.11	0.11	-0.12	IMI
t2_enj4	20	0.73	0.13	-0.07	-0.2	0.01	IMI
t2_enj6	22	0.51	-0.01	0.15	0.33	0.02	IMI
t2_enj7	23	0.49	0.11	-0.15	-0.07	0.02	IMI
t2_com3	13	-0.02	0.9	0.04	0.01	-0.04	KIM
t2_com2	12	0.07	0.82	-0.04	0.02	0.05	KIM
t2_com4	14	0	0.81	-0.04	0.05	-0.02	IMI
t2_com1	11	-0.02	0.72	0.11	-0.06	0.09	KIM
t2_com5	15	0.27	0.64	-0.12	0.04	-0.06	IMI
t2_com6	16	-0.07	0.56	0.08	-0.16	-0.03	IMI
t2_press3	26	0.21	-0.34	-0.16	0.33	-0.03	KIM
t2_choi4	5	-0.02	0.07	0.79	0.05	0.03	IMI
t2_choi7	8	0.09	-0.07	0.75	-0.03	-0.03	IMI
t2_choi10	2	0.01	-0.04	0.75	-0.14	-0.06	IMI
t2_choi9	10	-0.03	0.1	0.58	0	0.11	IMI
t2_choi8	9	-0.16	0.06	0.51	-0.1	0.07	IMI
t2_choi5	6	0.31	0	0.51	0.04	0.18	IMI
t2_choi6	7	0	0.08	0.44	-0.12	0.1	IMI
t2_press2	25	-0.01	0.03	0	0.86	-0.02	KIM
t2_press5	28	-0.16	-0.07	0.03	0.72	-0.1	IMI
t2_press1	24	-0.02	-0.01	-0.24	0.63	0	KIM
t2_press4	27	-0.01	-0.01	-0.08	0.62	0.02	IMI
t2_choi2	3	-0.02	-0.03	-0.04	-0.02	0.82	KIM
t2_choi3	4	-0.03	0.02	0	-0.01	0.76	KIM
t2_choi1	1	0.11	0	0.17	0.04	0.55	KIM

Note. Measures are sorted by loadings on the factors. Lines indicate change of the subscale.