Do Hexad User Types Matter? Effects of (Non-) Personalized Gamification on Task Performance and User Experience in an Image Tagging Task

MAXIMILIAN ALTMEYER, German Research Center for Artificial Intelligence (DFKI), Germany
BERINA ZENUNI, Saarland University, Germany
HANNE SPELT, Digital Engagement, Cognition and Behavior Group, Philips Research, Netherlands and Human-Technology Interaction Group, Eindhoven University of Technology, Netherlands
THIERRY JEGEN, German Research Center for Artificial Intelligence (DFKI), Germany
PASCAL LESSEL, German Research Center for Artificial Intelligence (DFKI), Germany
ANTONIO KRÜGER, German Research Center for Artificial Intelligence (DFKI), Germany

The perception of gamification elements differs across users, which is why personalizing gamified systems is important. Past research showed that the Hexad user types model is particularly suitable for this purpose by demonstrating correlations between user types and gamification elements. However, previous studies were mostly survey-based, i.e. relied on participants' rating of gamification elements based on e.g. textual descriptions or storyboards. Thus, the question whether personalization based on Hexad user types provides benefits in implemented gameful systems was neglected. We contribute to this by investigating the effects of (contra-) tailoring the set of gamification elements to a user's Hexad type on performance and user experience, assessed with survey and physiological measures, in an image tagging context. In a lab study (N=29), we found that gamification increases performance and affects psychophysiological measures of arousal. Moreover, we demonstrate that personalization increases enjoyment, positively-valenced affective experiences and participants' absorption in the task at hand.

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

Additional Key Words and Phrases: Gamification; Hexad; Personalization;

ACM Reference Format:
Maximilian Altmeyer, Berina Zenuni, Hanne Spelt, Thierry Jegen, Pascal Lessel, and Antonio Krüger. 2022. Do Hexad User Types Matter? Effects of (Non-) Personalized Gamification on Task Performance and User Experience in an Image Tagging Task. *Proc. ACM Hum.-Comput. Interact.* 6, CHI PLAY, Article 228 (October 2022), 27 pages. https://doi.org/10.1145/3549491

Authors' addresses: Maximilian Altmeyer, maximilian.altmeyer@dfki.de, German Research Center for Artificial Intelligence (DFKI), Saarland Informatics Campus, Saarbrücken, Germany; Berina Zenuni, zenuni.berina@gmail.com, Saarland University, Saarland Informatics Campus, Saarbrücken, Germany; Hanne Spelt, hanne.spelt@philips.com, Digital Engagement, Cognition and Behavior Group, Philips Research, Eindhoven, Netherlands, Human-Technology Interaction Group, Eindhoven University of Technology, Eindhoven, Netherlands; Thierry Jegen, jegenthierry@gmail.com, German Research Center for Artificial Intelligence (DFKI), Saarland Informatics Campus, Saarbrücken, Germany; Pascal Lessel, pascal.lessel@dfki.de, German Research Center for Artificial Intelligence (DFKI), Saarland Informatics Campus, Saarbrücken, Germany; Antonio Krüger, krueger@dfki.de, German Research Center for Artificial Intelligence (DFKI), Saarland Informatics Campus, Saarbrücken, Germany.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.
© 2022 Copyright held by the owner/author(s). Publication rights licensed to ACM.
2573-0142/2022/10-ART228 $15.00
https://doi.org/10.1145/3549491

1 INTRODUCTION

Gamification, *the use of game design elements in non-game contexts* [18], has been investigated in a broad range of domains and contexts, including e.g. education, health, crowdsourcing and commerce [25]. Over the past decade, it has become a thriving field in HCI research, and an established, ever growing practice in industry [56, 77]. For instance, gamification has been used to increase people's motivation to engage in certain behaviors, to enhance the user experience, or to turn rather unpleasant tasks into more engaging ones [25, 56, 77]. At its beginning, gamification research primarily focused on investigating *if* gamification works [56]. Here, many studies found positive effects of gamification when using a "one-size-fits-all" approach (i.e., using a static, pre-selected set of gameful design elements) [25, 77]. However, research has also revealed inconclusive or even negative outcomes of using such a static approach [1, 25, 77]. As a result, more recent work focused on understanding *why* gamification works [56]. It was found that there are interpersonal differences in the perception of gameful design elements [86], which could threaten static gamification approaches. To account for this, personalization of gameful systems is important.

Consequently, research focused on finding and investigating factors that moderate the perception of gameful design elements. Besides demographic factors such as gender [60] or age [5], personality traits [31] and other factors [34] were also found to be relevant in this context. This led gamification research to increasingly focus on investigating user preferences and individual differences in gameful systems [34]. As part of this development, Marczewski [44] proposed the Hexad user types model. This model has been developed to conceptualize and explain user preferences in gameful systems [62, 84]. It consists of six user types, which differ in the degree to which they are driven by their needs for autonomy, relatedness, and competence (the basic psychological needs postulated by Self-Determination Theory [73]). Tondello et al. [86] developed and refined a questionnaire for the assessment of Hexad user types, and demonstrated its reliability and validity [83]. This enabled researchers to use the Hexad user types model to explain user preferences in gameful systems across various domains, including education [53], physical activity [2], energy conservation [37] or health [62]. These studies revealed connections between the *perception* of gameful design elements and Hexad user types. When comparing the Hexad model to other factors and player typologies, it was found to be advantageous in explaining interpersonal differences in the perception of gameful design elements [23].

However, previous research investigating correlations between preferences for gameful design elements and Hexad user types mostly relied on non-interactive materials [34]. This means that participants had no chance to interact with gameful applications, but instead rated their *perception* based on e.g. textual descriptions or storyboards. Consequently, the actual *effects* of personalization based on Hexad user types on behavioral or psychological outcomes, when allowing users to interact with gameful applications, are not well-studied. More specifically, the practical relevance of personalizing gameful design elements based on Hexad user types in gamification research has not been demonstrated thus far, i.e. the question of whether personalizing gameful design elements based on Hexad user types leads to measurable effects on performance and user experience has not been extensively studied.

We aim to contribute answers to these intertwined questions and thereby advance gamification research. We implemented a gamified image tagging platform which allows personalizing the set of activated gameful design elements to the user. In a lab experiment (N=29), we studied the effects of tailored and contra-tailored versions of the platform on task performance, enjoyment, flow, affective experiences and psychophysiological reactions, compared to a non-gamified control condition. Our contribution is three-fold. First, we replicate previous research [41, 50, 75] by showing that gamification – independent of personalization – leads to positive effects on the performance of

users, i.e. an increase in tag counts and quality of tags. Second, our findings show that gamification affects psychophysiological measures. Considering that past gamification research has mostly relied on self-reported measures to analyze the user experience, combining both self-reported and psychophysiological measures to assess the user experience contributes a potentially more holistic perspective to the existing body of knowledge. Third, our main contribution is to show that personalizing the set of gameful design elements based on Hexad user types does not seem to affect immediate task performance, but has positive effects on the users' enjoyment, affective experience and can lead to participants being more absorbed in the task at hand. Summing up these findings, we provide empirical evidence for the positive impact of Hexad-based personalization in gameful systems. This supports the practical relevance of the Hexad model and – on a more abstract level – personalization of gameful systems in general.

2 RELATED WORK

In this section, we present related work in the fields of personalization in gameful systems, shed light on findings related to psychophysiological reactions to gameful stimuli and present results in the context of using gamification to motivate solving microtasks (since we use image tagging, in line with previous research [41, 50, 75], which can be seen as a form of microtask). We conclude this section by summarizing the main findings and framing our contribution.

2.1 Personalization in Gameful Systems

Although past gamification research has found mostly positive outcomes [77] when adopting a "one-size-fits-all" approach, neutral or even negative results have also been found [7, 25, 77]. This is not surprising, given that the motivational impact of gamification elements differs substantially across users [6, 84]. To achieve positive outcomes, understanding which factors influence the perception and effectiveness of gameful applications is important and has been investigated in the past. For instance, Jia et al. [31] found that the personality trait "extraversion" positively affects the perception of certain gameful design elements, such as points and levels, in a video-based setup, i.e. a researcher was filmed when interacting with gameful design elements and participants were asked to rate their perception in a survey. In a follow-up study [30], the authors also demonstrated that personality traits can affect the perception of several leaderboard representations, i.e. extroverts perceived leaderboards more positively, independent of their ranking. In this study, storyboards were used to explain the different types of leaderboards. Additionally, Orji et al. [61] studied whether personality traits play a role in explaining interpersonal differences in the perception of persuasive strategies. Participants were asked to rate the perceived persuasiveness of storyboards deploying various persuasive strategies, including virtual rewards, social competition and social cooperation. In line with the findings by Jia et al. [30, 31], the results show that extraversion, agreeableness, and openness explain most of the variance in the perceived persuasiveness of the deployed persuasive strategies. Besides personality traits, demographic factors also have been studied. Birk et al. [5] found that gaming habits and preferences among older adults change with age, i.e. older participants focus more on enjoyment instead of performance. Moreover, Kappen et al. [32] found that personalizing gameful applications for older adults is important, since age-specific challenges and barriers have an impact on the relevance of gameful design elements. In contrast to the aforementioned works, Lavoué et al. [39] used a web platform teaching French spelling to learners to investigate the effect of personalization. They investigated the effectiveness of tailoring gamification elements based on BrainHex [55] player types. In one condition, participants received adapted gamification elements. In a second condition, participants received gamification elements which were counter-adapted, i.e. not suitable for the participant's player type. In a third group, no gamification elements were integrated. They found that among the learners using the platform

particularly frequently, those who received adapted gamification elements spent significantly more time in the learning environment. Furthermore, participants receiving counter-adapted gamification elements reported higher levels of amotivation. However, it should be noted that the BrainHex model was created to be used in games; not in gamified systems. Moreover, it should be noted that the BrainHex model has severe issues regarding its psychometric properties [82]. For instance, Busch et al. [9] found that only two types–Socialiser and Achiever–could be discriminated as part of a confirmatory factor analysis and that the results are not stable over time, i.e. issues related to the test-retest reliability. Although the aforementioned results are useful to personalize gamified systems, none of the factors considered there were specifically developed for this purpose. The Hexad user type model [44, 86] bridges this gap since its purpose is to explain user preferences in gamified systems (instead of games). The model consists of six user types that differ in how much they are driven by their needs for autonomy, relatedness, competence and purpose (from Self-Determination Theory (SDT) [73]):

Philanthropists ("PH") are socially-minded, like to take responsibility, and share their knowledge with other users. Their main motivation is *purpose*.
Socialisers ("SO") are also socially-minded but are more interested in interacting with other users. Therefore, they are mainly driven by *relatedness*.
Free Spirits ("FS") like to explore and act without external control, with *autonomy* being most important for them.
Achievers ("AC") enjoy overcoming challenging obstacles and mastering difficult tasks. They are motivated by *competence*.
Players ("PL") are focused on their own benefits, and are driven by the will to win and earn external rewards. Hence, *extrinsic rewards* are most important for them.
Disruptors ("DI") like to test a system's boundaries and are motivated by triggering *change*, either positive or negative.

Tondello et al. [86] developed a questionnaire to assess Hexad user types, refined it and demonstrated its reliability and validity [83]. Subsequently, the Hexad user types model has been used successfully in various domains. For instance, it was used in the physical activity domain, where a storyboards-based approach was followed to explain gameful design elements [2]. Supporting the relevance of Hexad in explaining user preferences, it was found that a considerable majority of correlations between the Hexad user types and preferences for gameful design elements established in [86] could be replicated. Also, the Hexad model was successfully used in educational settings. Here, Mora et al. [53] found that personalizing learning experiences to the users' Hexad types led to higher engagement, underlining the usefulness of the Hexad model. Besides physical activity and education, the Hexad model has also been used in the context of unhealthy alcohol consumption by Orji et al. [62]. In line with the aforementioned findings, the results show that Hexad user types explain the perceived persuasiveness of strategies, and the reported effects align well to the user type definitions, thus supporting Hexad's applicability. Also, Kotsopoulos et al. [37] investigated the perception of certain gamification elements and correlations to Hexad user types in the context of energy savings at the workplace. The study revealed that the user types can be used to explain preferences towards gameful design elements, since similar correlations as reported by Tondello et al. [86] were found. Importantly, Hallifax et al. [23] compared the suitability of three models for explaining user preferences for gameful design elements, i.e. the BrainHex [55] model, the Hexad model, and the Big-5 personality model [47]. In a user study, they presented storyboards explaining gameful design elements to users and let them rate their perceptions. They concluded that the Hexad model is the most suitable typology for tailoring gameful systems, as most of the results that were found align with the definitions of the Hexad user types.

Recently, the effectiveness of personalizing gameful applications by utilizing the Hexad model has been investigated. For instance, Lopez and Tucker [43] studied the impact of using a recommender system to personalize a gameful application which required participants to perform full body motions to complete physical activity tasks. They found that participants who interacted with the adapted version of the application increased their performance, compared to others. In contrast to our work, only three gameful design elements were implemented. Consequently, e.g.. the Philanthropist user type, representing the most widespread user type, was not covered. Also, the gameful design elements that were implemented (avatar, points, content unlocking) do not holistically represent the underlying needs and preferences of all Hexad user types (e.g. there is no gameful design element covering social relatedness, which is important for Socializers and Philanthropists [86]). Also, the selected task is inherently challenging, which potentially confounds the results, since some user types are particularly motivated by mastering challenges (e.g. Achievers [86]). Moreover, the authors did not measure effects of personalization on user experience. Thus, the question whether personalization increases enjoyment and motivation, the prevalence of flow experiences, and the presence of positively valenced affective experiences, remains open. Reyssier et al. [69] analyzed the impact of single gameful design elements on the motivation of adolescent learners in secondary schools. In particular, the authors investigated whether the effect on motivation is moderated by the initial motivation of learners in the subject studied (mathematics) as well as the impact of Hexad user types. They conclude that both factors – initial motivation and Hexad user types – are important to explain how a gameful design element will affect motivation. In contrast to our work, the study targeted adolescents, for whom the Hexad questionnaire has been shown to perform sub-standard [59]. Also, the study did not specifically focus on comparing tailoring against counter-tailoring the set of gameful design elements based on Hexad user types, which is what we investigate in this paper.

2.2 Psychophysiological Reactions to Persuasive or Gameful Systems

Typically, gamification research relies on self-report measures to capture related variables such as immersion, flow or enjoyment. These measures are administered post-task, and are subjective; therefore they are often subject to introspection [35, 80]. Studying physiology measures can be seen as a complementary approach, since psychological experiences have physiological substrates [12] and affective experiences occur during task execution [35]. People are themselves not always (immediately) aware of these psychology-induced changes in physiology. Thus, assessing physiology might yield a less disturbed measurement of a player's psychological processes. Contemporary technology allows for a high temporal resolution and real-time assessment. Therefore, physiological measures provide a continuous measurement of mental state, without the need to interrupt the user [35, 57, 81].

Physiological assessment might yield additional information, as it is known to reflect both conscious and unconscious affective and cognitive processes [10, 12]. Physiology has proven to reflect (parts of) general psychological processes [12], such as the valence-arousal levels of an affective state [10, 68]: Emotions are not singular states but exist along two continuums with different neurophysiological bases, i.e. a valence neural circuit and an arousal neural circuit. Consequently, each affective reaction results in a unique physiological response. Related psychophysiology work on psychological reactance and persuasion effectiveness can also be relevant for gameful applications. An interaction loses its effectiveness when a person rejects or feels reactant to it [52]. Sittenthaler et al. [78] found that illegitimate restrictions caused an immediate but sustained increase in heart rate ("HR") and skin conductance levels ("SCL"), whereas a delayed physiological response was found in legitimate restrictions, and no response in the control condition. Moreover, Spelt et al. [79]

found that psychological reactance coincided with cardiovascular activity. Susceptibility to a persuasive communication can also be derived from physiological measures [11, 21, 81]. As such, an information system can potentially use physiological responses to understand the user, i.e. affective computing [67]. As arousal levels fluctuate with susceptibility, it is suggested that physiological state or reactivity can be used to personalize technologies by adapting to the user.

Consequently, researchers have used psychophysiological measures in the context of gameful systems, in addition to traditional self-report measures. For instance, physiology has been used to study game experience in several fast-paced first-person shooter games: Nacke et al. [57] demonstrated correlations between electrodermal and facial muscle activity with the scale-based game experience questionnaire. Similarly, Drachen et al. [20] found that self-reported gameplay experience and physiological arousal, i.e. HR and SCL, are consistently correlated across three different first-person shooter games. For instance, the results show that low HR is related to positive affect and achieving the flow state and that SCL is negatively correlated to flow and positive affect. This indicates a general relationship between game experience and psychophysiological reactions when playing first-person shooter games. The relationship between psychophysiological measures and flow experience has been investigated in the context of the popular game Tetris by Harmat et al. [26]. In a laboratory study, participants were asked to play three versions of the game, which differed in their difficulty level (easy, difficult and a condition in which the difficulty level was adapted to the user called "optimal"). Similar to our approach, the authors decided to use a combination of self-reported measures and psychophysiological measures of flow (HR, HR variability, respiratory depth). They found that self-reported flow was negatively associated with low frequency power of heart rate variability and positively associated with respiratory depth. Van Reekum et al. [89] investigated whether appraisal in games has an effect on physiology. In the user study, the authors investigated the effect of goal conduciveness and intrinsic pleasantness on skin temperature, and on electrodermal, cardiovascular and muscle activity. They used a game in which players had to control a spaceship in order to collect crystals while destroying enemies. While it was found that goal conduciveness had an effect on interbeat intervals, skin conductance level and skin temperature, intrinsic pleasantness had less impact on physiological responses. Korn and Rees [36] used gamification elements to guide workers in assembly tasks. The authors found that gamification led to an increased work speed. Moreover, the authors made use of biosignals (electrodermal activity ("EDA") and facial expressions) to assess affective states. They found that joy was detected significantly more frequently in the gamification condition than in the control condition. This finding was also backed by the EDA data, since participants in the gamification condition were constantly aroused while those in the control condition drifted towards boredom. Barathi et al. [4] investigated which measures are suitable to recognize affective experiences using psychophysiological measures in the domain of high intensity exergaming in virtual reality. Also, they were interested in understanding the relationships between affect and these measures. Based on two experiments, the authors identified relationships between eye blinks, gaze fixations, pupil diameter, skin conductivity and affective experiences. In addition, Passalacqua et al. [64] compared the effect of self-set versus assigned goals and feedback on user engagement and performance in a gamified item picking task. In line with our approach, engagement was assessed by combining self-reported and psychophysiological measures (HR, SCL, and electroencephalography) to increase results validity. The authors found that gamification, independent of the type of goal-setting, increased self-reported and psychophysiological measures of engagement and performance. When comparing the type of goal, no significant differences were found in the psychophysiological nor self-reported measures, and only task performance differed significantly.

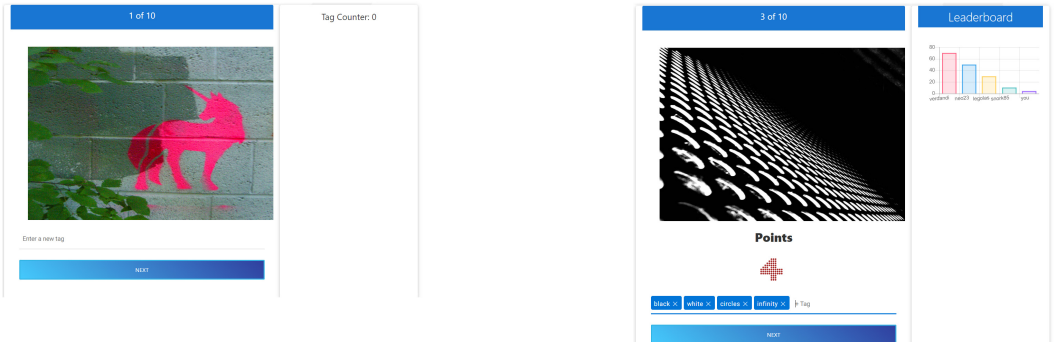
2.3 Gamification to Encourage Solving Microtasks

Solving microtasks, such as tagging images, is a context in which gamification is often employed to motivate users [77]. This context has been frequently used in gamification research, especially for basic research aiming at understanding the effect of specific gameful design elements or gamification strategies [40, 41, 49–51, 75]. This is because this task (image tagging) is considered as sufficiently tedious to benefit from gamification, and it allows one to assess and measure task performance (the amount and quality of tags) in an objective way [51]. For instance, Mekler et al. [49] investigated whether points, levels and leaderboards affect intrinsic motivation negatively, but could not find supporting evidence for this assumption. Their results revealed positive effects of these gameful design elements on the tagging performance (number and quality of tags) of participants, especially for levels and the leaderboard condition. In a follow-up publication [50], which investigates the effects of these gameful design elements on intrinsic motivation and need satisfaction, similar results on the performance are reported, whereas no significant differences regarding motivational aspects were found. Image tagging was also used as a context to investigate the effects of customizable gamification, i.e. allowing users to select which game elements are activated, on task performance in a study by Lessel et al. [40]. The findings revealed beneficial effects on task performance when allowing users to customize the set of gameful design elements. In a follow-up study [41], a similar microtask setting was used to study the effect of choice, i.e. allowing users to turn gamification on or off on the platform, on the amount and quality of tags. Again, beneficial effects on task performance were found. Related to this, Schubhan et al. [75] allowed participants to create their own gamification concepts from scratch and implemented them on an image tagging platform to study whether user-created gamification affects task performance and user experience related measures positively. In line with previous research, they found positive effects on the performance of users when their self-created concepts were implemented for them. Similarly, Tondello and Nacke [85] used an image tagging context to investigate whether users, when offered the choice to select which game elements they would like to activate on the platform, select game elements according to their Hexad user type, and whether this enhances task performance and engagement. Again, the image tagging context proved to be suitable to investigate the main research questions of the paper, and partially positive effects on both performance and engagement measures were found.

2.4 Summary

Related work showed that personalization is beneficial for gameful systems, since individual differences in the perception of gameful design elements influence the likelihood of positive outcomes. The Hexad user types model has proven to be able to explain these individual differences across various domains. Related work has shown that psychophysiological measures may help to enhance the understanding of (affective) user experiences by providing an additional layer to using questionnaires. Furthermore, the context of image tagging was shown to be frequently used for gamification research, due to its suitability for gamification (because of its rather tedious nature) and since it allows measurement of task performance in an objective way.

Based on the aforementioned findings, we contribute to ongoing efforts in understanding the impact of personalization in gameful systems. In contrast to previous works using mostly textual descriptions or storyboards, we implement an actual system where users are able to experience tailored and contra-tailored gamification setups, based on the Hexad model. This enables us to analyze whether the preferences, found using survey-based methods, actually affect the users' behavior and experience. This constitutes the main contribution of this paper. Second, we analyze



(a) Image tagging task without any gameful design elements

(b) Gamified image tagging task, using a leaderboard and points

Fig. 1. Image tagging platform

whether gamification, independent of whether it is tailored or not, leads to increased task performance. This is done to replicate previous research in the image tagging context and thereby back up the validity of our image tagging platform as such. Third, we investigate whether (tailored and contra-tailored) gamification affects psychophysiological reactions. Combining both self-reported and psychophysiological measures of enjoyment and arousal allows us to more holistically analyze the effect of gamification and personalization on the user experience, adding further to the existing body of knowledge.

3 RESEARCH DESIGN

We used an image tagging platform to investigate the effects of personalization based on Hexad user types on task performance and user experience. We decided to use image tagging as a task due to similar reasons as provided by Mekler et al. [49, 51]: First, people engage in human computation tasks (such as image tagging) voluntarily, reducing the threat of contextual factors confounding the study results (compared to e.g. the workplace, a potentially more controlled setting). Second, performance can be easily measured in this setting by counting the number of tags. Third, the task itself is sufficiently tedious to benefit from gamification. Additionally, image tagging has been frequently used in the past to investigate basic research questions in the field of gamification (see Section 2.3), thus allowing us to replicate and build upon these existing findings. User experience entailed enjoyment, affective experiences and flow and was measured by both questionnaires and psychophysiological reactions. Moreover, we aimed to replicate previous research [41, 49–51] by investigating the effect of gamification on task performance. Our user study has a within-subjects repeated measures design, in which participants received three conditions: Control, Tailored Gamification, and Contra-Tailored Gamification. The selection of (un-) suitable gamification elements for the gamified conditions was based on Hexad user types. The study has been reviewed and received ethics clearance through an institutional Research Ethics Committee (blinded for review).

3.1 Apparatus

We implemented a web-based study platform to investigate the effectiveness of personalization of gameful systems based on Hexad user types. We followed the approach of Mekler et al. [49–51].

3.1.1 Image Tagging Task. To ensure comparability, the general task and platform were similar to those used by Mekler et al. [49–51]. To begin with, the platform allowed participants to get

familiar with the tagging task in a tutorial, i.e. allowing participants to add tags for three consecutive images. After completing the tutorial, participants were asked to tag ten images in each of the three conditions (30 in total), appearing one at a time and in a random order. We did not use the images employed by Mekler et al. [49–51]. The authors noted in their most recent study using this platform [50] that utilizing abstract paintings and asking participants to tag emotions makes it hard to objectively assess tag quality. Instead, we decided to consider images that are used for object detection. This allows us to assess the quality of tags in a more objective way, since rating whether a certain object is present on an image is less subjective than rating whether an abstract painting might elicit certain emotions. The participants were shown images from the MIRFLICKR-25000 image collection [28]. This collection consists of 25,000 images downloaded from the social photography site Flickr and has been widely used in machine learning research to train object detection algorithms. Participants were asked to type anything they thought of when seeing the image, and could provide tags in a free text field, separating them by pressing enter. Above every image there was a brief description on how to tag the image. Figure 1a shows the image tagging task.

3.1.2 Gameful Design Elements. We implemented the image tagging platform in a modular way, such that gameful design elements could be activated or deactivated on an individual basis. This allowed for ad-hoc adaptations of the set of gameful design elements depending on the Hexad user type of the participant. We realized five gameful design elements, i.e. badges, points and leaderboard, virtual character, and unlockables, which are described in the following. We made sure that each Hexad user type has at least one suitable gameful design element, based on positively correlated gameful design elements described in the study by Tondello et al. [86]. One exception is the Disruptor, because it is negatively correlated (or not correlated at all) to most gameful design elements [62] making it difficult to find and include suitable gameful design elements. In fact, the Disruptor might also not be as practically relevant as the other user types, since a huge majority of users score lowest in this particular trait [2, 86]. Except for leaderboards, all gameful design elements used three score thresholds, which led to a state change of the corresponding element (e.g. unlocking a badge, changing the mood of the virtual character). These thresholds were based on previous gamification research about image tagging [41, 75]. The thresholds were the same across all gameful design elements to avoid a bias in the tag quantity depending on which elements are activated (the first state change happens after adding 20 tags, the second after adding 45 tags and the third after adding 70 tags across all images). To ensure comparability to Mekler et al. [49–51], who showed five users on their leaderboard, we slightly adapted these thresholds for the leaderboard, without changing the maximum amount to reach the first rank so as not to introduce ceiling effects (in line with all other gameful design elements, the first rank had 70 tags). The second rank had 50, the third 30 and the fourth 10 tags.

Badges. This gameful design element is especially suitable for **Achievers** as it builds on the concept of mastery [44]. Previous research has shown that the perception of Badges is positively correlated to the Achiever user type [86]. On the platform, three different Badges (using the score thresholds mentioned before) can be unlocked: A *bronze badge* could be unlocked after adding 20 tags, a *silver badge* after adding 45 tags and a *golden badge* after adding 70 tags. The badges are shown on the right side of the screen. A progress bar indicates the progress towards unlocking the next badge (see Figure 2).

Points and Leaderboard. Points have been shown to positively affect **Players** [44, 86] and **Socialisers** [2, 62]. Similarly, both user types have been shown to be particularly driven by social competition on leaderboards [44, 86]. In line with Mekler et al. [50], the leaderboard on our platform

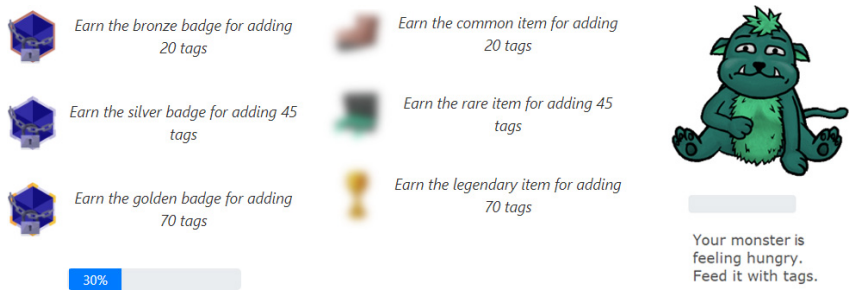


Fig. 2. The gameful design elements Badges, Unlockables and Virtual Character

shows fictitious users with scores similar to the thresholds established before, to ensure that all participants have equal chances to rise in the ranks. For each tag, users received one point. The leaderboard is shown on the right side of the screen, while the user's current amount of points is shown right below the image (see Figure 1b).

Virtual Character. **Philanthropists** are driven by purpose and like to care for others [44, 86]. Although past research has not revealed consistent correlations between the Philanthropist factor and the perception of gameful design elements, we expect that a virtual character should be particularly relevant for Philanthropists. A virtual character may induce feelings of care-taking and stimulate striving for purpose [3]. We used an animated virtual monster whose emotional state is coupled to the amount of tags. The three shifts in its emotional state are based on the score thresholds described before. A progress bar indicates the progress towards reaching the next emotional state of the monster (see Figure 2). The virtual character field was placed on the left side of each image.

Unlockables. Unlockables, i.e. unlocking unknown virtual items, are expected to motivate **Free Spirits** because they are mainly driven by autonomy and curiosity [44, 86]. To realize Unlockables, we provided virtual items on the image tagging platform that can be unlocked by adding tags. Reflecting the score thresholds, there were three items differing in rarity (common, rare, epic). The virtual items were blurred and gradually became more visible when adding tags, with the intention to make users curious and more motivated to explore (which is particularly interesting for Free Spirits [86]). The more tags the user added, the clearer the virtual item would become and the closer the user would get to unlocking it. A progress bar indicates progress towards unlocking a certain item. Unlockables were placed on the right side of each image, as can be seen in Figure 2.

3.2 Conditions

The user study had three different conditions, which all participants took part in. They differed in the type of feedback provided to users while tagging images. The conditions are explained in the following:

Control ("CO"): In this condition, participants were asked to complete the image tagging task while no gameful design elements were activated.

Tailored Gamification ("TG"): In this condition, we activated gameful design elements that correspond to the Hexad user type of the user (as described in Section 3.1.2). If the user scored (equally) high in two or more Hexad user types, all related gameful design elements were activated.

Contra-Tailored Gamification ("CG"): In this condition, gameful design elements were activated that correspond to the Hexad user types that the user scored lowest on (as described in

Section 3.1.2). These elements should be least relevant to the user. If two or more user type scores were equally low, all related gameful design elements were activated.

The Disruptor user type was not considered for assigning suitable or unsuitable gameful design elements (see Section 3.1.2). This is in line with previous research, excluding this user type due to a lack of practical relevance [53]. We decided to activate multiple gameful design elements when a participant had an equal score on their highest or their lowest user type to reflect the traits-based nature of the Hexad model. In case of a conflict, i.e. when a participant scored highest on Player and lowest on Socialiser (or vice-versa), we activated the gameful elements corresponding to their second-lowest score, since Players and Socialisers are both motivated by points and competition [86] and have been shown to be positively correlated [83].

3.3 Procedure

The user study was conducted in a laboratory. 30 participants were recruited via social media and flyers on the university campus. This number of participants was chosen based on an a-priori calculated power analysis, assuming a medium effect size of $\eta_p^2 = .06$, a power of 80% and a correlation among repeated measures of .5, revealing a minimum number of 27 participants. The expected effect size was informed by analyzing the effect sizes of previous research in the same context [50], which were between $\eta_p^2 = .02$ (medium-small) and $\eta_p^2 = .10$ (medium-large) on relevant measures. The study took approximately 60 minutes to complete. Participants were compensated with an 8 Euro Amazon gift card. Upon their arrival at the study site, the procedure was explained to the participants. After giving consent to participate, participants were asked to take a seat in front of a desktop computer. Next, the Empatica E4 wristband [46], a medical-grade wearable device to measure physiological data[1], was put on participants' non-dominant wrist. The Empatica E4 has a photoplethysmography sensor to measure blood volume pulse, an electrode to measure electrodermal activity, a temperature sensor and 3-axis accelerometer. Its validity and reliability has been demonstrated in previous studies [46, 76]. It has also been used in previous studies in the context of games. For instance, Dey et al. [19] used the E4 to share physiological states of players during gameplay. In line with the validation studies, the authors stated that the E4 provided reliable data, which was in line with physiological data from different sources. However, they noted that movement led to noise in the measurements, which was also found in the validation study by Schuurmans et al. [76]. Therefore, it was important for us to choose a task that does not require a lot of movement.

In the task explanation partial deception was used, since we did not want to reveal that the gamification elements were (contra-) tailored to the participants' Hexad user types. We told them that the purpose of the study was to advance the field of image classification and investigate the perception of different feedback mechanisms in this context. This was done to avoid introducing a potential bias due to participants trying to figure out which condition was being presented to them (which might affect their behavior or flow experiences). After the introduction, participants were asked to complete an initial survey, consisting of demographical data and the validated Hexad user types questionnaire [83]. After they filled out this survey, we assessed a baseline of psychophysiological measures. For this, participants were asked to relax while watching a 5-minute video of sea life [63], in the absence of any discrete environmental event/external stimulus. This video has been successfully used in previous research for the purpose of getting baseline measurements of physiology [63, 79]. While participants were watching the video, we prepared the *Tailored Gamification* and the *Contra-Tailored Gamification* conditions based on the results of the Hexad user types questionnaire by activating suitable and unsuitable gameful design elements on

[1] https://www.empatica.com/en-eu/research/e4/, last accessed September 23, 2022

the study platform. Next, participants completed the tutorial consisting of three image tagging tasks (as described in Section 3.1.1). After they completed the tutorial and became familiar with the task itself, the main part of the study followed. Here, participants were asked to tag ten images (one by one) in each of the three conditions. The order of the conditions as well as the order of the images shown to the user were randomized. After tagging ten images, participants were administered a set of questionnaires in each condition, to assess enjoyment, affective experiences and flow. In order to distinguish psychophysiological measures between conditions in the analysis, the study platform stored the current action of the user (e.g. starting/completing an image, watching the relaxation video etc.) in the physiological recording. This allowed us to consider solely the physiological data stored while the user was performing the task and exclude all other measures. After completing all three conditions, the participants were debriefed and the full purpose of the study was revealed.

3.4 Hypotheses

We investigated the following hypotheses:

H1: Task performance differs across conditions
- **H1a:** Tag quantity is higher in gamified conditions than in *Control*
- **H1b:** Tag quantity is higher in *Tailored Gamification* than in *Contra-Tailored Gamification*
- **H1c:** Tag quality is higher in gamified conditions than in *Control*
- **H1d:** Tag quality is higher in *Tailored Gamification* than in *Contra-Tailored Gamification*

H2: User enjoyment differs across conditions
- **H2a:** User enjoyment is higher in gamified conditions than in *Control*
- **H2b:** User enjoyment is higher in *Tailored Gamification* than in *Contra-Tailored Gamification*

H3: The strength of affective experiences differs across conditions
- **H3a:** Positive affective experiences are stronger in gamified conditions than in *Control*
- **H3b:** Positive affective experiences are stronger in *Tailored Gamification* than in *Contra-Tailored Gamification*

H4: The prevalence of flow experiences differs across conditions
- **H4a:** Experiences of flow are more prevalent in gamified conditions than in *Control*
- **H4b:** Experiences of flow are more prevalent in *Tailored Gamification* than in *Contra-Tailored Gamification*

In general, **H1** is motivated by previous research showing that gamification has an impact on the performance of users when tagging images [41, 49, 50]. Specifically, we considered both tag quality as well as tag quantity as indicators of task performance (based on [13]) and expected that gamification (independent of whether it is tailored or not) should increase both (**H1a, H1c**). For tag quantity (**H1a**), this assumption is based on previous research showing that gamification increases the number of tags in an image-tagging context [41, 50]. We further hypothesized that gamification should lead to an enhanced tag quality (**H1c**) since a meta-analysis on performance predictors came to the conclusion that motivation (especially intrinsic motivation, but also extrinsic incentives), which should be positively affected by gamification and goal-setting [42, 49], predicts quality [13]. **H1b** and **H1d** build on the assumption that gameful design elements which are tailored to a users' Hexad type lead to an additional increase on both performance measures, due to previous research showing correlations between user preferences for gameful design elements and their Hexad user type [86].

H2 refers to the enjoyment of tagging images. Based on literature reviews by Seaborn and Fels [77] and Hamari et al. [25], we expected that enjoyment should be improved by gamification (whether it is tailored or not) (**H2a**). We also hypothesized that a set of gameful design elements which is tailored to a user's Hexad type should lead to an increased enjoyment when compared to a contra-tailored set of gameful design elements (**H2b)**, since user preferences [86] should have an impact on the user experience of a gameful system. **H3** follows the same argumentation. We expect that an increased enjoyment is related to positive affective experiences and thus assume that gamification should lead to an increase in positive affective experiences (**H3a**), especially when tailored to the user's Hexad type (**H3b**).

Lastly, flow experiences, which can be defined as "the holistic sensation that people feel when they act with total involvement" [14], are related to optimal task performance [66]. This includes being completely focused on the task and an increased engagement [66]. We deemed analyzing flow experiences as important, since it is seen as "one of the main characteristics of the users experience that might be influenced by gamification" [58] and has been shown to be affected by gamification in the past [58]. Thus, following from **H1–H3**, we expect that flow experiences are more frequent in the gamified conditions (**H4a**) and that personalizing gameful design elements to a user's Hexad type further increases the prevalence of flow experiences (**H4b**).

3.5 Subjective Measures and Analysis

As explained before, the study involved both behavioral and psychological measures. To measure the behavior of users, task performance was used (tag quantity and tag quality). For psychological measures, we were mainly interested in measures to assess the experience of users. *Enjoyment* and intrinsic motivation were measured using the task evaluation questionnaire of the Intrinsic Motivation Inventory ("IMI") [45, 72]. We used the IMI to operationalize motivation in our study, enabling us to investigate whether this measure is affected by gamification and by receiving a tailored versus a counter-tailored set of gamification elements. This is an important psychological measure, since enhancing motivation is commonly seen as the ultimate goal of gamification [74, 77]. The IMI is one of the most widely used instruments in gamification and games research [87]. The task evaluation questionnaire of the IMI consists of four factors: interest/enjoyment, perceived competence, perceived choice, and pressure/tension. While the interest/enjoyment subscale is seen as the self-report measure of intrinsic motivation and the main measure to investigate **H2**, the perceived competence and perceived choice factors are considered positive predictors of intrinsic motivation. Pressure/tension is considered a negative predictor of intrinsic motivation. *Affective experiences* were assessed by using the Positive and Negative Affect Schedule ("PANAS") [90], consisting of two factors: positive affect and negative affect. Measuring affective experiences is motivated by previous research highlighting the importance of measuring both enjoyment and affect in gamification studies [27] and enables us to analyze the effect of gamification and tailoring on emotional responses, adding further insights on the user experience. The Activity Flow State Scale ("AFSS") [65, 66] was used to measure *flow*. It consists of nine factors, each measuring a different dimension of flow. These nine dimension relate back to the work by Csikszentmihalyi [15], who defined them to characterize the flow state. The flow state, which is described as a "state that people report when they are completely involved in something to the point of forgetting time, fatigue, and everything else but the activity itself" [17], is another measure which is deemed an important aspect of the user experience in gamified systems [24]. We used the validated Hexad user type questionnaire [83] to assess participants' *user type*. All questionnaires can be found in the supplementary materials and were analyzed as instructed by the authors of the corresponding instruments [45, 72, 83, 90].

3.6 Physiological Measures and Analysis

Subjective measures relying on the self-report method (asking users to self-report on their perceptions and preferences) are very commonly used [71], also in gamification research [34]. However, self-reports have the drawback that respondents' answers may be inaccurate [71], e.g. because of problems such as social desirability bias (participants answering in a socially desirable manner), acquiescent responding (participants tending to agree with statements) or constraints on self-knowledge [71]. Therefore, we decided to complement the aforementioned subjective measures by psychophysiological measures to get a more holistic picture on how gamification and personalization affect the user experience.

We used features of the electrodermal and cardiovascular system to analyze psychophysiological responses, i.e. flow, enjoyment and affective experiences. The electrodermal system, or sweat gland activity, is solely innervated by the sympathetic "fight-or-flight" nervous system [12]. It was previously used to research games [20, 57], and persuasive messages [81] and is among the most commonly used measures to assess *flow* [35]. An important feature of the cardiovascular system is heart rate variability. The variability in time between heartbeats is caused by an interplay between the sympathetic and parasympathetic nervous system [12]. It is the most commonly used measure to detect flow states [35] and was used previously to investigate the persuasiveness of messages [79, 80] or assess arousal and affective responses in game-related contexts [89]. It was shown that increases in workload or arousal are associated with decreases in HRV [33]. Changes in blood flow beneath the skin, induced primarily by the sympathetic nervous system, lead to changes in ST [70]. It was found that increased ST at the hands is associated with positive arousal [70] and that skin temperature slopes were more positive for conducive events within games [89]. For a full review on the psychophysiology of emotions and cognition see Jänig [29], Kreibig [38], or Posner et al. [68]. Based on previous research, we included the following physiological measures:

Skin conductance level (SCL): includes the tonic component measured in micro Siemens, μS. In our analysis, SCL was operationalized by dividing the average skin conductance in each condition by the average skin conductance while watching the relaxation video [20].

Skin conductance responses (SCR): concerns the phasic component in electrodermal activity, i.e. the number of abrupt increases in the skin conductance (peaks) [8, 12]. SCR was operationalized by counting the number of skin conductance peaks during a condition and dividing it by the minutes it took to complete that condition. Thus, we analyzed the average number of peaks per minute. To count peaks, we used the scientific python package SciPy[2].

Heart rate variability (HRV): was measured based on the inter-beat interval in milliseconds, which was used to calculate the square root of the mean squared successive heart period differences ("RMSSD"), a commonly used HRV statistic [12], for each condition. The RMSSD was normalized by dividing it by the RMSSD measured when watching the relaxation video.

Skin temperature (ST): was measured in °C and normalized by dividing the average measure in each condition by the average measure when watching the relaxation video [89].

Similar to Drachen et al. [20] we used a simple form of normalization, in which we divided the average value of a measure by the average value of the respective measure while watching the relaxation video, except for SCR.

4 RESULTS

To investigate the aforementioned hypotheses, we used repeated measures ANOVAs to compare the dependent variables between the three conditions. When assumptions for the ANOVA were

[2] https://docs.scipy.org/doc/scipy/reference/generated/scipy.signal.find_peaks.html, last accessed September 23, 2022

Participant	PL	SO	AC	PH	FS	DI	Participant	PL	SO	AC	PH	FS	DI
1	26	24	27*	22*	27*	21	**16**	24	24	23*	26*	23*	13
2	17*	28	27	22	28*	28	**17**	25	27*	27*	27*	25*	15
3	19*	27	26	28*	23	12	**18**	28*	21	27	20*	22	16
4	21*	23	23	25*	25*	15	**19**	26*	24	26*	24*	25	20
5	25*	17	23	22	21*	12	**20**	21*	21*	23	24*	24*	19
6	28*	27	27*	28*	27*	16	**21**	19	23	25*	23	19*	12
7	21*	27	24	28*	21*	18	**22**	16*	17	18	23*	22	17
8	23	26*	24	23	20*	15	**23**	19	14*	19*	19*	19*	15
9	27*	19	23	21*	26	13	**24**	16*	24	26	28*	27	23
10	20*	23	24	27*	24	14	**25**	25*	24	24	22	21*	13
11	20*	24	26*	26*	25	21	**26**	26*	17	25	24*	25	13
12	25	27*	25	23	21*	8	**27**	19*	28	28	28*	28*	10
13	17*	18	19	21*	19	12	**28**	24*	18	23	23	22*	18
14	23	24*	22	23	21*	16	**29**	22	26	27*	26	22*	16
15	26*	22	23	15*	21	19							

Table 1. Hexad scores of all participants. Green cells represent particularly high scores in the respective user type, red cells represent particularly low scores. An asterisk marks which user type was considered for the personalization of the gameful design elements.

not met, Friedman tests were used as non-parametric counterparts. When using Friedman tests, the Durbin-Conover method was used for post-hoc analysis. The Bonferroni-Holm method was used in both cases to control the family-wise error rate.

4.1 Participants

Out of 30 participants, one had to be excluded due to technical problems during the study, leading to a total sample size of 29 which was considered for the analysis.

Out of these participants, 10 self-reported their gender as female and 19 as male. Regarding age, 10 participants were aged 18-24 years, 18 participants were aged 25-31 years, and 1 participant was aged 32-38 years. We assessed gaming familiarity with 3 items ("I consider myself as gaming-affine", "I frequently play video games", "I have a passion for video games") with 5-point scales (1=strongly disagree). The means were rather neutral: 3.12, 2.82 and 2.92, respectively. The Hexad user types average scores are similar to the averages reported in the validation study of the Hexad questionnaire by Tondello et al. [83]. Achievers showed the highest average scores (M=24.28, SD=2.52), followed by Philanthropists (M=23.83, SD=3.10), Free Spirits (M=23.20, SD=2.71) and Socialisers (M=22.90, SD=3.80). Players (M=22.34, SD=3.58) and Disruptors (M=15.86, SD=4.13) followed with lower average scores. Table 1 shows the Hexad scores of all participants as well as which user types were highest (marked green) or lowest (marked red). An asterisk marks which user type was considered for the personalization of the gameful design elements (sometimes, the highest/lowest score could not be used for the selection of suitable/unsuitable gameful design elements due to a conflict in the reported preferences in the literature; see Section 3.2).

4.2 Task Performance

Overall, participants provided 3,967 individual tags (1,114 in Control ("CO"), 1,402 in Tailored Gamification ("TG") and 1,451 in Contra-Tailored Gamification ("CG")). Table 2 provides an overview of the mean and median tag count per condition. We compared the average number of tags per

condition and found they differed significantly ($F(2, 56) = 13.56, p<.001, \eta_p^2 = .33$). Pairwise comparisons revealed result **R1: The number of tags in both gamified conditions is significantly higher than in CO** ($p_{\text{holm}}<.001$ each; Cohen's $d_{\text{TG}}=.43$, $d_{\text{CG}}=.50$). When comparing the TG and CG conditions, no significant result was found ($p_{\text{holm}} =.48$).

To analyze tag quality, we followed a qualitative coding process, similar to Mekler et al. [50]. The coding process was conducted by two independent raters who manually inspected each of the 3,967 individual tags provided for the images and rated whether the tag was: neither related to any given object in the image nor captures a specific mood, or was just nonsense (value 1); describes a mood or color scheme that was present in the pictures but not a specific object (value 2) or describes a concrete object in the picture (value 3). After both raters rated all tags, the inter-rater agreement was calculated using Cohen's Kappa κ. The result was $\kappa=.66$, which is considered as substantial agreement, according to Cohen [48]. A more conservative interpretation of this result would indicate a moderate agreement [48]. When the rating differed between both raters, the mean of their ratings was calculated and used for the analysis. The average quality of tags for each condition is shown in Table 2. We compared the average rating per participant across conditions and found that it differed significantly ($F(2, 56) = 3.97, p =.024, \eta_p^2 = .02$). As part of the post-hoc procedure, we found that **R2: The average tag quality is significantly higher in both gamified conditions than in CO** ($p_{\text{holm}}<.05$ each; $d_{\text{TG}}=.35$, $d_{\text{CG}}=.27$). However, no significant difference was found between TG and CG ($p_{\text{holm}} =.82$).

Also, the amount of tags per minute differed significantly ($F(2, 56) = 16.64, p<.001, \eta_p^2 = .37$). We found that **R3: The amount of tags per minute is significantly higher in both gamified conditions than in CO** ($p_{\text{holm}}<.001$ each; $d_{\text{TG}}=.49$, $d_{\text{CG}}=.68$). Also, when comparing the tailored (TG) and contra-tailored (CG) conditions, we found that **R4: The amount of tags per minute in the Contra-Tailored Gamification condition is significantly higher than in the Tailored Gamification condition** ($p_{\text{holm}} =.049$, $d=.23$). Together with **R2**, **R3** suggests that gamification might have helped participants to come up with good tags, since the average time per tag decreased.

4.3 Subjective User Experience

To analyze the user experience, we considered enjoyment or intrinsic motivation, affective and flow experiences. In this section, we report the results of the survey-based measures. An overview of descriptive data can be found in Table 2.

Regarding the IMI factors, we did not find a significant effect for the competence ($F(2, 56) = .65, p =.52$), choice ($F(2, 56) = .32, p =.73$) nor pressure ($F(2, 56) = 1.51, p =.23$) factors. However, the enjoyment factor differed significantly across the conditions ($F(2, 56) = 3.45, p =.039, \eta_p^2 = .11$). While there were no significant differences between the gamified conditions and CO ($p_{\text{holm}} =.28$ each), we found that **R5: Enjoyment is significantly higher in the Tailored Gamification condition than in Contra-Tailored Gamification** ($p_{\text{holm}} =.003$, $d=.29$).

When analyzing the positive and negative affect factors of the PANAS, we found that positive affect differed significantly between the three conditions ($F(2, 56) = 6.39, p =.003, \eta_p^2 = .19$). Post-hoc comparisons showed that positive affect in TG was not significantly higher than in CO ($p_{\text{holm}} =.71$). However, positive affect was significantly higher in TG than in CG ($p_{\text{holm}} =.006$, $d=.50$), leading to **R6: Positive affect is significantly higher in Tailored Gamification than in Contra-Tailored Gamification**. In addition, we found that **R7: Positive affect is significantly lower in Contra-Tailored Gamification than in Control** ($p_{\text{holm}} =.001$, $d=.46$). These results not only show that selecting gameful design elements matching the users' Hexad types leads to increased positive affective experiences, but also that choosing unsuitable gameful design elements

			Control			Tailored Gamification			Contra-Tailored Gamification		
			Mean	SD	Median	Mean	SD	Median	Mean	SD	Median
Perfor.	H1	**Tag Count***	38.41	21.44	31.00	48.34	24.71	46.00	50.03	24.99	45.00
		Tag Qual.*	2.66	0.21	2.71	2.73	0.19	2.79	2.72	0.23	2.79
		Tags/min* .	3.17	1.53	2.92	3.97	1.73	3.98	4.40	2.06	4.25
IMI [summed up]	H2	**Enjoyment***	26.72	10.36	26.00	28.03	10.80	27.00	24.97	9.98	26.00
		Competence	19.69	4.74	19.00	20.76	6.90	20.00	19.90	5.17	20.00
		Choice	23.03	5.33	23.00	23.10	4.88	23.00	22.45	5.21	22.00
		Pressure	13.03	5.05	14.00	14.55	6.49	14.00	14.38	5.45	14.00
PANAS [summed up]	H3	**Pos. Affect***	27.07	8.15	27.00	27.62	8.94	27.00	22.79	10.39	18.00
		Neg. Affect	16.59	4.79	16.00	15.90	5.03	15.00	15.00	3.32	14.00
AFSS [summed up]	H4	**MAA***	9.52	2.72	9.00	10.48	2.68	11.00	9.17	3.11	9.00
		CG	11.03	2.34	12.00	11.59	2.51	12.00	10.52	2.49	10.00
		CO	15.41	2.28	16.00	15.03	3.17	15.00	14.28	3.32	14.28
		UF	6.59	1.82	6.00	7.48	1.72	8.00	6.83	2.28	8.00
		CS	9.28	2.37	9.00	9.79	2.72	10.00	9.34	2.35	9.00
		TT	9.07	2.81	9.00	9.66	3.25	10.00	8.86	3.03	9.00
		CN	7.72	1.36	8.00	7.66	1.59	8.00	7.48	1.62	8.00
		SC	10.52	2.57	11.00	10.59	2.85	10.00	9.83	2.52	9.00
		AE	8.97	2.74	9.00	10.17	3.12	11.00	9.28	3.28	9.00
Psychophys. [normalized]	H2–H4	RMSSD	1.01	0.04	1.00	1.01	0.05	1.00	1.02	0.07	1.00
		SCL	1.22	0.39	1.13	1.33	0.56	1.18	1.34	0.76	1.10
		SCR* [peaks/min]	44.27	14.26	50.34	49.42	9.92	52.40	50.86	7.28	51.85
		ST*	1.02	0.03	1.01	1.04	0.04	1.03	1.04	0.05	1.03

Table 2. Mean, standard deviation ("SD") and median for each dependent variable and condition. Bold, colored entries with * represent dependent variables for which a significant difference across conditions was found.

is worse than having no gameful design elements at all, regarding affective experiences. Concerning negative affect, no significant differences were found ($F(2, 56) = 2.06, p = .14$).

To measure self-assessed flow states, we relied on the AFSS, having nine factors, representing each of the nine dimensions of flow (see Table 2). We found a significant effect on the Merging Actions and Awareness dimension ($F(2, 56) = 3.31 p = .044, \eta_p^2 = .11$). This dimension assesses the extent to which people are absorbed in the task [17]. Based on the pairwise comparisons, we found that there were no significant differences between CO and TG ($p_{\text{holm}} = .15$) nor between CO and CG ($p_{\text{holm}} = .52$) regarding the MAA factor. However, similar to **R5** and **R6**, we found that **R8: Participants seem to be more absorbed in the task at hand in Tailored Gamification than in Contra-Tailored Gamification** ($p_{\text{holm}} = .048$, $d = .45$). This shows that selecting gameful design elements can have an effect on dimensions of flow experiences [16] and that personalizing the gameful design elements to a user's Hexad type can positively affect these experiences. For the remaining factors of the AFSS, no significant differences were found on the clear goals ("CG", $F(2, 56) = 2.36, p = .10$), concentration on task at hand ("CO", $F(2, 56) = 1.98, p = .15$), unambiguous feedback ("UF", $F(2, 56) = 2.45, p = .10$), autotelic experience ("AE", $F(2, 56) = 2.80, p = .07$), challenge skill balance ("CS", $F(2, 56) = .67, p = .52$), transformation of time ("TT", $F(2, 56) = 1.09, p = .34$), sense of control ("CN", $F(2, 56) = .35, p = .71$) nor on the loss of self-consciousness ("SC", $F(2, 56) = .94, p = .40$) factors.

4.4 Physiological User Experience

To complement the survey-based measures, we used physiological measures to assess the user experience in a multi-faceted way. These measures were analyzed using a Friedman test instead of an ANOVA, because the assumption of normality and/or the assumption of sphericity were violated. Table 2 provides an overview of the descriptive data across conditions.

For RMSSD, we found no effects between conditions ($\chi^2(2) = 1.10, p = .58$). Also, no effects were found regarding potential changes in SCL ($\chi^2(2) = .48, p = .79$). However, we found that the number of SCRs differed between the conditions ($\chi^2(2) = 18.83, p < .001$). As revealed by post-hoc comparisons, both gamified conditions showed an increase in SCRs compared to CO, leading to **R9: The number of peaks in the EDA (SCRs) is significantly higher in both gamified conditions than in Control** (p_{holm}<.001 each; rank-biserial correlation r_{TG}=.68, r_{CG}=.72). This shows an increased sympathetic arousal [12] in both gamified conditions and hints at either increased flow states [54] (assuming that the arousal is positively valenced) or increased pressure or tension (assuming that the increased arousal is negatively valenced) [12]. When comparing TG and CG, no significant difference was found regarding SCR ($p_{\text{holm}} = .64$). We also found a significant difference in skin temperature ($\chi^2(2) = 37.72, p$<.001). In line with **R9**, both gamified conditions differed from the CO condition, i.e. **R10: Skin temperature is significantly higher in both gamified conditions than in Control** (p_{holm}<.001 each; r_{TG}=.99, r_{CG}=.78). No effects were found between TG and CG ($p_{\text{holm}} = .83$).

5 DISCUSSION

Our results show that both gamified conditions led to an increase in the amount of tags, compared to the Control condition (**R1**). This supports **H1a: The number of tags is higher in gamified conditions than in *Control***. It shows that gamification, independent of whether it is tailored or not, increases the number of tags in an image tagging context. This is in line with previous research by Mekler et al. [49–51] as well as Lessel et al. [41] and therefore contributes a replication of previous results using a static set of gameful design elements. When comparing the number of tags between the Tailored Gamification and the Contra-Tailored Gamification conditions, we did not find a significant difference. Thus, **H1b: The number of tags is higher in *Tailored Gamification* than in *Contra-Tailored Gamification*** is not supported, given our data. This might be explainable by the fact that all gameful design elements, regardless of their suitability, introduce goals. According to goal-setting theory [42], goals motivate people by introducing a state of tension that activates actions. Also, the experimental setting and the fact that participants were compensated for participating might have led to participants feeling obligated to meet these established goals, independent of the gameful design elements that were activated and their user experience. Therefore, this aspect needs further research and should be investigated in in-the-wild studies over a longer time-span.

Related to tag quality, we found that the average quality of tags was significantly higher in both gamified conditions (**R2**), and that the time users took to create a tag was significantly lower in the gamified conditions (**R3**), both adding support for **H1c: Tag quality is higher in gamified conditions than in *Control***. This finding is explainable by the fact that increases in a user's motivation to perform a task (which likely occur due to the gamification that was used [42, 49]) have been shown to lead to increases in the quality of the task outcome [13]. Moreover, the fact that the mean time to add a tag to an image was significantly lower in both gamified conditions than in CO (**R3**) suggests that participants had to think less about which tags to provide, which might have been caused by the potentially stimulating gamification environment. However, it should be considered that previous work in the same context did not find significant effects regarding tag

quality [41, 50, 75]. In contrast to these previous studies, we used images showing actual real-world objects which participants had to tag, instead of using abstract paintings and asking participants to tag the mood that the images might evoke. This allowed us to assess tag quality in a more objective way and might be the reason why we were able to find an effect of gamification on tag quality. Also, the results are in line with the findings by Van Berkel et al. [88], who showed that gamification of experience sampling led to higher quality responses, and an increased response rate. Additionally, we found that the average time per tag was significantly lower in the CG than in the TG condition (**R4**). Since no difference between the CG and the TG condition could be found regarding tag quality, the reason for this result needs further investigation in the future. Also, based on these results, we cannot support **H1d: Tag quality is higher in *Tailored Gamification* than in *Contra-Tailored Gamification***.

Regarding the enjoyment of tagging images, we found no significant difference between the gamified conditions and CO. Thus, **H2a: User enjoyment is higher in gamified conditions than in *Control*** is not supported. This is similar to previous research in the same context, which also did not find an effect on the IMI enjoyment factor [41, 75]. A potential reason might be that the task itself, i.e. tagging images, was perceived as unexciting or boring. However, we found a significant difference between the Tailored Gamification and the Contra-Tailored Gamification condition regarding enjoyment. Our results show that participants in the Tailored Gamification condition enjoyed tagging images significantly more than in the Contra-Tailored Gamification condition (**R5**). Thus, **H2b: User enjoyment is higher in *Tailored Gamification* than in *Contra-Tailored Gamification*** is supported. This shows that personalizing a gameful application based on Hexad user types can lead to an increased task enjoyment.

Related to this, we investigated whether positive or negative affect differs across conditions. Again, no significant difference between CO and both gamified conditions was found. Therefore, **H3a: Positive affective experiences are stronger in gamified conditions than in *Control*** is not supported. Similar to the absence of an effect regarding enjoyment between gamified conditions and CO, the repetitive nature of the task itself could be the reason here again. However, in line with **R5**, a significant difference was found between Tailored Gamification and Contra-Tailored Gamification. Positive affect was significantly higher when participants were exposed to gameful design elements that were suitable for their highest-scored Hexad type (**R6**), adding support for **H3b: Positive affective experiences are stronger in *Tailored Gamification* than in *Contra-Tailored Gamification***. **R7**, i.e. the fact that positive affect was even significantly lower in CG than it was in CO, further supports **H3b** and underlines the importance of personalization for the users' experience in gameful systems.

Regarding self-reported flow experience, we found a significant difference between the TG and CG conditions on the Merging Actions and Awareness (MAA) factor of the AFSS questionnaire (**R8**), whereas no significant effects were found for the remaining eight factors of the AFSS. Since the MAA dimension of flow was characterized by attentional resources being fully invested in the task at hand [17], **R8** indicates that participants were more absorbed in the image tagging task when receiving tailored gamification elements, compared to when receiving contra-tailored ones. Thus, one dimension out of nine dimensions of flow was positively affected by providing tailored gameful design elements. Although this can be seen as a first indicator for flow experiences being more prevalent in TG [16], we consider this evidence as too weak to derive that the holistic flow experience differed between TG and CG. Therefore, **H4b: Experiences of flow are more prevalent in *Tailored Gamification* than in *Contra-Tailored Gamification*** is only partially supported and needs further research in the future. Since no significant effects were found between CO and both gamified conditions, we cannot support **H4a: Experiences of flow are more prevalent in**

gamified conditions than in *Control* using the survey-based flow assessment. Table 3 provides an overview of all hypotheses and whether they were supported by the results.

Taking together the aforementioned results related to the user experience covering self-reported enjoyment, affective experiences and flow (**R5–R8**), we see that measures of these factors differ between Tailored Gamification and Contra-Tailored Gamification. Thus, our results show that selecting a suitable set of gamification elements, based on the users' Hexad type, can lead to improvements in enjoyment, positive affect and the extent to which users are absorbed in the task at hand.

Lastly, we analyzed physiological reactions (RMSSD, SCL, SCR, ST) to complement the survey-based measures. While we did not find significant effects for heart rate variability nor skin conductance level, a significant effect was found for skin conductance responses as well as skin temperature. Both the number of peaks in the skin conductance signal as well as the skin temperature were significantly higher in the gamified conditions (**R9, R10**), indicating that participants were more aroused when interacting with a gameful system than when there were no gameful design elements at all. This is an interesting finding, considering that previous research rarely combined psychophysiological and self-reported measures to investigate the user experience of gamified systems. However, these psychophysiological measures do not allow us to assess whether participants were positively or negatively aroused. Therefore, to interpret these findings, we consider the results of the survey-based instruments measuring flow, affective experiences and enjoyment, since both skin temperature and skin conductance responses were shown to be linked to these measures [35]. Combining them suggests that the significant increase in SCR and ST seems to be related to positive experiences in the TG condition (supported by the increase in positive affect, enjoyment and the MAA factor of the AFSS) whereas it seems to be related to negatively valenced arousal in the CG condition (supported by the fact that positive affect, enjoyment and the MAA factor are rated lower in CG than in the CO condition, and significant effects were found between TG and CG).

5.1 Limitations

Although the selection of suitable gameful design elements is based on previous research, certain design decisions when realizing these gameful design elements are inherently a matter of interpretation, which might affect the external validity of our results. The fact that we investigated a specific context (image tagging) and considered a certain set of images adds to this as it potentially affects the generalizability of our findings to other tasks and contexts. In addition, study outcomes may be affected by participants noticing the deception, i.e., noticing that conditions differed in whether the gameful design elements were tailored to them. Also, regarding task performance, we analyzed tag quantity and tag quality. While the former is easy to measure, the latter is more prone to subjectivity. To counter this, we analyzed all created tags with two independent raters. When interpreting the results of this analysis, it should be considered that their agreement was moderate. Also, the method of considering the mean rating to resolve conflicts is built upon the assumption that the rating scheme is interval-like, which is debatable. Future work, which specifically focuses on assessing tag quality in a similar fashion as was done in this paper, should consider resolving conflicts by discussing them between the raters. Another limitation concerns the approach we followed to select which gameful design elements to activate in the TG and CG conditions. Here, we decided to activate the gameful design elements, which were shown to be particularly relevant for the participant's Hexad type having the highest (TG) or lowest (CG) score. While this approach was straightforward to implement for most participants, we had two special cases that should be considered: First, since the Hexad model is a traits model, it could happen that participants scored highest/lowest on multiple user types. In this case, we considered these multiple user types equally

Hypothesis	Supp.?	Why?
H1: Task performance differs across conditions	Yes	ANOVA sig. for both #tags and tag quality
H1a: Tag quantity is higher in gamified conditions than in CO	Yes	#tags sig. higher in TG and CG than in CO
H1b: Tag quantity is higher in TG than in CG	No	no sig. diff. in #tags betw. TG and CG
H1c: Tag quality is higher in gamified conditions than in CO	Yes	tag quality sig. higher in TG/CG than in CO
H1d: Tag quality is higher in TG than in CG	No	no sig. diff. in tag quality betw. TG and CG
H2: User enjoyment differs across conditions	Yes	ANOVA sig. for IMI Enjoyment
H2a: User enjoyment is higher in gamified conditions than in CO	No	no sig. diff on IMI factors betw. TG/CG and CO
H2b: User enjoyment is higher in TG than in CG	Yes	IMI Enjoyment sig. higher in TG than in CG
H3: The strength of affective experiences differs across conditions	Yes	ANOVA sig. for PANAS Positive Affect
H3a: Positive affective experiences are stronger in gamified conditions than in CO	No	no sig. diff. on PANAS Positive Affect betw. TG/CG and CO
H3b: Positive affective experiences are stronger in TG than in CG	Yes	PANAS Positive Affect sig. higher in TG than in CG
H4: The prevalence of flow experiences differs across conditions	No	ANOVA sig. on MAA factor, no effects on other 8 factors
H4a: Experiences of flow are more prevalent in gamified conditions than in CO	No	no sig. diff. on any AFSS factor betw. TG/CG and CO
H4b: Experiences of flow are more prevalent in TG than in CG	Partially	AFSS MAA sig. higher in TG than in CG; no effect on other 8 factors

Table 3. Overview of hypotheses, whether they are supported or not ("Supp."), and reasons

to select which gameful design elements to use. Second, due to the fact that the mapping between relevant gameful design elements and Hexad user types is not one to one, it could happen that the set of suitable gameful design elements overlaps with the set of unsuitable gameful design elements (i.e. Socialisers and Players both have a strong preference for leaderboards [86], which means that participants scoring highest on Socialiser and lowest on Player would get the leaderboard in both TG and CG conditions). To avoid this and ensure that participants actually are presented with irrelevant gameful design elements, we selected the user type where participants had the second lowest score to decide which gameful design elements to activate in the CG condition. These two decisions need to be considered when replicating and interpreting our results. Also, following this procedure, the number of activated gamification elements might differ between participants, which could have an effect on the results. Related to this, it must be noted that we excluded the Disruptor type (similar to previous research by Mora et al. [53]), since no clear relationships to gameful design elements have been shown previously. However, since the Disruptor is by far the least common user type [86], we do not see a major limitation in terms of the practical relevance of our findings.

Also, since the Hexad is a traits model, the scores in the lowest and the highest user type were close for some participants (especially for participant 6). In these cases, the question of whether participants perceived CT actually as contra-tailored remains open. To counter this in future work, considering more than one factor (in our case the Hexad user type) to create tailored and contra-tailored gamification setups, as was done by Hallifax et al. [22], seems a promising direction. Lastly, we would like to acknowledge that the validity of the psychophysiological measures is tightly coupled to the technical specification of the Empatica E4 wristband which was used. Although the validity of the band has been demonstrated [46], and participants were not moving a lot (due to the task itself), a certain level of noise in the measurements is unavoidable.

6 CONCLUSION

To better understand the effects of personalizing gameful systems based on Hexad user types, we implemented an image tagging platform which allowed us to dynamically activate or deactivate gameful design elements. Using an implemented gameful system allows participants to experience gameful design elements and enables us to investigate behavioral and psychological effects, thus moving a step further than past research, which relied mostly on survey-based methods to study perceptual differences of gameful design elements.

Based on the modularity of the system, we could compare three conditions in a user study using a within-subjects design. In the *Control* condition, participants were asked to tag images without receiving any gameful feedback. In the *Tailored-Gamification* condition, we activated gameful design elements which were particularly suitable for the Hexad user type in which participants had the highest score. In the *Contra-Tailored-Gamification* condition, gameful design elements were activated which were suitable for the Hexad user type in which participants had the lowest score, and thus should be least interesting.

In a lab study (N=29), we showed that in general (independent of personalization), gamification leads to an increased task performance. Thus, we replicate previous research using gamification in an image-tagging context [41, 49–51, 75]. Also, we demonstrate that gamification affects psychophysiological reactions and thus contribute complementary findings to existing gamification research. However, the main contribution of this paper is demonstrating that personalization based on Hexad user types positively affects the user experience. For instance, we found that enjoyment was significantly higher when activating gameful design elements which are suitable based on the participants' Hexad types, compared to the enjoyment of the gameful system when activating unsuitable gameful design elements. Similarly, we found that activating suitable gameful design elements led to stronger, positively-valenced affective experiences. Also, the feeling of being fully absorbed in the task at hand was more pronounced when using a suitable set of gameful design elements.

Based on these findings, the short answer to the question *Do Hexad User Types Matter?*, which focuses on the practical relevance of the Hexad user types model, is *yes, they do.* Although it seems like immediate task performance is not affected by personalization, we found evidence that the user experience is affected (enjoyment, affective experience and the MAA dimension of flow are positively affected by personalization). Therefore, gamified systems should be personalized to the users' Hexad type to provide a pleasurable experience. We assume that an improved user experience will lead to positive effects on the task performance in the long run, i.e. we expect that the chance of interacting with a gameful system again is higher when users have a better experience with the system.

Therefore, future work should investigate the long-term effects of personalization based on Hexad user types to investigate whether the improved user experience adds positively to the task performance of users. Also, replicating our findings in different contexts besides image tagging is an important next step to get a more holistic picture of the effects of personalized gameful systems. Furthermore, the study should be replicated with a larger sample size to find smaller effects which we might have missed. Future work should also consider in-the-wild studies, to minimize potential observer effects and study the impact of personalization in a more natural setting.

REFERENCES

[1] Noora Aldenaini, Felwah Alqahtani, Rita Orji, and Sampalli Srinivas. 2020. Trends in Persuasive Technologies for Physical Activity and Sedentary Behavior: A Systematic Review. *Frontiers in Artificial Intelligence Journal for Human Learning and Behavior Change* 21 Febraury 2020 (2020), 85. https://doi.org/10.3389/frai.2020.00007

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

[3] Maximilian Altmeyer, Gustavo F. Tondello, Antonio Krüger, and Lennart E. Nacke. 2020. HexArcade: Predicting Hexad User Types By Using Gameful Applications. *Proceedings of the the Annual Symposium on Computer-Human Interaction in Play (CHI Play-2020)* (2020).

[4] Soumya C. Barathi, Michael Proulx, Eamonn O'Neill, and Christof Lutteroth. 2020. Affect Recognition using Psychophysiological Correlates in High Intensity VR Exergaming. *Conference on Human Factors in Computing Systems - Proceedings* (2020), 1–15. https://doi.org/10.1145/3313831.3376596

[5] Max V. Birk, Maximilian A. Friehs, and Regan L. Mandryk. 2017. Age-Based Preferences and Player Experience: A Crowdsourced Cross-sectional Study. *Proceedings of the Annual Symposium on Computer-Human Interaction in Play - CHI PLAY '17* (2017), 157–170. https://doi.org/10.1145/3116595.3116608

[6] Martin Böckle, Isabel Micheel, and Markus Bick. 2018. A Design Framework for Adaptive Gamification Applications. *Proceedings of the 51st Hawaii International Conference on System Sciences (HICSS '18)* (2018), 1227–1236.

[7] Martin Böckle, Jasminko Novak, and Markus Bick. 2017. Towards Adaptive Gamification: A Synthesis of Current Developments. *Proceedings of the 25th European Conference on Information Systems (ECIS '17)* (2017). http://aisel.aisnet.org/ecis2017{_}rp/11

[8] Wolfram Boucsein. 2012. *Electrodermal Activity* (2nd ed.). Springer Science+Business Media, New York, NY. https://doi.org/10.1007/978-1-4614-1126-0

[9] Marc Busch, Elke Mattheiss, Rita Orji, Peter Fröhlich, Michael Lankes, and Manfred Tscheligi. 2016. Player Type Models - Towards Empirical Validation. *Conference on Human Factors in Computing Systems - Extended Abstracts Proceedings* 07-12-May-, October 2017 (2016), 1835–1841. https://doi.org/10.1145/2851581.2892399

[10] John T. Cacioppo, Gary G Berntson, Jeff T Larsen, Kirsten M Poehlmann, and Tiffany A Ito. 2000. The Psychophysiology of Emotion. In *Handbook of emotions* (first edit ed.), M Lewis, Jeanette M Haviland-Jones, and Lisa Feldman Barrett (Eds.). Vol. 2. Guilford Publications, Chapter Chapter 11, 173–191. https://doi.org/10.1097/00005768-200405001-00432

[11] John T. Cacioppo, Stephanie Cacioppo, and Richard E. Petty. 2017. The neuroscience of persuasion: A review with an emphasis on issues and opportunities. *Social Neuroscience* (2017), 1–44. https://doi.org/10.1080/17470919.2016.1273851

[12] John T. Cacioppo, Louis G Tassinary, and Gary G Berntson. 2007. *The Handbook of Psychophysiology* (third edit ed.). Vol. 44. Cambridge University Press, New York. 914 pages. https://doi.org/10.1017/CBO9780511546396 arXiv:arXiv:1011.1669v3

[13] 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* 140, 4 (2014), 980–1008. https://doi.org/10.1037/a0035661

[14] Mihaly Csikszentmihalyi. 1975. Beyond Boredom and Anxiety. *Jossey-Bass* (1975).

[15] M Csikszentmihalyi. 1990. Flow. The Psychology of Optimal Experience. New York (HarperPerennial) 1990. (1990).

[16] Mihaly Csikszentmihalyi. 1997. *Finding Flow: The Psychology of Engagement with Everyday Life*. Basic Books, New York, NY, US. ix, 181–ix, 181 pages.

[17] Mihaly Csikszentmihalyi, Sami Abuhamdeh, and Jeanne Nakamura. 2014. Flow. *Flow and the Foundations of Positive Psychology: The Collected Works of Mihaly Csikszentmihalyi* (2014), 1–298. https://doi.org/10.1007/978-94-017-9088-8

[18] Sebastian Deterding, Dan Dixon, Rilla Khaled, and Lennart Nacke. 2011. From Game Design Elements to Gamefulness: Defining Gamification. *Proceedings of the 15th International Academic MindTrek Conference. ACM, 2011.* (2011), 9–15. https://doi.org/10.1145/2181037.2181040

[19] Arindam Dey, Thammathip Piumsomboon, Youngho Lee, and Mark Billinghurst. 2017. Effects of sharing physiological states of players in a collaborative virtual reality gameplay. In *Proceedings of the 2017 CHI conference on human factors in computing systems*. 4045–4056.

[20] Anders Drachen, Georgios Yannakakis, Lennart E. Nacke, and Anja Lee Pedersen. 2010. Correlation Between Heart Rate, Electrodermal Activity and Player Experience in First-Person Shooter Games. *Proceedings of the 5th ACM SIGGRAPH Symposium on Video Games* (2010), 49–54. https://doi.org/10.1145/1836135.1836143

[21] Emily B Falk and Christin Scholz. 2018. Persuasion, Influence, and Value: Perspectives from Communication and Social Neuroscience. *Annual Review of Psychology* 69 (2018), 329–356. https://doi.org/10.1146/annurev-psych-122216-011821

[22] Stuart Hallifax, Elise Lavoué, and Audrey Serna. 2020. To tailor or not to Tailor Gamification? An Analysis of the Impact of Tailored Game Elements on Learners' Behaviours and Motivation. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)* 12163 LNAI, July (2020), 216–227. https://doi.org/10.1007/978-3-030-52237-7_18

[23] Stuart Hallifax, Audrey Serna, Jean-charles Marty, Guillaume Lavoué, and Elise Lavoué. 2019. Factors to Consider for Tailored Gamification. *Proceedings of the Annual Symposium on Computer-Human Interaction in Play - CHI PLAY '19* (2019).

[24] Juho Hamari and Jonna Koivisto. 2014. Measuring flow in gamification: Dispositional Flow Scale-2. *Computers in Human Behavior* 40 (2014), 133–143. https://doi.org/10.1016/j.chb.2014.07.048

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

[26] László Harmat, Örjan de Manzano, Töres Theorell, Lennart Högman, Håkan Fischer, and Fredrik Ullén. 2015. Physiological correlates of the flow experience during computer game playing. *International Journal of Psychophysiology* 97, 1 (2015), 1–7. https://doi.org/10.1016/j.ijpsycho.2015.05.001

[27] Johan Högberg, Juho Hamari, and Erik Wästlund. 2019. Gameful Experience Questionnaire (GAMEFULQUEST): An Instrument for Measuring the Perceived Gamefulness of System Use. *User Modeling and User-Adapted Interaction* 29, 3 (2019), 619–660. https://doi.org/10.1007/s11257-019-09223-w

[28] Mark J. Huiskes and Michael S. Lew. 2008. The MIR Flickr Retrieval Evaluation. *Proceedings of the 1st International ACM Conference on Multimedia Information Retrieval, MIR2008, Co-located with the 2008 ACM International Conference on Multimedia, MM'08* (2008), 39–43. https://doi.org/10.1145/1460096.1460104

[29] Wilfrid Jänig. 2003. The Autonomic Nervous System and its Coordination by the brain. In *Handbook of Affective Sciences* (1 eds ed.), Richard J Davidson, Klaus R Scherer, and Goldsmith H (Eds.). Oxford University Press, Chapter 9, 135–186.

[30] Yuan Jia, Yikun Liu, Xing Yu, and Stephen Voida. 2017. Designing Leaderboards for Gamification: Perceived Differences based on User Ranking, Application domain, and Personality traits. *Conference on Human Factors in Computing Systems - Proceedings* 2017-May, May (2017), 1949–1960. https://doi.org/10.1145/3025453.3025826

[31] Yuan Jia, Bin Xu, Yamini Karanam, and Stephen Voida. 2016. Personality-Targeted Gamification: A Survey Study on Personality Traits and Motivational Affordances. *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems - CHI '16* (2016), 2001–2013. https://doi.org/10.1145/2858036.2858515

[32] Dennis L Kappen, Lennart E Nacke, Kathrin M Gerling, and Lia E Tsotsos. 2016. Design Strategies for Gamified Physical Activity Applications for Older Adults. *Hawaii International Conference on System Sciences* (2016), 1309–1318. https://doi.org/10.1109/HICSS.2016.166

[33] Johannes Keller, Herbert Bless, Frederik Blomann, and Dieter Kleinböhl. 2011. Physiological aspects of flow experiences: Skills-demand-compatibility effects on heart rate variability and salivary cortisol. *Journal of Experimental Social Psychology* 47, 4 (2011), 849–852. https://doi.org/10.1016/j.jesp.2011.02.004

[34] Ana Carolina Tomé Klock, Isabela Gasparini, Marcelo Soares Pimenta, and Juho Hamari. 2020. Tailored Gamification: A Review of Literature. *International Journal of Human Computer Studies* 144, September 2019 (2020). https://doi.org/10.1016/j.ijhcs.2020.102495

[35] Michael T. Knierim, Raphael Rissler, Verena Dorner, Alexander Maedche, and Christof Weinhardt. 2017. The Psychophysiology of Flow: A Systematic Review of Peripheral Nervous System Features. *Lecture Notes in Information Systems and Organisation* 25 (2017), 109–120. https://doi.org/10.1007/978-3-319-67431-5_13

[36] Oliver Korn and Adrian Rees. 2019. Affective effects of gamification. Using biosignals to measure the effects on working and learning users. *ACM International Conference Proceeding Series* (2019), 1–10. https://doi.org/10.1145/3316782.3316783

[37] Dimosthenis Kotsopoulos, Cleopatra Bardaki, Stavros Lounis, and Katerina Pramatari. 2018. Employee Profiles and Preferences towards IoT-enabled Gamification for Energy Conservation. *International Journal of Serious Games* 5, 2 (2018), 65–85. https://doi.org/10.17083/ijsg.v5i2.225

[38] Sylvia D. Kreibig. 2010. Autonomic nervous system activity in emotion: A review. *Biological Psychology* 84, 3 (2010), 394–421. https://doi.org/10.1016/j.biopsycho.2010.03.010

[39] Élise Lavoué, Baptiste Monterrat, Michel Desmarais, and Sébastien George. 2018. Adaptive Gamification for Learning Environments. *IEEE Transactions on Learning Technologies* 12, 1 (2018), 16–28. https://doi.org/10.1109/TLT.2018.2823710

[40] 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 21st International Academic Mindtrek Conference* (2017), 63–72. https://doi.org/10.1145/3131085.3131086

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

[42] Edwin A. Locke and Gary P. Latham. 2002. Building a Practically Useful Theory of Goal Setting and Task Motivation: A 35-Year Odyssey. *American Psychologist* 57, 9 (2002), 705–717. https://doi.org/10.1037/0003-066X.57.9.705

[43] Christian E Lopez and Conrad S Tucker. 2021. Adaptive Gamification and Its Impact on Performance. In *International Conference on Human-Computer Interaction.* Springer, 327–341.

[44] Andrzej Marczewski. 2015. *Even Ninja Monkeys Like to Play: Gamification, Game Thinking and Motivational Design.* CreateSpace Independent Publishing Platform.

[45] Edward D. McAuley, Terry Duncan, and Vance V. Tammen. 1989. Psychometric Properties of the Intrinsic Motivation Inventoiy in a Competitive Sport Setting: A Confirmatory Factor Analysis. *Research Quarterly for Exercise and Sport* 60, 1 (1989), 48–58. https://doi.org/10.1080/02701367.1989.10607413

[46] Cameron McCarthy, Nikhilesh Pradhan, Calum Redpath, and Andy Adler. 2016. Validation of the Empatica E4 Wristband. *2016 IEEE EMBS International Student Conference: Expanding the Boundaries of Biomedical Engineering and Healthcare, ISC 2016 - Proceedings* (2016), 4–7. https://doi.org/10.1109/EMBSISC.2016.7508621

[47] Robert R. McCrae and Oliver P. John. 1992. An Introduction to the Five-Factor Model and Its Applications. *Journal of Personality* 60, 2 (1992), 175–215. https://doi.org/10.1111/j.1467-6494.1992.tb00970.x

[48] Mary L McHugh. 2012. Interrater Reliability: The Kappa Statistic. *Biochemia Medica* (2012), 276–282.

[49] Elisa D. Mekler, Florian Brühlmann, Klaus Opwis, and Alexandre N. Tuch. 2013. 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* (2013), 66–73. https://doi.org/10.1145/2583008.2583017

[50] Elisa D. Mekler, Florian Brühlmann, Alexandre N. Tuch, and Klaus Opwis. 2017. Towards Understanding the Effects of Individual Gamification Elements on Intrinsic Motivation and Performance. *Computers in Human Behavior* 71 (2017), 525–534. https://doi.org/10.1016/j.chb.2015.08.048

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

[52] Anca M. Miron and Jack W. Brehm. 2006. Reactance Theory - 40 Years Later. *Zeitschrift für Sozialpsychologie* 37, 1 (2006), 9–18. https://doi.org/10.1024/0044-3514.37.1.9 arXiv:arXiv:1011.1669v3

[53] Alberto Mora, Gustavo F. Tondello, Lennart E. Nacke, and Joan Arnedo-Moreno. 2018. Effect of Personalized Gameful Design on Student Engagement. *IEEE Global Engineering Education Conference, EDUCON* 2018-April (2018), 1925–1933. https://doi.org/10.1109/EDUCON.2018.8363471

[54] Lennart Nacke and Craig A. Lindley. 2008. Flow and Immersion in First-Person Shooters. (2008), 81. https://doi.org/10.1145/1496984.1496998

[55] Lennart E. Nacke, Chris Bateman, and Regan L. Mandryk. 2014. BrainHex: A Neurobiological Gamer Typology Survey. *Entertainment Computing* 5, 1 (2014), 55–62. https://doi.org/10.1016/j.entcom.2013.06.002

[56] Lennart E. Nacke and Sebastian Deterding. 2017. The Maturing of Gamification Research. *Computers in Human Behavior* 71 (2017), 450–454. https://doi.org/10.1016/j.chb.2016.11.062

[57] Lennart E. Nacke, Mark N. Grimshaw, and Craig A. Lindley. 2010. More than a Feeling: Measurement of Sonic User Experience and Psychophysiology in a First-Person Shooter Game. *Interacting with Computers* 22, 5 (2010), 336–343. https://doi.org/10.1016/j.intcom.2010.04.005

[58] Wilk Oliveira, Olena Pastushenko, Luiz Rodrigues, Armando M Toda, Paula T Palomino, Juho Hamari, and Seiji Isotani. 2021. Does gamification affect flow experience? A systematic literature review. In *5th International GamiFIN Conference, GamiFIN 2021*. 110–119.

[59] Jeroen Ooge, Robin De Croon, Katrien Verbert, and Vero Vanden Abeele. 2020. Tailoring Gamification for Adolescents: a Validation Study of Big Five and Hexad in Dutch. In *Proceedings of the Annual Symposium on Computer-Human Interaction in Play*. ACM, New York, NY, USA, 206–218. https://doi.org/10.1145/3410404.3414267

[60] Rita Orji, Regan L Mandryk, and Julita Vassileva. 2015. Gender, Age, and Responsiveness to Cialdini's Persuasion Strategies. *International Conference on Persuasive Technology* 9072, June (2015). https://doi.org/10.1007/978-3-319-20306-5

[61] 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

[62] Rita Orji, Gustavo F Tondello, and Lennart E Nacke. 2018. Personalizing Persuasive Strategies in Gameful Systems to Gamification User Types. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems - CHI '18* (2018). https://doi.org/doi.org/10.1145/3173574.3174009

[63] Thérèse J.M. Overbeek, Anton van Boxtel, and Joyce H.D.M. Westerink. 2012. Respiratory Sinus Arrhythmia Responses to Induced Emotional States: Effects of RSA Indices, Emotion Induction Method, Age, and Sex. *Biological Psychology* 91, 1 (2012), 128–141. https://doi.org/10.1016/j.biopsycho.2012.05.011

[64] Mario Passalacqua, Pierre Majorique Léger, Lennart E. Nacke, Marc Fredette, Élise Labonté-Lemoyne, Xinli Lin, Tony Caprioli, and Sylvain Sénécal. 2020. Playing in the Backstore: Interface Gamification Increases Warehousing Workforce Engagement. *Industrial Management and Data Systems* 120, 7 (2020), 1309–1330. https://doi.org/10.1108/IMDS-08-2019-0458

[65] B. R. Payne, J. J. Jackson, S. R. Noh, and E. A. L. Stine-Morrow. 2011. Activity Flow State Scale. *APA PsycTests* (2011). https://doi.org/10.1037/t06855-000

[66] B. R. Payne, J. J. Jackson, S. R. Noh, and E. A. L. Stine-Morrow. 2011. In The Zone: Flow State and Cognition in Older Adults Brennan. *Psychol Aging* 23, 1 (2011), 1–7. https://doi.org/10.1037/a0022359

[67] Rosalind W. Picard, Elias Vyzas, and Jennifer Healey. 2001. Toward machine emotional intelligence: Analysis of affective physiological state. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 23, 10 (2001), 1175–1191. https://doi.org/10.1109/34.954607

[68] J Posner, James A. Russell, and Bradley S. Peterson. 2005. The circumplex model of affect: An integrative approach to affective neuroscience, cognitive development, and psychopathology. *Development and Psychopathology* 17, 3 (2005), 715–734. https://doi.org/10.1017/S0954579405050340

[69] Stephanie Reyssier, Stuart Hallifax, Audrey Serna, Jean-Charles Marty, Simonian Stephane, and Elise Lavoué. 2022. The impact of game elements on learner motivation: influence of initial motivation and player profile. *IEEE Transactions on Learning Technologies* (2022).

[70] Sara E. Rimm-Kaufman and Jerome Kagan. 1996. The Psychological Significance of Changes in Skin Temperature. *Motivation and Emotion* 20, 1 (1996), 63–78. https://doi.org/10.1007/BF02251007

[71] Richard W Robins, Chris R. Fraley, and Robert F. Krueger. 2007. *Handbook of Research Methods in Personality Psychology*. The Guilford Press.

[72] Richard M. Ryan. 1982. Control and Information in the Intrapersonal Sphere: An Extension of Cognitive Evaluation Theory. *Journal of Personality and Social Psychology* 43, 3 (1982), 450–461.

[73] Richard M Ryan and Edward L Deci. 2000. Self-Determination Theory and the Facilitation of Intrinsic Motivation, Social Development, and Well-Being. *American Psychologist* 55, 1 (2000), 68–78. https://doi.org/10.1037/0003-066X.55.1.68

[74] Michael Sailer, Jan Ulrich Hense, Sarah Katharina Mayr, and Heinz Mandl. 2017. How Gamification Motivates: An Experimental Study of the Effects of Specific Game Design Elements on Psychological Need Satisfaction. *Computers in Human Behavior* 69 (2017), 371–380. https://doi.org/10.1016/j.chb.2016.12.033 arXiv:arXiv:1011.1669v3

[75] Marc Schubhan, Maximilian Altmeyer, Dominic Buchheit, and Pascal Lessel. 2020. Investigating User-Created Gamification in an Image Tagging Task. *CHI Conference on Human Factors in Computing Systems Proceedings* (2020), 1–12. https://doi.org/10.1145/3313831.3376360

[76] Angela AT Schuurmans, Peter de Looff, Karin S Nijhof, Catarina Rosada, Ron HJ Scholte, Arne Popma, and Roy Otten. 2020. Validity of the Empatica E4 Wristband to Measure Heart Rate Variability (HRV) Parameters: A Comparison to Electrocardiography (ECG). *Journal of medical systems* 44, 11 (2020), 1–11.

[77] Katie Seaborn and Deborah Fels. 2015. Gamification in Theory and Action: A Survey. *International Journal of Human-Computer Studies* 74 (2015), 14–31. https://doi.org/10.1016/j.ijhcs.2014.09.006

[78] Sandra Sittenthaler, Christina Steindl, and Eva Jonas. 2015. Legitimate vs. illegitimate restrictions - a motivational and physiological approach investigating reactance processes. *Frontiers in Psychology* 6, May (2015), 1–11. https://doi.org/10.3389/fpsyg.2015.00632

[79] Hanne Spelt, Elisabeth Kersten-van Dijk, Jaap Ham, Joyce Westerink, and Wijnand IJsselsteijn. 2019. Psychophysiological Measures of Reactance to Persuasive Messages Advocating Limited Meat Consumption. *Information* 10, 10 (2019). https://doi.org/10.3390/info10100320

[80] Hanne Spelt, Joyce Westerink, Jaap Ham, and Wijnand Ijsselsteijn. 2018. Cardiovascular Reactions During Exposure to Persuasion Principles. *PERSUASIVE 2018; 13th International Conference on Persuasive Technology* 2, 2018 (2018), 315. https://doi.org/10.1007/978-3-319-78978-1

[81] Hanne A. A. Spelt, Joyce H. D. M. Westerink, Jaap Ham, and Wijnand A. IJsselsteijn. 2019. Psychophysiological Reactions to Persuasive Messages Deploying Persuasion Principles. *IEEE Transactions on Affective Computing* (2019), 1–13. https://doi.org/10.1109/TAFFC.2019.2931689

[82] Gustavo F. Tondello, Karina Arrambide, Giovanni Ribeiro, Andrew Jian lan Cen, and Lennart E. Nacke. [n.d.]. "I don't fit into a single type": A trait model and scale of game playing preferences. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)* 11747 LNCS, May ([n. d.]), 375–395. https://doi.org/10.1007/978-3-030-29384-0_23

[83] Gustavo F. Tondello, Alberto Mora, Andrzej Marczewski, and Lennart E. Nacke. 2018. Empirical Validation of the Gamification User Types Hexad Scale in English and Spanish. *International Journal of Human-Computer Studies* (2018). https://doi.org/10.1016/j.ijhcs.2018.10.002

[84] Gustavo F. Tondello, Alberto Mora, and Lennart E. Nacke. 2017. Elements of Gameful Design Emerging from User Preferences. *Proceedings of the Annual Symposium on Computer-Human Interaction in Play - CHI PLAY '17* (2017), 129–142. https://doi.org/10.1145/3116595.3116627

[85] Gustavo F. Tondello and Lennart E. Nacke. 2020. Validation of User Preferences and Effects of Personalized Gamification on Task Performance. *Frontiers in Computer Science* 2, August (2020). https://doi.org/10.3389/fcomp.2020.00029

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

[87] April Tyack and Elisa D. Mekler. 2020. Self-Determination Theory in HCI Games Research: Current Uses and Open Questions. *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems* (2020), 1–22. https://doi.org/10.1145/3313831.3376723

[88] Niels Van Berkel, Jorge Goncalves, Simo Hosio, and Vassilis Kostakos. 2017. Gamification of mobile experience sampling improves data quality and quantity. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 1, 3 (2017), 1–21.

[89] Carien M. van Reekum, Tom Johnstone, Rainer Banse, Alexander Etter, Thomas Wehrle, and Klaus R. Scherer. 2004. Psychophysiological Responses to Appraisal Dimensions in a Computer Game. *Cognition and Emotion* 18, 5 (2004), 663–688. https://doi.org/10.1080/02699930341000167

[90] David Watson, Lee Anna Clark, and Auke Tellegen. 1988. Development and Validation of Brief Measures of Positive and Negative Affect: The PANAS Scales. *Journal of Personality and Social Psychology* 54, 6 (1988), 1063–1070. https://doi.org/10.1037/0022-3514.54.6.1063 arXiv:arXiv:1011.1669v3

Received February 2022; revised June 2022; accepted July 2022