frontiers in Computer Science

ORIGINAL RESEARCH
published: 06 August 2020
doi: 10.3389/fcomp.2020.00029

Validation of User Preferences and Effects of Personalized Gamification on Task Performance

Gustavo F. Tondello[1,2]* *and Lennart E. Nacke*[2,3]

[1] Cheriton School of Computer Science, University of Waterloo, Waterloo, ON, Canada, [2] HCI Games Group & the Games Institute, University of Waterloo, Waterloo, ON, Canada, [3] Stratford School of Interaction Design and Business & Department of Communication Arts, University of Waterloo, Waterloo, ON, Canada

Personalized gamification is the tailoring of gameful design elements to user preferences to improve engagement. However, studies of user preferences have so far relied on self-reported data only and few studies investigated the effects of personalized gameful systems on task performance. This study shows that personalized gamification works in practice as predicted by survey studies and leads to higher task performance. We asked 252 participants in two studies to interact with a customized (experimental) or a generic (control) online gameful application to classify images. In the customized version, they could select the game elements that they wanted to use for their experience. The results showed significant correlations between participants' choice of gameful design elements and their Hexad user type scores, which partly support existing user preference models based on self-reported preferences. On the other hand, user type scores were not correlated with participants' preferred game elements rated after interacting with the gameful system. These findings demonstrate that the Hexad user types are a viable model to create personalized gameful systems. However, it seems that there are other yet unknown factors that can influence user preferences, which should be considered together with the user type scores. Additionally, participants in the experimental condition classified more images and rated their experience of selecting the game elements they wanted to use higher than in the control, demonstrating that task performance improved with personalization. Nonetheless, other measures of task performance that were not explicitly incentivized by the game elements did not equally improve. This contribution shows that personalized gameful design creates systems that are more successful in helping users achieve their goals than generic systems. However, gameful designers should be aware that they must balance the game elements and how much they incentivize each user behavior, so that the business goals can be successfully promoted. Finally, we analyzed participants' qualitative answers about their experience with the generic and the customized gameful applications, extracting useful lessons for the designers of personalized gameful systems.

Keywords: gamification, gameful design, personalization, adaptation, customization, Hexad user types

OPEN ACCESS

Edited by:
Marc Herrlich,
University of Kaiserslautern, Germany

Reviewed by:
Patrick Dickinson,
University of Lincoln, United Kingdom
Athanasios Mazarakis,
University of Kiel, Germany

****Correspondence:***
Gustavo F. Tondello
gustavo@tondello.com

Specialty section:
This article was submitted to
Human-Media Interaction,
a section of the journal
Frontiers in Computer Science

Received: *31 January 2020*
Accepted: *29 June 2020*
Published: *06 August 2020*

Citation:
Tondello GF and Nacke LE (2020)
Validation of User Preferences and
Effects of Personalized Gamification
on Task Performance.
Front. Comput. Sci. 2:29.
doi: 10.3389/fcomp.2020.00029

1. INTRODUCTION

Gamification is now a established design approach in human-computer interaction (HCI) to create engaging gameful systems (Seaborn and Fels, 2015; Landers et al., 2018; Koivisto and Hamari, 2019). Gamification or gameful design is the use of gameful design elements in non-game contexts (Deterding et al., 2011). In the past 5 years, gamification research has been maturing. Recent publications have been developing the theories that inform the gameful design practice and providing detailed empirical evidence of the effects of specific gameful design elements, for specific users, in specific contexts (Nacke and Deterding, 2017; Landers et al., 2018; Rapp et al., 2019).

One of the approaches to improve the design of gameful systems is personalized (or adaptive) gamification, meaning the tailoring of the gameful design elements, the interaction mechanics, the tasks, or the game rules according to the preferences or skills of each user (Lessel et al., 2016; Böckle et al., 2017; Tondello et al., 2017b; Klock et al., 2018; Tondello, 2019). Recent advances in the study of personalized gamification include the development of personalized gameful design methods (see section 2.1), the development of user preferences models and taxonomies of game elements (see section 2.2), and the evaluation of the effects of personalized gameful systems (see section 2.3).

Nonetheless, studies of user preferences have so far mostly relied on self-reported data instead of observation of actual user behavior. In addition, only a few studies investigated the effects of personalized gameful systems in comparison to generic alternatives. In the present work, we contribute to the literature on personalized gamification by observing user interaction with an online gameful system to study their game element preferences and the effects of personalization on their behavior and performance. In two studies, we observed 252 participants who interacted with either a customized (experimental) or a generic (control) version of a gameful image classification platform and reported on their experiences. Participants on the experimental condition were allowed to select the gameful design elements for their interaction with the platform, whereas participants in the control condition had all the gameful design elements available without the possibility of customization. This research answers two questions:

RQ1: If allowed to choose the gameful design elements they prefer, do user choices correspond to the theoretical relationships with user types, personality, gender, and age reported in previous survey-based studies?

RQ2: Are user performance and engagement better for a personalized gameful system than a generic system?

The results show several significant correlations between participants' choices of gameful design elements in the personalized condition with their Hexad user type scores, congruent to the expected relationships between elements and types according to the existing literature (Tondello et al., 2016b, 2017a). However, the results were less conclusive for personality traits, gender, and age. In addition, participants in the experimental condition classified more images and rated the experience of selecting which game elements to use higher than participants in the control condition. This new empirical evidence based on user behavior supports the user preference models previously devised based on Hexad user types and self-reported preferences. It also adds to the growing body of knowledge on personalization in gamification research demonstrating that user performance can be improved with personalized gameful design.

This contribution is important to the HCI and gamification communities because it provides evidence of the validity of personalized gameful design methods based on the selection of gameful design elements considering the different Hexad user types (such as Lessel et al., 2018; Marczewski, 2018; Mora Carreño, 2018; Tondello, 2019). Therefore, gamification designers can use the insights from this and the related works to create personalized gameful systems that are more effective than generic systems in helping users achieve their goals, such as improved learning, engagement, health, or well-being.

2. RELATED WORK

2.1. Methods for Personalized Gameful Design

Personalized gamification (or gameful design) is the tailoring of the gameful design elements, the interaction mechanics, the tasks, or the game rules for each user, according to their preferences. The tailoring is usually based on some knowledge about the users and their preferences and aims to boost the achievement of the goals of the gameful system (Tondello, 2019, chapter 3). Personalization in gamification is inspired by the reported positive results with other digital applications in general (Adomavicius and Tuzhilin, 2005; Sundar and Marathe, 2010), and more specifically in closely related applications such as games (e.g., Bakkes et al., 2012; Orji et al., 2013, 2014) and persuasive technologies (e.g., Nov and Arazy, 2013; Kaptein et al., 2015; Orji and Moffatt, 2018). Personalization can be implemented in two ways (Sundar and Marathe, 2010; Orji et al., 2017; Tondello, 2019):

- as a **customization** (also referred as user-initiated personalization), where the user selects the elements that they wish to use;
- as a (semi-)**automatic adaptation** (also referred as system-initiated personalization), where the system takes the initiative to select the gameful design elements for each user—with or without some user input in the process.

In previous work, we proposed a method for personalized gameful design (Tondello, 2019) based on three steps: (1) classification of user preferences using the Hexad user types (Tondello et al., 2016b, 2019b), (2) classification and selection of gameful design elements, where the user selects what elements they want to use (customization) or the system (semi-)automatically selects elements based on the user's Hexad scores and the classification of gameful design elements (Tondello et al., 2017a), and (3) a heuristic evaluation (Tondello et al., 2016a, 2019a) to verify if all the dimensions of motivational affordances are potentially integrated into the design.

Mora Carreño (2018) employs a similar approach based on the Hexad user types and a selection of gameful design elements for different groups of users. His work is more focused on the design of educational gamification services.

Lessel et al. (2016) also present a similar approach that is based on letting users customize their gameful experience by deciding when to use gamification and what elements to use. However, it is more focused on letting users freely choose from a defined (Lessel et al., 2016) or undefined (Lessel et al., 2018) set of gameful design elements, instead of relying on user types to aid in the selection. They have named this approach "bottom-up gamification."

Böckle et al. (2018) also propose a framework for adaptive gamification. It is based on four main elements, which can be applied to the gameful design process in diverse orders: (1) the purpose of the adaptivity, which consists on defining the goal of the adaptation, such as support of learning of participation, (2) the adaptivity criteria, such as user types or personality traits, which serve as an input for the adaptation, (3) the adaptive game mechanics and dynamics, which is the actual tailoring of game elements to each user, and (4) adaptive interventions, such as suggestions and recommendations, which represent the adaptation in the front-end layer.

In the gamification industry, Marczewski (2018) uses the Hexad user types to select gameful design elements for different users or as design lenses to design for different audiences. Furthermore, Chou (2015) considers different user profiles in one of the levels of the Octalysis Framework. The specific user model to be employed is not specified, with common examples being Bartle's player types (Bartle, 1996) and the Hexad user types.

Looking at these personalized gameful design methods together, there are some commonalities between them. All these methods suggest some means of understanding the user (e.g., user types or personality traits), some means of selecting gameful design elements for different users, and some mechanism to allow users to interact with the adaptation (e.g., customization or recommendation). In the present work, we build upon our previous publications by evaluating the user experience with a gameful application created using our personalized gameful design method (Tondello, 2019) and comparing the results with related works.

2.2. User Preference Models

The Hexad framework (Tondello et al., 2016b, 2019b; Marczewski, 2018) is the most used model of user preferences in gamification (Klock et al., 2018; Bouzidi et al., 2019). Monterrat et al. (2015) also developed a mapping of gamification elements to BrainHex player types (Nacke et al., 2014). However, Hallifax et al. (2019) compared the Hexad user types with the BrainHex and the Big-5 personality traits (Goldberg, 1993; Costa and McCrae, 1998). They concluded that the Hexad is the most appropriate for use in personalized gamification because the results with the Hexad were the most consistent with the definitions of its user types and it had more influence on the perceived user motivation from different gameful design elements than the other two models.

Although there are studies of the relationships between the Hexad user types and different variables in the literature, the relationship with participants' preferred gameful design elements is of particular interest for our study because our personalized gameful application relies on element selection. Publications that provide data about these relationships include the works of Tondello et al. (2016b, 2017a), Marczewski (2018), Orji et al. (2018), Mora et al. (2019), and Hallifax et al. (2019).

Studies that investigate user preferences in gamification by personality traits, gender, and age are also abundant in the literature. Again, we are interested in the publications that establish relationships between these variables and participants' preferred gameful design elements, so we could validate the relationships in the present study. Publications that provide these relationships with personality traits include the works of Butler (2014), Jia et al. (2016), Tondello et al. (2017a), Orji et al. (2017), and Hallifax et al. (2019); relationships with gender are provided by Tondello et al. (2017a) and Codish and Ravid (2017); and relationships with age are provided only by Tondello et al. (2017a).

These findings suggest that if allowed to choose the gameful design elements for their experience, participants' choices would be influenced by their user type scores, personality trait scores, gender, and age. Therefore, our first research question (**RQ1**) aims to validate these relationships.

2.3. Evaluation of Personalized Gameful Systems

We previously conducted a pilot study of personalized gamification (Tondello, 2019, chapter 7) using the same gameful application that we use in this study. We asked 50 participants to select four gameful design elements to customize their experience. The goal of that pilot study was to test the personalized gameful design method and gather participants' impressions regarding how they customize their experience. Progress feedback was the game element that was selected more often by participants: 36 times. It was followed by levels (30), power-ups (30), leaderboards (23), chance (23), badges (20), unlockable content (16), challenges (16), and moderating role (6 times).

The user types and personality trait scores were generally not good predictors of game element selection in the pilot study. However, there were some significant relationships: participants who chose challenges scored lower in conscientiousness; participants who chose unlockable content scored higher in the user type achiever and in emotional stability; participants who chose leaderboards scored lower in conscientiousness; participants who chose levels scored higher in the user type achiever and in openness to experiences; and participants who chose progress feedback scored lower in the user types socialiser and achiever, as well as emotional stability.

In the qualitative analysis, around 80% of participants expressed a positive experience, 10% expressed a negative experience, and 10% were neutral. The answers highlighted how participants enjoyed the variety of elements offered and the perceived control over their own experience. This shows that participants generally appreciated the customization options. Participants who expressed neutral or negative experiences would have preferred no gamification at all, rather than having an issue with the customization. Therefore, we concluded that participants can understand, carry out, and comment on the gamification customization task. Therefore, we suggested that more studies should be carried out with more participants and comparing personalized with non-personalized conditions to better understand the effects of personalized gamification, which is precisely what we do in the present study.

In the educational context, Mora et al. (2018) compared a generic with a personalized gameful learning experience with 81 students of computer network design. The descriptive statistics suggested that personalization seems to better engage students behaviorally and emotionally. However, the characteristics of the sample did not lead to any statistically significant result, suggesting that additional studies would be needed to confirm the preliminary findings. Herbert et al. (2014) observed that learner behavior on a gameful application varied according to their user types. Araújo Paiva et al. (2015) created a pedagogical recommendation system that suggested missions to students according to their most common and least common interactions, to balance their online behavior. Roosta et al. (2016) evaluated a gamified learning management system for a technical English course and demonstrated that student participation increased in a personalized version in comparison with a control version. Barata et al. (2017) conducted an extensive study to classify student behavior with a gameful interactive course. Based on their results, they presented a model that classifies students in four clusters and provided design lessons for personalized gameful education systems.

Evaluating their "bottom-up gamification" approach, Lessel et al. (2017) conducted a study with 106 participants in which they had to complete several image classification, article correction, or article categorization tasks. Several conditions where tested, from a fully generic gameful system (in which all elements were enabled) to a fully customizable system (in which participants could combine the elements in any way), and a control condition with no gamification. Participants who could customize their experience performed significantly better, solving more tasks faster without a decrease in correctness. The authors conclude that "bottom-up gamification" can lead to a higher motivational impact than fixed gamification.

In another study with 77 participants, Lessel et al. (2019) tested the impact of allowing participants to enable or disable gamification for an image tagging task. They found out that the choice did not affect participants who used gamification, but it improved the motivation of participants who were not attracted by the elements when they had the choice. Therefore, allowing users to enable or disable gamification seems to be a simple, but useful customization option when more sophisticated personalization is not available.

Böckle et al. (2018) employed their adaptive gameful design method to gamify an application for knowledge exchange in medical training. They compared application usage in the 6 months directly after introduction of adaptive gamification and in the period preceding it and noted an increase in overall system activity. However, they did not explicitly test if the effect was due to the adaptive nature of the implementation, or just due to the introduction of gamification itself.

Altogether, these related works show promising evidence that personalized gameful systems can be more engaging and lead to better task performance than generic systems with fixed gameful design elements. However, additional studies are required to replicate these initial findings and expand the available evidence to different applications and contexts. In response, we seek to provide additional evidence that personalized gamification increases user engagement and task performance (**RQ2**) in a context that was previously tested before: image classification tasks. Therefore, we provide additional evidence of the benefits of personalized gamification by replicating the positive effects of previous studies in a similar context, but with a different personalized design.

3. METHODS

3.1. Gameful Application

The two studies reported here were carried out using a gameful online application developed by the first author. The platform was designed as a customizable system that uses a variety of gameful design elements implemented around a central task, which was an image classification task for these studies. Thus, each task consisted on listing all the classification tags that the participant could think of for a stock image. Royalty-free stock images were randomly downloaded from Pexels[1]. The gameful design elements can be activated or deactivated by the researcher or the user, allowing experiments to be conducted in which participants interact with different sets of elements.

The use of classification tasks was already reported on previous studies of customizable gamification (Altmeyer et al., 2016; Lessel et al., 2017). Therefore, this is an interesting type of task to allow for comparisons with previous results. Moreover, these tasks are similar to brainstorming tasks, which have also been used in previous empirical studies of gamification (Landers et al., 2017) because they have been found to provide a good opportunity to investigate task performance in relation to goal setting. By combining these two types of tasks in our study, we implemented gameful design elements with the goal of motivating participants on two levels: (1) to complete more tasks and (2) to perform better in each task by listing a higher number of tags.

Following our proposed method for personalized gameful design (Tondello, 2019), we employed gameful design elements that would be appealing to users with different preferences. This design method suggests trying to include at least one or two game elements from each of the eight groups identified by Tondello et al. (2017a). The rationale for the design elements selected from each group for inclusion in the application is as follows:

- **Progression elements:** *Levels* are a common choice of progression element because they are easy to implement and are generally engaging. Therefore, it was our chosen progression element for the application.

[1] https://www.pexels.com/

- **Altruism elements:** This group includes elements that promote social interactions in which one user helps the other. In our application, direct help was not possible because users did not interact with each other directly. Therefore, we chose the element *moderating role*, as we anticipated that by feeling they could help moderate the tags entered into the platform, users could feel they were somehow being helpful.
- **Incentive:** This group includes elements that reward the user for completing tasks. We selected two types of incentives that we could easily implement in the application: *badges* and *unlockable content* (additional avatar choices).
- **Socialization:** Similar to the altruism group, social interaction was limited in the application because users did not have direct contact with each other. Therefore, we decide to implement only a *leaderboard* because it is a social element that works without the need for direct user interaction.
- **Risk/Reward:** This group includes elements that reward the user for taking chances or challenges. Together with elements from the Incentive group, these elements can be very engaging in short-term experiences. Therefore, we selected two elements from this group: *challenges* and earnings moderated by *chance*.
- **Assistance:** This group includes elements that help the user accomplish their goals. We selected *power-ups* as the assistance element for our platform because it is generally easy to implement and well-received by users.
- **Customization:** We chose to let users change their *avatar* in the platform as an element of the customization group.
- **Immersion:** We did not find any suitable immersion element that we could easily implement. The tasks that users had to complete (image tagging) were not very immersive on their own, unless users decided to focus on taking some time to appreciate the images that they were tagging. Other elements that could provide additional immersion, such as a narrative or theme, could not be easily integrated into the application in the available time for development. Therefore, we did not select any element from this group.

The gameful design elements included in the application are listed in **Table 1**. **Figure 1** shows the user interface of the application. In addition to the elements listed in the table, four features were implemented to support the gameful elements: points, progress feedback, avatars, and customization.

Points are used by the following elements: levels, to decide when the user should level up; unlockable content, so users can spend points to unlock additional avatars; leaderboards, which allow users to compare the amount of points they earned with other users; chance, which applies a random modifier to the amount of points earned after each task; and power-ups, which apply a fixed modifier to the amount of points earned. Points are automatically enabled when any of these elements are also enabled, otherwise they are disabled. Users earn 10 points each time they submit tags for an image, with an additional one point per tag provided.

Progress feedback is implemented in form of a progress bar that shows how many of the total available images the user has already completed and how many are left to be completed. It was always enabled. An avatar can be selected by the user to represent them in the system. It is always possible to select an avatar, but the available options are limited unless the game element unlockable content is enabled. Customization allows the user to select what gameful design elements they want to use in the application. In this study, customization was enabled for participants in the customized (experimental) condition and disabled for the generic (control) condition.

3.2. Study Design

3.2.1. Experimental Conditions

Participants were divided into two conditions:

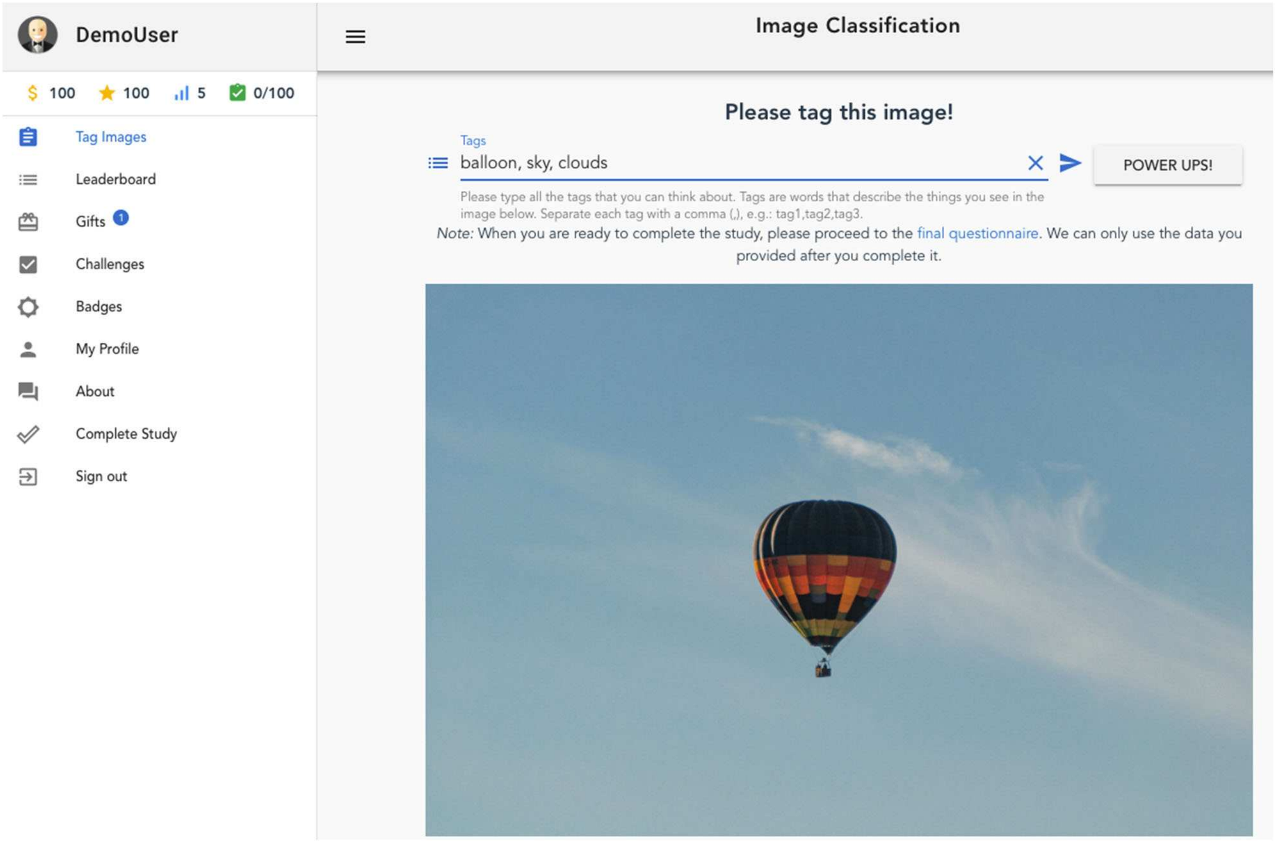
- Participants in the **generic** (control) condition were presented the list of game design elements for information only and all elements were automatically enabled for them. This conditions represents a generic (or one-size-fits-all) system because all participants should have similar experiences as they all have the same game elements in the interface. This mimics the current approach in gamification (without personalization), which consists in including different elements into the system to please different users, but without offering any mechanism for adaptation. We believe that this may overwhelm the user with too many elements to interact with, lead them to just ignore the game elements, or force users to select the elements they want to use just by directing their attention, i.e., by using the desired elements and ignoring the others in the interface.
- Participants in the **customized** (experimental) condition were asked to select as many game elements they wanted to use from the eight available options (see **Table 1**). **Figure 2** shows the user interface for customization, including the description of each game element provided to users before their selection. This is an example of user-initiated personalization (customization). The goal of this customization is to allow the users to improve their experience by removing the elements they do not want from the interface. In other words, the game elements that users do not select will not appear while they are working in the image classification tasks. Therefore, it should be easier for users to interact with the selected elements on a cleaner interface, potentially improving their experience and engagement. While answering our second research question, we will evaluate if these expectations will indeed correspond to the experience reported by the participants.

3.2.2. RQ1: Influence of User Characteristics on Element Selection

Our first research question is "If allowed to choose the gameful design elements they prefer, do user choices correspond to the theoretical relationships with user types, personality, gender, and age reported in previous survey-based studies?" The values for these four demographic variables were obtained from a survey presented to participants at the start of the experiment. We used the 24-item Hexad user types scale from Tondello et al. (2019b) and the 10-item Big-5 personality traits scale from Rammstedt and John (2007). The dependent variables were boolean values representing if the user selected each game element or not when given the choice in the customized condition. Therefore, data from

TABLE 1 | Gameful design elements implemented in the application.

Element (type)	Description
Levels (Progression)	After submitting the tags for each image, users would see a popup dialog informing if they leveled up as they earned points. The current level is also always displayed in the menu bar.
Moderating role (Altruism)	After writing tags for an image, the user can check a list of tags given by other people on a popup dialog and flag the unrelated tags.
Badges (Incentive)	Users earn badges as they complete tasks. When this element is selected, a new menu option appears that allows users to check the acquired and available badges and select one of the acquired badges to display in their profile besides their nickname.
Unlockable content (Incentive)	When this element is selected, additional customization options for the avatar are displayed, which are initially unlocked. Users can spend virtual coins (points) to unlock and use them.
Leaderboards (Socialization)	When this element is selected, a new option appears in the menu. Users can then see how they compare to others (points and level) in the leaderboard.
Challenges (Risk/Reward)	When this element is selected, a new menu option appears that allow users to see the available challenges, such as tagging a certain number of images or writing a certain number of tags for an individual image. Users earn additional points by completing any of the challenges.
Chance (Risk/Reward)	After each completed task, the amount of points received will be decided by luck. When this element is selected, a value between 5 and 1/5 is randomly selected, the earned points are multiplied by this value, and the results are displayed to the user in the popup dialog.
Power-ups (Assistance)	A power-up boosts the number of points received by the user for a few tasks (e.g., double the points earned for the next five images). When this element is selected, users will randomly earn a power-up after submitting the tags for an image. This power-up can be activated at any time in the image classification interface and will apply the boost for the next classified images.



Please tag this image!

Tags

balloon, sky, clouds

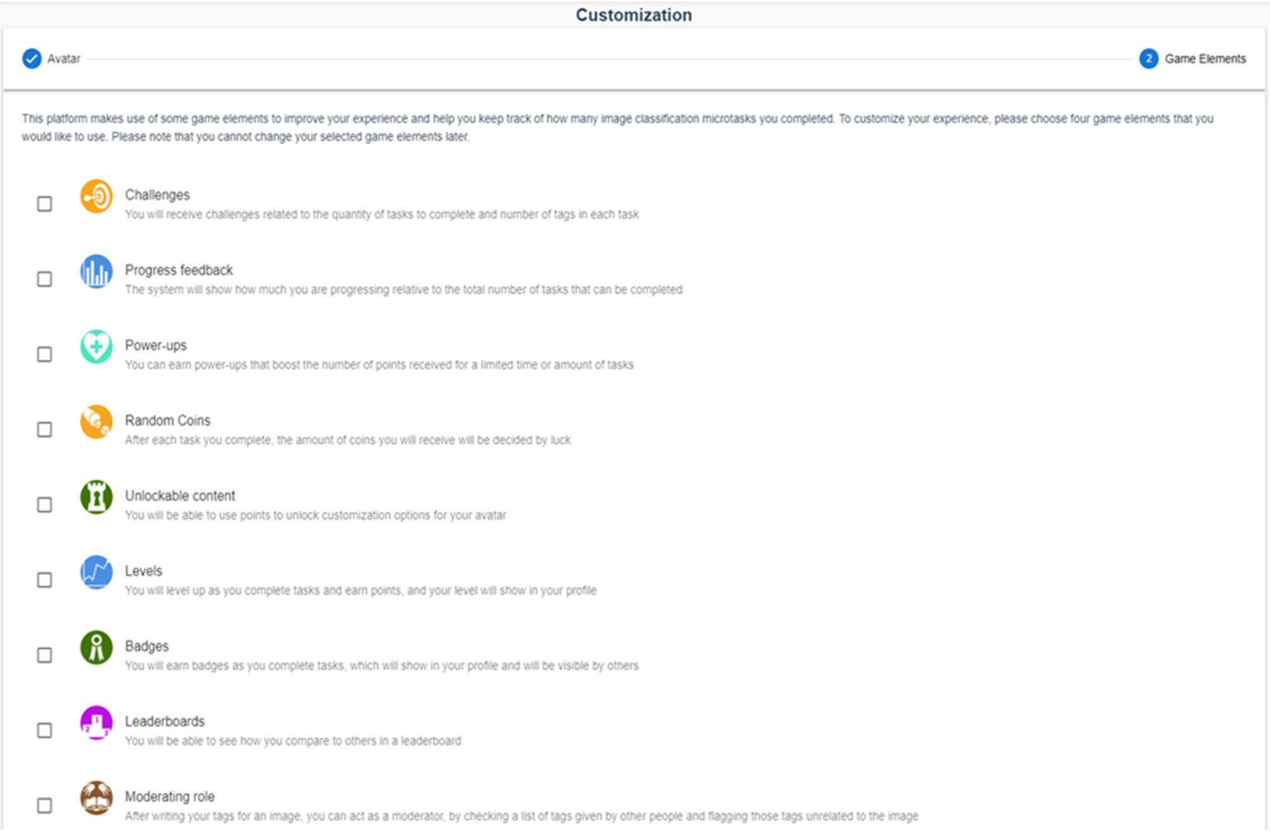
Please type all the tags that you can think about. Tags are words that describe the things you see in the image below. Separate each tag with a comma (,), e.g.: tag1,tag2,tag3.

Note: When you are ready to complete the study, please proceed to the final questionnaire. We can only use the data you provided after you complete it.

FIGURE 1 | Gameful image classification application used in this study.

participants in the generic (control) condition were not used to answer RQ1 as they were not given the chance to select game elements.

Based on the significant relationships between Hexad user type scores and game elements preferences observed by Tondello et al. (2016b,



Customization

✓ Avatar — 2 Game Elements

This platform makes use of some game elements to improve your experience and help you keep track of how many image classification microtasks you completed. To customize your experience, please choose four game elements that you would like to use. Please note that you cannot change your selected game elements later.

□ Challenges
You will receive challenges related to the quantity of tasks to complete and number of tags in each task

□ Progress feedback
The system will show how much you are progressing relative to the total number of tasks that can be completed

□ Power-ups
You can earn power-ups that boost the number of points received for a limited time or amount of tasks

□ Random Coins
After each task you complete, the amount of coins you will receive will be decided by luck

□ Unlockable content
You will be able to use points to unlock customization options for your avatar

□ Levels
You will level up as you complete tasks and earn points, and your level will show in your profile

□ Badges
You will earn badges as you complete tasks, which will show in your profile and will be visible by others

□ Leaderboards
You will be able to see how you compare to others in a leaderboard

□ Moderating role
After writing your tags for an image, you can act as a moderator, by checking a list of tags given by other people and flagging those tags unrelated to the image

FIGURE 2 | User interface for game element selection. Participants could customize their experience by selecting any number of elements from this interface, based on the descriptions depicted in the image.

2017a) and Orji et al. (2018), we formulated the following hypotheses:

H1: The user type scores are different between participants who selected or not each game element in the application.

- **H1.1:** Participants who select Levels have higher Achiever and Player scores than those who do not select it.
- **H1.2:** Participants who select Moderating role have higher Philanthropist and Socializer scores than those who do not select it.
- **H1.3:** Participants who select Badges have higher Achiever and Player scores than those who do not select it.
- **H1.4:** Participants who select Unlockable content have higher Free Spirit and Player scores than those who do not select it.
- **H1.5:** Participants who select Leaderboards have higher Socializer and Player scores than those who do not select it.
- **H1.6:** Participants who select Challenges have higher Achiever, Player, and Disruptor scores than those who do not select it.
- **H1.7:** Participants who select Chance have higher Achiever and Player scores than those who do not select it.

Based on the significant relationships between personality trait scores and game element preferences observed by Jia et al. (2016) and Tondello et al. (2017a), we formulated the following hypotheses:

H2: The personality trait scores are different between participants who selected or not each game element in the application.

- **H2.1:** Participants who select Levels have higher Extraversion and Conscientiousness scores than those who do not select it.
- **H2.2:** Participants who select Moderating role have higher Extraversion scores than those who do not select it.
- **H2.3:** Participants who select Badges have lower Emotional Stability scores than those who do not select it.
- **H2.4:** Participants who select Leaderboards have higher Extraversion scores than those who do not select it.
- **H2.5:** Participants who select Challenges have higher Agreeableness scores than those who do not select it.

Based on the significant relationships between gender and game element preferences observed by Tondello et al. (2017a) and Codish and Ravid (2017), we formulated the following hypotheses:

H3: The frequency that each game element is selected is different by gender.

- **H3.1:** Men select Leaderboards and Moderating role more often than women.
- **H3.2:** Women select Badges, Unlockable content, and Power-ups more often than men.

Based on the significant relationships between age and game element preferences observed by Tondello et al. (2017a), we formulated the following hypothesis:

H4: The average participant age is lower for those who select Moderating role, Badges, Unlockable content, Challenges, and Chance than those who do not select it.

3.2.3. RQ2: Task Performance and User Engagement

Our second research question is "Are user performance and engagement better for a personalized gameful system than a generic system?" Because image tagging is the main user task, the quantity of images tagged, total number of tags for all images, and average number of tags per image are the direct measures of user performance in the task. Additionally, we wanted to evaluate if user performance would also improve for the measures generated by the game elements, which are total points earned and final level achieved. Although these are not direct indicators of performance in the image tagging task, they may represent how much the user was invested in the application. Finally, another measure that helps understand user involvement is the total amount of time spent in the application.

To measure user engagement, we employed the Intrinsic Motivation Inventory (IMI; McAuley et al., 1989) because it has been previously used in similar gamification studies. Additionally, we asked participants to directly rate their overall game selection experience on a Likert scale (see **Q2** in the next subsection), as this seemed a more direct form of participant feedback regarding their perceived engagement than the IMI questions. Therefore, direct participant rating is a more direct but non-standardized measure of engagement, whereas the IMI scale is a less direct but standardized measure.

As the literature reviewed in section 2.3 showed that user performance and engagement was generally better for personalized gameful applications than generic ones, we formulated the following hypotheses:

H5: User Performance measures are higher for participants in the experimental condition than in the control condition.
H6: User Engagement measures are higher for participants in the experimental condition than in the control condition.

3.3. Procedure

After following the link to the application, participants had to read and accept the informed consent letter. It described the image tagging tasks and framed the study as image classification research, without mentioning that we were actually studying gameful design elements. This initial deception was done to ensure that participants would interact naturally with the gameful elements without any bias.

Next, participants answered a short demographic information form that asked about their gender, age, Hexad user types, and Big-5 personality traits scale. Then, they were invited to customize their profile by selecting a nickname and an avatar. For the final step of the initial part, participants were assigned to one of the experimental conditions in counter-balanced order. Participants in the control condition were presented with a list of game elements for information only, whereas participants in the customized (experimental) condition were also able to select which game elements they wanted to use for the image classification task.

Upon completion of the initial part, participants were left to interact with the platform freely. Logically, the image tagging tasks were the focus point of the platform. In the first study, participants were recruited via Mechanical Turk and could complete as many tasks as they wanted (with no lower limit) up to the limit of 50 available images. The tasks were to be completed in one sitting. During this period, they could also interact with the features provided by the gameful design elements that they selected (experimental condition) or all elements (control condition). On the other hand, participants were recruited via social media for the second study and could interact with the application as many times as they wanted for 7 days. They could complete as many tasks as they wanted (with no lower limit) up to the limit of 100 available images. These participants also received a daily email reminder (sent by one of the researchers) that they needed to go back to the platform and complete the study by filling out the final survey.

When they felt they had tagged enough images, participants clicked the option "Complete Study" in the menu. At this point, they were asked to complete a questionnaire that included the Intrinsic Motivation Inventory (IMI) and the following free-text questions:

- **Q1:** Overall, how do you describe your experience with the image classification activities you just completed?
- **Q2:** How do you describe and rate the experience of selecting game elements to customize the platform for you? *(Likert scale with very negative, negative, neutral, positive, and very positive, in addition to the free-text answer)*
- **Q3:** Were you satisfied with the selection of game elements provided by the system? Why?
- **Q4:** Were you able to select game elements that matched your preferences? Why?
- **Q5:** How much do you feel that the selection of game elements you used to customize the platform for you influenced your enjoyment of the image classification tasks? Why?
- **Q6:** Now that you have used this system, which one was your preferred game element to use? Please explain why it was your preferred element. *(selection box with the eight game elements, in addition to the free-text answer)*
- **Q7:** Now that you have used this system, which game element do you feel most influenced how you tagged images? *(selection box with the eight game elements)*
- **Q8:** Which game element motivated you more to tag images? *(selection box with the eight game elements)*

After completing the post-study questionnaire, participants were presented with a post-study information letter and additional consent form. This additional letter debriefed participants about the deception used in the study. Thus, the letter explained that participants were initially told that we were interested in the tags to help us develop image classification systems; however, we were actually interested in studying their experience with the gameful design elements. It also explained that this was done to avoid bias in the participant's interaction with the game elements and their responses about their experiences. Participants were then given the chance to accept or to decline having their study data used after knowing the real purpose of the study and were instructed to contact the researchers by email if they had any question about the deception employed in the study. These procedures followed the guidelines for ethical participant recruitment established by the Office of Research Ethics at the University of Waterloo. Upon completion of this last step, the software then generated a completion code for participants recruited via Mechanical Turk, which they used to complete the task on the platform and receive their payment.

3.4. Participants

We planned to collect two data sets to answer our research questions. For the first study, we recruited participants through Amazon's Mechanical Turk, which is being increasingly used for HCI experiments (Buhrmester et al., 2018). This form of recruitment allowed us to determine the number of participants we wanted to recruit. Therefore, we planned to recruit a total of 200 participants (100 per condition). However, for **RQ2**, one concern was if participant motivation would have any effect on their performance. As Mechanical Turk participants were paid a fixed amount for completion of the task, it would be reasonable to assume that some of them might want to complete the task as quickly as possible to maximize their earnings. Therefore, we also collected a second data set only with volunteers that were not receiving a fixed payment for participation (although they were offered a chance to enter a draw as an incentive). This allowed us to also analyze data from participants that were presumably more willing to collaborate with the study without being too concerned with maximizing their time usage. For this second data set, we recruited participants through social media. Thus, it was hard to control how many participants would voluntarily complete the study. We aimed to recruit at least 100 participants and ended with 127 people creating an account, but in the end only 54 completed the study (27 per condition). Nonetheless, we considered that this sample size was sufficient to test hypotheses **H5** and **H6**. These two hypotheses were tested separately for each data set.

To answer **RQ1** and test the associated hypotheses, we used only the data from participants in the customized condition because participants in the generic condition were not allowed to select their game elements. Thus, only the customized condition contained data that we could use to test **H1–H4**. Considering that the number of participants in this condition was 99 per condition in the first study and 27 per condition in the second, we combined the data from the two studies because the groups of participants who selected or not each game element would otherwise be too small to carry out reliable statistical analyses, especially in the second study. Additionally, we have no theoretical reason to believe that the recruitment source (Mechanical Turk or social media) would make any difference in participants' preferred game elements according to their demographic characteristics. Even if their motivation to complete image tagging tasks was different depending on if they were being paid or not, we assumed that their gaming preferences would not be affected by it. Although **Table 2** shows that there were some differences in the user type scores, personality trait scores, and average age between the two datasets, these are the independent variables being analyzed in the statistical tests. Therefore, they are not confounding variables in the analyses. Therefore, we consider that combining the two datasets does not create a confounding factor in the analyses.

TABLE 2 | Description of participants' user type scores and personality trait scores.

	Study 1				Study 2			
User type	**Med**	**Mean**	**SD**	**α**	**Med**	**Mean**	**SD**	**α**
Philanthropist	5.75	5.44	1.10	0.879	6.00	6.00	0.66	0.633
Socialiser	4.75	4.68	1.31	0.893	5.62	5.26	1.26	0.887
Achiever	5.75	5.61	0.97	0.848	6.00	5.94	0.70	0.710
Free spirit	5.50	5.47	0.98	0.762	5.75	5.60	0.65	0.260
Player	5.75	5.64	0.96	0.786	5.75	5.43	1.02	0.713
Disruptor	3.25	3.42	1.24	0.783	3.50	3.67	1.10	0.630
	Study 1				**Study 2**			
Personality trait	**Med**	**Mean**	**SD**	**α**	**Med**	**Mean**	**SD**	**α**
Extraversion	3.25	3.42	1.69	0.694	4.00	3.79	1.39	0.723
Agreeableness	4.50	4.68	1.49	0.532	5.00	4.85	1.02	0.352
Conscientiousness	6.00	5.45	1.37	0.649	4.50	4.62	1.27	0.570
Emotional stability	4.50	4.59	1.69	0.762	3.50	3.57	1.59	0.855
Openness to experiences	5.50	5.32	1.35	0.463	4.50	4.57	1.46	0.572

Study 1: N = 198. Study 2: N = 54. Median and Mean values based on a 7-point Likert scale (range: 1.0–7.0). Cronbach's α calculated with 4 items per user type and 2 items per personality trait.

As mentioned above, we recruited a total of 200 participants through Amazon Mechanical Turk for the first study, with 100 per condition in counter-balanced order. Participants were required to have a HIT (high intelligence task) approval rate greater than 97%, a number of HITs approved higher than 5,000, and reside in the United States of America. This was done to ensure that only workers with a good history in the platform accepted our task. The HIT description on Mechanical Turk contained a brief description of the image classification task without mentioning the gameful elements and a link to the online system. Participants were informed that the estimated duration of the task was between 30 min and 1 h and were paid a fixed amount of $4.00 (four US dollars) after completion of the task. This remuneration was paid to all participants who submitted a completion code for the HIT, even if they did not complete all the steps of the study procedure, congruent to the ethical participant recruitment guidelines.

After verification, we had to remove two participants who did not complete the final survey with the final participation agreement. Therefore, the final dataset contained 198 participants (99 per condition). The sample contained answers from 90 women and 106 men (2 not disclosed), with ages varying from 19 to 72 years old ($M = 36.9$, $SD = 10.6$). They spent an average of 26.2 min on the platform ($SD = 23.4$), tagged 25.4 images on average ($SD = 19.2$) with a total of 118.9 tags on average ($SD = 135.6$), and earned a total of 873 points on average ($SD = 1,054$). Participants in the customized condition selected between zero and eight game elements ($M = 3.5$, $SD = 2.2$, $Med = 3.0$, $Mod = 1.0$, $N = 99$).

For the second study, we recruited participants through social media (Facebook, Twitter, and Reddit) and email lists of people interested in our research. They did not receive any direct compensation, but were offered the opportunity to enter a draw for one out of two $200 (two hundred US dollars) international gift cards. They could interact with the platform freely for a suggested limit of 7 days, but this limit was not enforced. However, the study actually ended when each participant decided to complete the final survey.

In total, 127 participants created an account and started interacting with the application. They were assigned to one of the two conditions in counter-balanced order. However, only 54 participants completed the study by filling out the end survey (27 per condition), which constitutes our final data set. The sample contained answers from 25 women and 28 men (1 not disclosed), with ages varying from 18 to 50 years old ($M = 25.8$, $SD = 5.8$). They were from Canada (17), China (7), India (6), France (5), United States of America (4), Iran (4), Nigeria (2), and nine other countries (only 1 participant each). They tagged 45.4 images on average ($SD = 37.3$) with a total of 345.8 tags on average ($SD = 391.1$), and earned a total of 2,243 points on average ($SD = 2,369$). Participants interacted with the platform between two and 13 different days ($M = 4.8$, $SD = 2.5$) and completed 1,089 action on average ($SD = 1,988$). Participants in the customized condition selected between zero and eight game elements ($M = 4.0$, $SD = 2.3$, $Med = 4.0$, $Mod = 3.0$, $N = 27$).

Table 2 summarizes the descriptive statistics for the Hexad user types and personality trait scores for all participants. Although there are some differences in the average values for these demographic variables between the two data sets, these were the independent variables being analyzed in the tests for hypotheses **H1–H4**. Therefore, we do not consider that this difference may have affected our results.

4. RESULTS

We present the results in this section for each one of the research questions. All statistical analyses were performed using SPSS v. 23 (IBM, 2015).

4.1. RQ1: Influence of User Characteristics on Element Selection

To answer **RQ1**, we carried out several splits of the data set according to whether the participant selected a specific element or not. For example, we compared participants who selected leaderboards with those who did not select it, participants who selected levels with those who did not select it, and so on. As explained in section 3.4, we combined the data from both samples and used only data from participants in the experimental (customized) condition because this was the only condition in which participants were given the chance to select the game elements they wanted to use.

Table 3 presents the results of the statistical tests comparing the Hexad and personality trait scores between participants who selected or did not select each element. Because the scores were not parametric, we employed the Mann–Whitney U-test. We also calculated the effect size $r = Z \div \sqrt{N}$, as suggested by Field (2009, p. 550).

There are several significant differences in the Hexad user type scores in relation to element selection:

- **H1.1: not supported.** Participants who selected **Levels** did not have higher **Achiever** ($p = 0.1595$, $r = 0.135$) and **Player** ($p = 0.160$, $r = 0.125$) scores than those who did not select it.
- **H1.2: not supported.** Participants who selected **Moderating role** did not have higher **Philanthropist** ($p = 0.449$, $r = 0.067$) and **Socializer** ($p = 0.333$, $r = 0.086$) scores than those who did not select it.
- **H1.3: partially supported.** Participants who selected **Badges** had higher **Achiever** scores than those who did not select it ($p = 0.015$, $r = 0.216$). However, they did not have higher **Player** scores ($p = 0.765$, $r = 0.027$).
- **H1.4: not supported.** Participants who selected **Unlockable content** did not have higher **Free Spirit** ($p = 0.787$, $r = 0.024$) and **Player** ($p = 0.641$, $r = 0.042$) scores than those who did not select it.
- **H1.5: partially supported.** Participants who selected **Leaderboards** had higher **Player** scores than those who did not select it ($p = 0.006$, $r = 0.244$). However, they did not have higher **Socializer** scores ($p = 0.116$, $r = 0.140$)
- **H1.6: partially supported.** Participants who selected **Challenges** had higher **Achiever** ($p = 0.005$, $r = 0.249$) and **Player** ($p = 0.045$, $r = 0.179$) scores than those who did not select it. However, they did not have higher **Disruptor** scores ($p = 0.682$, $r = 0.036$).
- **H1.7: partially supported.** Participants who selected **Chance** had higher **Achiever** scores than those who did not select it ($p = 0.005$, $r = 0.175$). However, they did not have higher **Player** scores ($p = 0.266$, $r = 0.099$).

On the other hand, the following significant differences were not predicted by the existing literature and were not part of our hypotheses, but appeared in the results:

- **Philanthropist** scores are higher for participants who selected **Badges** ($p = 0.027$, $r = 0.196$). This relationship was not suggested in any previous research.
- **Free Spirit** scores are higher for participants who selected **Chance** ($p = 0.050$, $r = 0.175$). This relationship was also not suggested in previous research.
- **Player** scores are higher for participants who selected **Power ups** ($p = 0.029$, $r = 0.194$). This makes sense because

TABLE 3 | Non-parametric tests (Mann–Whitney U) comparing the differences in Hexad user type scores and personality trait scores between users who selected or not each gameful design element.

Elements					**Hexad user types**					**Personality traits**		
		Phil	**Soc**	**Ach**	**Free**	**Play**	**Dis**	**Ext**	**Agr**	**Con**	**Emo**	**Ope**
Levels	$\widetilde{\mathbf{No}}$ (42)	5.75	5.50	5.75	5.75	5.75	3.50	3.50	4.75	5.50	4.50	5.00
	$\widetilde{\mathbf{Yes}}$ (84)	6.00	5.00	6.00	5.75	6.00	3.25	3.50	5.00	6.00	4.50	6.00
	U	1580.5	1577.5	1472.5	1489.0	1494.0	1693.0	1762.5	1529.5	1379.0	1706.5	1451.5
	Z	0.956	0.968	1.519	1.432	1.404	0.368	0.008	1.019	1.827	0.299	1.435
	p	0.339	0.333	0.129	0.152	0.160	0.713	0.994	0.308	0.068	0.765	0.151
	r	0.085	0.086	0.135	0.128	0.125	0.033	0.001	0.091	0.163	0.027	0.128
Moderation	$\widetilde{\mathbf{No}}$ (99)	5.75	5.25	6.00	5.75	6.00	3.25	3.50	5.00	6.00	4.50	6.00
	$\widetilde{\mathbf{Yes}}$ (27)	5.75	5.25	5.63	5.75	5.75	3.75	4.00	4.50	6.00	4.50	4.75
	U	1210.0	1321.0	1132.5	1125.0	1014.0	1119.0	1125.0	1280.0	1210.0	1258.5	1025.5
	Z	0.757	0.092	1.221	1.265	1.926	1.296	1.263	0.260	0.687	0.466	1.800
	p	0.449	0.926	0.222	0.206	0.054	0.195	0.207	0.795	0.492	0.641	0.072
	r	0.067	0.008	0.109	0.113	0.172	0.115	0.112	0.023	0.061	0.042	0.160
Badges	$\widetilde{\mathbf{No}}$ (60)	5.75	5.13	5.75	5.75	6.00	3.63	3.00	4.50	5.75	4.50	5.00
	$\widetilde{\mathbf{Yes}}$ (66)	6.00	5.25	6.00	5.75	5.75	3.25	3.50	5.25	6.00	4.50	6.00
	U	1531.5	1746.5	1487.5	1652.5	1919.0	1640.0	1738.5	1488.0	1598.0	1909.0	1529.5
	Z	2.205	1.144	2.422	1.609	0.299	1.665	1.184	2.286	1.748	0.348	2.083
	p	**0.027**	0.252	**0.015**	0.108	0.765	0.096	0.236	**0.022**	0.080	0.728	**0.037**
	r	**0.196**	0.102	**0.216**	0.143	0.027	0.148	0.106	**0.204**	0.156	0.031	**0.186**
Unlockables	$\widetilde{\mathbf{No}}$ (78)	5.75	5.25	5.88	5.75	5.75	3.38	3.50	5.00	6.00	4.50	5.50
	$\widetilde{\mathbf{Yes}}$ (48)	5.75	5.00	6.00	5.75	6.00	3.38	3.25	5.00	6.00	4.50	5.50
	U	1805.0	1853.5	1622.5	1818.5	1779.5	1757.0	1655.5	1778.0	1813.5	1765.0	1774.0
	Z	0.339	0.093	1.262	0.270	0.467	0.579	1.092	0.358	0.177	0.540	0.379
	p	0.735	0.926	0.207	0.787	0.641	0.562	0.275	0.720	0.859	0.589	0.705
	r	0.030	0.008	0.112	0.024	0.042	0.052	0.097	0.032	0.016	0.048	0.034
Leaderboards	$\widetilde{\mathbf{No}}$ (65)	5.75	5.00	5.75	5.75	5.75	3.50	3.50	5.00	6.00	4.50	5.00
	$\widetilde{\mathbf{Yes}}$ (61)	6.00	5.25	6.00	5.75	6.00	3.25	4.00	5.00	6.00	5.00	6.00
	U	1863.0	1661.5	1613.5	1874.0	1424.0	1976.5	1734.0	1940.0	1869.5	1693.5	1539.5
	Z	0.587	1.572	1.813	0.533	2.739	0.029	1.218	0.060	0.413	1.417	2.055
	p	0.557	0.116	0.070	0.594	**0.006**	0.977	0.223	0.952	0.680	0.156	**0.040**
	r	0.052	0.140	0.162	0.047	**0.244**	0.003	0.109	0.005	0.037	0.126	**0.183**
Challenges	$\widetilde{\mathbf{No}}$ (71)	5.75	5.25	5.75	5.75	5.75	3.25	3.00	5.00	6.00	4.50	5.50
	$\widetilde{\mathbf{Yes}}$ (55)	6.00	5.25	6.00	5.75	6.00	3.50	3.50	5.00	6.00	4.50	6.00
	U	1718.5	1914.0	1387.0	1816.0	1546.5	1869.5	1625.5	1875.0	1770.5	1846.5	1684.0
	Z	1.158	0.190	2.800	0.675	2.006	0.409	1.615	0.250	0.778	0.524	1.209
	p	0.247	0.849	**0.005**	0.499	**0.045**	0.682	0.106	0.802	0.436	0.600	0.227
	r	0.103	0.017	**0.249**	0.060	**0.179**	0.036	0.144	0.022	0.069	0.047	0.108
Chance	$\widetilde{\mathbf{No}}$ (79)	5.75	5.00	5.75	5.75	5.75	3.38	3.50	4.50	5.75	4.50	5.00
	$\widetilde{\mathbf{Yes}}$ (47)	5.75	5.25	6.13	6.00	6.00	3.38	3.75	5.50	6.00	4.50	6.00
	U	1673.0	1677.5	1309.5	1469.5	1637.0	1772.5	1715.5	1258.0	1505.5	1623.0	1426.0
	Z	0.931	0.906	2.778	1.964	1.112	0.425	0.714	2.882	1.615	1.183	2.019
	p	0.352	0.365	**0.005**	**0.050**	0.266	0.671	0.475	**0.004**	0.106	0.237	**0.043**
	r	0.083	0.081	**0.247**	**0.175**	0.099	0.038	0.064	**0.257**	0.144	0.105	**0.180**
Power-ups	$\widetilde{\mathbf{No}}$ (57)	5.75	5.25	5.75	5.75	5.75	3.25	3.50	5.00	6.00	4.50	5.50
	$\widetilde{\mathbf{Yes}}$ (69)	5.75	5.25	6.00	5.75	6.00	3.50	3.50	5.00	6.00	4.50	6.00
	U	1806.5	1964.0	1638.0	1843.0	1524.0	1916.5	1815.0	1854.0	1789.0	1846.0	1898.5
	Z	0.789	0.012	1.621	0.609	2.179	0.246	0.746	0.390	0.719	0.593	0.168
	p	0.430	0.990	0.105	0.543	**0.029**	0.806	0.456	0.697	0.472	0.553	0.867
	r	0.070	0.001	0.144	0.054	**0.194**	0.022	0.066	0.035	0.064	0.053	0.015

N = 126.
Bolded values are significant at the 0.05 level.
$\widetilde{\mathbf{No}}$*: median scores for users who did not select each element (range: 1.0–7.0).*
$\widetilde{\mathbf{Yes}}$*: median scores for users who selected each elements (range: 1.0–7.0).*
The numbers in brackets following $\widetilde{\mathbf{No}}/\widetilde{\mathbf{Yes}}$ *are the number of participants for each row.*
U*/****Z****/****p****: results of the Mann–Whitney U-tests.*
r*: effect sizes, calculated as* $r = Z \div \sqrt{N}$.
The absolute values of ***Z*** *and* ***r*** *are displayed for improved readability.*

TABLE 4 | Distribution of users who selected or not each gameful design element by gender.

	Levels	Moderation	Badges	Unlockables	Leaderboard	Challenges	Chance	Power-ups
N-M/F	20/22	50/49	34/26	39/38	31/34	36/35	42/36	29/28
Y-M/F	44/39	14/12	30/35	25/23	33/27	28/26	22/25	35/33
χ^2	0.325	0.092	1.380	0.024	0.667	0.016	0.581	0.004
p	0.569	0.762	0.240	0.876	0.414	0.899	0.446	0.947

N = 126.
N-M/F: proportion of men and women who did not select each game element.
Y-M/F: proportion of men and women who selected each game element.
χ^2/p: results of Pearson's Chi-square tests comparing the proportions above (Crosstabs option on SPSS).

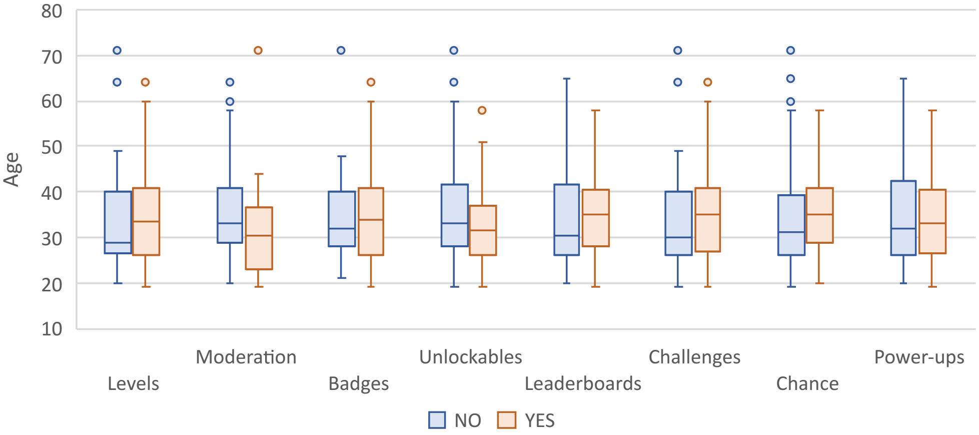


FIGURE 3 | Differences in age between users who selected or not each gameful design element.

power-ups allowed users to easily earn more points, which would be appealing to people with high scores on this user type.

Regarding participants' personality trait scores, none of the hypotheses were supported:

- **H2.1: not supported.** Participants who selected **Levels** did not have higher **Extraversion** ($p = 0.994$, $r = 0.001$) and **Conscientiousness** ($p = 0.068$, $r = 0.163$) scores than those who did not select it.
- **H2.2: not supported.** Participants who selected **Moderating role** did not have higher **Extraversion** scores than those who did not select it ($p = 0.207$, $r = 0.112$).
- **H2.3: not supported.** Participants who selected **Badges** did not have lower **Emotional Stability** scores than those who did not select it ($p = 0.728$, $r = 0.031$).
- **H2.4: not supported.** Participants who selected **Leaderboards** did not have higher **Extraversion** scores than those who did not select it ($p = 0.223$, $r = 0.109$).
- **H2.5: not supported.** Participants who selected **Challenges** did not have higher **Agreeableness** scores than those who did not select it ($p = 0.802$, $r = 0.022$).

On the other hand, there were some significant differences, which were not predicted by the existing literature and were not part of our hypotheses:

- **Agreeableness** scores are higher for participants who selected **Badges** ($p = 0.022$, $r = 0.204$) and **Chance** ($p = 0.004$, $r = 0.257$).
- **Openness** scores are higher for participants who selected **Badges** ($p = 0.037$, $r = 0.186$), **Leaderboards** ($p = 0.040$, $r = 0.183$), and **Chance** ($p = 0.043$, $r = 0.180$).

There were no significant relationships between the participants' selection of game elements and their genders (see **Table 4**). Therefore, **H3.1 and H3.2 are not supported**.

Regarding age, there was just one significant difference (see **Figure 3** and **Table 5**): participants who selected **Moderating role** were younger (*Med* = 30.5) than participants who did not select it (*Med* = 33.0, $p = 0.040$, $r = 0.183$; note that this is the absolute value of r because SPSS does not consider the direction of the relationship on the output of the Mann–Whitney *U*-test). However, age was not significantly different between participants who selected Badges, Unlockable content, Challenges, and Chance and the participants who did not select them. Therefore, **H4 is only partially supported**.

4.2. RQ2: Task Performance and User Engagement

To answer **RQ2**, we compared the participants' task performance between both conditions across the seven measures: total points earned, final level achieved, total images tagged, total tags entered

TABLE 5 | Non-parametric tests (Mann–Whitney U) comparing the differences in age between users who selected or not each gameful design element.

	Levels	Moderation	Badges	Unlockables	Leaderboard	Challenges	Chance	Power-ups
N/Y	42/84	99/27	60/66	78/48	65/61	71/55	79/47	57/69
$\widetilde{\text{No}}$	29.00	33.00	31.50	33.00	30.00	30.00	31.00	31.00
$\widetilde{\text{Yes}}$	33.50	30.50	34.00	31.50	35.00	35.00	35.00	33.00
U	1557.5	950.5	1882.5	1534.0	1810.0	1665.0	1457.0	1882.5
Z	0.866	2.049	0.319	1.595	0.702	1.294	1.918	0.246
p	.387	**0.040**	0.750	0.111	0.483	0.196	0.055	0.806
r	.077	**0.183**	0.028	0.142	0.063	0.115	0.171	0.022

N = 126.
Bolded values are significant at the 0.05 level.
***N/Y**: number of participants who did not select/did select each element.*
***$\widetilde{\text{No}}$**: median age for users who did not select each element.*
***$\widetilde{\text{Yes}}$**: median age for users who selected each element.*
***U/Z/p**: results of the Mann-Whitney U tests.*
***r**: effect sizes, calculated as $r = Z \div \sqrt{N}$.*
*The absolute values of **Z** and **r** are displayed for improved readability.*

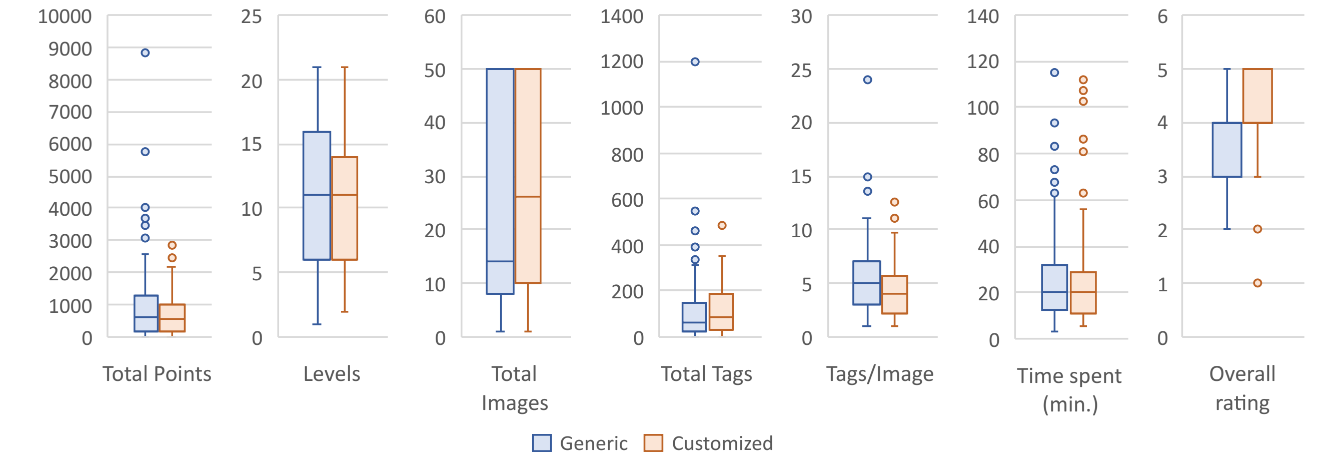


FIGURE 4 | Differences in performance and engagement variables between conditions (Study 1).

for all images, average tags per image, and time spent in the application (measured in minutes on study 1 and in days active in the application on study 2). User engagement was compared between both condition across the seven dimensions of the intrinsic motivation inventory (IMI). Participants' rating of their experience of selecting game elements (from their answer to **Q2**) was also compared as a measure of user engagement. Because the measures were not parametric, we employed the Mann–Whitney U-test and like in the previous subsection, we also calculated the effect size $r = Z \div \sqrt{N}$. We analyzed the data from each study separately to avoid the recruitment method (Mechanical Turk vs social media) as a confounding variable.

4.2.1. Study 1

Figure 4 displays the box plots comparing the performance and engagement variables between participants who selected or did not select each element. **Table 6** presents the results of the statistical tests.

Regarding task performance, users in the customized (experimental) condition classified more images ($Med = 26.0$) than in the generic (control) condition ($Med = 14.5$, $p = 0.013$, $r = 0.177$, a weak effect size). In this application, classifying more images means that participants contributed more to the systemic goal that was presented to them (collecting tags for images), and is therefore a relevant performance improvement. Nonetheless, the total number of tags did not change significantly between conditions. Because participants tagged more images in the customized condition, but wrote approximately the same total number of tags, the number of tags per image dropped significantly from $Med = 5.0$ in the generic condition to $Med = 4.0$ tags per image in the customized condition ($p = 0.008$, $r = 0.188$, a weak effect size). The other measures of task performance were not significantly different between conditions. Therefore, **H5 is partially supported in study 1**.

Regarding engagement, there were no statistically significant differences for any of the IMI measures. On the other hand, the experience rating was significantly higher in the customized condition than in the generic condition: $p = 0.025$, $r = 0.160$ (a weak effect size). Although the calculated median rating was 4.0 in both conditions, the boxplot in **Figure 4** shows that 50% of the

TABLE 6 | Comparison of performance and engagement variables and IMI scores between conditions (Study 1).

Variables	Median (generic)	Median (customized)	U	Z	p	r
Total points	611.5	551.0	4526.5	0.928	0.354	0.066
Level	11.0	11.0	4553.5	0.863	0.388	0.061
Total images	14.5	26.0	3915.5	2.495	**0.013**	**0.177**
Total tags	62.5	83.0	4548.5	0.873	0.383	0.062
Tags per image	5.0	4.0	3836.0	2.642	**0.008**	**0.188**
Time Spent (min.)	20.0	20.0	4807.5	0.231	0.817	0.016
Experience rating	4.0	4.0	3924.0	2.241	**0.025**	**0.160**
IMI scores	**Median (generic)**	**Median (customized)**	**U**	**Z**	***p***	**r**
Interest	5.00	5.17	4393.0	1.146	0.252	0.081
Competence	5.50	5.75	4184.5	1.782	0.075	0.127
Effort	5.50	6.00	4375.5	1.193	0.233	0.085
Pressure	2.00	2.25	4696.5	0.388	0.698	0.028
Choice	5.50	5.63	4388.5	1.159	0.246	0.082
Value	5.00	5.00	4362.0	1.225	0.220	0.087
Relatedness	3.83	4.00	4479.5	1.046	0.296	0.075

N = 198 (99 per condition).
Bolded values are significant at the 0.05 level.
Overall rating is a 5-point scale (range: 1–5).
***U/Z/p**: results of the Mann–Whitney U-tests.*
***r**: effect sizes, calculated as $r = Z \div \sqrt{N}$.*
*The absolute values of **Z** and **r** are displayed for improved readability.*

ratings in the generic condition were between 3 and 4, whereas 50% of the ratings in the customized condition were between 4 and 5. Therefore, **H6 is partially supported in study 1**.

4.2.2. Study 2

Figure 5 displays the box plots comparing the performance and engagement variables between participants who selected or did not select each element. **Table 7** presents the results of the statistical tests.

Regarding task performance, the number of images classified in the customized (experimental) ($Med = 51.0$) is higher than in the generic (control) condition ($Med = 25.0$); however, the difference is not significant: $p = 0.064$, $r = 0.132$. Although this effect is not significant in study 2, it is interesting to note on the box plot that only participants in the experimental condition classified all the available 100 images, but none in the control condition. Similarly to study 1, the total number of tags did not change significantly between conditions, but differently from the first study, this time the number of tags per image also did not change significantly between conditions. The other measures of task performance were once more not significantly different between conditions. Therefore, **H5 is not supported in study 2**.

Regarding engagement, participants scored higher in the IMI measure for competence in the customized condition ($Med = 5.25$) than the generic condition ($Med = 4.50$, $p = 0.022$, $r = 0.163$). The other IMI scores were not significantly different between conditions. In addition, the experience rating was significantly higher in the customized condition ($Med = 4.0$) than in the generic condition ($Med = 3.0$, $p = 0.012$, $r = 0.179$). This effect size is slightly larger than in the first study, and the difference in the medians is more pronounced, but the effect still has a similar order of magnitude (weak). Therefore, **H6 is partially supported in study 2**.

4.3. Participants' Perceived Usefulness of Each Element

Table 8 presents the number of times that each gameful design element was listed as the participant's preferred element, the element that most influenced them, or the element that most motivated them (in response to **Q6**, **Q7**, and **Q8**). The differences in the frequency distributions between conditions are significant for all three variables: $p = 0.003$ for preferred element, $p = 0.005$ for most influential element, $p = 0.001$ for most motivational element (Pearson's chi-square test; $N = 227$).

Although this was not one of the original research questions for this study, an analysis of this table provides interesting insights for personalized gameful design.

First, it is noticeable that the number of times that participants mentioned each element as preferred, influential, or motivating is similar, meaning that participants probably enjoy an element when they perceive it as influential or motivational. An interesting exception is that a few participants in the generic condition perceived Levels as the most motivating element even if it was not their preferred or most influential element.

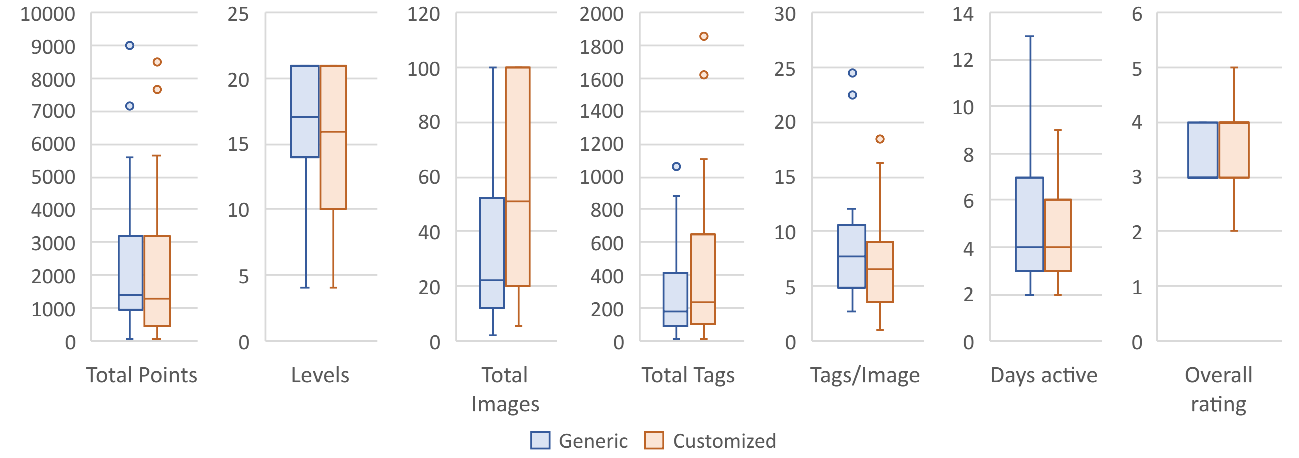


FIGURE 5 | Differences in performance and engagement variables between conditions (Study 2).

TABLE 7 | Comparison of performance and engagement variables and IMI scores between conditions (Study 2).

Variables	Median (generic)	Median (customized)	U	Z	*p*	r
Total Points	1622.0	1359.0	341.0	0.407	0.684	0.029
Level	18.0	16.0	331.0	0.588	0.556	0.042
Total images	25.0	51.0	258.5	1.854	0.064	0.132
Total tags	216.0	242.0	323.5	0.709	0.478	0.050
Tags per image	7.8	6.7	287.0	1.341	0.180	0.095
Days active	4.0	5.0	361.5	0.053	0.958	0.004
Experience rating	3.0	4.0	195.0	2.506	**0.012**	**0.179**
IMI Scores	**Median (generic)**	**Median (customized)**	**U**	**Z**	***p***	**r**
Interest	4.33	4.50	321.5	0.745	0.456	0.053
Competence	4.50	5.25	232.5	2.291	**0.022**	**0.163**
Effort	4.00	4.25	354.0	0.182	0.855	0.013
Pressure	2.25	2.50	344.5	0.347	0.728	0.025
Choice	5.75	6.00	274.0	1.574	0.116	0.112
Value	4.00	4.50	303.5	1.059	0.289	0.075
Relatedness	3.50	3.33	337.5	0.468	0.640	0.034

N = 54 (27 per condition).
Bolded values are significant at the 0.05 level.
Overall rating is a 5-point scale (range: 1–5).
***U/Z/p**: results of the Mann–Whitney U-tests.*
***r**: effect sizes, calculated as $r = Z \div \sqrt{N}$.*
*The absolute values of **Z** and **r** are displayed for improved readability.*

Another insightful observation is that Challenges and Power-ups were mentioned more often than any other element as the preferred and most influential elements, and as the second/third most motivating elements by participants in the generic condition. However, they were mentioned less often by participants in the customized condition, to the point that they are not mentioned more often than some of the other elements. In particular, Power-ups showed an accentuated decline.

On the other hand, Levels was only the third more cited element as preferred and most influential in the generic condition, but it appears as the sole element most often mentioned as preferred, most influential, and most motivating by participants in the customized condition. It was also the element selected more often by participants in the customization: 84 times. Similarly, Leaderboards, and Chance also received more interest by being mentioned more often as preferred,

TABLE 8 | Comparison of preferred, most influential, and most motivating elements per condition.

Elements	Generic			Customized			
	Pref	Infl	Mot	Sel	Pref	Infl	Mot
Levels	20	21	32	84	37	37	38
Moderating role	2	6	1	27	2	8	0
Badges	8	6	7	66	10	9	10
Unlockable content	2	2	1	48	2	2	2
Leaderboards	13	11	11	61	17	17	22
Challenges	29	32	28	55	23	16	14
Chance	2	5	2	47	11	13	14
Power-ups	35	28	30	69	14	14	14
N/A	15	15	14	–	10	10	12

N = 252 (126 per condition).

Sel, Number of times that each element was selected by participants in the customization step. Pref, Number of times that each element was listed as the participant's preferred element (Q6 in the end survey). Infl, Number of times that each element was listed as the participant's most influential element (Q7 in the end survey). Mot, Number of times that each element was listed as the participant's most motivating element (Q8 in the end survey).

most influential, and most motivating by participants in the customized condition than in the generic condition.

There were no significant relationships between participants' preferred element, most influential element, and most motivating element with their user type scores, personality traits, age (Kruskal–Wallis *H*-test), and gender (Pearson's chi-square test). However, with a sample of 252 participants distributed across eight gameful design elements, and some of the elements being mentioned very few times (e.g., moderating role and unlockable content), the sample was probably not large enough to detect any relationship.

4.4. Thematic Analysis

In this subsection, we examine participants' responses to the open-ended questions in our post-study survey. Three of the questions (**Q1**, **Q2**, and **Q3**) were meant to just obtain the general impressions about the use of the platform from participants in both conditions. The goal of this part of the analysis is to better understand the context in which participants' experience with the application occurred.

On the other hand, two questions specifically asked participants if the elements they selected matched their preferences and how they influenced the enjoyment of the task (**Q4** and **Q5**). While these questions make more sense in the customized condition, we also analyzed participants' responses in the generic condition to understand their experience. By having all the elements available to them, participants in the generic condition had to select elements for their experience by just deciding when to interact with them and when to ignore them, i.e., just by shifting their attention focus. Differently, participants in the customized condition were allowed to pre-select the elements they wanted to use, so their user interface was cleaner because only the selected elements were shown. The goal of this part of the analysis is to understand how participants experienced the customization and how their experiences differed by having all elements available to them (control condition) or being able to pre-select the desired elements (experimental condition).

These analyses were carried out by the first author using thematic analysis. The focus of our analysis was to identify themes that represented recurrent answers to the open questions answered by participants in the end survey. For example, **Q1** is "Overall, how do you describe your experience with the image classification activities you just completed?" Therefore, we focused our analysis in summarizing the themes frequently used by participants to describe their experience. Our analysis procedure was similar to reflexive thematic analysis (Braun and Clarke, 2006, 2019). Thus, the coding process was flexible, without a code book, and carried out by a single researcher. The process consisted on four steps: (1) *familiarization with the data*, i.e., an initial reading to become familiar with the content, (2) *coding*, i.e., labeling each participant's response with words extracted from the content of their answer, (3) *theme generation*, i.e., summarizing the themes from the codes that appeared more frequently, and (4) *writing up*, i.e., reporting the identified themes along with quotes from participants. These steps were carried out separately for each question in the survey (**Q1–Q5**). We combined the data from both studies for the analyses.

In the remainder of this subsection, we also present selected quotes from participants' responses to illustrate the identified themes.

4.4.1. Overall Experience

In response to **Q1**, some participants mentioned that they enjoyed their experience with the applications, but others did not. Participants who enjoyed the experience mentioned that it was fun, unique, easy, and interesting. Some specifically mentioned that the game elements contributed to making the experience fun or unique, whereas others mentioned that the photos were enjoyable, and some did not explicitly explain the reason for their enjoyment. For example:

> "I really enjoyed it more than I expected. The game elements captured my attention and made me want to do more of the tasks to earn more badges, complete challenges, etc." (P17, study 1, control condition)

> "It was interesting and enjoyable to come up with tags for the images. They were also nice photographs so it was fun to look at them." (P191, study 1, experimental condition)

Participants who reported less positive experiences mentioned that the task was boring or difficult, they had trouble understanding some of the instructions, or they felt that the game elements were not useful, for example, because the in-game rewards could not be carried out to the real world.

> "It was very dull, there was no real tangible reward outside your gamification systems. Without some kind of bonus this felt very 'Meh'." (P97, study 1, control condition)
> "It started out interesting and a bit exciting, but got boring after the first dozen or so images." (P3, study 2, control condition)

Regarding the experience of selecting elements, responses to **Q2** in the control condition were varied, which was expected because those participants did not actually customize their elements. Some participants just mentioned that interacting with the game elements was enjoyable, others said that they were not interested in the game elements, and some participants said that they did not actually select any game element:

> "I thought the game activities added a benefit to the classification task. It made it more fun and interesting." (P2, study 1, control condition)
> "I did explore the various game elements, but none of them were very interesting to me. I made use of the power-ups and claimed the challenges, but was a bit weirded out by the gifts feature and didn't really care about the levels, badges, or leaderboard. Also there were so many different elements that it was a bit confusing/hard to keep track of, so I mostly just stuck with the actual tagging." (P3, study 2, control condition)
> "I did not really do much in the way of customizing besides the avatar." (P103, study 1, control condition)

On the other hand, participants in the experimental condition did actually select game elements and so were able to explicitly comment about this experience. Participants said that the customization was easy, that they felt in control, and they tried to select the elements that matched their style or would help them in the task. Some participants enjoyed the possibility of customization because it is generally not offered or because they recognize that people may have different preferences. For example:

> "They were akin to filters on a shopping website in that I could choose the data that was most important/relevant to me and what I wanted to best assist me in my assessment of my progress." (P24, study 1, experimental condition)
> "I felt like I had control and like what I was doing mattered." (P44, study 1, experimental condition)
> "It was interesting because not many games allow you to do this." (P89, study 1, experimental condition)
> "It's a good idea, everyone can choose what they prefer, so every can play and be motivated with something they are interesting in." (P40, study 2, experimental condition)

However, there were also some participants who disliked the customization because it was not necessary or did not add much to their experience, they felt that the description of the elements was not enough for an informed choice, or that the application should allow them to modify their initial selection.

> "I thought it wasn't really necessary. I always try my best." (P10, study 1, experimental condition)
> "A bit arbitrary and there was little information given for each choice. I went in blind and I was stuck with what I chose." (P4, study 2, experimental condition)

With regards to the game elements offered by the system (**Q3**), participants who were satisfied mentioned that the elements made the task more fun or gameful, that they were varied enough, they were easy to choose, and provided a personalized experience. For example:

> "I was very satisfied. I felt like there was a good variety of options that I was familiar with. I liked some and disliked others, so I liked that I was able to pick." (P14, study 1, experimental condition)
> "I was more than satisfied by all the game elements provided. I knew that I could take any one of them and make the game more fun, but having more than one to choose from made it even more exciting." (P37, study 1, control condition)
> "Yes, lots of variety to cater to different personalities and improve user experience." (P53, study 2, experimental condition)

Some participants also reported not paying attention to the game elements, not interacting with them, or just feeling that they did not change anything. It seems that these participants had no specific issue with the offered elements, they just preferred to focus on the image classification task and were not interested in using the game elements. For example:

> "None of them make the task more interesting. The points mean nothing." (P35, study 1, experimental condition)
> "It really did not change anything for me." (P82, study 1, control condition)
> "Neutral, because I didn't use them." (P48, study 2, experimental condition)

4.4.2. Preference Matching and Task Enjoyment

When asked if they were able to select game elements that matched their preferences (**Q4**), some participants in the control condition responded that they could not select anything, which was to be expected as it was really the case. Some participants also mentioned that they were not aware of or did not understand what the game elements were. Echoing some of the responses in the previous subsection, there were also some participants who just did not care about the game elements or did not have any preference. But it is also interesting to note that some participants felt that they could select elements just because they could take a look at all of them and choose the ones they wanted to use and those they wanted to ignore. Other participants interpreted the ability to use some elements (for example, activating a power-up) as if it was an ability to select the game elements they wanted, which is understandable because they were not given a mechanism to better customize their experience like participants in the experimental condition.

> "No, I was just given game elements that would be in place with no options to choose." (P17, study 1, control condition)
> "Sort of. I played around with a whole bunch of them and they were all available to me as far as I could tell. My preferences are to unlock things which were available, so I would say the preferences were met." (P168, study 1, control condition)
> "Yes, it was mostly easy to ignore the ones I didn't care about (except the moderation feature, selecting yes/no for other people's tags, which got kind of annoying after a while since it popped up after each image)." (P3, study 2, control condition)
> "Didn't have a strong feeling with game elements. So no preferences really. I think it might be because that these techniques have been used too many times in a lot of applications, so people (or at least me) learn to ignore this and get to the core." (P8, study 2, control condition)

As expected, participants in the experimental (customized) condition responded more specifically about the task of selecting the game elements in the customization interface. Most participants said they were satisfied with the task of selecting game elements, mentioning that they were able to choose the elements that they preferred or that they thought would motivate them more. Only a few participants said that they did not appreciate the customization task because they would prefer to focus on the image classification task. Specifically, some participants on study 1 said they wanted to just classify the images and avoid interacting with the game elements so they would not decrease their hourly earnings. Logically, this reason did not appear on study 2 as they were participating voluntarily, not for payment like the Mechanical Turk workers from study 1.

> "Yes. I didn't want to examine other people's work, so it was nice that we had choices. If I was doing this long term, the game elements I chose would have added something to the activity." (P26, study 1, experimental condition)
> "Yes, I was able to find and select game elements that matched my preferences that would motivate me." (P78, study 1, experimental condition)
> "Not really—the only thing I really cared about was increasing my hourly earnings." (P91, study 1, experimental condition)
> "Yes because you can choose among a large set of game elements so you can easily find the one(s) that suit(s) you the best." (P42, study 2, experimental condition)

Finally, we asked participants if their selection of game elements influenced their enjoyment of the image classification task (**Q5**). A few participants in the control condition said that the game elements made the experience more enjoyable to them, but they did not relate this effect to the possibility of a customized experience, which was expected as they did not have a choice. However, many participants said that the game elements did not influence their enjoyment of the task. Explanations for this fact suggest that the task was already enjoyable enough without the game elements, or it was boring and the game elements could not change this fact.

> "The game elements made this a lot more enjoyable than a simple image classification task. I could see doing this for fun in my spare time." (P28, study 1, control condition)
> "I don't know if it influenced it too much. I was content doing the task without much customization, although I didn't explore it too deeply. I think if I had it would have become more enjoyable." (P55, study 1, control condition)
> "It didn't really. The task would've been the same without them." (P197, study 1, control condition)
> "Not that much. I mean, of course getting one badge made me feel accomplished and want to collect as many of them as possible but I did enjoy simply tagging the images without any gaming elements." (P5, study 2, control condition)

Responses from participants in the experimental condition generally followed the same themes, with some participants mentioning that the game elements made the experience more enjoyable, whereas others said that they did not make much difference. We were particularly interested in how participants felt that having customized their experience influenced their enjoyment; however, only a few participants specifically mentioned this aspect. Those who did said that customizing the game elements helped shape their experience and made them feel in control, or allowed them to choose their own goals or rewards.

> "I felt like I had control over the game." (P13, study 1, experimental condition)
> "I feel that my selection was important and really shaped my experience. I was motivated by the star rewards." (P51, study 1, experimental condition)
> "The selection of game elements allowed me to make the image classification suit my needs. It allowed me to make the classification more enjoyable and try to earn the highest score." (P77, study 1, experimental condition)
> "I don't think so because the task itself remained the same." (P89, study 1, experimental condition)
> "I think being able to choose rewards for myself made them more meaningful, choosing the elements that made me want to keep on going. Achieving those levels/badges/leaderboard spots/etc because I had decided that was the cool thing in this game made it more interesting than if all of those elements had been hardcoded and set for me by the game masters." (P2. study 2, experimental condition)
> "Not at all, I kind of forgot the game elements were there." (P14, study 2, experimental condition)
> "Like most people, I enjoyed being rewarded for my progress which allowed me to set specific goals and I felt accomplished when I was able to reach them. The game elements allowed me to be a little competitive with myself which is a good motivator for me." (P22, study 2, experimental condition)

5. DISCUSSION

5.1. Influence of User Characteristics on Element Selection

After analyzing the relationships between Hexad user type scores and gameful element selections, we found eight significant ones. From these, five were expected according to **H1** (Achiever-Badges, Achiever-Challenges, Achiever-Chance, Player-Leaderboards, and Player-Challenges); one was not expected, but is clearly understandable considering the description of the user type (Player-Power-ups); and two were not expected and cannot be easily explained (Philanthropist-Badges and Free Spirit-Chance). As reported in section 4.1, these results partially support **H1**, i.e., some of the expected differences in user type scores between participants who selected or not each game element were observed, but not all of them.

These results further support previous statements (such as Tondello et al., 2017a; Hallifax et al., 2019; Tondello, 2019) about the suitability of the Hexad user types as an adequate model of user preferences for the selection of gameful design elements in personalized gamification. Therefore, our work adds to the existing evidence that users with higher scores in specific user types are more likely to select specific game elements when given the choice, according to the eight pairs of user types and gameful elements listed above. By extension, we can assume that other relationships between user types and gameful elements proposed in the literature but not tested in this study may likely also hold true when tested in practice.

This contribution is important because the literature had relied so far on survey studies with only self-reported answers to establish relationships between Hexad user types and gameful design elements. Thus, the question remained if users would behave in an actual gameful system like they stated in their self-reported responses. The present work is the first one, to the best of our knowledge, to answer this question by demonstrating that participants' behavior (selection of gameful design elements) indeed correspond to their self-reported Hexad user type scores. While previous studies had compared two types of self-reported measures (user type scores and hypothetical game element preferences), we compared a self-reported measure (user type scores) with participant's actual behavior (their choice of game elements). This reinforces the confidence of gamification designers when using personalized gameful design methods that rely on selecting gameful design elements based on user types (such as Marczewski, 2018; Mora Carreño, 2018; Tondello, 2019).

On the other hand, some relationships between user types and gameful elements that were expected were not significant in this study (Philanthropist-Levels, Philanthropist-Moderating role, Socializer-Leaderboards, and Disruptor-Challenges). We believe that this happened because the context of the task was not favorable to create the type of experience that these users would enjoy. For example, the way that moderating role was implemented in our application did not seem very engaging as very few participants selected and enjoyed it; the leaderboard may have looked underwhelming because it was a very short experience and participants did not know and interact with each other. Better designs for these elements might have led to a higher appreciation by these participants. Additional studies will need to better evaluate these relationships.

Our results differ from those of Lessel et al. (2018) because they were not able to observe clear relationships between Hexad user types and gameful design elements like we did. But in their study, they asked participants to consider a few scenarios and try to design a gameful system for each one, which they thought they would enjoy. Although it provided many insights about how participants approached this task of designing a gameful experience for themselves, we believe that it speaks more about their capacity as designers than users because the designs were not implemented and tested. In contrast, our study allowed participants to actually use the gameful design elements, effectively testing how well each element worked for each participant.

Regarding the relationship between personality traits and gameful design elements, we found five significant ones (Agreeableness-Badges, Agreeableness-Chance, Openness-Badges, Opennes-Leaderboards, and Opennes-Chance). However, none of them were expected according to previous research or are not explained by the available literature. Therefore, **H2** was not supported. These results mirror previous literature, which also noticed inconsistent results when analyzing gameful design element preferences by personality traits (such as Tondello et al., 2017a; Lessel et al., 2018; Hallifax et al., 2019). Due to these variations in results across studies, it is hard to suggest how gamification designers could use this information in their practice. Therefore, we echo the existing literature in arguing that the Hexad user types are a better model for user preferences in personalized gamification than the Big-5 personality traits.

Finally, we found only one significant relationship between participants' age and their gameful element choices (moderating role was generally selected by younger participants) and none between gender and element choices. Thus, **H3** was only partially supported and **H4** was not supported. It is not clear why the differences identified in the existing literature were not observed in this study. More research will be needed to specifically try to observe in practice these different preferences by age and gender identified in the previous survey studies.

In summary, our response to **RQ1** "If allowed to choose the gameful design elements they prefer, do user choices correspond to the theoretical relationships with user types, personality, gender, and age reported in previous survey-based studies?" is that we found evidence that user choices do indeed correspond to their Hexad user type scores as reported in previous studies, at least partially. However, clear correspondences between element choices and participants' personalities, genders, and ages were not observed.

5.2. Task Performance and User Engagement

The results showed a significant improvement on the number of images tagged per participant in the experimental condition in study 1. Thus, **H5** was partially supported in study 1, but it was not supported in study 2. Additionally, results showed a higher rating for the experience of selecting game elements in both studies. Thus, **H6** was partially supported in both studies. However, participants spent approximately the same amount of time and wrote approximately the same number of tags for all images in both conditions. In addition, participants on study 1 did not want to lower their hourly rate of earnings in the Mechanical Turk platform, so they compensated the incentive to tag more images by writing less tags per image in the experimental condition. This effect was not observed on study 2, as the number of tags per image was not significantly different between conditions.

Therefore, it seems that personalization encouraged participants to achieve a higher task performance by classifying each image faster in order to complete more images in total. In a real application, this could be what designers wanted or not. In our application, this can be easily understood as a result of our design. Our application gave participants 10 points for each image classified and one additional point for each tag written for the image. It is reasonable to assume that participants quickly realized that they could earn more points by classifying more different images instead of spending time writing additional tags for the same image. If instead the design goal was to have participants adding more tags for each image, we could modify the design so that more points would be awarded for additional tags and less points for each classified image. We suppose that the performance change would have occurred in the opposite direction then, i.e., that participants would have classified less images, but provided more tags for each image.

This is evidence that personalization or customization can lead to higher task performance than generic gamification. Nonetheless, the design and incentives of the system must be well adjusted by the designers to achieve the intended goal. Our results showed that performance increased for the activity that was better rewarded by the system (classifying more images), even by perhaps decreasing the performance of other elements of the activity (e.g., adding more tags for each image). However, this should not be understood as an issue of personalization; it is just important to realize that personalization may not be able to automatically improve performance in all aspects of the task. It is part of the designer's job to fine tune the mechanics of gameplay to incentivize better performance where it is more important.

The intrinsic motivation measures did not differ significantly between conditions, except that perceived challenge was higher for participants in the customized condition on study 2. Looking at participants' free-text responses summarized in the thematic analysis, it is clear that some participants were already intrinsically motivated by the task and said that the game elements were not needed, whereas others said that they were bored by the task and the game elements could not change it. Considering this, it seems that the observed effects on task performance and engagement due to personalization did not occur because of changes in participants' intrinsic motivation. Therefore, future studies could consider different engagement or experience measures instead of the IMI to try and identify what are the mediators of these effects.

These findings are consistent with the evidence by Lessel et al. (2017, 2019), in which task performance was also higher for personalized than generic gameful systems. Even though our study is not the first to demonstrate the positive effects of personalized gamification for task performance, evidence of these effects is still scarce and additional studies are still needed to reinforce the preliminary findings. Our work contributes with additional empirical evidence of performance improvement with personalized gamification on an application context that is similar to that of Lessel et al. (image classification), but with a different application design and study design.

The analysis of participants' qualitative answers showed that the customization task was generally well received. However, designers should note that some participants asked for better descriptions of the game elements, for the possibility of changing the initial selection, or disabling all the game elements entirely. These are all features that should be included in the design of a customized gameful application. Moreover, Lessel et al. (2019) had already suggested that offering the possibility of disabling all the game elements may be desirable for some users, which is supported by some of the free-text answers from our participants.

In summary, our response to **RQ2** "Are user engagement and performance better for a personalized gameful system than a generic system?" is yes, user engagement and performance can be improved by adopting a personalized instead of a generic gamification design. However, designers must pay attention to clearly incentivize the behaviors that they want to improve in the gameful system, as providing more incentives for one type of behavior can lead to increased performance for that behavior in detriment of performance for different behaviors. Nonetheless, these findings are important because they demonstrate that it is worthy investing in personalized gameful design, which is undoubtedly more complex than generic gameful design, because it can lead to better achievement of the goals of the gameful system.

5.3. Participants' Perceived Usefulness of Each Element

The results from the analysis of participants' preferred, most influential, and most motivating elements suggest that users may perceive and experience some gameful design elements differently depending on whether they selected those elements themselves, or had no choice. Also considering the findings by Lessel et al. (2018), we can also suppose that participants would similarly experience elements differently if they were designing a system instead of just using a system previously built for them. This suggests an interesting line of investigation for future work because so far the relationships between user types and gameful design elements have been presented as universal. Future studies could investigate if the differences in user perception of each gameful design element depending whether they are designing, customizing, or just using the elements without modification can be replicated and mapped.

It is also noteworthy that we found no relationship between participants' preferred elements and their Hexad user type scores, even though there were relationships between those scores and the frequency of selection of specific game elements. In line with the comment above, it may be that the user type scores are currently better in capturing users' desire and intention regarding the use of specific game elements, rather than their perceived preferences after actually using the elements. It is possible that other factors may be in play during the actual user experience with the elements. For example, there are multiple ways of designing and implementing the same game element and Mora Carreño (2018, chapter 3) suggested that different designs can make each element more or less appealing for different user types. This is a question that requires more studies in future work.

5.4. Limitations and Future Work

Our study provided valuable findings about the correspondence between user types and gameful design elements in participant preferences, as well as the potential effect of personalized gamification on task performance. However, it was limited to one application context, which was image classification. We expect that similar results will be observed in different contexts and with different types of tasks, but this must be verified in future work. Therefore, we plan to conduct additional studies replacing image classification with different types of tasks.

Furthermore, we evaluated task performance considering only the number of tagged images and tags, but not the quality of tags. In future studies, it would be interesting to also consider tag quality by evaluating if the tags provided by participants corresponded to the presented images, to confirm that the quality of the tags remained the same or improved together with the improvement in the number of tagged images.

Additionally, participants in the first study were all Mechanical Turk workers residing in the United States of America. On the other hand, the second study had a more varied participation, with similar results to the first one, which suggest that the findings can probably be replicated with more diverse samples. Nonetheless, the difference in the number of images classified between conditions was significant in the first but not in the second study, despite a similar median difference. We believe that this was due to the smaller sample size in the second study. However, **Table 2** showed a few differences in the mean user type and personality trait scores between the two data sets. These differences may also have had any influence in the different results between the two studies. However, testing if the user type or personality trait scores would moderate the performance increase in **H5** was not one of the goals of this study. Therefore, we plan to carry out additional studies with participants from different countries to verify if our findings are similar for people with different cultural backgrounds. These additional studies may also test if demographic variables, such as user types, personality traits, age, and gender, may act as moderators of the performance difference between participants using a generic or a customized gameful application.

Finally, the personality traits inventory used in this study (Rammstedt and John, 2007) is very short, with just two items per trait. Although it has been validated and used frequently in HCI studies, its reliability is lower than longer scales, as the α values in **Table 2** show. This can have contributed to the inconsistent results in our analysis of the relationship between personality trait scores and element preferences. Thus, we plan to conduct additional studies using longer and more reliable personality trait scales to obtain more consistent results in the future.

6. CONCLUSION

In the present work, we showed that participants' choice of gameful design elements in a customizable gameful application partly corresponded to their Hexad user type scores, as predicted by models previously established from survey-based studies. This is the first study to demonstrate these relationships based on the actual observation of participants' experiences with a gameful application. This shows that personalized gameful design methods based on the selection of gameful design elements by user types can work in practice as suggested in the current literature.

On the other hand, these significant relationships were of weak effect sizes. Additionally, participants' user type scores were not related to their preferred, most influential, or most motivational game elements after they had interacted with the platform. This suggests that gameful designers can use the Hexad user types as one of the factors for personalization, but not the only one. There are yet other factors to be discovered in future work to determine with more precision what the preferences of a specific user will be in a gameful system.

Moreover, participants achieved a higher task performance and a better experience of selecting which game elements to use in a customizable version of our gameful application than a generic version with the same gameful design elements. These results show that personalization or customization of gameful design elements is a viable solution to increase task performance and improve the user experience. Nonetheless, the design of our application encouraged users to improve the number of images classified without at the same time improving the number of tags per image. This means that personalization may be more effective in increasing user behaviors that are more explicitly incentivized, and not necessarily all user behaviors in the application. This is something that designers should take in consideration when creating any gameful system, and especially personalized ones.

This contribution is valuable to the HCI and gamification communities because several personalized gameful design methods have been recently suggested in the literature. Our work shows that they are a promising approach to improve the design of gameful applications and make them more successful in achieving their goals.

DATA AVAILABILITY STATEMENT

The datasets generated for this study are available on request to the corresponding author.

ETHICS STATEMENT

The studies involving human participants were reviewed and approved by Office of Research Ethics, University of Waterloo.

The patients/participants provided their written informed consent to participate in this study.

AUTHOR CONTRIBUTIONS

GT designed the study, developed the digital application, collected and analyzed data, and wrote the manuscript draft. LN revised and approved the study design. All authors contributed to manuscript revision, read, edited, and approved the submitted version.

FUNDING

This work was supported by the CNPq Brazil, SSHRC (895-2011-1014, IMMERSe), NSERC Discovery (RGPIN-2018-06576), NSERC CREATE SWaGUR, and CFI (35819).

ACKNOWLEDGMENTS

The authors would like to thank GT’s thesis committee (Daniel Vogel, Mark Hancock, Edith Law, James Wallace, and Juho Hamari) for the feedback on prior work leading to this contribution and Stuart Hallifax for helping with data collection.

REFERENCES

Adomavicius, G., and Tuzhilin, A. (2005). Personalization technologies: a process-oriented perspective. *Commun. ACM* 48, 83–90. doi: 10.1145/1089107.1089109

Altmeyer, M., Lessel, P., and Krüger, A. (2016). “Expense control: a gamified, semi-automated, crowd-based approach for receipt capturing,” in *Proceedings of the 21st International Conference on Intelligent User Interfaces - IUI ’16* (Sonoma, CA: ACM), 31–42. doi: 10.1145/2856767.2856790

Araújo Paiva, R. O., Bittencourt, I. I., Da Silva, A. P., Isotani, S., and Jaques, P. (2015). “Improving pedagogical recommendations by classifying students according to their interactional behavior in a gamified learning environment,” in *Proceedings of the ACM Symposium on Applied Computing - SAC ’15* (Salamanca: ACM), 233–238. doi: 10.1145/2695664.2695874

Bakkes, S., Tan, C. T., and Pisan, Y. (2012). “Personalised gaming,?? in *Proceedings of The 8th Australasian Conference on Interactive Entertainment Playing the System - IE ’12* (Auckland: ACM), 1–10. doi: 10.1145/2336727.2336731

Barata, G., Gama, S., Jorge, J., and Gonçalves, D. (2017). Studying student differentiation in gamified education: a long-term study. *Comput. Hum. Behav.* 71, 550–585. doi: 10.1016/j.chb.2016.08.049

Bartle, R. (1996). Hearts, clubs, diamonds, spades: players who suit MUDs. *J. MUD Res.* 1.

Böckle, M., Micheel, I., Bick, M., and Novak, J. (2018). “A design framework for adaptive gamification applications,” in *Proceedings of the 51st Hawaii International Conference on System Sciences (HICSS)* (Manoa: University of Hawaii), 1227–1236. doi: 10.24251/HICSS.2018.151

Böckle, M., Novak, J., and Bick, M. (2017). “Towards adaptive gamification: a synthesis of current developments,” in *Proceedings of the 25th European Conference on Information Systems (ECIS)* (Guimarães).

Bouzidi, R., Nicola, A. D., Nader, F., Chalal, R., and Laboratoire, M. (2019). “A systematic literature review of gamification design,” in *20th annual European GAME-ON Conference (GAME-ON’2019)* (Breda), 89–93.

Braun, V., and Clarke, V. (2006). Using thematic analysis in psychology. *Qual. Res. Psychol.* 3, 77–101. doi: 10.1191/1478088706qp063oa

Braun, V., and Clarke, V. (2019). Reflecting on reflexive thematic analysis. *Qual. Res. Sport Exerc. Health* 11, 589–597. doi: 10.1080/2159676X.2019.1628806

Buhrmester, M., Talaifar, S., and Gosling, S. D. (2018). An evaluation of Amazon’s mechanical turk, its rapid rise, and its effective use. *Perspect. Psychol. Sci.* 13, 149–154. doi: 10.1177/1745691617706516

Butler, C. (2014). “A framework for evaluating the effectiveness of gamification techniques by personality type,” in *HCI in Business. HCIB 2014. Lecture Notes in Computer Science, Vol. 8527*, ed F. Nah (Cham: Springer), 381–389. doi: 10.1007/978-3-319-07293-7_37

Chou, Y.-K. (2015). *Actionable Gamification - Beyond Points, Badges, and Leaderboards*. Octalysis Media, Kindle Edition.

Codish, D., and Ravid, G. (2017). “Gender moderation in gamification: does one size fit all?,” in *Proceedings of the 50th Hawaii International Conference on System Sciences - HICSS 2017* (Waikoloa Village, HI), 2006–2015. doi: 10.24251/HICSS.2017.244

Costa, P. T. Jr, and McCrae, R. R. (1998). “Trait theories of personality,” in *Advanced Personality*, eds D. F. Barone, M. Hersen, and B. Van Hasselt (Boston, MA: Springer), 103–121. doi: 10.1007/978-1-4419-8580-4_5

Deterding, S., Dixon, D., Khaled, R., and Nacke, L. E. (2011). “From game design elements to gamefulness: defining “gamification”,” in *Proceedings of the 15th International Academic MindTrek Conference* (Tampere: ACM), 9–15. doi: 10.1145/2181037.2181040

Field, A. (2009). *Discovering Statistics Using SPSS, 3rd Edn*. London: Sage Publications.

Goldberg, L. R. (1993). The structure of phenotypic personality traits. *Am. Psychol.* 48, 26–34. doi: 10.1037/0003-066X.48.1.26

Hallifax, S., Serna, A., Marty, J.-C., Lavoué, G., and Lavoué, E. (2019). “Factors to consider for tailored gamification,” in *Proceedings of the 2019 Annual Symposium on Computer-Human Interaction in Play - CHI PLAY ’19* (Barcelona: ACM), 559–572. doi: 10.1145/3311350.3347167

Herbert, B., Charles, D., Moore, A., and Charles, T. (2014). “An investigation of gamification typologies for enhancing learner motivation,” in *2014 International Conference on Interactive Technologies and Games (iTAG 2014)* (Nottingham, UK), 71–78. doi: 10.1109/iTAG.2014.17

Jia, Y., Xu, B., Karanam, Y., and Voida, S. (2016). “Personality-targeted gamification: a survey study on personality traits and motivational affordances,” in *Proceedings of the 34th Annual ACM Conference on Human Factors in Computing Systems - CHI ’16* (San Jose, CA), 2001–2013. doi: 10.1145/2858036.2858515

Kaptein, M., Markopoulos, P., De Ruyter, B., and Aarts, E. (2015). Personalizing persuasive technologies: explicit and implicit personalization using persuasion profiles. *Int. J. Hum. Comput. Stud.* 77, 38–51. doi: 10.1016/j.ijhcs.2015.01.004

Klock, A. C. T., Pimenta, M. S., and Gasparini, I. (2018). “A systematic mapping of the customization of game elements in gamified systems,” in *XVII Brazilian Symposium on Computer Games and Digital Entertainment (SBGames 2018)* (Foz do Iguaçu).

Koivisto, J., and Hamari, J. (2019). The rise of motivational information systems: a review of gamification research. *Int. J. Inform. Manage.* 45, 191–210. doi: 10.1016/j.ijinfomgt.2018.10.013

Landers, R. N., Auer, E. M., Collmus, A. B., and Armstrong, M. B. (2018). Gamification science, its history and future: definitions and a research agenda. *Simul. Gaming* 49, 315–337. doi: 10.1177/1046878118774385

Landers, R. N., Bauer, K. N., and Callan, R. C. (2017). Gamification of task performance with leaderboards: a goal setting experiment. *Comput. Hum. Behav.* 71, 508–515. doi: 10.1016/j.chb.2015.08.008

Lessel, P., Altmeyer, M., and Krüger, A. (2018). “Users as game designers: analyzing gamification concepts in a "bottom-up" setting,” in *Proceedings of the 22nd International Academic Mindtrek Conference - Mindtrek ’18* (Tampere: ACM), 1–10. doi: 10.1145/3275116.3275118

Lessel, P., Altmeyer, M., Müller, M., Wolff, C., and Krüger, A. (2016). ““Don’t whip me with your games”: Investigating “Bottom-Up” gamification,” in *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems - CHI ’16* (San Jose, CA: ACM), 2026–2037. doi: 10.1145/2858036.2858463

Lessel, P., Altmeyer, M., Müller, M., Wolff, C., and Krüger, A. (2017). "Measuring the effect of "bottom-up" gamification in a microtask setting," in *Proceedings of the 21st International Academic Mindtrek Conference - AcademicMindtrek '17* (Tampere: ACM), 63–72. doi: 10.1145/3131085.3131086

Lessel, P., Altmeyer, M., Schmeer, L. V., and Krüger, A. (2019). ""Enable or Disable Gamification?": Analyzing the Impact of Choice in a Gamified Image Tagging Task," in *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems (CHI 2019)* (Glasgow: ACM), 150. doi: 10.1145/3290605.3300380

Marczewski, A. (2018). *Even Ninja Monkeys Like to Play: Unicorn Edition*. Gamified.

McAuley, E., Duncan, T., and Tammen, V. V. (1989). Psychometric properties of the intrinsic motivation inventory in a competitive sport setting: a confirmatory factor analysis. *Res. Quart. Exerc. Sport* 60, 48–58. doi: 10.1080/02701367.1989.10607413

Monterrat, B., Desmarais, M., Lavoué, É., and George, S. (2015). "A player model for adaptive gamification in learning environments," in *Artificial Intelligence in Education. AIED 2015. Lecture Notes in Computer Science, Vol. 9112* (Cham: Springer), 297–306. doi: 10.1007/978-3-319-19773-9_30

Mora Carre no, A. (2018). *A framework for agile design of personalized gamification services* (Ph.D. Dissertation). Universitat Oberta de Catalunya, Barcelona, Spain.

Mora, A., Tondello, G. F., Calvet, L., González, C., Arnedo-Moreno, J., and Nacke, L. E. (2019). "The quest for a better tailoring of gameful design: an analysis of player type preferences," in *Proceedings of the XX International Conference on Human Computer Interaction - Interacción '19* (Donostia Gipuzkoa: ACM), 1–8. doi: 10.1145/3335595.3335625

Mora, A., Tondello, G. F., Nacke, L. E., and Arnedo-Moreno, J. (2018). "Effect of personalized gameful design on student engagement," in *Proceedings of the IEEE Global Engineering Education Conference - EDUCON 2018* (Tenerife). doi: 10.1109/EDUCON.2018.8363471

Nacke, L. E., Bateman, C., and Mandryk, R. L. (2014). BrainHex: a neurobiological gamer typology survey. *Entertain. Comput.* 5, 55–62. doi: 10.1016/j.entcom.2013.06.002

Nacke, L. E., and Deterding, S. (2017). The maturing of gamification research. *Comput. Hum. Behav*. 71, 450–454. doi: 10.1016/j.chb.2016.11.062

Nov, O., and Arazy, O. (2013). "Personality-targeted design: theory, experimental procedure, and preliminary results," in *Proceedings of the 2013 Conference on Computer Supported Cooperative Work - CSCW '13* (San Antonio, TX: ACM), 977–984. doi: 10.1145/2441776.2441887

Orji, R., Mandryk, R. L., Vassileva, J., and Gerling, K. M. (2013). "Tailoring persuasive health games to gamer type," in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems - CHI '13* (Paris), 2467–2476. doi: 10.1145/2470654.2481341

Orji, R., and Moffatt, K. (2018). Persuasive technology for health and wellness: state-of-the-art and emerging trends. *Health Inform. J*. 24, 66–91. doi: 10.1177/1460458216650979

Orji, R., Oyibo, K., and Tondello, G. F. (2017). "A comparison of system-controlled and user-controlled personalization approaches," in *UMAP 2017 - Adjunct Publication of the 25th Conference on User Modeling, Adaptation and Personalization* (Bratislava), 413–418. doi: 10.1145/3099023.3099116

Orji, R., Tondello, G. F., and Nacke, L. E. (2018). "Personalizing persuasive strategies in gameful systems to gamification user types," in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems - CHI '18* (Montreal, QC: ACM), 435. doi: 10.1145/3173574.3174009

Orji, R., Vassileva, J., and Mandryk, R. L. (2014). Modeling the efficacy of persuasive strategies for different gamer types in serious games for health. *User Model. User-Adapt. Interact*. 24, 453–498. doi: 10.1007/s11257-014-9149-8

Rammstedt, B., and John, O. P. (2007). Measuring personality in one minute or less: a 10-item short version of the Big Five Inventory in English and German. *J. Res. Pers*. 41, 203–212. doi: 10.1016/j.jrp.2006.02.001

Rapp, A., Hopfgartner, F., Hamari, J., Linehan, C., and Cena, F. (2019). Strengthening gamification studies: current trends and future opportunities of gamification research. *Int. J. Hum. Comput. Stud*. 127, 1–6. doi: 10.1016/j.ijhcs.2018.11.007

Roosta, F., Taghiyareh, F., and Mosharraf, M. (2016). "Personalization of gamification elements in an e-learning environment based on learners' motivation," in *2016 8th International Symposium on Telecommunications (IST)* (Tehran), 637–642. doi: 10.1109/ISTEL.2016.7881899

Seaborn, K., and Fels, D. I. (2015). Gamification in theory and action: a survey. *Int. J. Hum. Comput. Stud*. 74, 14–31. doi: 10.1016/j.ijhcs.2014.09.006

Sundar, S. S., and Marathe, S. S. (2010). Personalization versus customization: the importance of agency, privacy, and power usage. *Hum. Commun. Res*. 36, 298–322. doi: 10.1111/j.1468-2958.2010.01377.x

Tondello, G. F. (2019). *Dynamic personalization of gameful interactive systems* (Ph.D. thesis). University of Waterloo, Waterloo, ON, Canada.

Tondello, G. F., Kappen, D. L., Ganaba, M., and Nacke, L. E. (2019a). "Gameful design heuristics: a gamification inspection tool," in *Human-Computer Interaction. Perspectives on Design. Proceedings of HCI International 2019. LNCS 11566* (Springer), 224–244. doi: 10.1007/978-3-030-22646-6_16

Tondello, G. F., Kappen, D. L., Mekler, E. D., Ganaba, M., and Nacke, L. E. (2016a). "Heuristic evaluation for gameful design," in *Proceedings of the 2016 Annual Symposium on Computer-Human Interaction in Play Extended Abstracts - CHI PLAY EA '16*, 315–323. doi: 10.1145/2968120.2987729

Tondello, G. F., Mora, A., Marczewski, A., and Nacke, L. E. (2019b). Empirical validation of the gamification user types hexad scale in English and Spanish. *Int. J. Hum. Comput. Stud*. 127, 95–111. doi: 10.1016/j.ijhcs.2018.10.002

Tondello, G. F., Mora, A., and Nacke, L. E. (2017a). "Elements of gameful design emerging from user preferences," in *Proceedings of the 2017 Annual Symposium on Computer-Human Interaction in Play - CHI PLAY '17* (Amsterdam: ACM), 129–142. doi: 10.1145/3116595.3116627

Tondello, G. F., Orji, R., and Nacke, L. E. (2017b). "Recommender systems for personalized gamification," in *Adjunct Publication of the 25th Conference on User Modeling, Adaptation and Personalization - UMAP'17* (Bratislava: ACM), 425–430. doi: 10.1145/3099023.3099114

Tondello, G. F., Wehbe, R. R., Diamond, L., Busch, M., Marczewski, A., and Nacke, L. E. (2016b). "The gamification user types hexad scale," in *Proceedings of the 2016 Annual Symposium on Computer-Human Interaction in Play - CHI PLAY '16* (Austin, TX: ACM), 229–243. doi: 10.1145/2967934.2968082

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.