UNIVERSIDADE DE SÃO PAULO
Instituto de Ciências Matemáticas e de Computação

Tailoring gamified educational systems to users and context

Luiz Antonio Lima Rodrigues

Tese de Doutorado do Programa de Pós-Graduação em Ciências de Computação e Matemática Computacional (PPG-CCMC)

ICMC USP
SÃO CARLOS

SERVIÇO DE PÓS-GRADUAÇÃO DO ICMC-USP

Data de Depósito:

Assinatura: ______________________

Luiz Antonio Lima Rodrigues

Tailoring gamified educational systems to users and context

Doctoral dissertation submitted to the Instituto de Ciências Matemáticas e de Computação – ICMC-USP, in partial fulfillment of the requirements for the degree of the Doctorate Program in Computer Science and Computational Mathematics. *FINAL VERSION*

Concentration Area: Computer Science and Computational Mathematics

Advisor: Prof. Dr. Seiji Isotani

USP – São Carlos
January 2023

Ficha catalográfica elaborada pela Biblioteca Prof. Achille Bassi
e Seção Técnica de Informática, ICMC/USP,
com os dados inseridos pelo(a) autor(a)

L732t

Lima Rodrigues, Luiz Antonio
Tailoring gamified educational systems to users and context / Luiz Antonio Lima Rodrigues; orientador Seiji Isotani. -- São Carlos, 2023.
267 p.

Tese (Doutorado - Programa de Pós-Graduação em Ciências de Computação e Matemática Computacional) -- Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, 2023.

1. Gamification. 2. Personalization. 3. Persuasion. 4. Education. 5. Learning. I. Isotani, Seiji, orient. II. Título.

Bibliotecários responsáveis pela estrutura de catalogação da publicação de acordo com a AACR2:
Gláucia Maria Saia Cristianini - CRB - 8/4938
Juliana de Souza Moraes - CRB - 8/6176

Luiz Antonio Lima Rodrigues

Personalizando sistemas educacionais gamificados para usuários e o contexto

Tese apresentada ao Instituto de Ciências Matemáticas e de Computação – ICMC-USP, como parte dos requisitos para obtenção do título de Doutor em Ciências – Ciências de Computação e Matemática Computacional. *VERSÃO REVISADA*

Área de Concentração: Ciências de Computação e Matemática Computacional

Orientador: Prof. Dr. Seiji Isotani

USP – São Carlos
Janeiro de 2023

ACKNOWLEDGEMENTS

I would not have been able to complete this journey without the much appreciated help of so many people: family, friends, lab colleagues, coworkers, study participants, and so on. I'm deeply thankful to each and everyone that helped and supported me throughout this path, and each of these people knows how much I appreciate them! Although I am not good at doing this, I have to make some special acknowledgments here.

I'm deeply thankful to God, my wife Jheniffer, my parents Lucia and Toninho, and my brother Antenor, who supported, inspired, and helped me in everything since ever. Without you, nobody would be able to read this thesis because it would not exist. You are the best!

I am also profoundly thankful to Seiji Isotani, my supervisor, who accepted to supervise me without even knowing me, helped me at all times, supported me every time I needed, and always made clear how much he appreciated all of the work we did during these almost four years. Seiji, thank you very much for believing in me and treating me as a friend, you are amazing!

I cannot express how grateful I am to Armando Toda, who also believed in me to the point of helping me become a Ph.D. student with Seiji despite the short time we (virtually) knew each other. Also, I am very thankful to Armando and Paula Toledo Palomino for supporting me at all times and in all aspects since day one of this Ph.D. journey. You are some of the best friends someone can have, and I am so fortunate to have you as my friends!

I must also highlight how grateful I am to Ana Klock, Filipe Pereira, Anderson Avila-Santos, and Marcela Pessoa, who helped and were close to me during a big part of this journey. We met as students seeking for a degree, but became good friends because you are amazing people. Thank you very much for being on my side!

I have also been fortunate to have had the support of professors Isabela Gasparini and Elaine Harada for a couple of years. Since we met, you supported and believe in me, you helped with everything I needed and more, and you always treated me like a friend. You are great friends, I am very thankful for everything!

Finally, I am thankful that this study was financed in part by the Brazilian National Council for Scientific and Technological Development (CNPq) - 141859/2019-9 - and the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES) - Finance Code 001.

ABSTRACT

RODRIGUES, L. A. L. **Tailoring gamified educational systems to users and context**. 2023. 267 p. Tese (Doutorado em Ciências – Ciências de Computação e Matemática Computacional) – Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, São Carlos – SP, 2023.

Gamification refers to adding game elements to contexts other than games. As its success is assumed to depend on many moderators (e.g., user characteristics) and the gamification design itself, research has investigated how to tailor gamification to users to improve its effectiveness; that is, personalized gamification. However, personalization approaches i) are predominantly focused on user data, whereas contextual factors (e.g., task to be done and geographic location) also are important for gamification success; ii) often do not take into account multiple factors simultaneously, neither their interactions, which is necessary given that many of those are relevant moderators; iii) have not been properly validated, consequently, whether they improve one-size-fits-all gamification (i.e., no tailoring) remains unknown. That is, there is no empirically validated approach for personalizing gamification designs to the users' and context's characteristics within the educational domain. To address this problem, this research aims to develop and validate a personalization approach for tailoring gamification designs of educational systems to contextual and user characteristics. Three main steps were devised to achieve that goal: i) studying moderators of gamification's success to understand factors to be considered in the tailoring process; ii) creating an approach that simultaneously considers multiple contextual and user characteristics to tailor educational systems' gamification designs; and iii) empirically validating our approach compared to one-size-fits-all gamification. Based on 10 studies, this research contributes i) empirical evidence suggesting contextual and user characteristics that moderate gamification's success, ii) two recommender systems that guide on the multidimensional personalization of gamification, and iii) empirical evidence on how personalized gamification compares to the one-size-fits-all approach in terms of student motivation. Thus, this research informs the design of gamified educational systems, presenting empirical evidence on which information to consider, recommendations on how to personalize them to multiple characteristics simultaneously, and evidence on how such personalization contributes to gamification applied to education.

Keywords: Gamification, Personalization, Persuasion, Education, Learning.

RESUMO

RODRIGUES, L. A. L. **Personalizando sistemas educacionais gamificados para usuários e o contexto**. 2023. 267 p. Tese (Doutorado em Ciências – Ciências de Computação e Matemática Computacional) – Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, São Carlos – SP, 2023.

A gamificação diz respeito a adicionar elementos de jogos em contextos que não sejam de jogos. Como se considera que seu sucesso depende de vários moderadores (e.g., características do usuário) e do design de gamificação em si, pesquisas têm investigado como tornar a gamificação sob medida para os usuários para melhorar sua eficiência; isto é, gamificação personalizada. No entanto, abordagens de personalização i) concentram-se predominantemente nos dados do usuário, enquanto fatores contextuais (e.g., tarefa a ser executada e localização geográfica) também são importantes para o sucesso da gamificação; ii) muitas vezes não levam em consideração múltiplos fatores simultaneamente, nem suas interações, o que é necessário, dado que muitos deles são moderadores relevantes; iii) não foram devidamente validadas, consequentemente, se elas melhoram ou não a gamificação um-para-todos (*one-size-fits-all*) não é sabido. Ou seja, não existe uma abordagem para personalizar os designs de gamificação para as características dos usuários e do contexto no domínio educacional empiricamente validada. Para resolver este problema, essa pesquisa tem como objetivo desenvolver e validar uma abordagem de personalização para adequar designs de gamificação de sistemas educacionais às características contextuais e do usuário. Três etapas principais foram planejadas para atingir esse objetivo: i) estudar moderadores do sucesso da gamificação para entender os fatores a serem considerados no processo de personalização; ii) criar uma abordagem que considere simultaneamente múltiplas características contextuais e do usuário para personalizar os design de gamificação de sistemas educacionais; e iii) validar empiricamente a abordagem desenvolvida em comparação com a gamificação um-para-todos. Baseado em 10 estudos, essa pesquisa contribui i) evidências empíricas sugerindo características de usuário e contexto que moderam o sucesso da gamificação, ii) dois sistemas de recomendação que guiam a personalização multidimensional da gamificação, e iii) evidências empíricas sobre como a gamificação personalizada se compara à abordagem um-para-todos com base na motivação de estudantes. Portanto, essa pesquisa informa o design de sistemas educacionais gamificados, apresentando evidências empíricas sobre quais informações considerar, recomendações sobre como personalizá-los para várias características simultaneamente, e evidências sobre como tal personalização contribui para a gamificação aplicada à educação.

Palavras-chave: Gamificação, Personalização, Persuasão, Educação, Aprendizado..

LIST OF FIGURES

LIST OF TABLES

CONTENTS

CHAPTER

1

INTRODUCTION

Criticism regarding the traditional learning format suggests it is boring and unable to motivate students to engage with educational activities (LEE; HAMMER, 2011; LO; HEW, 2020; BOUWMEESTER *et al.*, 2019). That is problematic because research shows motivation is prominent for learning (HANUS; FOX, 2015; VANSTEENKISTE *et al.*, 2009). Hence, the traditional learning format's inability to motivate students to engage in educational activities needs to be addressed (PALOMINO *et al.*, 2019).

Aiming to address that need, this thesis explored the use of gamification applied to education, or gamified learning. Gamification is often defined as the use of game elements outside their original context (DETERDING *et al.*, 2011). It has been widely applied to educational contexts (KOIVISTO; HAMARI, 2019) aiming to improve student's learning outcomes (e.g., learning gains, motivation to learn, and tasks completed) (BORGES *et al.*, 2014; KLOCK *et al.*, 2018). Whereas its overall application has been successful in positively affecting those outcomes (SAILER; HOMNER, 2020), studies have also found uncertain and negative results, such as demotivation and disengagement (TODA; VALLE; ISOTANI, 2018; ANDRADE; MIZOGUCHI; ISOTANI, 2016).

The gamification design process and the resulting design have been discussed as an explanation for uncertain and negative results (TODA *et al.*, 2019b; LOUGHREY; BROIN, 2018; MORSCHHEUSER *et al.*, 2018). Gamification designs are often the same for all users (i.e., the one-size-fits-all approach) (ORJI; TONDELLO; NACKE, 2018). However, the literature states that people with, for instance, different demographic characteristics and cultural backgrounds have distinct preferences (YEE, 2016; ORJI; OYIBO; TONDELLO, 2017), behave, and are motivated differently (ORJI; VASSILEVA; MANDRYK, 2014; RODRIGUES; BRANCHER, 2018). Consequently, they will likely experience and respond to the same conditions in distinct ways (KNUTAS *et al.*, 2019; RODRIGUES; BRANCHER, 2019). Overall, this context suggests the shortcomings of developing one-size-fits-all gamified systems (LIU; SANTHANAM; WEBSTER, 2017).

Accordingly, researchers started to hypothesize that the lack of personalization (i.e., tailoring the gamification designs to specific aspects) was playing a role in gamified systems effectiveness (KNUTAS *et al.*, 2019; BÖCKLE; NOVAK; BICK, 2017; MEKLER *et al.*, 2017). When designing gamified systems, selecting which game elements to be added from a list aiming to replicate game patterns is common practice (KOIVISTO; HAMARI, 2019; TONDELLO; MORA; NACKE, 2017). Consequently, scholars invested in providing guidelines on which game elements suit better different users (e.g., (OLIVEIRA; BITTENCOURT, 2019; ORJI; VASSILEVA; MANDRYK, 2014)).

However, the literature on personalized gamification, at the start of this Ph.D. research, was uncertain on its effectiveness. A literature review (HALLIFAX *et al.*, 2019b) suggested that, overall, personalization results were positive, especially in the short run. On the other hand, the same review suggested findings were mixed for applications lasting more than two weeks. Importantly, many of the reviewed studies did not compare personalized to properly defined one-size-fits-all gamification designs. Instead, they mostly compared personalized gamification to random designs (ROOSTA; TAGHIYAREH; MOSHARRAF, 2016; LAVOUÉ *et al.*, 2018; MONTERRAT; LAVOUÉ; GEORGE, 2017)). Thus, showing personalized gamification needs further studies to ground its effectiveness as well as to find how to reduce mitigated outcomes.

Before deepening into how we sought to advance the personalized gamification field study, this section provides background information on two fundamental topics of this research: personalization and context.

1.1 Personalization and Gamification

Personalizing information systems is important to enhance these systems' relevance to users (LIU; SANTHANAM; WEBSTER, 2017). As discussed in Fan and Poole (2006), however, the exact definition of personalization is varied. The authors present a literature review demonstrating that different disciplines see personalization differently and that, even within a discipline, there are varied views. Then, they proceed to summarize previous research and define personalization as *a process that aims to increase the system's personal relevance to (a group of) users by changing aspects such as its functionality and interface* (FAN; POOLE, 2006). Within gamification research, that definition relates to what is called tailored gamification (KLOCK *et al.*, 2020). Therefore, the remainder of this section discusses tailored gamification as an umbrella term that encompasses personalization, as well as the factors and steps involved in this process, to facilitate understanding the scope of this research.

1.1.1 Tailored Gamification

Broadly speaking, tailoring gamification refers to designing it so that it equally suits the many users of a given gamified system in a given context (KLOCK *et al.*, 2020). Nevertheless, we can further understand it in terms of *who*, *when*, and *how*.

Tailoring gamification involves both personalization and customization of gamified systems, definitions which concern *who* is in charge of the tailoring. On the one hand, *customization* concerns situations wherein the final users are in charge (TONDELLO, 2019), an approach previously referred to as user-initiated personalization (ORJI; OYIBO; TONDELLO, 2017). On the other hand, *personalization* attributes that responsibility to designers or the system itself (TONDELLO, 2019), an approach that has been referred to as system-initiated personalization (ORJI; OYIBO; TONDELLO, 2017) and static adaptation (HALLIFAX *et al.*, 2019b).

Additionally, personalization can also refer to *when* the tailoring happens. For instance, personalization refers to when the gamification design changes (right) before usage. Such an approach is especially valuable to address the cold-start problem, wherein there is little to no information about the users, allowing the system to provide a tailored design since one's first interaction with the gamified system (TONDELLO; ORJI; NACKE, 2017). In contrast, *adaptive* gamification (or dynamic adaptation) refers to having a system change its gamification design as usage occurs (HALLIFAX *et al.*, 2019b; KLOCK; PIMENTA; GASPARINI, 2018). For instance, that can be used to adapt the gamification design when the user is returning to the system after a first-time experience or going from one page to another.

Furthermore, *how* to tailor can be accomplished through two alternatives. One of them is to change *which game elements are available*. For instance, research has widely explored this approach, seeking to understand which game elements suit a user better to turn elements on or off when a person uses a gamified system (Mora *et al.*, 2018; BOVERMANN; BASTIAENS, 2020). Differently, one might change *how game elements work*. For instance, that can be implemented by changing the criteria for completing a goal or leveling up (HALLIFAX *et al.*, 2019b).

In light of that overview, we highlight those are not exclusive properties. For instance, one can employ tailoring by changing both the game elements available as well as their functioning. Similarly, a system can personalize its gamification design at one's first usage and, then, adapt it as new usage occurs. Regardless of how the tailoring is characterized, that process often relies on a three-step process: i) acquiring user and contextual information, ii) passing it to a tailoring model, and iii) using the model to define the tailored gamification design aiming to improve users' experiences; often, student motivation in educational domains (KLOCK *et al.*, 2020; ZAINUDDIN *et al.*, 2020). Thus, we discuss what information to consider, how to create tailoring models, and tailoring's effectiveness next.

1.1.2 Personalization of Gamification

In the scope of this paper, we consider personalization as the system or designers tailoring game elements to users to improve the system's goals (TONDELLO, 2019). Accordingly, within the scope of gamified systems, a common practice to personalization has been to tailor the gamification design (set of game elements) to specific user's characteristics (e.g., (OLIVEIRA *et al.*, 2020; ORJI; TONDELLO; NACKE, 2018)). In other words, gamified systems have been personalized by performing static adaptations (i.e., before, not during system usage) on the game elements it features, based on pre-defined characteristics (i.e., gender), to tailor the gamification designs (HALLIFAX *et al.*, 2019b).

Research addressing personalization has mostly focused on exploiting the user's behavioral profiles. A recent literature review (HALLIFAX *et al.*, 2019b) has found that information used to drive personalization are, predominantly, user types (TONDELLO *et al.*, 2016; NACKE; BATEMAN; MANDRYK, 2014) and personality traits (MCCRAE; JOHN, 1992). The review also found most results are positive in the short-run, but mixed in long-term applications. Additionally, some of those studies' baseline comparison was not identical version differing only by featuring a properly defined one-size-fits-all gamification design instead of the personalized one (e.g., Lavoué *et al.* (2018), Monterrat, Lavoué and George (2017), Roosta, Taghiyareh and Mosharraf (2016)). With this practice, whether the latter approach is an improvement of the former remains unclear.

Furthermore, it has been shown that other user characteristics, such as gaming habits (Denden *et al.*, 2017) and gender (TODA *et al.*, 2019a), also impact their preferences. Similarly, the relationship between user demographics (i.e., age and gender) and user types (TONDELLO *et al.*, 2016) suggests the impact of those aspects. Nevertheless, these aspects have been rarely explored in guidelines for tailoring gamification designs in education, which relates to the fact that personalized gamification still is in its infancy, with many theoretical but few empirical studies (TONDELLO; ORJI; NACKE, 2017).

Moreover, the user is not the only factor to be considered when defining gamification designs. A factor that has been often discussed as relevant for gamification effectiveness, which is rarely considered by tailoring guidelines, is the context (KOIVISTO; HAMARI, 2019; MORA *et al.*, 2019; DETERDING, 2015). Specifically for educational systems, aspects researchers have recently argued as relevant and recommended to consider when tailoring gamified systems are geographic location and the learning activities (KLOCK *et al.*, 2020; RODRIGUES *et al.*, 2019; HALLIFAX *et al.*, 2019b). The latter additionally relates to the recommendation that gamified designs should match the task to be done (LIU; SANTHANAM; WEBSTER, 2017). Given that tasks in educational systems are mostly learning activities, personalizing the gamified designs based on these activities should be considered. However, few works towards personalizing gamified designs based on contextual factors are available, and none of those approaches have been empirically validated to our best knowledge (BOVERMANN; BASTIAENS, 2020; DICHEV; DICHEVA; IRWIN, 2019; BALDEóN; RODRÍGUEZ; PUIG, 2016).

1.1.3 Operationalization of the Context

In operationalizing context, this research follows Savard and Mizoguchi (2019). These authors define context from two perspectives: internal and external. The former fundamentally relates to users' mental representations that might impact their learning process. The latter also holds the potential to impact the learning process but interprets context as environmental and circumstantial. Furthermore, external context might be defined as circumstances that shape an event/object, being composed of four main elements: agent(s), which are participants, such as students or users; environment(s), such as educational systems or classrooms; event(s), which concern a participant and an action, such as a student (participant) learning (action) in a class (event); and a focus entity, the one that either holds the participants' role or the event. Figure 1 exemplifies the distinction between internal and external contexts, as well as the elements of the latter.

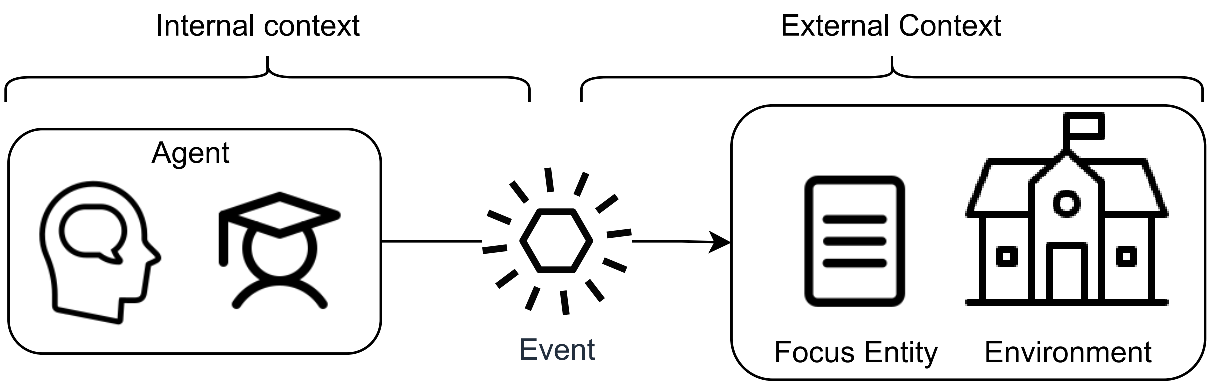

Figure 1 – Example illustrating the distinction between internal and external context along with the identification of elements of the latter in this case.

Source: Elaborated by the author.

Based on this context, we discuss how factors to be studied as moderators of gamification's success concern contextual characteristics. For instance, studies have called for research to include learning activities on the personalization process (e.g., (HALLIFAX *et al.*, 2019b)). When using a gamified educational system (GES), completing learning activities is related to the event and the action a participant performs. Similarly, the participant's geographic location relates to the environment in which the event and the action take part. Thereby, those aspects are related to the external context. On the other hand, we can also explore the internal context. In this case, we can explore learners' mental representation in terms of, for instance, their previous knowledge of the task's topic. Hence, we would be investigating how internal context, operationalized through users' previous knowledge/familiarity/affinity with the to-be-learned content, moderates gamification's effects.

1.1.4 Operationalization of Learning Tasks

To generally describe a task, one might rely on its desirable outcomes, behavioral requirements, and/or complexity (WOOD, 1986; LIU; SANTHANAM; WEBSTER, 2017). Similarly, from the human-computer interaction perspective, a task refers to the activities required to achieve a specific goal (DIAPER; STANTON, 2003). Consequently, given the context of this research, a learning task refers to a set of activities that aim at some educational outcome. From this definition, it is possible to note that numerous tasks might be found in GES, which makes developing a specific personalization approach for each one unfeasible. An alternative to that limitation is categorizing the activities, which can substantially reduce their quantity, enabling the recommendation of game elements to each category.

To overcome the numerous learning tasks and categorize them, we opted to rely on the revision of Bloom's taxonomy of educational objectives (KRATHWOHL, 2002), which contributes to the learning process by matching the educational activities' gamification designs to a cognitive taxonomy (BALDEóN; RODRÍGUEZ; PUIG, 2016). Although there are other options available, the revision of Bloom's taxonomy is a widely cited, well-accepted taxonomy, similar to its original version (BLOOM, 1956). It acts as a framework that can be used to classify what is expected from an educational activity (outcome), as well as its complexity. The revised version is composed of two dimensions: knowledge (concerned with *what*; e.g., the subject of matter) and cognitive process (concerned with *how* to learn) (KRATHWOHL, 2002).

In the scope of this research, we explore the second dimension, similar to Baldeón, Rodríguez and Puig (2016). By categorizing learning activities based on the cognitive domain of such a taxonomy, we avoid having the gamification focused on the activity itself (e.g., completing a quiz or answering a forum) and allow it to be aligned with the activity's expected learning outcome, addressing the recommendation that gamification should match the task (LIU; SANTHANAM; WEBSTER, 2017). Moreover, as many GES feature tasks of varied subjects, the second dimension choice makes the guideline subject-independent, focusing the tailoring process on the activities' objectives while allowing it to be used regardless of the educational topic.

The structure of the cognitive process dimension is split into six categories: remember (to retrieve knowledge from long-term memory), understand (to determine instructional messages' meaning), apply (to apply a procedure in some situation), analyze (to decompose material and how its parts relate, both to one another as well as generally), evaluate (to judge based on well-defined criteria), and create (to form something from its elements or scratch) (KRATHWOHL, 2002). Here, we consider each dimension a different Learning Activity Type (LAT), wherein their complexity increases following the order in which they were introduced. Hereafter, we refer to those as LAT1 to LAT6, also following the introduced order. Furthermore, although an activity might fit in more than one LAT, our approach considers every activity will have a predominant, main objective to be achieved. Hence, the personalization process should be based on that main goal. It is worth noting that those LAT might be further split. However, we opted to work with the high-level abstraction because the similarities within these sub-categories might be even higher.

1.2 Literature Gaps and Research Questions

Understanding which factors influence gamification's success is important to informing which aspects to consider when personalizing gamification. The gamification literature has suggested a number of such factors, such as users' characteristics (e.g., age, genre, traits, expertise, and skill) (MORA *et al.*, 2019; ROY; ZAMAN, 2018; DENDEN *et al.*, 2018; BUCKLEY; DOYLE, 2017; TONDELLO; MORA; NACKE, 2017; BORGES *et al.*, 2017), the context (MORSCHHEUSER *et al.*, 2018; NICHOLSON, 2012; MARACHE-FRANCISCO; BRANGIER, 2013), intervention duration (KOIVISTO; HAMARI, 2014), and the game elements used (HUANG *et al.*, 2020; BAI; HEW; HUANG, 2020). However, most of those aspects lack empirical tests (LANDERS *et al.*, 2019), which corroborates the claim that factors influencing gamification's success are not defined (SAILER; HOMNER, 2020). Thus, regarding which factors to consider when personalizing gamification, our first research question asks:

RQ1: What factors impact the success of gamified systems?

To personalize gamification, one needs guidance on how to tailor gamification to factors relevant to its success, but guidelines/recommendations for tailoring gamified systems rarely explore factors related to the context (e.g., the activities and geographic location). On the one hand, approaches for personalizing gamification designs are mainly based on users' behavioral profiles (HALLIFAX *et al.*, 2019b; TONDELLO *et al.*, 2016; NACKE; BATEMAN; MANDRYK, 2014). On the other hand, gamification studies have yet not given the attention context deserves (MORSCHHEUSER *et al.*, 2018), even though it has been advocated that it impacts gamification's effectiveness (KOIVISTO; HAMARI, 2019; HALLIFAX *et al.*, 2019a) and that the gamification designs must be aligned to the context and tasks, besides the users (LIU; SANTHANAM; WEBSTER, 2017).

In the scope of this study, we opted for the education domain to address the shortcoming of the traditional learning format mentioned before. Given the domain, users will perform learning activities when using the gamified systems, therefore we focused on learning tasks, which have been recommended for personalization by previous research (HALLIFAX *et al.*, 2019b; RODRIGUES *et al.*, 2019). Furthermore, tailoring approaches often consider a single aspect (e.g., (OLIVEIRA *et al.*, 2020; MONTERRAT; LAVOUÉ; GEORGE, 2017; ROOSTA; TAGHIYAREH; MOSHARRAF, 2016)). However, it has been suggested that multiple factors play a role when users are interacting with gamified systems, as previously discussed. Consequently, those should be considered simultaneously (KLOCK *et al.*, 2020). Thus, related to improving gamification personalization approaches, our second research question asks:

> **RQ2:** How to tailor gamified educational systems to the context and the user's characteristics?

Evidence on the effectiveness of personalized gamification, compared to one-size-fits-all gamification, is important to understand whether the approaches for the former are, as expected, improving the latter. Generally, studies performing such comparisons are still needed and there is no clear understanding established in the literature, as found in Hallifax *et al.* (2019b). Furthermore, these authors report that studies considered users' profiles, performance, and behaviors to tailor gamification, whereas none of those relied on aspects related to the context (e.g. the task). Therefore, due to the lack of personalization guidelines that consider both the user and the activities, their role remains unknown. Moreover, the few personalization approaches that consider learning activities have not been empirically validated in terms of how they compare to one-size-fits-all gamification (BOVERMANN; BASTIAENS, 2020; DICHEV; DICHEVA; IRWIN, 2019; BALDEóN; RODRÍGUEZ; PUIG, 2016). Thus, concerning the effectiveness of personalized gamification, our third research question asks:

> **RQ3:** Are gamification designs tailored to the context and the user's characteristics more effective than one-size-fits-all designs, in the context of educational systems?

1.3 Problem Statement

Summarizing the research gaps and questions discussed before, we define our problem according to the following: Many factors moderate gamification's effectiveness, such as multiple user characteristics and aspects related to the context; there is a lack of guidance on how to simultaneously tailor gamification to more than one factor; there is the need for empirical studies to understand the effects of personalized gamification. Thus, the problem we tackle is:

> *There is no empirically validated approach for personalizing gamification designs to the users' and context's characteristics within the educational domain.*

1.4 Objectives

The main objective of this research is *to develop and validate a personalization approach for tailoring gamification designs of educational systems to contextual and user characteristics*. To achieve this goal, the following specific objectives will be accomplished:

1. To identify factors that moderate gamification's effectiveness (Chapter 3);

2. To develop an approach for tailoring gamification designs of educational systems to contextual and user characteristics (Chapters 4 and Chapter 6);

3. To empirically validate the developed approach (Chapter 5);

4. To provide the developed approach as a resource to help those interested in tailoring the gamification designs of educational systems (Chapters 4 and Chapter 6).

1.5 Hypothesis

According to the literature, we might assume that by tailoring gamification designs to multiple factors that moderate gamification's success, its effectiveness will be improved. Thus, upon the gamification literature problem identified previously, our thesis is that:

> *Gamification designs tailored to the context and the user's characteristics are more effective in affecting learning outcomes than one-size-fits-all designs, in the context of educational systems.*

1.6 Method

To achieve our objective, this research followed an iterative method, with Figure 2 providing an overview of our proposal, illustrating that the game elements to be added depend on user characteristics, the context (i.e., the task they will perform and their geographic location), and the domain (i.e., education in our case), following the literature recommendations as previously discussed.

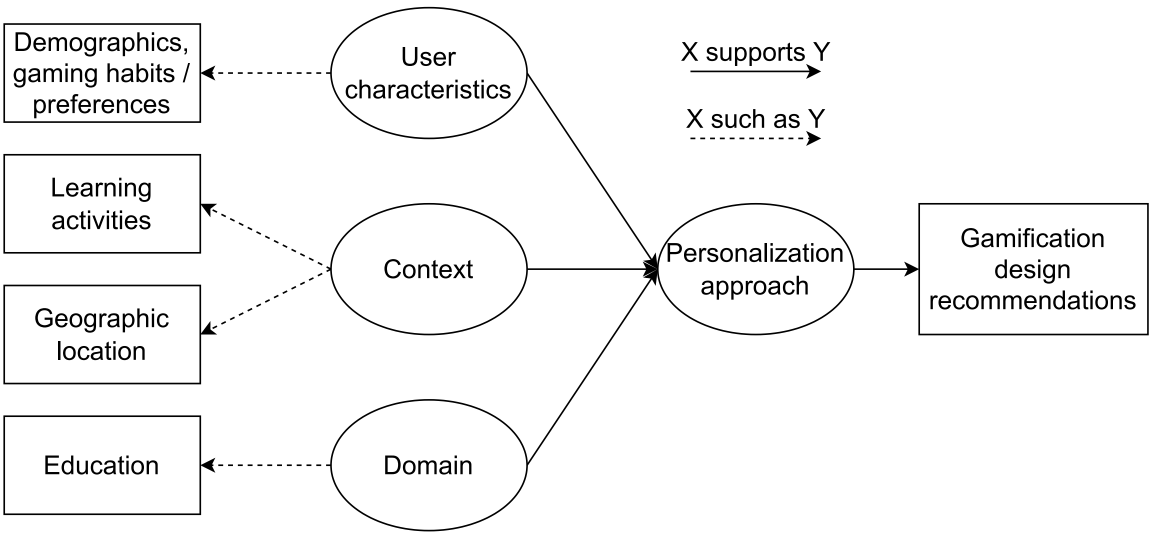

Figure 2 – Overview of our proposal for personalizing gamification.

Source: Elaborated by the author.

Developing and validating our proposal explores the gamification literature in implementing and defining its inputs, as well as planning its validation. The proposal concerns three main inputs (see Figure 2). First, we opted for the education domain, which is the one gamification research has focused the most (KOIVISTO; HAMARI, 2019) and both positive (SAILER; HOMNER, 2020) and negative (TODA; VALLE; ISOTANI, 2018) outcomes have been found, showing the need for further research. Second, there is the context, in which we follow recent recommendations by considering the geographic location as well as the task (KLOCK *et al.*, 2020; LIU; SANTHANAM; WEBSTER, 2017). According to the domain, the tasks users will perform when using the gamified systems are learning activities. As one might create numerous of those activities, we propose to consider LAT, based on the revised Bloom's taxonomy (KRATHWOHL, 2002), an established, well-accepted taxonomy within the educational domain. Third, we chose to focus on users' demographic characteristics and gaming habits and preferences, deepening into aspects that have been discussed as relevant factors (MORA *et al.*, 2019; ROY; ZAMAN, 2018; Denden *et al.*, 2017) but have received less attention from the academic community compared to the most used ones (see (HALLIFAX *et al.*, 2019a) and Chapter 2).

An overview of the four main steps that we followed to develop and validate our proposal, which is in line with our specific objectives, is shown in Figure 3. For each step, the figure shows what will be answered/done, the output that will be achieved, and why we need that step. A brief description of each of those is as follows:

1. The *first step* relates to the first specific objective: Identifying factors that moderate gamification's effectiveness. Accomplishing this objective will provide evidence on which factors influence gamification's success, which is important to understand success' predeterminants once these factors still demand research (KOIVISTO; HAMARI, 2019; SAILER; HOMNER, 2020). Also, this step will guide what to consider (e.g., which user demographic) as input in step two.

2. The *second step* relates to the second specific objective: Developing a guideline for tailoring gamification designs of educational systems to contextual and user characteristics. Achieving this goal will lead to recommendations on which game elements to select depending on the characteristics of the user and their geographic location, as well as the task they will perform, addressing a gap in gamification literature (see Chapter 2).

3. The *third step* relates to the third specific objective: Empirically validating the developed approach. This step's result will provide evidence of the effect of our proposal, advancing the scant literature on the impact of personalized gamification (see Chapter 2) and showing whether our approach improves one-size-fits-all gamification.

4. The *fourth step* relates to the fourth specific objective: Refining the approach based on the empirical validation's findings. This approach will aid in grounding our findings as well as improving the developed approach.

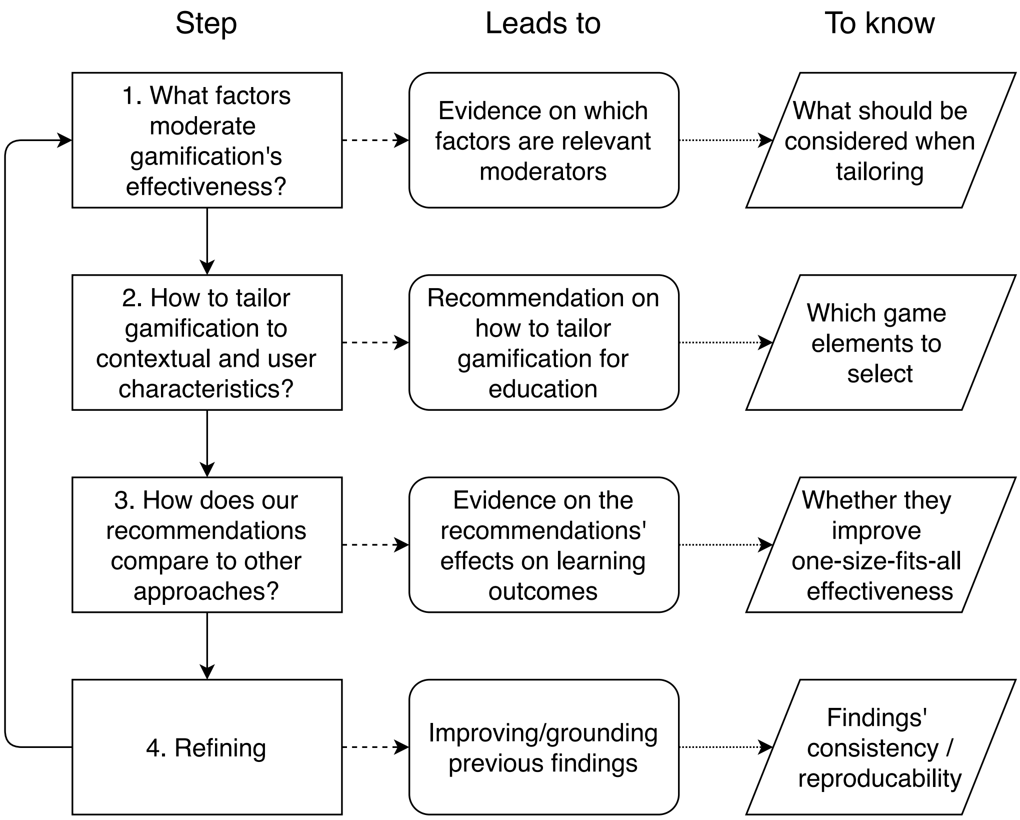

Figure 3 – Research development process.

Source: Elaborated by the author.

1.7 Outline

In following the method described before, we conducted 10 empirical studies. First, we revised the personalized gamification literature (Chapter 2). Then, to address RQ1 and step one, we conducted four (quasi-)experimental studies to understand the moderators of gamification's success (Chapter 3). Next, to address RQ2 and step two, we first conducted two studies to develop a personalization approach based on user preference (Chapter 4). Subsequently, we sought to validate that approach (RQ3; step three) with two experimental studies comparing it to one-size-fits-all gamification (Chapter 5). Lastly, we addressed step four by creating a data-driven personalization approach (Chapter 6) to refine the one developed in step two. This document compiles all of these studies, which are the main publications (articles and papers) of this Ph.D.

CHAPTER

2

PERSONALIZED GAMIFICATION: A LITERATURE REVIEW OF OUTCOMES, EXPERIMENTS, AND APPROACHES

Personalized gamification has attracted the attention of several researchers, which led to a number of secondary studies aiming to summarize this field study (e.g., (HALLIFAX *et al.*, 2019b; KLOCK *et al.*, 2020)). However, prior secondary studies i) did not present a broad analysis of how primary studies assessed personalized gamification's effects, ii) nor analyzed the development as well as the selection of which game elements should be considered by personalization approaches. Thus, we conducted a literature review on personalized gamification (RODRIGUES *et al.*, 2020) to investigate **how empirical studies employing personalized gamification have been conducted to answer whether personalization improves one-size-fits-all gamification effectiveness**. Specifically, we sought to answer the following:

- **SRQ2.1**: How empirical studies employing personalized gamification have been conducted?
- **SRQ2.2**: Does personalized gamification improve one size fits all gamification effectiveness?
- **SRQ2.3**: How approaches for personalizing gamification have been developed?

Considering the number of related works, as well as their publishing year, at the time of planning our review, we did not conduct a standard, systematic literature review (e.g., (KITCHENHAM *et al.*, 2009)). Following recommendations on when to update those studies (GARNER *et al.*, 2016), we conducted an exploratory study by analyzing primary studies included in related work (KLOCK *et al.*, 2020; KLOCK; PIMENTA; GASPARINI, 2018; HALLIFAX *et al.*, 2019b; ALJABALI; AHMAD, 2018). In that process, we screened 95 primary studies and selected

39 for inclusion because they either i) presented an empirical study employing personalized gamification or ii) presented an approach for personalizing gamification; that is, recommendations/guidance on how to tailor gamification designs to some specific criteria. For details on the study's protocol, please see Annex A.

For **SRQ2.1**, we found that most studies compare personalized gamification to counter-tailored and random gamification, use few game elements in both the experimental and baseline conditions, analyze quantitative data collected in classrooms and in the wild, and rely on sample sizes and time frames comparable to general gamification studies. For **SRQ2.2**, we found that the existing literature provides inconclusive evidence on how personalized and one-size-fits-all gamification compare as few studies did so. For **SRQ2.3**, we found that approaches were mainly developed based on data (e.g., from surveys) rather than theory, using literature reviews and deliberate choices for selecting the game elements to consider, predominantly focusing on user characteristics and with few approaches focused on specific domains.

Overall, the implications of our findings are twofold. First, they provide a general overview of different approaches for personalizing gamification. Hence, one can use this as a starting point toward understanding how gamification has been personalized. Second, we provide guidelines on how future empirical research in this field might be conducted by showing previous studies' settings. For instance, one can follow those guidelines to plan future research. Furthermore, we presented a research agenda with five gaps that emerged from our findings: i) comparing personalized and one-size-fits-all gamification, ii) developing personalization approaches that consider contextual characteristics, iii) performing qualitative user studies; iv) building personalization approaches from validated taxonomies of game elements; v) providing and establishing resources to be reused in future empirical studies. That agenda not only guided our subsequent steps; it informs others on valuable research lines needing attention.

CHAPTER

3

MODERATORS OF GAMIFIED LEARNING'S EFFECT

As discussed in Chapter 2, most research personalizes gamification to user characteristics. However, research shows that contextual factors also play a significant role in gamification's effectiveness (KOIVISTO; HAMARI, 2019) and that the complete understanding of which factors moderate such effectiveness remains unknown (SAILER; HOMNER, 2020). Therefore, this chapter presents four empirical studies (RODRIGUES *et al.*, 2020; RODRIGUES *et al.*, 2021; RODRIGUES *et al.*, 2022a; RODRIGUES *et al.*, 2022b) we conducted to understand which user and contextual factors moderate (i.e., increase/decrease) gamification's effects. Thus, addressing this research's first specific objective (see Figure 3, step 1).

3.1 Exploring the influences of competition and task-related factors in gamified learning environments

In this initial study (RODRIGUES *et al.*, 2020), we worked with a single game element, Competition. This choice aimed to allow a clear understanding of the game element that would potentially affect participants, as opposed to most studies that apply multiple elements simultaneously (BAI; HEW; HUANG, 2020). As we were interested in addressing students' motivation issues, we sought to answer (**SRQ3.1**) *how does Competition affect learners' motivation?* Furthermore, given that contextual data are assumed to moderate gamification's effect (KOIVISTO; HAMARI, 2019), we also sought to answer (**SRQ3.2**) *how do task-related factors moderate Competition's impact on learners' motivation?*

To answer those SRQ3s, we performed a within-subject (gamified vs non-gamified) quasi-experimental study. As we aimed to answer them within a real learning setting (i.e., a real class), we recruited students (15 Brazilian males with an average age of 31 ($\pm$8.43) years) from a graduate course where we were allowed to perform the study by convenience sampling.

In the experimental task, considering Artificial Intelligence was the class topic, participants worked with a console-based fight simulator, which was developed specifically to be used as a learning resource in the course this quasi-experiment was performed. In summary, participants i) created a reflexive intelligent agent (RUSSELL, 2010) for the fight simulator (non-gamified), ii) completed the Situational Motivation Scale - SIMS (GUAY; VALLERAND; BLANCHARD, 2000), iii) worked to improve their agent with the goal of beating peers' agents (i.e., gamifying by deploying a *player-to-player* unplugged competition (TODA *et al.*, 2019c)), and iv) completed SIMS again.

Then, we analyzed paired measures' difference, calculated by subtracting the pre- from the post-measure, based on descriptive measures (mean - M, standard deviation - SD, and effect size - ES). Note that ES is sample size-independent, unlike inferential tests, which would be underpowered due to our sample size (WOHLIN *et al.*, 2012). Additionally, we evaluated task-dependent factors (i.e., self-reported familiarity with AI, programming, and Python) as possible moderators, considering that analyzing users' previous knowledge contributes to understanding the effect of context-related moderators as the activity at hand is fundamental to the context (SAVARD; MIZOGUCHI, 2019). See Annex B for details of the study design.

Concerning **SRQ3.1**, we found Competition has an overall small positive effect on learners' motivation, decreasing their amotivation whilst increasing intrinsic, identified, and external motivations. Concerning **SRQ3.2**, we found the task-dependent factors we assessed moderated Competition's effect on learners' motivation in varied ways: AI familiarity appears to have no overall impact; more familiarity with general programming appears to slightly improve motivation, except that it decreased external motivation for those with high familiarity; and the familiarity with Python programming language showed a U function impact on autonomous motivation and, for external and amotivation, a positive effect up to moderate familiarity with a decrease thereafter.

These findings inform both researchers and practitioners. They contribute to the knowledge of how a specific game element impacts different types of motivation, which is the main psychological aspect scholars argue gamification seeks to affect (LANDERS *et al.*, 2018). Hence, these results can be used to inform the design of gamified interventions by suggesting that even the isolated use of competition can lead to motivation gains, extending the external validity of previous research (LANDERS; COLLMUS; WILLIAMS, 2019). Furthermore, the insights can be used to inform research on personalized gamification in terms of how Competition worked for our sample and the way context-related factors moderate its effect.

3.2 Gamification Works, but How and to Whom?

In this study (RODRIGUES *et al.*, 2021), we expanded our prior research by enriching the gamification design and increasing the intervention duration. Additionally, we studied students learning introductory programming instead of AI. Accordingly, we sought to answer *how does using gamification for half semester affect Brazilian undergraduates' programming learning?* In doing so, based on a literature review (HANUS; FOX, 2015; LANDERS *et al.*, 2018; SANCHEZ; LANGER; KAUR, 2020; KOIVISTO; HAMARI, 2014) and our prior research, we tested the following hypotheses: (**H1**) Intrinsic motivation and completing quizzes positively affect learning gains; (**H2**) Gamification improves intrinsic motivation, an effect moderated by intervention duration and users' familiarity with class' topics; and (**H3**) The more the learners' intrinsic motivation, the more quizzes they complete.

To test those hypotheses, we conducted an experimental study. Its context is a Brazilian undergraduate *Algorithms* class focused on programming lessons. Students (19 Brazilian males with an average age of 20.32 ($\pm$3.64) years) were randomly assigned to complete quizzes, which were delivered weekly, in one of two Moodle versions (gamified vs non-gamified) for half semester. For the gamification design, we implemented gamification heuristics focused on SDT (ROY; ZAMAN, 2017), aiming to affect intrinsic motivation, which led to adding unannounced badges and presenting quizzes as graphic-enriched missions, similar to (ROY; ZAMAN, 2018). We analyzed how gamification affected programming learning based on cognitive (learning gains based on pre-post-testing), motivational (intrinsic motivation measured with the *interest/enjoyment* sub-scale of the Intrinsic Motivation Inventory version validated by Pedro (2016)), and behavioral (number of tasks completed, similar to Sanchez, Langer and Kaur (2020)) learning outcomes; the last two were weekly measured.

Notice that we understand that context involves someone, with their mental state, performing an activity in a given environment for a given period (SAVARD; MIZOGUCHI, 2019). An example would be a learner, with their current knowledge, completing quizzes in a gamified system. Accordingly, we analyzed context's moderating role based on intervention duration and students' self-reports of their familiarity with general course topics (i.e., algorithms, C programming language, programming, and pseudo-language). For data analysis, we used multiple linear regression to test **H1** and multilevel regressions to test **H2** and **H3** as they involve repeated-measures data. For details on the study method, please see Anexx C.

In terms of how gamification affected Brazilian undergraduates' programming learning in a six-week period, we found that was accomplished by influencing their motivation. Specifically, gamification influenced learners' motivation (**H2**) that, in turn, positively predicted learning gains (**H1**). We also predicted gamification would indirectly affect learning via users' behaviors, which would be affected by intrinsic motivation (**H3**). However, we found no support for that. Furthermore, **H2** predicted intervention duration and context-related factors would moderate gamification's effect on intrinsic motivation. Intervention duration and only one (i.e., familiarity with algorithms) out of the four context-related factors were significant moderators: over time, gamification was positive when participants had at least some familiarity with algorithms but negative otherwise. Hence, the way gamification affected participants learning was through intrinsic motivation, positively or negatively affecting it over time depending on learners' previous familiarity with algorithms.

From these findings, there are two main implications. First, supporting **H1** implies instructors should seek to improve this psychological state of learners, as well as having them self-tested (e.g., completing quizzes), as this likely enhances their learning. Second, partially supporting **H2** implies research on tailored gamification should investigate contextual factors, such as intervention duration and familiarity with the topic under study, as tailoring gamification to such familiarity as time passes might be crucial to improve its effectiveness.

3.3 Moderator conditions of gamification's success in programming classrooms

In this study (RODRIGUES *et al.*, 2022a), we advanced our prior research by analyzing another gamification design based on 14 instead of six weeks. Additionally, we analyzed the moderator role of both user and contextual factors, but considering them simultaneously instead of independently as in the previous studies. On the other hand, we similarly worked within the context of introductory programming lessons. Accordingly, we sought to answer *how do user and contextual factors influence the effect of gamification on the academic achievement of CS1 students?* To answer that question, based on the relevant literature (SANCHEZ; LANGER; KAUR, 2020; ROWLAND, 2014; KOIVISTO; HAMARI, 2019), we posed and tested three hypotheses: (**H1**) Practicing positively affects academic achievement in CS1; (**H2**) Gamification enhances the testing effect and, consequently, academic achievement; and (**H3**) Gamification's effects will be moderated by a) user and b) contextual characteristics.

To test those hypotheses, we conducted a retrospective, quasi-experimental study. Its context is introductory programming classes from seven STEM[1] majors of the Federal University of Amazonas (UFAM). Participants (399 Brazilian students; 64.2% males, 35.8% females; average age: 22.1 ($\pm$3.6) years) completed programming assignments in one of two versions (gamified vs non-gamified) of Codebench[2] during the whole semester. The gamification design was planned according to the proposal of Wangenheim and Wangenheim (2012), which combined aspects of the ADDIE (Analyze, Design, Develop, Implement, and Evaluate) approach Branch (2009) with the Instructional Systematic Design model Dick, Carey and Carey (2005), resulting in a mix of immersive-, social-, and challenge-based gamification (TODA *et al.*, 2019c). As measures, we analyzed *academic achievement* (i.e., a student's final grade) as the dependent variable and *practicing* to code (i.e., the extent to which learners practiced programming) as a proxy to the former.

Table 1 presents each moderator, along with a brief description that indicates its possible

[1] Science, Technology, Engineering, and Mathematics
[2] <https://codebench.icomp.ufam.edu.br/>

values, an alias that will be used hereafter, and the category it fits in. We consider age and gender as user characteristics because they are basic user information. Because the circumstance and environment wherein participants completed the activities differ depending on their major, we understand it as part of the external context. The others are classified as internal context, as they are characteristics that influence users' learning experiences, such as previous experience, currently working/interning, or having internet access. Note that despite covariates being selected by convenience, studying these addresses literature limitations as what are the possible moderators of gamification's effects have not been defined by gamification frameworks or empirical evidence yet (LANDERS, 2014; LANDERS *et al.*, 2018; SAILER; HOMNER, 2020; KOIVISTO; HAMARI, 2019). For data analysis, we used multilevel regressions to account for the hierarchical structure of our data (i.e., students nested within majors). See Annex D for details on the study method.

Table 1 – Possible moderators of gamification's success analyzed.

Description	Alias	Category
Student's age (numeric)	Age	User characteristic
Whether one is male (1) or female (0)	Male*	User characteristic
Whether one has a PC at home (1) or not (0)	PC	Internal context
Whether one share a PC at home (1) or not (0)	SharesPC	Internal context
Whether one has internet at home (1) or not (0)	Internet	Internal context
Whether one has previous experience with any programming language (1) or not (0)	Exp	Internal context
Whether one has worked/interned (1) or not (0)	Worked	Internal context
Which major one is enrolled at	Major	External context

* Because this information was dichotomous (male or female), we dummy coded it as 1 for Male and 0 for Female to facilitate the interpretation of the regression coefficient.

From our findings, the answer to *how user and contextual factors influence gamification's effect on the academic achievement of CS1 students* is that gender significantly moderated that influence - positively for females and nonsignificant for males - whereas user age and contextual factors (i.e., having a PC at home, sharing a PC at home, having internet, having previous experience with some programming language, having already worked/interned, and major enrolled at) had nonsignificant roles on that effect. Furthermore, we found that for those who practiced more than two standard deviations above the average (i.e., 1094 attempts), gamification's effect shifted from positive to negative. Our results also revealed additional insights concerning moderators of academic achievement as well as the testing effect, which are detailed in Annex D.

These findings hold three main implications we discuss next. First, we have shown that students who practice more are likely to yield higher academic achievement at the end of the semester. Instructors can explore this finding by providing their students with opportunities to practice programming as much as possible, which can be accomplished by making several programming assignments available during the course. Second, we found gamification's impact was different depending on whether users were males or females. This further supports the claim that one size does not fit all (NACKE; DETERDING, 2017), highlighting the need for developing and providing gamification designs tailored to specific audiences. Third, we found no support for the moderator effect of various user and contextual characteristics, which is contrary to the overall discussion from previous studies (e.g., (HALLIFAX *et al.*, 2019a; TODA *et al.*, 2019a)) and our initial expectations. Hence, pointing out that more research is needed to understand what moderates gamification's success.

3.4 Gamification suffers from the novelty effect but benefits from the familiarization effect

This study (RODRIGUES *et al.*, 2022b) is similar to our previous research, mainly differing in the data analysis perspective. Both were conducted in the same context, but here we focused on understanding *how does gamification's effects on CS1, STEM students' behavioral outcomes change over time?*

To answer that question, we conducted another quasi-experimental study involving students from STEM majors of the UFAM. Similarly, participants (756 Brazilians; 62.3% males, 37.7% females; average age: 22.2 ($\pm$4.2) years) completed programming assignments in one of two versions (gamified vs non-gamified) of CodeBench throughout the whole semester. The gamified version featured the design described before (see Section 3.3). To analyze students' behavioral learning outcomes, the measures were *attempts* (the number of attempts that a student made in the assignments), *IDE usage* (the time - in minutes - that a student used the IDE integrated into CodeBench), and *system access* (how many times a student accessed CodeBench). For data analysis, we first used Robust ANOVAs, based on trimmed means (WILCOX, 2017), to test whether there is an interaction between the two factors (gamification: yes or no; time point: 1 to 7), which would indicate the effect of gamification changes over time. Then, in cases of significant interactions, we performed posthoc tests using the Yuen-Welch test (YUEN, 1974) to how that effect changed.

How the effect of gamification changes over time, based on the effect sizes estimated as previously described, is shown in Figure 4. In this figure, time points are shown on the x-axis, the magnitude of the effect size is shown on the y-axis, each measure is represented by a different line type/color, and error bars are based on the confidence intervals estimated in the analyses described previously. Overall, the results suggest that, over time, the effect of gamification on behavioral learning outcomes changes, following, to some extent, a U-shaped curve. Note that we intentionally chose to first analyze variations in gamification effect based on effect sizes estimated during posthoc testing. Then, we analyzed such results by inspecting plots featuring confidence intervals, given that the literature discourages using such a quantitative approach in the context of exploratory analyses (VORNHAGEN; TYACK; MEKLER, 2020; CAIRNS, 2019; DRAGICEVIC, 2016). Thus, summarizing our results, we have i) empirical evidence that the gamification effect does change over time and ii) a model (i.e., the somewhat U-shaped function) of how such variation occurs during the course of a 14-week period, which must be further explored and validated in future research.

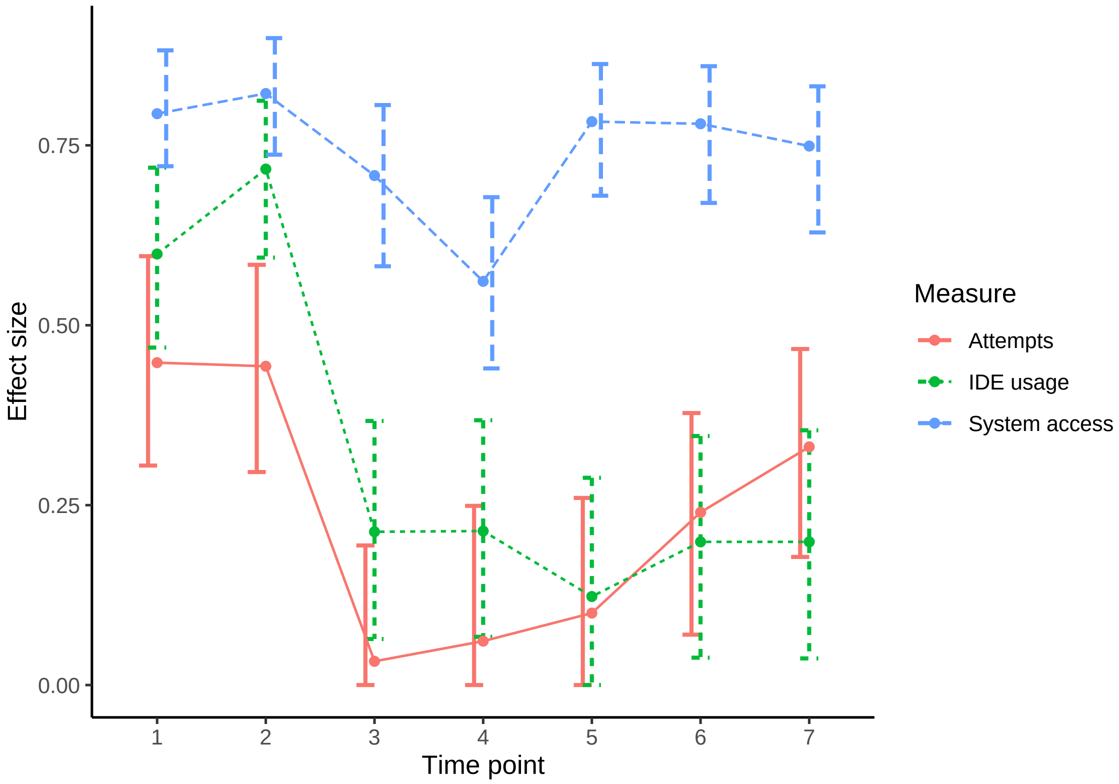

Figure 4 – Gamification's effect (compared to no gamification) over time, based on effect sizes estimated from differences in behavioral outcomes. Error bars represent the 95% confidence intervals estimated from effect sizes.

These findings have several implications. For instance, they inform practitioners on how long it takes for the novelty effect to start acting; revealing that although gamification's impact might decrease over time, it will likely recover after a familiarization period. Furthermore, our findings have raised concerns that demand further research, such as confirming for how long the novelty effect decreases gamification's impact and for how long (and to what extent) the familiarization effect recovers that effectiveness loss. Hence, our main contribution is advancing the understanding of how gamification impacts students' behavior over time by providing empirical evidence on how the novelty effect acts, as well as supporting the familiarization effect. Thus, we provide interesting time-related recommendations on when to tailor gamification to prevent its effect start to decrease.

3.5 Summary

This chapter addressed our first specific objective: understanding moderators of gamification's effects. In Section 3.1, we found task-related factors moderate such an effect, but limited to a one-time usage of a single game element (RODRIGUES *et al.*, 2020). In Section 3.2, we found similar results based on a six-week intervention that deployed an enriched gamification design, which also revealed time's role on gamification's effect (RODRIGUES *et al.*, 2022). Section 3.3 analyzed multiple possible moderators simultaneously, after a 14-week usage period, yielding results that question our previous findings (RODRIGUES *et al.*, 2022a). In contrast, findings from Section 3.4 extended those from Section 3.2 based on a 14-week study of another gamification design.

Overall, those findings corroborate previous literature from two perspectives. On the one hand, they provide evidence that user and contextual factors might moderate gamification's effects. On the other hand, they demonstrate the challenge of understanding which (and how those) factors affect gamification's success. As the main takeaway, these studies suggested that multiple user and contextual factors, as well as their interaction, play a significant role in gamification's success. Thus, supporting the view that, if one wants to personalize gamification, they should do so by considering user and contextual factors simultaneously.

CHAPTER

4

PREFERENCE-BASED PERSONALIZATION OF GAMIFIED EDUCATIONAL SYSTEMS

In reviewing the literature, we found there are several strategies for personalizing gamification (see Chapter 2). Two main limitations of those strategies, however, are i) only considering user information, while contextual factors also play a key role in gamification's effectiveness (see Chapter 3), and ii) not taking into account the several moderators of gamification's success simultaneously (RODRIGUES *et al.*, 2020). Moreover, most of those strategies are conceptual (KLOCK *et al.*, 2020), implying that they cannot be readily deployed into a gamified educational system. Based on that context, this chapter presents two studies we conducted to address theses limitations, along with the tool resulting from them (RODRIGUES *et al.*, 2019; RODRIGUES *et al.*, 2022). Thus, addressing this research's second and fourth specific objective (see Figure 3, step 2).

4.1 How to Tailor Gamified Educational Systems Based on Learning Activities Types

In this initial study (RODRIGUES *et al.*, 2019), we proposed the idea of thinking about what users would do in a gamified system (inside the box) to address the limitations of personalization strategies that focused outside the box (only thinking about the user). The main motivation was that existing strategies were often not as effective as expected (RODRIGUES *et al.*, 2020). To operationalize that idea within the educational context, we proposed that one should consider the learning activity type as the input for defining personalized gamification designs. Our rationale was that different game elements have different functions and effects (HUANG *et al.*, 2020), as well as different types of learning activities (TOETENEL; RIENTIES, 2016). Hence, we hypothesized that selecting the most appropriate game elements for each of these activity types, based on their specific aims, holds the potential to maximize gamification's success. Thus, the study tackled the following RQ: *how to tailor gamified educational systems designs based on learning activity type?*

To start answering that question, we conducted a conceptual study split into two parts. First, it introduced and compared our proposal to similar personalization strategies. Fundamentally, our approach considered users provided information from outside, whereas knowing the learning activity type they would do was information from inside. Based on that, we proposed that personalization should be performed from inside the box given that each activity type has different aims. Consequently, they will likely benefit from a specific gamification design because, similarly, different game elements often lead to different outcomes.

Second, the study discussed three case studies exemplifying our approach. In one of those, we considered the situation wherein the learning activity concerns attending to information (TOETENEL; RIENTIES, 2016). Specifically, we mentioned having to watch a video lesson in an educational system, which aimed at providing learners with explanations regarding a specific subject (e.g. balanced search trees) (CHANG, 2016). In this case, if the system featured the user-based approach, the following would be necessary. First, users have to provide some demographic data (e.g., genre and age) or complete a questionnaire that would, for instance, allow the identification of their gamer type to the system (e.g., Daredevil). Next, the personalization strategy would update the system according to each user and, then, present the video player with the personalized gamification design. For instance, a user classified as Daredevil would be provided with Levels, Acknowledgment, and Competition (OLIVEIRA; BITTENCOURT, 2019). Similarly, when users of other gamer types accessed the system, other gamification elements would be available accordingly.

If that system featured the "inside-out" personalization approach, users would not be required to complete a questionnaire or provide any demographic information because this strategy relies on data from inside the box (system), thus preventing the dependence and need for users' data. Hence, when the user accesses the video player screen, the system recognizes the type of the activity (i.e., attending to information) and the personalization strategy would select the gamification elements to be available on the screen accordingly. Following that approach, in the user-based method, different users will be provided with varied gamification designs based on their personal characteristics, regardless of the learning activity they are performing. In contrast, in the "inside-out" approach, different users will be provided with gamification designs selected based on the LATs, regardless of their personal characteristics. For further details, please refer to F.

The main implication of this research was proposing and advocating for the importance of considering LATs when personalizing gamified educational systems, instead of being limited to user information. Thereby, incorporating contextual information into the personalization strategy. However, by itself, this approach was limited due to i) the lack of empirical evidence on which game elements are appropriate for which LAT and ii) the propensity to yield a personalized design appropriate for the LAT but inadequate for a specific user. In light of these limitations, we conducted the study presented next.

4.2 Automating Gamification Personalization to the User and Beyond

This study (RODRIGUES *et al.*, 2022) is based on the fact that gamification is mainly personalized to users' profiles (HALLIFAX *et al.*, 2019b; TONDELLO *et al.*, 2016), while the application context is also relevant for gamification's success (KOIVISTO; HAMARI, 2019; HALLIFAX *et al.*, 2019a) and gamification designs should be aligned to it (LIU; SANTHANAM; WEBSTER, 2017). Furthermore, multiple factors moderate users' experience, either positively or negatively. Nevertheless, tailoring approaches often consider a single one, highlighting the following needs in the field of personalized gamification (RODRIGUES *et al.*, 2020; KLOCK *et al.*, 2020; HALLIFAX *et al.*, 2019b): i) personalization models should consider more than users' characteristics, such as encompassing the learning activities and geographic locations, and ii) personalization methods should consider multiple aspects simultaneously, as well as their interactions.

Consequently, we aimed to address those needs by answering the following: (**SRQ4.1**) *Does users' preferences differ depending on (a) their characteristics, (b) geographic location, and (c) the type of the learning activity to be performed?*; and (**SRQ4.2**) *What is the most useful game elements set, from users' preferences, according to their characteristics, geographic location, and Learning Activity Type (LAT)?* Note, however, that the interactions from multiple characteristics might lead to several combinations. For instance, five binary characteristics would lead to 25 combinations (i.e., 25 recommendations); the number of recommendations for five three-valued characteristics would exponentially increase. Thereby, providing a way to automate such recommendations becomes imperative, which corroborates another challenge of personalized gamification: automating the personalization process (KLOCK *et al.*, 2020). Thus, we also aimed to answer (**SRQ4.3**) *How to automate gamification personalization?*

To answer those RQs, we performed a survey-based research asking participants to indicate their preferred game elements for each LAT. Up to date, this methodology is the most used by similar works (KLOCK *et al.*, 2020; RODRIGUES *et al.*, 2020) and has been widely accepted given the number of related research following it (TONDELLO; MORA; NACKE, 2017; BOVERMANN; BASTIAENS, 2020). Therefore, we considered it the most adequate approach to adopt. Specifically, we achieved our goals through the following five-step procedure (for further information regarding the procedure, please refer to G):

1. **Survey development**: defining the survey design and sections and the game elements and LAT to consider;

2. **Data collection**: disclosing the survey online, through Amazon's Mechanical Turk[1] (MTurk), to collect participants' opinions.

3. **Data analysis**: running analyses to identify which characteristics impact users' preferences.

4. **Users preferences analysis**: investigating our findings to identify how to tailor educational systems' gamification designs to users, geographic location, and LAT.

5. **RS design**: developing a free, ready-to-use resource, based on our findings, to aid those who want to tailor their educational systems' gamified designs.

By conducting this procedure, our findings provided evidence that users' preferences differ depending on their characteristics, geographic location, and the LAT to be performed (**SRQ4.1**). Also, we were able to develop a Recommender System (RS) that recommends a gamification design given a LAT to be performed by a set of user characteristics and a defined geographic location (**SRQ4.2**). The main contribution of this research is, therefore, providing a free RS for personalized gamification, built upon a state-of-the-art approach, that aids in automating the tailoring of gamification designs by suggesting which game elements to use (**SRQ4.3**). This RS is based on three aspects of personalization: domain, user, and task (LIU; SANTHANAM; WEBSTER, 2017), implemented as the educational domain, demographics and gaming characteristics, and LAT and geographic location, respectively. Additionally, we shed light on which context and user characteristics impact their preferences, and which of those are more or less relevant, contributing to expanding and grounding knowledge from previous studies (e.g., (TONDELLO; MORA; NACKE, 2017; Denden *et al.*, 2017)).

Overall, those findings imply that: i) demographics and gaming-related characteristics are moderators of user preference that should be prioritized differently, ii) rather than just thinking on what users generally prefer, aspects of the task that will be performed and the user's geographic location should be taken into account, iii) the interaction between relevant characteristics cannot be ignored, iv) when designing GES, two people might prefer the same game elements, but with different priorities, and v) one can use our RS to automate gamified systems' personalization process as well as be informed on how to tailor gamification designs of educational systems. Thus, our contributions are twofold. First, we provided practitioners with a ready-to-use resource able to guide them on how to design GES that are tailored to users' characteristics, as well as geographic location, according to the tasks they will perform. Second, we expanded the literature on how to tailor gamification designs to any learning activity (based on its type) by presenting recommendations that might be empirically tested in future research, providing empirical evidence on which demographics and game-related user characteristics impact their preferences, as well as which one is more important than the others, and supporting literature suggestions by showing that LAT and geographic location do affect user preference.

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

4.3 Summary

This chapter addressed our second specific objective: to develop an approach for tailoring gamification designs of educational systems to contextual and user characteristics. In Section 4.1, we started to address this objective by proposing a personalization approach that would take into account the task users of a gamified educational system would do (RODRIGUES *et al.*, 2019). Next, the research presented in Section 4.2 advanced that approach by acknowledging both user and contextual information should be considered simultaneously, modeling user preference as such, and providing a concrete personalization strategy implemented as a free-to-use RS (RODRIGUES *et al.*, 2022).

Mainly, these results advanced the literature by i) providing empirical evidence on the importance of considering contextual information as well as interactions between relevant factors and ii) offering technological support for those interested in deploying personalized gamification to their educational activities. This can be achieved by either consulting the RS to receive recommendations on which game elements to use or using it as a service to automate the personalization of the gamification design of an educational system. Nevertheless, these findings have one main limitation that must be taken into account: recommendations are based on user preferences, not true experiences. Consequently, they do not provide empirical evidence that gamification designs our RS suggests are more effective than, for instance, the one-size-fits-all approach. We address this issue in Chapter 5.

CHAPTER

5

PERSONALIZED VERSUS ONE-SIZE-FITS-ALL GAMIFICATION

The recommender system we developed and introduced in Chapter 4 is based on user preference. Hence, it reflects potential, instead of true experiences. Additionally, up to this point, we have no empirical evidence to support (or refute) the hypothesis that using our recommender system to personalize gamified educational systems will lead to improvements compared to the standard, one-size-fits-all (OSFA) approach. Thereby, this chapter presents two studies comparing those approaches to address that need. Thus, addressing this research's third specific objective (see Figure 3, step 3).

5.1 Personalization Improves Gamification: Evidence from a Mixed-methods Study

In this study (RODRIGUES *et al.*, 2021), we built upon the context that little research compared personalized to standard, OSFA gamification empirically (e.g., (OLIVEIRA *et al.*, 2020; Mora *et al.*, 2018)) and that those are unclear on whether the former improves the latter, especially in the educational domain. We assumed a possible reason for the inconclusive findings was that prior research (e.g., (OLIVEIRA *et al.*, 2020; Mora *et al.*, 2018)) applied personalization strategies that consider a single user dimension for the personalization, whereas a dual personalization approach resulted in positive motivational outcomes in (STUART; LAVOUÉ; SERNA, 2020). Then, we expected multidimensional personalization[1] to be effective because users are not reduced to a single dimension. However, the only study (to our best knowledge) applying multidimensional personalization of gamification with users did not compare it to the OSFA approach (STUART; LAVOUÉ; SERNA, 2020).

[1] That is, personalizing to multiple dimensions (criteria).

Accordingly, the goal of this study was to understand the effect of gamification personalized to multiple dimensions, compared to the OSFA approach, on users' motivations in assessment learning tasks. Therefore, we conducted a mixed-methods sequential explanatory study (CRESWELL; CRESWELL, 2017). In the first phase, we compared OSFA and personalized gamification through a 2x2 mixed factorial experiment. We manipulated *gamification design* (between-subject) to create two versions of the system where students would complete the assessments. Those versions featured either an OSFA or the personalized gamification design. Participants (n = 26) engaged in two sessions that differed by the assessment *discipline* (within-subject): Programming Techniques and Object-Oriented Analysis and Design. Thus, we were able to compare the gamification designs based on two applications. In the second phase, we conducted semi-structured interviews to understand participants' motivations to use and engage with the gamified system. For further information on the method, please refer to Annex H.

In summary, our findings suggested that i) multidimensional personalization improved students' experiences with the gamified educational system, compared to the OSFA approach, in a degree above the hinge-point for educational interventions and OSFA gamification when compared to no gamification; and ii) the personalized design overcame the OSFA approach, in terms of autonomous motivation, by providing game elements suitable to learners' preferences that supported their needs and mitigated drawbacks from regular assessment activities. Accordingly, we derived two main implications. First, our empirical evidence suggests that gamifying educational systems with the strategy for multidimensional personalization we used can improve students' autonomous motivation (intrinsic and identified) compared to using the standard, OSFA approach. Thus, given that increased motivation is related to learning gains, applying multidimensional personalization is likely to enhance students' learning more than OSFA gamification. Second, our findings provide promising evidence that multidimensional personalization can improve OSFA gamification, a result that has not been found in other studies personalizing gamification using a single dimension. Thereby, contributing indication that personalizing to a single criterion might explain why related research found inconclusive results. Thus, suggesting the need for more complex personalization strategies such as the one used in this paper, and providing an initial empirical validation of that strategy.

5.2 How Personalization Affects Motivation in Gamified Review Assessments

Next, we conducted another experiment (RODRIGUES *et al.*, 2022), acknowledging the need for replications to increase the reliability of experimental studies (CAIRNS, 2019). In this study, the goal was to test whether the effect of gamification personalized to the learning task and users' gaming habits/preferences and demographics generalizes to other samples/contexts, as well as investigate possible moderators[2] of that effect. We accomplished that goal with an experimental study conducted in three institutions. Compared to the experiment presented in Section 5.1, the main differences are i) involving three, instead of one institution, ii) sampling northwestern Brazilian students instead of southwesterns, and iii) capturing repeated measures with four to six weeks of spacing instead of a one-day interval. Additionally, this study performed exploratory analyses to understand variations in multidimensional personalization's effect. This is important to advance the field from *whether* to *when/to whom* personalization works. Differently, (RODRIGUES *et al.*, 2021) is limited to confirmatory analyses. For further description of the method and differences between these studies, please refer to Annex I.

In summary, this study yielded four main findings. First and surprisingly, results do not confirm the personalization's positive effect on autonomous motivation; instead, indicating a nonsignificant difference. Second, exploratory analyses suggested *gender* and *education* positively moderated personalization's effect, in contrast to *preferred game genre* and *preferred playing setting*. Personalization was positive for *females* and those holding a *technical* degree, but negative for people who prefer either the *adventure* game genre or the *singleplayer* setting. Third, exploratory analyses also revealed motivation varied according to six characteristics for students who used the OSFA design: *performance, preferred game genre, preferred playing setting, education, assessment's subject*, and *usage interval*. The analyses indicated the motivation of students who used personalized gamification varied according to only four factors: *education*, *assessment's subject* (common to OSFA), *age* and *gender* (uncommon). Fourth, qualitative results indicated the gamified assessments provided positive experiences that students considered well designed and good for their learning, although a few of them mentioned gamification demands improvement and perceived the assessments as complex and poorly presented.

From those findings, the main implications for the design of GES are that: i) designers might use multidimensional personalization to offer more even experiences to their systems' users while paying attention to possible moderators of its effect; ii) gamified designs with more than three game elements might improve users' perceptions about gamification; and iii) successful personalization of gamification requires tailoring the game elements' mechanics as well as their availability. Additionally, we provide a number of theoretical contributions, such as i) questioning what is the exact mechanism through which personalization contributes to gamification, ii) discussing that the recommender system we provided in Section 4.2 might benefit from considering other information and further modeling some characteristics it already considers, and iii) reflecting on whether data-driven personalization strategies are more effective than preference-based ones. Overall, those results reveal a new way of seeing personalization's role in gamification and inform designers, instructors, and researchers by: i) showing whereas personalization might not increase the learning outcome's average, it likely improves gamification by reducing its outcome variation; ii) showing gamified review assessments provide positive

[2] Moderators are factors that increase/decrease an intervention's effect (LANDERS *et al.*, 2018).

experiences considered good learning means from students' perspectives; and iii) providing several design and research directions towards advancing the field study.

5.3 Summary

This chapter addressed our third specific objective: To empirically validate our developed approach. In Section 5.1, we presented a study that started to address this objective with an initial experimental study comparing personalized designs, suggested by the recommender system we presented in Section 4.2, to the standard, OSFA gamification (RODRIGUES *et al.*, 2021). Although the results were promising, we found no support for those encouraging findings when we conducted a replication study with a new sample (RODRIGUES *et al.*, 2022), as we discussed in Section 5.2. In contrast, confirmatory analyses revealed no significant differences between the approaches tested. On the other hand, when we conducted exploratory analyses and related them to the relevant literature, we were able to raise several hypotheses to be tested in future research. Due to time and resource constraints, we were not able to test many of those within the scope of this Ph.D. research. Nevertheless, we started to address concerns regarding the usage of data-driven approaches along with improving the recommender system we provided in Section 4.2, which we present in Chapter 6.

CHAPTER 6

GARFIELD: A RECOMMENDER SYSTEM TO PERSONALIZE GAMIFIED LEARNING

Based on the experimental studies reported in Chapter 5, we identified the need for further modeling the personalization strategy presented in Chapter 4. Additionally, the findings reported in Chapter 5 raised the possibility that relying on a personalization strategy based on true, instead of potential experiences, could maximize personalized gamification's effectiveness. Accordingly, this chapter presents an initial study (RODRIGUES *et al.*, 2022c) that refines our previous approach and provides a new, data-driven recommender system. Thus, respectively addressing this research's fourth step and fourth specific objective (see Figure 3, step 4).

In this study (RODRIGUES *et al.*, 2022c), we built upon the following: researchers have developed strategies to improve gamification designs through personalization, but most of those are based on theoretical understanding of game elements and their impact on students (RODRIGUES *et al.*, 2020), instead of considering real interaction data (KLOCK *et al.*, 2020). Thereby, they are limited because *potential* experiences might not reflect *real* experiences (PALOMINO *et al.*, 2019). To address this gap, we developed GARFIELD[1], the Gamification Automatic Recommender for Interactive Education and Learning Domains (see Figure 5).

GARFIELD is a recommender system for personalizing gamification built upon data from real experiences. Our goal with GARFIELD was to indicate the most suitable gamification design according to students' intrinsic motivation due to its positive relationship with learning (RODRIGUES *et al.*, 2021), while also taking students' characteristics into account. For this, we followed a two-step reverse engineering approach, informed by the CRISP-DM methodology (WIRTH; HIPP, 2000): we collected self-reports of users' intrinsic motivations from *actually* using a gamification design, then, regressed from such data (N = 221) to obtain recommendations of which design is the most suitable to achieve the desired motivation level given the user's information. Note that GARFIELD's development relied on the dataset collected and made

[1] https://github.com/rodriguesluiz/GARFIELD/wiki

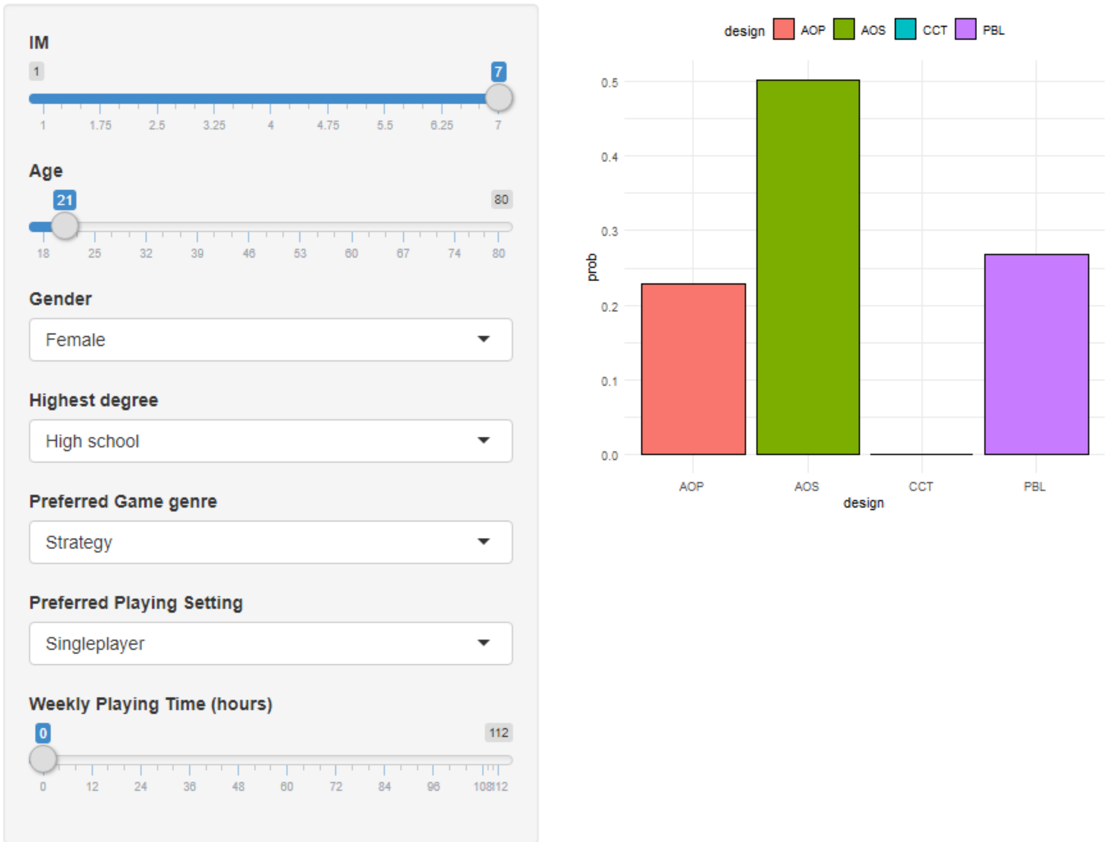

Figure 5 – Printscreen of GARFIELD, wherein recommendations (shown through the barplot) change in response to updating the sidebar.

available by Rodrigues *et al.* (2022), the second study we reported in Chapter 5. This dataset has data from students enrolled in STEM undergraduate courses at three Brazilian northwestern universities (ethical committee approval: 42598620. 0.0000.5464), wherein students self-reported their motivations to complete in-lecture assessments after using a gamified educational system.

After performing CRISP-DM's initial steps (understanding and preparation), we used Multinomial Logistic Regression (KWAK; CLAYTON, 2002) to achieve our goal of obtaining gamification design (dependent variable) recommendations according to students' intrinsic motivation and characteristics. Specifically, we used the following student characteristics: age, gender, educational background, preferred game genre, preferred game setting, and weekly playing time[2]. As independent variables, we started with all of those and, because recommendations should consider how students' intrinsic motivation from using a gamification design changes depending on their characteristics, our model assumes intrinsic motivation interacts with all other variables. Additionally, we defined the PBL design (points, badges, and leaderboard) as the reference value of the dependent variable because PBL is the most used gamification design in educational contexts (SAILER; HOMNER, 2020). For further details, please refer to Annex J.

In evaluating the regression model, we found that Cohen's Kappa for the agreement

2 Note that we selected those characteristics because we aimed to improve our previous personalization strategy. Hence, we used the same as before.

between its predictions and the ground truth is 0.43. This value is significantly different from zero ($p < 0.001$), with its 95% confidence interval ranging from 0.34 to 0.52, revealing a moderate agreement (VIERA; GARRETT *et al.*, 2005). Additionally, we found that the model performed the best for designs AOP (acknowledgments, objectives, and progression) and CCT (competition/leaderboard, chance, and time pressure), whereas its performance for designs PBL and AOS (acknowledgment, objectives, social pressure) were slightly worse. Following those results, we addressed CRISP-DM's last step - deployment - by developing GARFIELD, our interactive recommender system. Its interface receives user input and passes it to our model. Then, our model predicts the probability of recommending each possible design and presents it as a barplot. Accordingly, practitioners can use it to get transparent recommendations for personalizing their gamified designs in a simple, interactive way, based on a regression model that recommends gamified designs with moderate performance.

Based on these results, this research expands the literature by i) creating personalization guidelines from feedback collected after *real* experiences, in contrast to prior research that developed personalization guidelines based on potential experiences (e.g., (RODRIGUES *et al.*, 2022; STUART; LAVOUÉ; SERNA, 2020) and ii) providing a concrete, interactive recommender system unlike the conceptual tools related work has contributed (KLOCK *et al.*, 2020). With GARFIELD, practitioners have technological support to help them personalize their gamified practices aiming to overcome the limitations of the one-size-fits-all approach. For instance, instructors can use it to deploy user-centered gamification, based on the class's overall characteristics, and to define individualized designs according to each student's characteristics. Hence, bridging the gap between academic research and interested parties, such as companies and non-profit organizations, so they can benefit from GARFIELD to maximize learning everywhere, to everyone without the need for monetary investments.

To the best of our knowledge, GARFIELD is the first tool that guides practitioners and instructors on how to personalize gamification based on empirical data from *real* usage. While its recommendations are limited to moderate predictive power, compared to the ground truth, at the time of writing the paper, we believe it provides practitioners with a reliable starting point and paves the way for researchers to expand and improve it in future research.

CHAPTER 7

FINAL REMARKS

When I started my Ph.D. research, I found researchers exploring personalized gamification to mitigate the shortcomings of the one-size-fits-all approach. Ultimately, the goal was to achieve equitable gamification designs that would maximize the learning experiences of all learners. At that time, however, empirical studies showed inconclusive results on whether personalization leads to improvements compared to the one-size-fits-all approach. To understand that context, I reviewed the state of the art and noted that most empirical studies personalized gamification to a single input. In contrast, the overall gamification literature showed several user and contextual factors moderate gamification's effectiveness. Nevertheless, to my best knowledge, there was no empirically validated approach for personalizing gamification designs to the users' and context's characteristics within the educational domain.

In light of that context, my objective was to develop and validate a personalization approach for tailoring gamification designs of educational systems to contextual and user characteristics. Accordingly, I hypothesized that gamification designs tailored to the context and the user's characteristics were more effective in affecting learning outcomes than one-size-fits-all designs, in the context of educational systems. On the one hand, I successfully accomplished my goal. Following an iterative method, I developed and validated recommendations for the multidimensional personalization of gamified educational systems. On the other hand, I only found partial support for my thesis. The first validation study suggested personalization led to improvements in motivational learning outcomes. However, I could not replicate those findings in my second validation study, which suggested nonsignificant differences among the personalized and one-size-fits-all designs. Nevertheless, the latter study suggested personalized gamification designs were less sensitive to user characteristics than the one-size-fits-all counterpart. Thus, while I cannot confirm my thesis, based on empirical results I conclude that my strategy for personalizing gamification might improve the one-size-fits-all approach by making it more inclusive, if not increasing overall learning outcomes as initially expected.

Based on that context, the remainder of this chapter discusses my contributions that led to this research's conclusion, as well as limitations that should be taken into account and recommendations for future studies. Additionally, this chapter summarizes publications and the participation of the student author in the scientific community throughout this Ph.D. research.

7.1 Overall Contributions

In seeking my main objective, I asked three research questions to understand what to consider and how to personalize gamification within the educational domain, as well as its effectiveness. First, I asked *what factors impact the success of gamified systems*. Based on four empirical studies (see Chapter 3), I found that prior familiarity with the to-be-learned content, usage time, and gender affect the extent to which gamification improves learning outcomes compared to no-gamification. Results also revealed that when two factors are analyzed simultaneously, such as usage time and prior familiarity, their interaction plays a role on gamification success. Furthermore, based on a survey study (see Chapter 4), I found that demographic information, gaming habits, geographic location, and those interactions affect people's perception of the usefulness of different gamification designs. Thus, I contribute empirical evidence that multiple user and contextual characteristics, as well as their interaction, moderate gamification's effectiveness within the educational context.

Second, I asked *how to tailor gamified educational systems to the context and the user's characteristics*. In light of my first contribution, I aimed for a multidimensional approach that considered user and contextual information simultaneously. Then, I first developed a preference-based personalization strategy, which I made available as a recommender system (see Chapter 4). Additionally, I developed and made available another recommender system, GARFIELD, which in contrast was built based on user feedback following true gamified experiences (see Chapter 6). Thus, I contribute two recommender systems that can be used both to inform designers and instructors on how to personalize their gamified systems/lessons as well as to enable automatic personalization if integrated within gamified educational systems.

Third, I asked *whether gamification designs tailored to the context and the user's characteristics are more effective than one-size-fits-all designs, in the context of educational systems*. I addressed this question with two experimental studies, conducted within real classrooms from south and northwestern Brazilian institutions (see Chapter 5). In the first one, I found students who used gamification personalized based on my strategy reported higher autonomous motivation than those who used the one-size-fits-all approach. In the second study, I found nonsignificant differences, but exploratory analyses suggested personalization could be acting by making gamification more inclusive compared to the one-size-fits-all approach. Thus, I contribute empirical evidence on i) how my personalization strategy compares to the one-size-fits-all approach and what kind of benefits to expect from the multidimensional personalization of gamification.

Moreover, conducting this research allowed us to help hundreds of students while advancing scientific knowledge. To increase the validity of the experimental studies, I prioritized conducting them within real classroom environments. Therefore, I first contacted instructors willing to deploy gamified experiences within their classrooms. Second, I designed assessments tailored to each classroom based on careful feedback from its instructor. Hence, while I was collecting data to achieve this research's goal, participants were practicing/learning topics relevant for their graduation, besides helping advance scientific knowledge. Similarly, I contributed to the instructors who gave us the opportunity to run the experimental studies by deploying gamified activities in their classrooms. As a result, I contributed to the lessons of over 10 instructors as well as the learning experiences of hundreds of undergraduate students from SENAI Londrina, Federal University of Roraima, Federal University of Amazonas, and Amazonas State University.

Lastly, this Ph.D. project extended its contributions in an effort to contribute to open science. In both experimental studies reported in Chapter 5, I openly shared their materials (e.g., the questionnaire used to measure learning outcomes; assessments participants completed as part of the experimental task), data (e.g., answers to measures), and data analysis scripts. Note that I used a third-party system for data collection, which I cannot openly share in a similar way. Nevertheless, I provided videos and screenshots to transparently illustrate the gamification designs I studied and facilitate replications. Thus, I additionally contribute two datasets, data analysis plans, and materials for future research to exploit and build upon.

Accordingly, I can summarize my contributions as follows:

- Empirical evidence that user and contextual factors, along with their interaction, moderate gamification's effectiveness;
- Two recommender systems to inform the multidimensional personalization of gamification applied to education;
- Empirical evidence on how my personalization strategy compares to the one-size-fits-all approach;
- Applying gamified assessments to enhance students learning experiences and instructors' lessons in real classrooms;
- Datasets, data analysis plans, and materials for replications and exploratory analyses;

7.2 Limitations and Threats to Validity

This section discusses the main limitations readers should be aware of when interpreting my findings.

Concerning the literature review, the main limitations relate to the protocol of secondary studies. For instance, those involve search string definition, source selection, and selection bias, among others (KITCHENHAM *et al.*, 2009). Because I was aware of recent secondary studies when I started this research project, I opted to exploit them instead of conducting a new study from the scratch. While that reduced the time for reviewing the literature, it carried the limitations of each of the studies I relied upon. Nevertheless, I consider those are not major limitations because each secondary study had been through a rigorous peer-review process. Additionally, despite starting from already published secondary studies, I defined and followed a systematic protocol to mitigate further threats to validity.

Concerning the studies on moderators of gamification's effect, the main limitations relate to external and conclusion validity. Two of those studies were limited to less than 20 participants, which directly affect their findings' generalization and analyses' statistical power. To mitigate those, I used robust statistical tests and conducted replication studies aiming to increase external and conclusion validity, respectively.

Concerning the preference-based personalization strategy, its main limitation relates to being based on potential experiences. In empirically evaluating its recommendations' effectiveness, I found positive results in the first study, but nonsignificant overall improvements in the second. Despite that, the second study suggested that, instead of increasing overall learning outcomes, the personalization strategy might act by making gamification's effectiveness less sensitive to users with different characteristics. Hence, despite based on potential experiences, empirical findings from two experimental studies indicate it might make gamification more inclusive if not more effective in terms of overall learning experiences.

Concerning the empirical studies comparing my preference-based personalization strategy to one-size-fits-all gamification, their main limitation relates to external validity. The first study was limited to two short-term interactions of 26 students, which limits the findings' generalization to other samples and is prone to the novelty effect. Additionally, the sample size prevented us from conducting exploratory analyses to further understand the perceptions of users with different characteristics. To mitigate those limitations, I replicated this study in a different context (n = 58). As a result, findings on the strategy's overall effect were contrasting, as the latter revealed nonsignificant differences between personalization and one-size-fits-all gamification. Also, I was able to conduct exploratory analyses that revealed directions for improving the recommendations as well as a different perspective on how personalization might improve gamification. Overall, whereas the replication increases the results' external validity, one still should interpret these findings with parsimony, especially the insights from the second study's exploratory analyses.

Lastly, the data-driven recommender system - GARFIELD - shares limitations similar to those of the experimental study. GARFIELD is built upon data gathered in my second experimental study. Consequently, its recommendations are specific to samples similar to that of the experimental study. Similarly, GARFIELD's recommendations are limited to the four gamification designs participants could interact with in the second experimental study (see

Chapter 6). To mitigate those, I developed GARFIELD based on a systematic process and validated it according to the null hypothesis significance testing framework to increase conclusion validity. Nevertheless, I cannot extend it to other gamification designs without new data and its users should be aware of its target sample.

7.3 Recommendations for Future Research

Taking this research's findings, contributions, and limitations together, I mainly recommend the following directions for future studies:

- Future research should explore similar and new moderators of gamification's effectiveness, based on experimental studies, to ground which information should be taken into account when personalizing gamification;
- Future research should replicate the experimental studies analyzing my preference-based personalization strategy to ground whether and how it compares to one-size-fits-all gamification;
- Future research should conduct experimental studies to test how and whether GARFIELD's recommendations compare to one-size-fits-all gamification;
- Future research should conduct experimental studies comparing GARFIELD's recommendations to those of my preference-based strategy to understand how preference-based and data-driven personalization strategies compare;
- Future research should explore the openly shared datasets and data analysis plans to expand GARFIELD so that its recommendations encompass a broader sample and more gamification designs.

7.4 Publications

While conducting this research project, the student author published four journal articles as the first author:

- Rodrigues, L., Pereira, F. D., Toda, A. M., Palomino, P. T., Pessoa, M., Carvalho, L. S. G., ... & Isotani, S. (2022). Gamification suffers from the novelty effect but benefits from the familiarization effect: Findings from a longitudinal study. International Journal of Educational Technology in Higher Education, 19(1), 1-25.
- Rodrigues, L., Pereira, F., Toda, A., Palomino, P., Oliveira, W., Pessoa, M., ... & Isotani, S. (2022). Are they learning or playing? Moderator conditions of gamification's success in programming classrooms. ACM Transactions on Computing Education (TOCE).

- Rodrigues, L., Palomino, P. T., Toda, A. M., Klock, A. C., Oliveira, W., Avila-Santos, A. P., ... & Isotani, S. (2021). Personalization improves gamification: evidence from a mixed-methods study. Proceedings of the ACM on Human-Computer Interaction, 5(CHI PLAY), 1-25.
- Rodrigues, L., Toda, A. M., Oliveira, W., Palomino, P. T., Vassileva, J., & Isotani, S. (2022). Automating gamification personalization to the user and beyond. IEEE Transactions on Learning Technologies, 15(2), 199-212.

Another journal article, Rodrigues *et al.* (2022), is currently under the third round of reviews in the International Journal of Artificial Intelligence in Education.

Additionally, he published five conference papers as the first author:

- Rodrigues, L., Toda, A. M., Oliveira, W., Palomino, P. T., Avila-Santos, A. P., & Isotani, S. (2021, March). Gamification Works, but How and to Whom? An Experimental Study in the Context of Programming Lessons. In Proceedings of the 52nd ACM Technical Symposium on Computer Science Education (pp. 184-190).
- Rodrigues, L., Toda, A. M., Oliveira, W., Palomino, P. T., & Isotani, S. (2020, November). Just beat it: Exploring the influences of competition and task-related factors in gamified learning environments. In Anais do XXXI Simpósio Brasileiro de Informática na Educação (pp. 461-470). SBC.
- Rodrigues, L., Toda, A. M., Palomino, P. T., Oliveira, W., & Isotani, S. (2020, October). Personalized gamification: A literature review of outcomes, experiments, and approaches. In Eighth International Conference on Technological Ecosystems for Enhancing Multiculturality (pp. 699-706).
- Rodrigues, L., Oliveira, W., Toda, A., Palomino, P., & Isotani, S. (2019, November). Thinking inside the box: How to tailor gamified educational systems based on learning activities types. In Brazilian symposium on computers in education (Simpósio Brasileiro de Informática na Educação-SBIE) (Vol. 30, No. 1, p. 823).
- Rodrigues, L., Toda, A., Pereira, F., Palomino, P. T., Klock, A. C., Pessoa, M., ... & Isotani, S. (2022). GARFIELD: A Recommender System to Personalize Gamified Learning. In International Conference on Artificial Intelligence in Education (pp. 666-672). Springer, Cham.

Furthermore, the student author published seven journal articles as coauthor:

- Oliveira, W., Hamari, J., Shi, L., Toda, A. M., Rodrigues, L., Palomino, P. T., & Isotani, S. (2022). Tailored gamification in education: A literature review and future agenda. Education and Information Technologies, 1-34.

- Santos, A. C. G., Oliveira, W., Hamari, J., Rodrigues, L., Toda, A. M., Palomino, P. T., & Isotani, S. (2021). The relationship between user types and gamification designs. User Modeling and User-Adapted Interaction, 31(5), 907-940.

- Pereira, F. D., Fonseca, S. C., Oliveira, E. H., Cristea, A. I., Bellhäuser, H., Rodrigues, L., ... & Carvalho, L. S. (2021). Explaining Individual and Collective Programming Students' Behavior by Interpreting a Black-Box Predictive Model. IEEE Access, 9, 117097-117119.

- Oliveira, W., Toda, A. M., Palomino, P. T., Rodrigues, L., & Isotani, S. (2020). Which one is the best? A quasi-experimental study comparing frameworks for unplugged gamification. RENOTE, 18(1).

- Toda, A. M., Palomino, P. T., Oliveira, W., Rodrigues, L., Klock, A. C., Gasparini, I., ... & Isotani, S. (2019). How to gamify learning systems? an experience report using the design sprint method and a taxonomy for gamification elements in education. Journal of Educational Technology & Society, 22(3), 47-60.

- Palomino, P. T., Toda, A. M., Oliveira, W., Rodrigues, L., & Isotani, S. (2019). Teaching Interactive Fiction for Undergraduate Students with the Aid of Information Technologies: An Experience Report. RENOTE, 17(3), 527-536.

- Toda, A. M., Klock, A. C., Oliveira, W., Palomino, P. T., Rodrigues, L., Shi, L., ... & Cristea, A. I. (2019). Analysing gamification elements in educational environments using an existing Gamification taxonomy. Smart Learning Environments, 6(1), 1-14.

Lastly, he published 14 conference papers as coauthor:

- Toda, A., Palomino, P. T., Rodrigues, L., Klock, A. C., Pereira, F., Borges, S., ... & Cristea, A. I. (2022). Gamification Through the Looking Glass-Perceived Biases and Ethical Concerns of Brazilian Teachers. In International Conference on Artificial Intelligence in Education (pp. 259-262). Springer, Cham.

- Armando Toda, Ana Klock, Filipe Dwan Pereira, Luiz Rodrigues, Paula Toledo Palomino, Vinicius Lopes, Craig Stewart, Elaine H. T. Oliveira, Isabela Gasparini, Seiji Isotani, Alexandra Cristea. (2022). Towards the understanding of cultural differences in between gamification preferences: A data-driven comparison between the US and Brazil. Proceedings of the 15th International Conference on Educational Data Mining, 560–564.

- Pessoa, M., Melo, R., Haydar, G., de Oliveira, D. B., Carvalho, L. S., de Oliveira, E. H., ... & Isotani, S. (2021, November). Uma análise dos tipos de jogadores em uma plataforma de gamificação incorporada a um sistema juiz on-line. In Anais do XXXII Simpósio Brasileiro de Informática na Educação (pp. 474-486). SBC.

- Oliveira, W., Pastushenko, O., Rodrigues, L., Toda, A. M., Palomino, P. T., Hamari, J., & Isotani, S. (2021). Does gamification affect flow experience? a systematic literature review. arXiv preprint arXiv:2106.09942.

- Pereira, F. D., Junior, H. B., Rodriguez, L., Toda, A., Oliveira, E. H., Cristea, A. I., ... & Isotani, S. (2021, June). A recommender system based on effort: Towards minimising negative affects and maximising achievement in cs1 learning. In International Conference on Intelligent Tutoring Systems (pp. 466-480). Springer, Cham.

- Oliveira, W., Toda, A. M., Palomino, P. T., Rodrigues, L., Shi, L., & Isotani, S. (2020, November). Towards automatic flow experience identification in educational systems: A qualitative study. In Anais do XXXI Simpósio Brasileiro de Informática na Educação (pp. 702-711). SBC.

- Toda, A., Pereira, F. D., Klock, A. C. T., Rodrigues, L., Palomino, P., Oliveira, W., ... & Isotani, S. (2020, November). For whom should we gamify? Insights on the users intentions and context towards gamification in education. In Anais do XXXI Simpósio Brasileiro de Informática na Educação (pp. 471-480). SBC.

- Toda, A., Klock, A. C. T., Palomino, P. T., Rodrigues, L., Oliveira, W., Stewart, C., ... & Isotani, S. (2020, October). GamiCSM: relating education, culture and gamification-a link between worlds. In Proceedings of the 19th Brazilian Symposium on Human Factors in Computing Systems (pp. 1-10).

- Palomino, P., Toda, A., Rodrigues, L., Oliveira, W., & Isotani, S. (2020). From the Lack of Engagement to Motivation: Gamification Strategies to Enhance Users Learning Experiences. In 2020 19th Brazilian Symposium on Computer Games and Digital Entertainment (SBGames)-GranDGames BR Forum.

- Palomino, P., Toda, A., Oliveira, W., Rodrigues, L., Cristea, A., & Isotani, S. (2019, November). Exploring content game elements to support gamification design in educational systems: narrative and storytelling. In Brazilian symposium on computers in education (Simpósio brasileiro de informática na educação-SBIE) (Vol. 30, No. 1, p. 773).

- Toda, A., Palomino, P., Rodrigues, L., Oliveira, W., Shi, L., Isotani, S., & Cristea, A. (2020). Validating the effectiveness of data-driven gamification recommendations: An Exploratory study. arXiv preprint arXiv:2008.05847.

- Oliveira, W., Rodrigues, L., Toda, A., Palomino, P., & Isotani, S. (2019, November). Automatic game experience identification in educational games. In Brazilian Symposium on Computers in Education (Simpósio Brasileiro de Informática na Educação-SBIE) (Vol. 30, No. 1, p. 952).

- Palomino, P. T., Toda, A. M., Oliveira, W., Rodrigues, L., & Isotani, S. (2019). Gamification journey: A Novel approach for classifying gamer types for gamified educational systems. Simpósio Brasileiro de Jogos e Entretenimento Digital, 2019.

- Oliveira, W., Toda, A., Palomino, P., Rodrigues, L., Isotani, S., & Shi, L. (2019, October). Towards automatic flow experience identification in educational systems: A theory-driven approach. In Extended abstracts of the annual symposium on computer-human interaction in play companion extended abstracts (pp. 581-588).

BIBLIOGRAPHY

ALJABALI, R. N.; AHMAD, N. A review on adopting personalized gamified experience in the learning context. In: IEEE. **2018 IEEE Conference on e-Learning, e-Management and e-Services (IC3e)**. [S.l.], 2018. p. 61–66. Citation on page 31.

ANDRADE, F. R. H.; MIZOGUCHI, R.; ISOTANI, S. The bright and dark sides of gamification. In: MICARELLI, A.; STAMPER, J.; PANOURGIA, K. (Ed.). **Intelligent Tutoring Systems**. Cham: Springer International Publishing, 2016. p. 176–186. ISBN 978-3-319-39583-8. Citation on page 19.

BAI, S.; HEW, K. F.; HUANG, B. Is gamification "bullshit"? evidence from a meta-analysis and synthesis of qualitative data in educational contexts. **Educational Research Review**, Elsevier, p. 100322, 2020. Citations on pages 25 and 33.

BALDEóN, J.; RODRÍGUEZ, I.; PUIG, A. Lega: A learner-centered gamification design framework. In: **Proceedings of the XVII International Conference on Human Computer Interaction**. New York, NY, USA: ACM, 2016. (Interacción '16), p. 45:1–45:8. ISBN 978-1-4503-4119-6. Available: <http://doi.acm.org/10.1145/2998626.2998673>. Citations on pages 22, 24, and 26.

BLOOM, B. S. **Taxonomy of educational objectives: The classification of educational goals**. New York: Longman, 1956. Citation on page 24.

BÖCKLE, M.; NOVAK, J.; BICK, M. Towards adaptive gamification: a synthesis of current developments. In: **Proceedings of the 25th European Conference on Information Systems (ECIS)**. Guimarães, Portugal: AIS, 2017. p. 158–174. Citation on page 20.

BORGES, S.; DURELLI, V.; REIS, H.; BITTENCOURT, I. I.; MIZOGUCHI, R.; ISOTANI, S. Selecting effective influence principles for tailoring gamification-based strategies to player roles. In: **Brazilian Symposium on Computers in Education**. Recife, Brazil: SBC, 2017. v. 28, p. 857–866. Citation on page 25.

BORGES, S. d. S.; DURELLI, V. H.; REIS, H. M.; ISOTANI, S. A systematic mapping on gamification applied to education. In: ACM. **Proceedings of the 29th annual ACM symposium on applied computing**. [S.l.], 2014. p. 216–222. Citation on page 19.

BOUWMEESTER, R. A.; KLEIJN, R. A. de; BERG, I. E. van den; CATE, O. T. J. ten; RIJEN, H. V. van; WESTERVELD, H. E. Flipping the medical classroom: Effect on workload, interactivity, motivation and retention of knowledge. **Computers & Education**, Elsevier, v. 139, p. 118–128, 2019. Citation on page 19.

BOVERMANN, K.; BASTIAENS, T. J. Towards a motivational design? connecting gamification user types and online learning activities. **Research and Practice in Technology Enhanced Learning**, Springer, v. 15, n. 1, p. 1–18, 2020. Citations on pages 21, 22, 26, and 43.

BRANCH, R. M. **Instructional design: The ADDIE approach**. [S.l.]: Springer Science & Business Media, 2009. Citation on page 36.

BUCKLEY, P.; DOYLE, E. Individualising gamification: An investigation of the impact of learning styles and personality traits on the efficacy of gamification using a prediction market. **Computers & Education**, Elsevier, v. 106, p. 43–55, 2017. Citation on page 25.

CAIRNS, P. **Doing better statistics in human-computer interaction**. [S.l.]: Cambridge University Press, 2019. Citations on pages 38, 39, and 48.

CHANG, V. Review and discussion: E-learning for academia and industry. **International Journal of Information Management**, Elsevier, v. 36, n. 3, p. 476–485, 2016. Citation on page 42.

CRESWELL, J. W.; CRESWELL, J. D. **Research design: Qualitative, quantitative, and mixed methods approaches**. [S.l.]: Sage publications, 2017. Citation on page 48.

Denden, M.; Tlili, A.; Essalmi, F.; Jemni, M. An investigation of the factors affecting the perception of gamification and game elements. In: **2017 6th International Conference on Information and Communication Technology and Accessibility (ICTA)**. Muscat, Sultanate of Oman: IEEE, 2017. p. 1–6. ISSN 2379-4402. Citations on pages 22, 28, and 44.

DENDEN, M.; TLILI, A.; ESSALMI, F.; JEMNI, M. Does personality affect students' perceived preferences for game elements in gamified learning environments? In: IEEE. **2018 IEEE 18th International Conference on Advanced Learning Technologies (ICALT)**. [S.l.], 2018. p. 111–115. Citation on page 25.

DETERDING, S. The lens of intrinsic skill atoms: A method for gameful design. **Human–Computer Interaction**, Taylor & Francis, v. 30, n. 3-4, p. 294–335, 2015. Citation on page 22.

DETERDING, S.; DIXON, D.; KHALED, R.; NACKE, L. From game design elements to gamefulness: defining gamification. In: ACM. **Proceedings of the 15th international academic MindTrek conference: Envisioning future media environments**. [S.l.], 2011. p. 9–15. Citation on page 19.

DIAPER, D.; STANTON, N. **The handbook of task analysis for human-computer interaction**. [S.l.]: CRC Press, 2003. Citation on page 24.

DICHEV, C.; DICHEVA, D.; IRWIN, K. Towards activity-centered gamification design. In: **Proceedings of the IEEE International Conference on Teaching, Assessment, and Learning for Engineering (TALE 2019)**. [S.l.: s.n.], 2019. Citations on pages 22 and 26.

DICK, W.; CAREY, L.; CAREY, J. O. The systematic design of instruction. Citeseer, 2005. Citation on page 36.

DRAGICEVIC, P. Fair statistical communication in hci. In: **Modern statistical methods for HCI**. : Springer, 2016. p. 291–330. Citations on pages 38 and 39.

FAN, H.; POOLE, M. S. What is personalization? perspectives on the design and implementation of personalization in information systems. **Journal of Organizational Computing and Electronic Commerce**, Taylor & Francis, v. 16, n. 3-4, p. 179–202, 2006. Citation on page 20.

GARNER, P.; HOPEWELL, S.; CHANDLER, J.; MACLEHOSE, H.; AKL, E. A.; BEYENE, J.; CHANG, S.; CHURCHILL, R.; DEARNESS, K.; GUYATT, G. *et al.* When and how to update systematic reviews: consensus and checklist. **bmj**, British Medical Journal Publishing Group, v. 354, 2016. Citation on page 31.

GUAY, F.; VALLERAND, R. J.; BLANCHARD, C. On the assessment of situational intrinsic and extrinsic motivation: The situational motivation scale (sims). **Motivation and emotion**, Springer, v. 24, n. 3, p. 175–213, 2000. Citation on page 34.

HALLIFAX, S.; AUDREY, S.; JEAN-CHARLES, M.; GUILLAUME, L.; ELISE, L. Factors to Consider for Tailored Gamification. In: **CHI Play**. Barcelona, Spain: ACM, 2019. p. 559–572. Available: <https://hal.archives-ouvertes.fr/hal-02185647>. Citations on pages 25, 28, 38, and 43.

HALLIFAX, S.; SERNA, A.; MARTY, J.-C.; LAVOUÉ, É. Adaptive gamification in education: A literature review of current trends and developments. In: SCHEFFEL, M.; BROISIN, J.; PAMMER-SCHINDLER, V.; IOANNOU, A.; SCHNEIDER, J. (Ed.). **Transforming Learning with Meaningful Technologies**. Cham: Springer International Publishing, 2019. p. 294–307. ISBN 978-3-030-29736-7. Citations on pages 20, 21, 22, 23, 25, 26, 31, and 43.

HANUS, M. D.; FOX, J. Assessing the effects of gamification in the classroom: A longitudinal study on intrinsic motivation, social comparison, satisfaction, effort, and academic performance. **Computers & Education**, v. 80, p. 152 – 161, 2015. ISSN 0360-1315. Available: <http://www.sciencedirect.com/science/article/pii/S0360131514002000>. Citations on pages 19 and 35.

HUANG, R.; RITZHAUPT, A. D.; SOMMER, M.; ZHU, J.; STEPHEN, A.; VALLE, N.; HAMPTON, J.; LI, J. The impact of gamification in educational settings on student learning outcomes: a meta-analysis. **Educational Technology Research and Development**, Springer, p. 1–27, 2020. Citations on pages 25 and 41.

KITCHENHAM, B.; BRERETON, O. P.; BUDGEN, D.; TURNER, M.; BAILEY, J.; LINKMAN, S. Systematic literature reviews in software engineering–a systematic literature review. **Information and software technology**, Elsevier, v. 51, n. 1, p. 7–15, 2009. Citations on pages 31 and 58.

KLOCK, A. C. T.; GASPARINI, I.; PIMENTA, M. S.; HAMARI, J. Tailored gamification: A review of literature. **International Journal of Human-Computer Studies**, Elsevier, p. 102495, 2020. Citations on pages 20, 21, 22, 26, 28, 31, 41, 43, 51, and 53.

KLOCK, A. C. T.; OGAWA, A. N.; GASPARINI, I.; PIMENTA, M. S. Does gamification matter?: a systematic mapping about the evaluation of gamification in educational environments. In: ACM. **Proceedings of the 33rd Annual ACM Symposium on Applied Computing**. [S.l.], 2018. p. 2006–2012. Citation on page 19.

KLOCK, A. C. T.; PIMENTA, M. S.; GASPARINI, I. A systematic mapping of the customization of game elements in gamified systems. **Anais do Simpósio Brasileiro de Jogos e Entretenimento Digital**, 2018. Citations on pages 21 and 31.

KNUTAS, A.; ROY, R. V.; HYNNINEN, T.; GRANATO, M.; KASURINEN, J.; IKONEN, J. A process for designing algorithm-based personalized gamification. **Multimedia Tools and Applications**, Springer, v. 78, n. 10, p. 13593–13612, 2019. Citations on pages 19 and 20.

KOIVISTO, J.; HAMARI, J. Demographic differences in perceived benefits from gamification. **Computers in Human Behavior**, v. 35, p. 179 – 188, 2014. ISSN 0747-5632. Available: <http://www.sciencedirect.com/science/article/pii/S0747563214001289>. Citations on pages 25 and 35.

____. The rise of motivational information systems: A review of gamification research. **International Journal of Information Management**, Elsevier, v. 45, p. 191–210, 2019. Citations on pages 19, 20, 22, 25, 28, 33, 36, 37, and 43.

KRATHWOHL, D. R. A revision of bloom's taxonomy: An overview. **Theory into practice**, Taylor & Francis, v. 41, n. 4, p. 212–218, 2002. Citations on pages 24 and 28.

KWAK, C.; CLAYTON, A. Multinomial logistic regression. **Nursing research**, v. 51, 2002. Citation on page 52.

LANDERS, R. N. Developing a theory of gamified learning: Linking serious games and gamification of learning. **Simulation & gaming**, SAGE Publications Sage CA: Los Angeles, CA, v. 45, n. 6, p. 752–768, 2014. Citation on page 37.

LANDERS, R. N.; AUER, E. M.; COLLMUS, A. B.; ARMSTRONG, M. B. Gamification science, its history and future: Definitions and a research agenda. **Simulation & Gaming**, SAGE Publications Sage CA: Los Angeles, CA, v. 49, n. 3, p. 315–337, 2018. Citations on pages 34, 35, 37, and 49.

LANDERS, R. N.; COLLMUS, A. B.; WILLIAMS, H. The greatest battle is within ourselves: An experiment on the effects of competition alone on task performance. **International Journal of Human-Computer Studies**, Elsevier, v. 127, p. 51–61, 2019. Citation on page 34.

LANDERS, R. N.; TONDELLO, G. F.; KAPPEN, D. L.; COLLMUS, A. B.; MEKLER, E. D.; NACKE, L. E. Defining gameful experience as a psychological state caused by gameplay: Replacing the term 'gamefulness' with three distinct constructs. **International Journal of Human-Computer Studies**, Elsevier, v. 127, p. 81–94, 2019. Citation on page 25.

LAVOUÉ, E.; MONTERRAT, B.; DESMARAIS, M.; GEORGE, S. Adaptive gamification for learning environments. **IEEE Transactions on Learning Technologies**, IEEE, v. 12, n. 1, p. 16–28, 2018. Citations on pages 20 and 22.

LEE, J. J.; HAMMER, J. Gamification in education: What, how, why bother? **Academic exchange quarterly**, Chattanooga State Technical Community College, v. 15, n. 2, p. 146, 2011. Citation on page 19.

LIU, D.; SANTHANAM, R.; WEBSTER, J. Toward meaningful engagement: A framework for design and research of gamified information systems. **MIS quarterly**, v. 41, n. 4, p. 1011–1034, 2017. Citations on pages 19, 20, 22, 24, 25, 28, 43, and 44.

LO, C. K.; HEW, K. F. A comparison of flipped learning with gamification, traditional learning, and online independent study: the effects on students' mathematics achievement and cognitive engagement. **Interactive Learning Environments**, Taylor & Francis, v. 28, n. 4, p. 464–481, 2020. Citation on page 19.

LOUGHREY, K.; BROIN, D. Are we having fun yet? misapplying motivation to gamification. In: IEEE. **2018 IEEE Games, Entertainment, Media Conference (GEM)**. [S.l.], 2018. p. 1–9. Citation on page 19.

MARACHE-FRANCISCO, C.; BRANGIER, E. **Process of Gamification From The Consideration of Gamification To Its Practical Implementation**. 2013. Citation on page 25.

MCCRAE, R. R.; JOHN, O. P. An introduction to the five-factor model and its applications. **Journal of personality**, Wiley Online Library, v. 60, n. 2, p. 175–215, 1992. Citation on page 22.

MEKLER, E. D.; BRüHLMANN, F.; TUCH, A. N.; OPWIS, K. Towards understanding the effects of individual gamification elements on intrinsic motivation and performance. **Computers in Human Behavior**, v. 71, p. 525 – 534, 2017. ISSN 0747-5632. Available: <http://www.sciencedirect.com/science/article/pii/S0747563215301229>. Citation on page 20.

MONTERRAT, B.; LAVOUÉ, E.; GEORGE, S. Adaptation of gaming features for motivating learners. **Simulation & Gaming**, SAGE Publications Sage CA: Los Angeles, CA, v. 48, n. 5, p. 625–656, 2017. Citations on pages 20, 22, and 25.

MORA, A.; TONDELLO, G. F.; CALVET, L.; GONZÁLEZ, C.; ARNEDO-MORENO, J.; NACKE, L. E. The quest for a better tailoring of gameful design: An analysis of player type preferences. In: ACM. **Proceedings of the XX International Conference on Human Computer Interaction**. [S.l.], 2019. p. 1. Citations on pages 22, 25, and 28.

Mora, A.; Tondello, G. F.; Nacke, L. E.; Arnedo-Moreno, J. Effect of personalized gameful design on student engagement. In: **2018 IEEE Global Engineering Education Conference (EDUCON)**. [S.l.: s.n.], 2018. p. 1925–1933. Citations on pages 21 and 47.

MORSCHHEUSER, B.; HASSAN, L.; WERDER, K.; HAMARI, J. How to design gamification? a method for engineering gamified software. **Information and Software Technology**, Elsevier, v. 95, p. 219–237, 2018. Citations on pages 19 and 25.

NACKE, L. E.; BATEMAN, C.; MANDRYK, R. L. Brainhex: A neurobiological gamer typology survey. **Entertainment computing**, Elsevier, v. 5, n. 1, p. 55–62, 2014. Citations on pages 22 and 25.

NACKE, L. E.; DETERDING, C. S. The maturing of gamification research. **Computers in Human Behaviour**, York, p. 450–454, 2017. Citation on page 38.

NICHOLSON, S. A user-centered theoretical framework for meaningful gamification. In: **Games+Learning+Society 8.0**. [S.l.: s.n.], 2012. p. 1–7. Citation on page 25.

OLIVEIRA, W.; BITTENCOURT, I. I. **Tailored Gamification to Educational Technologies**. [S.l.]: Springer Nature, 2019. Citations on pages 20 and 42.

OLIVEIRA, W.; TODA, A.; TOLEDO, P.; SHI, L.; VASSILEVA, J.; BITTENCOURT, I. I.; ISOTANI, S. Does tailoring gamified educational systems matter? the impact on students' flow experience. In: **Proceedings of the 53rd Hawaii International Conference on System Sciences**. [S.l.]: ScholarSpace, 2020. p. 1226–1235. Citations on pages 22, 25, and 47.

ORJI, R.; OYIBO, K.; TONDELLO, G. F. A comparison of system-controlled and user-controlled personalization approaches. In: **Adjunct Publication of the 25th Conference on User Modeling, Adaptation and Personalization**. New York, NY, USA: ACM, 2017. (UMAP '17), p. 413–418. ISBN 978-1-4503-5067-9. Available: <http://doi.acm.org/10.1145/3099023.3099116>. Citations on pages 19 and 21.

ORJI, R.; TONDELLO, G. F.; NACKE, L. E. Personalizing persuasive strategies in gameful systems to gamification user types. In: ACM. **Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems**. [S.l.], 2018. p. 435. Citations on pages 19 and 22.

ORJI, R.; VASSILEVA, J.; MANDRYK, R. L. Modeling the efficacy of persuasive strategies for different gamer types in serious games for health. **User Modeling and User-Adapted Interaction**, Springer, v. 24, n. 5, p. 453–498, 2014. Citations on pages 19 and 20.

PALOMINO, P. T.; TODA, A. M.; OLIVEIRA, W.; RODRIGUES, L.; ISOTANI, S. Gamification journey: A novel approach for classifying gamer types for gamified educational systems. **Proceedings of SBGames**, p. 165–173, 2019. Citations on pages 19 and 51.

PEDRO, L. Z. **Uso de gamificação em ambientes virtuais de aprendizagem para reduzir o problema da externalização de comportamentos indesejáveis**. Phd Thesis (PhD Thesis) — Universidade de São Paulo, 2016. Citation on page 35.

RODRIGUES, L.; BRANCHER, J. D. Improving players' profiles clustering from game data through feature extraction. In: **Proceedings of SBGames 2018 - Computing Track**. [S.l.]: IEEE, 2018. p. 177–186. Citation on page 19.

____. Playing an educational game featuring procedural content generation: Which attributes impact players' curiosity? **Revista Novas Tecnologias (RENOTE)**, v. 17, n. 1, p. 254–263, 2019. Citation on page 19.

RODRIGUES, L.; OLIVEIRA, W.; TODA, A.; PALOMINO, P.; ISOTANI, S. Thinking inside the box: How to tailor gamified educational systems based on learning activities types. In: **Proceedings of the Brazilian Symposium of Computers on Education**. [S.l.]: SBC, 2019. p. 823–832. Citations on pages 22, 25, 41, and 45.

RODRIGUES, L.; PALOMINO, P. T.; M.TODA, A.; KLOCK, A. C. T.; PESSOA, M.; D.PEREIRA, F.; OLIVEIRA, E. H. T.; F.OLIVEIRA, D.; CRISTEA, A. I.; GASPARINI, I.; ISOTANI, S. How personalization affects motivation in gamified review assessments. **International Journal of Artificial Intelligence in Education (under review)**, OSF, Jun 2022. Available: <osf.io/ehm43>. Citations on pages 40, 48, 50, 52, and 60.

RODRIGUES, L.; PALOMINO, P. T.; TODA, A. M.; KLOCK, A. C.; OLIVEIRA, W.; AVILA-SANTOS, A. P.; GASPARINI, I.; ISOTANI, S. Personalization improves gamification: evidence from a mixed-methods study. **Proceedings of the ACM on Human-Computer Interaction**, ACM New York, NY, USA, v. 5, n. CHI PLAY, p. 1–25, 2021. Citations on pages 47, 49, and 50.

RODRIGUES, L.; PEREIRA, F.; TODA, A.; PALOMINO, P.; OLIVEIRA, W.; PESSOA, M.; CARVALHO, L.; OLIVEIRA, D.; OLIVEIRA, E.; CRISTEA, A. *et al.* Are they learning or playing? moderator conditions of gamification's success in programming classrooms. **ACM Transactions on Computing Education (TOCE)**, ACM New York, NY, 2022. Citations on pages 33, 36, and 40.

RODRIGUES, L.; PEREIRA, F. D.; TODA, A. M.; PALOMINO, P. T.; PESSOA, M.; CARVALHO, L. S. G.; FERNANDES, D.; OLIVEIRA, E. H.; CRISTEA, A. I.; ISOTANI, S. Gamification suffers from the novelty effect but benefits from the familiarization effect: Findings from a longitudinal study. **International Journal of Educational Technology in Higher Education**, SpringerOpen, v. 19, n. 1, p. 1–25, 2022. Citations on pages 33 and 38.

RODRIGUES, L.; TODA, A.; PEREIRA, F.; PALOMINO, P. T.; KLOCK, A. C.; PESSOA, M.; OLIVEIRA, D.; GASPARINI, I.; TEIXEIRA, E. H.; CRISTEA, A. I. *et al.* Garfield: A recommender system to personalize gamified learning. In: SPRINGER. **International Conference on Artificial Intelligence in Education**. [S.l.], 2022. p. 666–672. Citation on page 51.

RODRIGUES, L.; TODA, A. M.; OLIVEIRA, W.; PALOMINO, P. T.; ISOTANI, S. Just beat it: Exploring the influences of competition and task-related factors in gamified learning environments. In: SBC. **Anais do XXXI Simpósio Brasileiro de Informática na Educação**. [S.l.], 2020. p. 461–470. Citations on pages 33 and 40.

RODRIGUES, L.; TODA, A. M.; OLIVEIRA, W.; PALOMINO, P. T.; AVILA-SANTOS, A. P.; ISOTANI, S. Gamification works, but how and to whom? an experimental study in the context of programming lessons. In: **Proceedings of the 52nd ACM Technical Symposium on Computer Science Education**. [S.l.: s.n.], 2021. p. 184–190. Citations on pages 33, 34, and 51.

RODRIGUES, L.; TODA, A. M.; PALOMINO, P. T.; OLIVEIRA, W.; ISOTANI, S. Personalized gamification: A literature review of outcomes, experiments, and approaches. In: **Eighth International Conference on Technological Ecosystems for Enhancing Multiculturality**. [S.l.: s.n.], 2020. p. 699–706. Citations on pages 31, 41, 43, and 51.

RODRIGUES, L.; TODA, A. M.; SANTOS, W. O. dos; PALOMINO, P. T.; VASSILEVA, J.; ISOTANI, S. Automating gamification personalization to the user and beyond. **IEEE Transactions on Learning Technologies**, IEEE, 2022. Citations on pages 41, 43, 45, and 53.

ROOSTA, F.; TAGHIYAREH, F.; MOSHARRAF, M. Personalization of gamification-elements in an e-learning environment based on learners' motivation. In: IEEE. **2016 8th International Symposium on Telecommunications (IST)**. [S.l.], 2016. p. 637–642. Citations on pages 20, 22, and 25.

ROWLAND, C. A. The effect of testing versus restudy on retention: a meta-analytic review of the testing effect. **Psychological Bulletin**, American Psychological Association, v. 140, n. 6, p. 1432, 2014. Citation on page 36.

ROY, R. V.; ZAMAN, B. Need-supporting gamification in education: An assessment of motivational effects over time. **Computers & Education**, Elsevier, v. 127, p. 283–297, 2018. Citations on pages 25, 28, and 35.

ROY, R. van; ZAMAN, B. Why gamification fails in education and how to make it successful: introducing nine gamification heuristics based on self-determination theory. In: **Serious Games and edutainment applications**. [S.l.]: Springer, 2017. p. 485–509. Citation on page 35.

RUSSELL, S. J. **Artificial intelligence a modern approach**. [S.l.]: Pearson Education, Inc., 2010. Citation on page 34.

SAILER, M.; HOMNER, L. The gamification of learning: a meta-analysis. **Educ Psychol Rev**, v. 32, p. 77–112, 2020. ISSN 1573-336X. Available: <https://doi.org/10.1007/s10648-019-09498-w>. Citations on pages 19, 25, 28, 33, 37, and 52.

SANCHEZ, D. R.; LANGER, M.; KAUR, R. Gamification in the classroom: Examining the impact of gamified quizzes on student learning. **Computers & Education**, v. 144, p. 103666, 2020. ISSN 0360-1315. Available: <http://www.sciencedirect.com/science/article/pii/S0360131519302192>. Citations on pages 35 and 36.

SAVARD, I.; MIZOGUCHI, R. Context or culture: what is the difference? **Research and Practice in Technology Enhanced Learning**, SpringerOpen, v. 14, n. 1, p. 1–12, 2019. Citations on pages 23, 34, and 35.

STUART, H.; LAVOUÉ, E.; SERNA, A. To tailor or not to tailor gamification? an analysis of the impact of tailored game elements on learners' behaviours and motivation. In: **21th International Conference on Artificial Intelligence in Education**. [S.l.: s.n.], 2020. Citations on pages 47 and 53.

TODA, A.; OLIVEIRA, W.; SHI, L.; BITTENCOURT, I.; ISOTANI, S.; CRISTEA, A. Planning gamification strategies based on user characteristics and dm: A gender-based case study. In: **12th International Conference on Educational Data Mining**. [S.l.: s.n.], 2019. p. 438–443. Citations on pages 22 and 38.

TODA, A. M.; CARMO, R. M. do; SILVA, A. P. da; BITTENCOURT, I. I.; ISOTANI, S. An approach for planning and deploying gamification concepts with social networks within educational contexts. **International Journal of Information Management**, v. 46, p. 294 – 303, 2019. ISSN 0268-4012. Available: <http://www.sciencedirect.com/science/article/pii/S0268401218304614>. Citation on page 19.

TODA, A. M.; KLOCK, A. C.; OLIVEIRA, W.; PALOMINO, P. T.; RODRIGUES, L.; SHI, L.; BITTENCOURT, I.; GASPARINI, I.; ISOTANI, S.; CRISTEA, A. I. Analysing gamification elements in educational environments using an existing gamification taxonomy. **Smart Learning Environments**, Springer, v. 6, n. 1, p. 16, 2019. Citations on pages 34 and 36.

TODA, A. M.; VALLE, P. H. D.; ISOTANI, S. The dark side of gamification: An overview of negative effects of gamification in education. In: CRISTEA, A. I.; BITTENCOURT, I. I.; LIMA, F. (Ed.). **Higher Education for All. From Challenges to Novel Technology-Enhanced Solutions**. Cham: Springer International Publishing, 2018. p. 143–156. Citations on pages 19 and 28.

TOETENEL, L.; RIENTIES, B. Analysing 157 learning designs using learning analytic approaches as a means to evaluate the impact of pedagogical decision making. **British Journal of Educational Technology**, Wiley Online Library, v. 47, n. 5, p. 981–992, 2016. Citations on pages 41 and 42.

TONDELLO, G. F. **Dynamic Personalization of Gameful Interactive Systems**. Phd Thesis (PhD Thesis) — University of Waterloo, 2019. Citation on page 21.

TONDELLO, G. F.; MORA, A.; NACKE, L. E. Elements of gameful design emerging from user preferences. In: ACM. **Proceedings of the Annual Symposium on Computer-Human Interaction in Play**. [S.l.], 2017. p. 129–142. Citations on pages 20, 25, 43, and 44.

TONDELLO, G. F.; ORJI, R.; NACKE, L. E. Recommender systems for personalized gamification. In: ACM. **Adjunct Publication of the 25th Conference on User Modeling, Adaptation and Personalization**. [S.l.], 2017. p. 425–430. Citations on pages 21 and 22.

TONDELLO, G. F.; WEHBE, R. R.; DIAMOND, L.; BUSCH, M.; MARCZEWSKI, A.; NACKE, L. E. The gamification user types hexad scale. In: ACM. **Proceedings of the 2016 annual symposium on computer-human interaction in play**. [S.l.], 2016. p. 229–243. Citations on pages 22, 25, and 43.

VANSTEENKISTE, M.; SIERENS, E.; SOENENS, B.; LUYCKX, K.; LENS, W. Motivational profiles from a self-determination perspective: The quality of motivation matters. **Journal of educational psychology**, American Psychological Association, v. 101, n. 3, p. 671, 2009. Citation on page 19.

VIERA, A. J.; GARRETT, J. M. *et al.* Understanding interobserver agreement: the kappa statistic. **Fam med**, v. 37, n. 5, p. 360–363, 2005. Citation on page 53.

VORNHAGEN, J. B.; TYACK, A.; MEKLER, E. D. Statistical significance testing at chi play: Challenges and opportunities for more transparency. In: **Proceedings of the Annual Symposium on Computer-Human Interaction in Play**. [S.l.: s.n.], 2020. p. 4–18. Citations on pages 38 and 39.

WANGENHEIM, C. G. von; WANGENHEIM, A. von. Ensinando computação com jogos. **Bookess Editora, Florianópolis, SC, Brasil**, 2012. Citation on page 36.

WILCOX, R. **Introduction to robust estimation and hypothesis testing**. 4th. ed. : Elsevier, 2017. Citation on page 38.

WIRTH, R.; HIPP, J. Crisp-dm: Towards a standard process model for data mining. In: SPRINGER-VERLAG LONDON, UK. **Proceedings of the 4th international conference on the practical applications of knowledge discovery and data mining**. [S.l.], 2000. v. 1. Citation on page 51.

WOHLIN, C.; RUNESON, P.; HST, M.; OHLSSON, M. C.; REGNELL, B.; WESSLN, A. **Experimentation in Software Engineering**. [S.l.]: Springer Publishing Company, Incorporated, 2012. ISBN 3642290434, 9783642290435. Citation on page 34.

WOOD, R. E. Task complexity: Definition of the construct. **Organizational behavior and human decision processes**, Elsevier, v. 37, n. 1, p. 60–82, 1986. Citation on page 24.

YEE, N. **Gaming Motivations Align with Personality Traits**. 2016. Accessed in January 2020. Available: <quanticfoundry.com/2016/01/05/personality-correlates/>. Citation on page 19.

YUEN, K. K. The two-sample trimmed t for unequal population variances. **Biometrika**, Oxford University Press, v. 61, n. 1, p. 165–170, 1974. Citation on page 38.

ZAINUDDIN, Z.; CHU, S. K. W.; SHUJAHAT, M.; PERERA, C. J. The impact of gamification on learning and instruction: A systematic review of empirical evidence. **Educational Research Review**, Elsevier, p. 100326, 2020. Citation on page 21.

ANNEX

A

PERSONALIZED GAMIFICATION: A LITERATURE REVIEW OF OUTCOMES, EXPERIMENTS, AND APPROACHES

Personalized gamification: A literature review of outcomes, experiments, and approaches

Luiz Rodrigues
Institute of Mathematics and Computer Science, University of São Paulo, São Carlos, São Paulo, Brazil
lalrodrigues@usp.br

Armando M. Toda
Institute of Mathematics and Computer Science, University of São Paulo, São Carlos, São Paulo, Brazil
armando.toda@usp.br

Paula T. Palomino
Institute of Mathematics and Computer Science, University of São Paulo, São Carlos, São Paulo, Brazil
paulatpalomino@usp.br

Wilk Oliveira
Institute of Mathematics and Computer Science, University of São Paulo, São Carlos, São Paulo, Brazil
wilk.oliveira@usp.br

Seiji Isotani
Institute of Mathematics and Computer Science, University of São Paulo, São Carlos, São Paulo, Brazil
sisotani@icmc.usp.br

ABSTRACT

Personalized gamification has gained substantial interest due to the expectation that it can improve gamification's success. Considering some secondary studies on this topic, they lack to present the characteristics of empirical studies and some aspects on how personalization approaches were designed. In this paper, we present a literature review based on previous research to address these gaps. Based on our analysis, our results provide: insights on how experiments to compare personalized gamification and non-personalized gamification are designed and evaluated; evidence on the effectiveness of personalized gamification found in primary studies; and an overview of how personalization approaches were designed. Our analysis converged in possible guidelines and a research agenda revealing five main needs: i) empirical studies comparing one size fits all and personalized gamification; ii) qualitative user studies; iii) personalization approaches that consider contextual characteristics as well as iv) rely on a broader, unambiguous set of game elements; and v) a benchmark of established resources to increase research reproducibility.

CCS CONCEPTS

• **Applied computing** → Education; Interactive learning environments; • **Human-centered computing** → Human computer interaction (HCI); Interaction techniques.

KEYWORDS

Gamification, personalization, tailoring, literature review

ACM Reference Format:

Luiz Rodrigues, Armando M. Toda, Paula T. Palomino, Wilk Oliveira, and Seiji Isotani. 2020. Personalized gamification: A literature review of outcomes, experiments, and approaches. In *Eighth International Conference on Technological Ecosystems for Enhancing Multiculturality (TEEM'20), October 21–23, 2020, Salamanca, Spain.* ACM, New York, NY, USA, 8 pages. https://doi.org/10.1145/3434780.3436665

TEEM'20, October 21–23, 2020, Salamanca, Spain

ACM ISBN 978-1-4503-8850-4/20/10…$15.00
https://doi.org/10.1145/3434780.3436665

1 INTRODUCTION

It has been argued that, for gamified applications, one size does not fit all [41]. Given that the same gamification design is unlikely to work for all users, researchers started to investigate personalized gamification. That is, tailoring game elements based on information about the users aiming to improve gamification's effectiveness [59]. This approach is based on the fact that people with different characteristics have different preferences, perceptions, and experiences [47, 51, 61]. Consequently, by offering gamification designs tailored to users' characteristics, providers expect to improve their experiences and, thus, gamification's success [59].

Due to such expectations, personalized gamification has substantially attracted researchers' attention. A significant number of studies have been published, which have been recently summarized in a few literature reviews (e.g., [1, 27, 28]). However, two points were not addressed in previous research. First, despite some literature reviews analyzed the results from empirical studies [1, 19], *they do not present a broad analysis of how personalized gamification's impact was assessed.* That is, aspects such as the control group that personalized gamification was compared to and the sample sizes, which would provide a valuable understanding of the validity of those experiments. Hence, although primary studies mostly report positive outcomes, as indicated in Hallifax et al. [19], whether using personalization improves one size fits all gamification remains unknown. To fill this gap, this paper investigates ***how empirical studies employing personalized gamification have been conducted to answer whether personalization improves one size fits all gamification effectiveness***.

Second, *some issues related to how personalization approaches were designed have not been discussed in previous research.* In this regard, none of the secondary studies discuss how the game elements at a given personalization approach were selected. This aspect is important because self-selection might lead the approach to ignore some game elements or use different names to represent elements with the same goal (e.g., both medals and trophies provide acknowledgment), leading to ambiguity in the experience they will provide to users. To illustrate such limitations, one might consider a personalization approach that uses game elements from the literature review by Nah et al. [42]. Although systematic, using this source would lead to an approach that does not consider the Collaboration game element, while featuring badges, prizes, and rewards, all of which have highly similar goals [58].

Additionally, the approach's development method is another relevant aspect that is discussed in a single related work. Klock et al. [27] discusses the algorithms used in that process (e.g., linear regression and factor analysis) as well as how those were evaluated, where survey and questionnaires were used most often. To provide a complementary point of view, we discuss whether those were developed based on theory or data. That is, whether they are theory- or data-driven. A clear understanding on which of those approaches has been used is important to shed light on development trends, especially considering the novelty of data-driven gamification, which has been advocated to enable the development of more tailored gamification designs than theory-driven alternatives [35]. To fill these gaps, this paper investigates ***how approaches for personalizing gamification have been developed, especially focusing on development choices, such as personalization criteria, the kind of source each approach relied on to select the game elements it uses, and whether it is theory- or data-driven.***

Therefore, our main objective is to conduct an exploratory study, through a literature review process, to expand the current body of knowledge on the impacts, approaches, and insights adopted to design personalized gamification systems. To achieve this goal, our study focuses on answering the following questions (RQ): (**RQ1**) *How empirical studies employing personalized gamification have been conducted?* (**RQ2**) *Does personalized gamification improve one size fits all gamification effectiveness?* and (**RQ3**) *How approaches for personalizing gamification have been developed?*. Thus, our main contributions are:

- Revealing the impact of personalized gamification;
- Showing an overview of how empirical studies have been conducted in this context; and
- Providing insights on how personalization approaches were developed.

Furthermore, we answer our RQ by extracting information from the studies included in previous literature reviews as they were recently published, a context in which conducting a new systematic process would be of little benefit [16]. Thereby, speeding up the process of communicating our findings while answering new research questions based on a state-of-the-art view.

2 METHOD

To achieve the objectives of this study, we opted to conduct an exploratory study since this kind of study is conducted to delve deep into an existing problem aiming to explicit it [32]. We choose to analyze existing works on the field that were found in previous literature reviews on the subject, since the studies were published recently [1, 19, 27, 28] and there is no need for updating the existing reviews [16]. Based on the exposed, we analyzed points that previous studies did not (i.e., how empirical studies have been conducted and how personalization approaches have been conducted).

To conduct this literature review, we defined two steps that are commonly encountered in other literature review processes, such as Kitchenham [26] and PRISMA checklist [64]: Studies selection and Data extraction. The first encompasses the process of selecting our studies based on predefined criteria, while the second step consists in defining and extracting information to answer our research questions.

2.1 Studies Selection

Figure 1 summarizes the study selection process. For study selection, our main criterion was secondary studies focused on personalized gamification. That is, we relied on recent secondary studies rather than conducting another systematic literature review. The secondary studies on personalized gamification that we screened have been published between three months and two years before the time of writing, covering state-of-the-art research on this topic. In addition, those have been published in varied venues, ranging from general human-computer studies [27] to the perspective of learning technologies [19]. Given that context, conducting a new systematic process would be of little benefit [16]. Thereby, our approach speeds up the process of communicating our findings while answering new research questions based on a state-of-the-art view.

Based on the secondary studies' results, our selection criteria for primary studies were that they should either i) present an empirical study employing personalized gamification or ii) present an approach for personalizing gamification; that is, providing recommendations/guidance on how to tailor gamification designs to some specific criteria. Nevertheless, we also included other primary studies through snowballing, by verifying which studies cited the analyzed study, allowing us to collect some recent studies that were recently published and, therefore, could not be added in previous literature reviews. Thus, primary studies included in this paper are those meeting the aforementioned inclusion criteria and either found during our literature searches or included in the following secondary studies, which were selected in an ad-hoc manner: [1, 19, 27, 28].

2.2 Data Extraction

We extracted data from the included studies following the categorization presented before (i.e., empirical studies and recommendations). Figure 2 summarizes data extracted from each study of each category, which are further described next.

In reviewing recommendations, we focus on four main characteristics. The *first* characteristic is the personalization criteria, which we further split in two: user (e.g., demographics) and context (e.g., gamified task). This is important to demonstrate if approaches are exploring all kinds of characteristics known as relevant for gamification success.

The *second* is how the recommendation was developed, which might be theory- (e.g., linking game elements to some characteristics based on theories behind those) or data-driven (e.g., exploring data from users' behavior, opinions, or preferences to determine

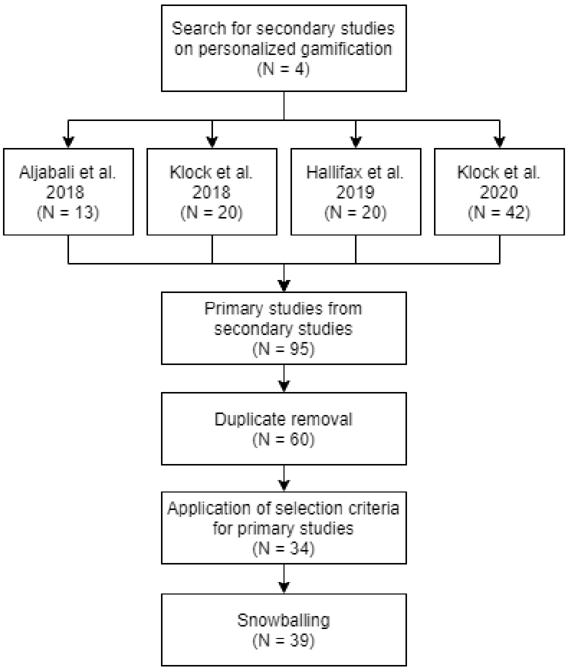

Figure 1: Study selection process. N represents the number of studies selected after each step.

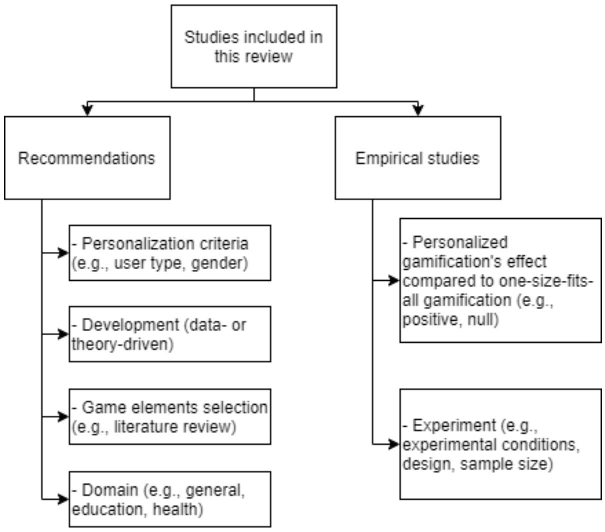

Figure 2: Data extracted from selected studies depending on their type. Recommendations provide guidance on how to personalize gamification whereas empirical studies perform user studies applying personalized gamification.

the recommendations). This is important to understand which development approach has been used more or less.

Third, we consider how the recommendation's game elements were selected (e.g., from a literature review or some taxonomy). The relevance of this choice is avoiding selection bias as well as preventing the recommendation of ambiguous game elements.

The *fourth* characteristic is to which domain the recommendation was built for (e.g., education or general; that is, no specific domain), which is important as the domain is another aspect that needs to be considered when designing gamified systems.

In reviewing empirical studies on the use of personalized gamification, we follow two main perspectives. One is analyzing the effect of personalized gamification overall, as well as that of different personalization approaches. To that end, we extracted the kind of outcome analyzed and whether personalization was positive, null, mixed, or negative for each kind. The other perspective concerns understanding how such experiments have been conducted, aiming to identify best practices and perspectives needing to be tackled. Then, we extracted the condition personalization was compared to, the personalization criteria, the number of game elements in each condition, the intervention duration, the sample size, the data analysis form, and the study context.

3 RESULTS AND DISCUSSION

This section presents and discusses the results of each RQ.

3.1 RQ1: How empirical studies employing personalized gamification have been conducted

Table 1 summarizes the empirical studies included in this review[1]. A key point to understanding those studies is the condition to which personalized gamification was compared. Frequently, the baseline comparison was random gamification; that is, randomly selected game elements [38, 53, 56]. There were also comparisons to no gamification [23, 25, 31, 38]. Additionally, there were cases in which comparisons were made to counter-tailored gamification; that is, when users receive the game element they (are expected to) prefer the less [31, 36, 38]. The main problem of such approaches is that personalized gamification emerged as a means to improve one size fits all gamification's effectiveness, and comparing it to random, counter-tailored, or no gamification does not contribute to understanding whether that objective was met.

On the other hand, some studies indeed compared personalized to one size fits all gamification. Hassan et al. [21] is one of these works, however, their experimental condition featured adaptive learning as well as gamification. Similarly, the personalization approach in Daghestani et al. [9] influenced the content's difficulty as well as the help (e.g., guidance on further readings) users received. Such approaches limit the understanding of whether personalized gamification was the source of the results as it was not experimentally controlled. Differently, three studies controlled personalized gamification experimentally, comparing it to one size fits all gamification in terms of students' flow experience [44], undergraduate learners' motivation and behavior [39], and social network users' behavior [17]. Note that there is a condition studied in Monterrat et al. [38] that might be considered one size fits all: providing all three

[1] [56] appears three times because authors compared three personalization

Table 1: Overview of empirical studies included in this review.

Ref	Compared to	Personalization Criteria	# GE	Intervention Duration	Sample size	Measured Outcome	Data analysis	Context
[9]	OSFA	Gamer type*	21/2-2	?	76	M~, C+	Qt	Undergraduate class
[44]	OSFA	Gamer type	9/?	<1 day	121	M~	Qt	Elementary students
[17]	OSFA	Likes	3/2-3	2 months	2102	B+/-	Qt	Social network users
[39]	OSFA	User type	?/2-3	14 weeks	81	B+/-, M+/-	Qt	Undergrad online class
[21]	OSFA	Learning style*	5/2-5	?	175	B+, M+	Qt	Undergrad class
[38]	OSFA, NG, CT, R	Gamer type	0-3/1	<1 day	59	B+/-, M+/-	Qt	Middle school
[23]	NG, Cmp, Col	Performance	0-1/2	<1 day	54	B+/-	Mx	Primary class
[53]	R	Motivation style	?/2-3	1 month	100	B+	Qt	Undergrad class
[56]	R	User type	1/1	6 weeks	258	B+/-, M~	Qt	Daily children
[56]	R	IM	1/1	6 weeks	258	B~, M+/-	Qt	Daily children
[56]	R	User type and IM	1/1	6 weeks	258	B+/-, M+/-	Qt	Daily children
[36]	CT	Gamer Type	2/2	3 weeks	280	B+, M+/-	Qt	Daily adults
[31]	NG, CT	Gamer type	2/2	3 weeks	266	B+/-, M+/-	Qt	Daily adults
[25]	NG	Age*	1/1	<1 day	40	B+/-	Qt	Primary class

* Personalized gamification was not controlled experimentally
#GE = Number of game elements in personalized/baseline condition; OSFA = one size fits all; NG = no gamification; Cmp = competitive gamification; Col = collaborative gamification; CT = counter-tailored gamification; R = random gamification; IM = Initial motivation; ? = information undefined in the study; M = motivational; C = cognitive; B = behavioral; + = positive; +/- = mixed; ~= null; Qt = quantitative; Mx = mixed.

game elements deliberately selected during the study. However, authors do not compare this condition to the one with tailored gamification.

Concerning the empirical setups, most studies experimented with few elements in both the personalized and the baseline conditions (i.e., around one and three game elements in each), with a few exceptions [9, 21, 44]. Furthermore, few interventions lasted less than a day [23, 25, 38, 44], while the majority lasted between three and eight weeks (or two months), most sample sizes ranged from 40 to 280, with a case of studying a social network that featured over 2000 participants [17], and the application contexts were varied, including primary, elementary, middle, and undergraduate classes, social networks, and adults and children daily tasks. This overview demonstrates empirical studies were conducted in varied settings, mostly using few game elements in both conditions and with sample sizes and time frames (interventions duration) comparable to overall gamification studies [2, 30, 54].

Another point is that all but one study employed a quantitative data analysis approach. As personalized gamification is a recent field [27, 62], performing qualitative analyses could substantially contribute to the field by shedding light on users' perceptions of personalized gamification designs. Those insights are likely to help designers to better understand how to personalize as well as reasons for unexpected findings.

Concerning the personalization criteria, all studies considered either user's characteristics, such as Brainhex gamer types [40] (e.g., [9, 44]) and Hexad user types [34, 63] (e.g., [39, 56]), or interactions (e.g., [17]). Hence, demonstrating a predominant focus on user characteristics. This predominance reveals a gap in studying the impact of personalization approaches exploring contextual criteria (e.g., [3, 5]), despite the literature has acknowledged and recommended research should explore this vein [19, 30, 33, 52].

3.2 RQ2: Effectiveness of Personalized gamification compared to one size fits all gamification

Primary studies suggest personalized gamification's effects are mixed. Overall, outcomes are mixed in terms of positive with null (e.g., [9, 25, 39]), although in some cases both positive and negative results have been reported (e.g., [17, 38]). In other cases, the reported findings were overall null [44, 56] or positive [21]. As shown in Table 1, these mixed findings appear to hold regardless of analyzing motivational, behavioral, or cognitive outcomes. Nevertheless, the baseline comparison for those findings was varied. Therefore, one can only assure that, for instance, personalized gamification is more effective than counter-tailored or random gamification. However, in seeking to understand how personalized gamification compares to one size fits all gamification, these results must be analyzed with caution. As discussed before, only three studies compared personalized and one size fits all gamification through a proper experimental setting. In this context, Oliveira et al. [44] found personalization did not affect elementary students' flow experience. Mora et al. [39] found positive but statistically insignificant results from personalization use based on undergraduate learners' motivation and behavior. Outside the educational context, Hajarian et al. [17] found both positive (e.g., in time-in-system) and negative (e.g., in gamification usage) results for their like-based personalization approach for social networks. Thereby, demonstrating that the few studies that contribute to understanding whether personalized gamification improves one size fits all gamification show inconclusive results. Thus, at the current point, the literature has insufficient evidence to confirm if available approaches for personalized gamification improves the general one size fits all approach.

3.3 RQ3: How approaches for personalizing gamification have been developed

Table 2 summarizes research on recommendations for personalizing gamification included in this review. Predominantly, these recommendations rely on user data criteria. More specifically, these criteria are user typologies [14, 15, 18, 22, 29, 36, 43, 47, 57, 60, 61, 63], personality [4, 6, 8, 10, 12, 15, 24, 46], demographics [11, 45, 49], and in-system behavior [7]. In contrast, the number of recommendations that consider contextual factors (i.e., four; [3, 5, 48, 50]) is substantially smaller compared to those considering some user aspect. Hence, demonstrating recommendations for personalizing gamification rarely consider contextual factors, which might explain the lack of empirical studies involving this kind of criteria.

Concerning recommendations' development, data-driven approaches were most often adopted than the theory-driven ones. On one hand, recommendations build from theoretical concerns are commonly inspired by definitions behind criteria and game elements' definitions, using those to link one to another (e.g., users with specific player type are more likely to enjoy specific game elements; [3, 6, 14, 15, 22, 37, 57]). On the other hand, data-driven development procedures are mostly based on surveys, asking users to indicate their preferences based on game elements definitions, storyboards, and prototypes (e.g., [5, 10–12, 18, 47, 60, 61]), as well as exploring users' interactions to implicitly identify which are likely to be the best game elements for them (e.g., [4, 7, 29]). There is also the recommendation by [63], which relies on both perspectives; that is, exploring both user data and theoretical foundations. This context demonstrates most approaches have been developed based on data, such as user preference. Thereby, showing that personalization approaches are mainly embracing the recent perspective of data-driven gamification [35].

In terms of the game elements selection, studies explored games literature [3, 15, 29], gamification literature [5, 7, 22, 60, 63], informal [4, 10–12, 14, 18, 61] and systematic [24, 43] literature reviews, and, in other cases, deliberately made self-selections [6, 8, 36, 57]. In addition, there were cases in which the recommendations were to personalize persuasive [45–47] and social influence strategies [48–50]. This context suggests most recommendations are built upon a set of game elements that was deliberately selected (e.g., self-selecting elements from games literature) or from literature reviews that summarize game elements used in, for instance, previous gamification research. A limitation is that these summaries are prone to featuring ambiguous game elements due to the lack of validation (e.g., experts assessing similarities) and might fail to consider less common ones because they are rarely studied. Thus, available recommendations likely will suffer from similar limitations.

Lastly, recommendations to only two specific domains were found. These are health [45, 46] and education [3–5, 7, 8, 10–12, 14, 22, 29, 43, 57]. The remaining ones did not target specific domains [6, 15, 18, 24, 36, 47–50, 60, 61, 63]. Although the general view from these recommendations is valuable to allow them to be used in any domain, the lack of specificity might be seen as limiting, though. It has been argued that gamification should consider not only the user but the context and the domain (e.g., [33]). Accordingly, recommendations for personalizing gamification that focus on a specific domain are expected to be more beneficial to that domain, compared to generic ones. From this perspective, the health and education domains are one step ahead as recommendations for them have been developed, while other domains would have to rely on the general approaches.

4 RESEARCH AGENDA

This section delineates five lines of research yet to be addressed based on our findings.

Comparing personalized and one size fits all gamification: Perhaps the most important step towards advancing the personalized gamification field is conducting empirical studies comparing personalized and one size fits all gamification. As the main goal of the former is to improve the effectiveness of the latter, such comparisons are of utmost importance. However, as we demonstrated, most studies compare personalized gamification to other conditions (e.g., counter-tailored and random gamification). Therefore, future studies should perform such comparative analyses to reveal which personalization approaches achieve the goal of personalized gamification. The planning of those studies can follow the characteristics revealed in this review (e.g., the number of game elements per condition and sample size).

Developing personalization approaches that consider contextual characteristics: Several researchers agree that gamification's success depends not only on user characteristics but also on contextual factors (e.g., [13, 18, 20, 33]). Accordingly, personalization approaches should also consider contextual aspects as personalization criteria. However, as we have shown, few approaches explore those characteristics. Therefore, we call for future research to expand beyond the use of user characteristics as personalization criteria, exploring contextual factors as well (e.g., culture). In doing so, the work by Savard and Mizoguchi [55] might provide valuable guidance in understanding and operationalizing the context.

Performing qualitative user studies: Personalized gamification is a recent field study, in which a deeper understanding of users' experiences would certainly contribute to shedding light on positive and negative aspects of personalization approaches available, as well as help to achieve a better understanding of the effectiveness of subjective, rarely used game elements such as storytelling and narrative. Performing qualitative analyses (e.g., interviews) is a way to gather such knowledge, however, our findings reveal that studies are predominantly based on quantitative data. Therefore, we urge for research capturing qualitative feedback to reveal users' subjective experiences with personalized gamification and, consequently, insights into how to improve personalization approaches. Such studies could be performed similarly to the quantitative studies analyzed in this paper. The difference would be in terms of data collection, though, capturing participants' feedback through structured interviews, focal groups, and/or observation. Consequently, it will be possible to analyze personalized gamification's impact based on the context of use, through subjects' emotions/perceptions, which is unfeasible through common quantitative approaches, such as questionnaires/scales and data log [32].

Building personalization approaches from validated taxonomies of game elements: When designing gamification, using well-defined game elements that provide the expected affordances is necessary. Accordingly, one expects that personalization approaches will be able to recommend game elements with the

Table 2: Recommendations/guidelines for personalizing gamification included in this review.

Ref	Game elements selection	User Criteria	Context Criteria	Development	Domain
[47]	Persuasive strategies	User type	None	Data-driven	General
[60]	Marczewski work	User type, personality trait, age, gender	None	Data-driven	General
[61]	Informal literature review	User types, personality traits, age, and gender	None	Data-driven	General
[18]	Informal literature review	User type, gamer type, personality trait	None	Data-driven	General
[63]	Marczewski work	User type, personality traits	None	Mixed	General
[22]	Marczewski work	User type	None	Theory-driven	Education
[8]	Self-selected	Personality trait	None	Data-driven	Education
[10]	Informal literature review	Personality trait	None	Data-driven	Education
[12]	Informal literature review	Personality trait	None	Data-driven	Education
[24]	Gamification literature review	Personality trait	None	Data-driven	General
[6]	Self-selected	Personality trait	None	Theory-driven	General
[46]	Persuasive strategies	Personality trait	None	Data-driven	Health
[15]	Games literature	Bartle's player type and personality trait	None	Theory-driven	General
[14]	Informal literature review	Bartle's profile	None	Theory-driven	Education
[57]	Self-selected	Bartle's profile	None	Theory-driven	Education
[29]	Games literature	Bartle's profile	None	Data-driven	Education
[36]	Self-selected	Gamer type	None	Theory-driven	General
[43]	Gamification literature review	Gamer type	None	Data-driven	Education
[45]	Persuasive strategies	Gender	None	Data-driven	Health
[11]	Informal literature review	Age, Gender, gaming frequency	None	Data-driven	Education
[49]	Social Influence Strategies	Age, Gender	None	Data-driven	General
[4]	Informal literature review	Player role	None	Theory-driven	Education
[7]	Gamification literature	Motivational and gameplay strategy	None	Theory-driven	Education
[48]	Social Influence Strategies	Gender, Age	Culture	Data-driven	General
[50]	Social Influence Strategies	None	Culture	Data-driven	General
[3]	Games literature	Personality traits and learning style	LAT	Theory-driven	Education
[5]	Gamification literature	User type	Moodle's LA	Data-driven	Education

purposes that best suit a user/circumstance (e.g., providing performance feedback, immersing the user in a ludic experience [58]) while preventing the suggestion of two or more of those that feature the same goal (unambiguity). As we found, most personalization approaches recommend game elements selected deliberately or from systematic studies, which makes them prone to such limitations as some elements might not be included or different names might be used for the same end. Therefore, we call for future research to develop personalization approaches upon validated taxonomies of game elements aiming to ensure a broader set of unambiguous game elements.

Providing and establishing resources to be reused in future empirical studies. To increase research reproducibility, the field study would benefit from a toolkit featuring resources such as systems and validated questionnaires to enable the execution of empirical studies in more similar conditions. For instance, a system that can be used to experiment with personalized gamification designs is presented by Tondello [59]. Therefore, we call for future research to develop, disclose, and discuss such resources towards establishing a benchmark for empirical studies on personalized gamification.

5 CONCLUSIONS AND LIMITATIONS

Personalized gamification emerged as a means to improve the effectiveness of one size fits all gamification. It has attracted the attention of several researchers, and much research has been conducted to understand how to personalize as well as to reveal its impact. In this paper, we presented a literature review that answered three RQ regarding personalized gamification.

We answered our first RQ by founding that most studies compare personalized gamification to counter-tailored and random gamification, use few game elements in both the experimental and baseline conditions, analyze quantitative data collected in classrooms and *in the wild*, and rely on sample sizes and time frames comparable to general gamification studies. For the second RQ, we found that the existing literature provides inconclusive evidence on how personalized and one size fits all gamification compare as few studies performed such comparison. In the third RQ, we found that approaches were mainly developed based on data (i.e., data-driven) rather than theory, using literature reviews and deliberate choices for selecting the game elements to consider, predominantly focusing on user characteristics and with few approaches focused on specific domains.

By answering those RQ, this study helps to fill up some previously cited gaps by providing a broader understanding of how empirical studies employing personalized gamification have been conducted, as well as the approach's impact, along with an overview of how personalization approaches have been developed. Thus, contributing by revealing the impact of personalized gamification, showing an overview of how empirical studies have been conducted in this context, and providing insights on how personalization approaches were developed.

The implications from our contribution are twofold. First, it provides a general overview of different approaches for personalizing gamification. Hence, one can use this overview as a starting point towards understanding how gamification has been personalized. Second, we provide guidelines of how future empirical research on this field might be conducted by showing previous studies' settings. Thereby, one can follow those guidelines to plan future research. Furthermore, we presented a research agenda with five gaps to be addressed in future studies, which emerged from our findings, guiding researchers on tracks needing attention.

Nevertheless, this paper has a main limitation that must be considered in interpreting our findings: the lack of following a systematic procedure for study selection. Studies were selected from a set of literature reviews published recently, reviews that were determined based on our described steps. While this jeopardizes future replications from this paper, it improves the chances of selecting a broader range of relevant primary studies as we exploited the selection process from four systematic studies. However, this does not exclude the limitations from the systematic procedures those studies followed, such as failing to find relevant studies due to string and search engine selection or not including a study due to interpretation or coding problems.

ACKNOWLEDGMENTS

The authors are thankful to the funding provided by CNPq, CAPES, and FAPESP (Projects 2016/02765-2; 2018/11180-3; 2018/15917-0; 2018/07688-1).

REFERENCES

[1] Rasha Najib Aljabali and Norasnita Ahmad. 2018. A Review on Adopting Personalized Gamified Experience in the Learning Context. In 2018 IEEE Conference on e-Learning, e-Management and e-Services (IC3e). IEEE, 61–66.
[2] Shurui Bai, Khe Foon Hew, and Biyun Huang. 2020. Is gamification "bullshit"? Evidence from a meta-analysis and synthesis of qualitative data in educational contexts. Educational Research Review (2020), 100322.
[3] Johan Baldeón, Inmaculada Rodríguez, and Anna Puig. 2016. LEGA: A LEarner-centered GAmification Design Framework. In Proceedings of the XVII International Conference on Human Computer Interaction (Salamanca, Spain) (Interacción '16). ACM, New York, NY, USA, Article 45, 8 pages. https://doi.org/10.1145/2998626.2998673
[4] Simone S Borges, Riichiro Mizoguchi, Vinicius HS Durelli, Ig I Bittencourt, and Seiji Isotani. 2016. A link between worlds: Towards a conceptual framework for bridging player and learner roles in gamified collaborative learning contexts. In Advances in Social Computing and Digital Education. Springer, 19–34.
[5] Klaudia Bovermann and Theo J Bastiaens. 2020. Towards a motivational design? Connecting gamification user types and online learning activities. Research and Practice in Technology Enhanced Learning 15, 1 (2020), 1–18. https://doi.org/10.1186/s41039-019-0121-4
[6] Charles Butler. 2014. A framework for evaluating the effectiveness of gamification techniques by personality type. In International Conference on HCI in Business. Springer, 381–389.
[7] Geiser Chalco Challco, Dilvan A Moreira, Riichiro Mizoguchi, and Seiji Isotani. 2014. An ontology engineering approach to gamify collaborative learning scenarios. In CYTED-RITOS International Workshop on Groupware. Springer, 185–198.
[8] David Codish and Gilad Ravid. 2014. Personality based gamification-Educational gamification for extroverts and introverts. In Proceedings of the 9th CHAIS Conference for the Study of Innovation and Learning Technologies: Learning in the Technological Era, Vol. 1. 36–44.
[9] Lamya F Daghestani, Lamiaa F Ibrahim, Reem S Al-Towirgi, and Hesham A Salman. 2020. Adapting gamified learning systems using educational data mining techniques. Computer Applications in Engineering Education 28, 3 (2020), 568–589.
[10] Mouna Denden, Ahmed Tlili, Fathi Essalmi, and Mohamed Jemni. 2017. Educational gamification based on personality. In 2017 IEEE/ACS 14th International Conference on Computer Systems and Applications (AICCSA). IEEE, 1399–1405.
[11] M. Denden, A. Tlili, F. Essalmi, and M. Jemni. 2017. An investigation of the factors affecting the perception of gamification and game elements. In 2017 6th International Conference on Information and Communication Technology and Accessibility (ICTA). IEEE, Muscat, Sultanate of Oman, 1–6. https://doi.org/10.1109/ICTA.2017.8336019
[12] Mouna Denden, Ahmed Tlili, Fathi Essalmi, and Mohamed Jemni. 2018. Does personality affect students' perceived preferences for game elements in gamified learning environments?. In 2018 IEEE 18th International Conference on Advanced Learning Technologies (ICALT). IEEE, 111–115.
[13] Christo Dichev and Darina Dicheva. 2017. Gamifying education: what is known, what is believed and what remains uncertain: a critical review. International journal of educational technology in higher education 14, 1 (2017), 9.
[14] Fernando Timoteo Fernandes and Plinio Thomaz Aquino Junior. 2016. Gamification aspects in the context of electronic government and education: A case study. In International Conference on HCI in Business, Government, and Organizations. Springer, 140–150.
[15] Lauren S Ferro, Steffen P Walz, and Stefan Greuter. 2013. Towards personalised, gamified systems: an investigation into game design, personality and player typologies. In Proceedings of The 9th Australasian Conference on Interactive Entertainment: Matters of Life and Death. ACM, 7.
[16] Paul Garner, Sally Hopewell, Jackie Chandler, Harriet MacLehose, Elie A Akl, Joseph Beyene, Stephanie Chang, Rachel Churchill, Karin Dearness, Gordon Guyatt, *et al.* 2016. When and how to update systematic reviews: consensus and checklist. bmj 354 (2016), i3507.
[17] Mohammad Hajarian, Azam Bastanfard, Javad Mohammadzadeh, and Madjid Khalilian. 2019. A personalized gamification method for increasing user engagement in social networks. Social Network Analysis and Mining 9, 1 (2019), 47.
[18] Stuart Hallifax, Serna Audrey, Marty Jean-Charles, Lavoué Guillaume, and Lavoué Elise. 2019. Factors to Consider for Tailored Gamification. In CHI Play. ACM, Barcelona, Spain, 559–572. https://hal.archivesouvertes. fr/hal-02185647
[19] Stuart Hallifax, Audrey Serna, Jean-Charles Marty, and Élise Lavoué. 2019. Adaptive Gamification in Education: A Literature Review of Current Trends and Developments. In Transforming Learning with Meaningful Technologies, Maren Scheffel, Julien Broisin, Viktoria Pammer-Schindler, Andri Ioannou, and Jan Schneider (Eds.). Springer International Publishing, Cham, 294–307.
[20] Juho Hamari, Jonna Koivisto, Harri Sarsa, *et al.* 2014. Does Gamification Work?- A Literature Review of Empirical Studies on Gamification.. In HICSS, Vol. 14. 3025–3034.
[21] Muhammad Awais Hassan, Ume Habiba, Fiaz Majeed, and Muhammad Shoaib. 2019. Adaptive gamification in e-learning based on students' learning styles. Interactive Learning Environments (2019), 1–21.
[22] Barryl Herbert, Darryl Charles, Adrian Moore, and Therese Charles. 2014. An investigation of gamification typologies for enhancing learner motivation. In 2014 International Conference on Interactive Technologies and Games. IEEE, 71–78.
[23] Tomislav Jagušt, Ivica Botički, and Hyo-Jeong So. 2018. Examining competitive, collaborative and adaptive gamification in young learners' math learning. Computers & education 125 (2018), 444–457.

[24] Yuan Jia, Bin Xu, Yamini Karanam, and Stephen Voida. 2016. Personality-targeted gamification: a survey study on personality traits and motivational affordances. In Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems. 2001–2013.

[25] Michael D Kickmeier-Rust, Eva-C Hillemann, and Dietrich Albert. 2014. Gamification and smart feedback: Experiences with a primary school level math app. International Journal of Game-Based Learning (IJGBL) 4, 3 (2014), 35–46.

[26] Barbara Kitchenham. 2004. Procedures for performing systematic reviews. Keele, UK, Keele University 33, 2004 (2004), 1–26.

[27] Ana Carolina Tomé Klock, Isabela Gasparini, Marcelo Soares Pimenta, and Juho Hamari. 2020. Tailored gamification: A review of literature. International Journal of Human-Computer Studies 144 (2020), 102495.

[28] Ana Carolina Tomé Klock, Marcelo Soares Pimenta, and Isabela Gasparini. 2018. A systematic mapping of the customization of game elements in gamified systems. Anais do Simpósio Brasileiro de Jogos e Entretenimento Digital (2018).

[29] Antti Knutas, Jouni Ikonen, Dario Maggiorini, Laura Ripamonti, and Jari Porras. 2016. Creating student interaction profiles for adaptive collaboration gamification design. International Journal of Human Capital and Information Technology Professionals (IJHCITP) 7, 3 (2016), 47–62.

[30] Jonna Koivisto and Juho Hamari. 2019. The rise of motivational information systems: A review of gamification research. International Journal of Information Management 45 (2019), 191–210. https://doi.org/10.1016/j.ijinfomgt.2018.10.013

[31] Elise Lavoué, Baptiste Monterrat, Michel Desmarais, and Sébastien George. 2018. Adaptive gamification for learning environments. IEEE Transactions on Learning Technologies 12, 1 (2018), 16–28.

[32] Jonathan Lazar, Jinjuan Heidi Feng, and Harry Hochheiser. 2017. Research methods in human-computer interaction. Morgan Kaufmann.

[33] De Liu, Radhika Santhanam, and Jane Webster. 2017. Toward Meaningful Engagement: A Framework for Design and Research of Gamified Information Systems. MIS quarterly 41, 4 (2017), 1011–1034.

[34] Andrzej Marczewski. 2018. Even Ninja Monkeys Like to Play: Unicorn Edition. Gamified UK (2018).

[35] Michael Meder, Till Plumbaum, and Sahin Albayrak. 2017. A Primer on Data-Driven Gamification Design. In Proceedings of the Data-Driven Gamification Design Workshop. CEUR-WS.org, 12–17.

[36] Baptiste Monterrat, Michel Desmarais, Elise Lavoué, and Sébastien George. 2015. A player model for adaptive gamification in learning environments. In International conference on artificial intelligence in education. Springer, 297–306.

[37] Baptiste Monterrat, Elise Lavoué, and Sébastien George. 2014. Toward an adaptive gamification system for learning environments. In International Conference on Computer Supported Education. Springer, 115–129.

[38] Baptiste Monterrat, Elise Lavoué, and Sébastien George. 2017. Adaptation of gaming features for motivating learners. Simulation & Gaming 48, 5 (2017), 625–656.

[39] A. Mora, G. F. Tondello, L. E. Nacke, and J. Arnedo-Moreno. 2018. Effect of personalized gameful design on student engagement. In 2018 IEEE Global Engineering Education Conference (EDUCON). 1925–1933. https://doi.org/10.1109/EDUCON.2018.8363471

[40] Lennart E Nacke, Chris Bateman, and Regan L Mandryk. 2014. BrainHex: A neurobiological gamer typology survey. Entertainment computing 5, 1 (2014), 55–62. https://doi.org/10.1016/j.entcom.2013.06.002

[41] Lennart E Nacke and Christoph Sebastian Deterding. 2017. The maturing of gamification research. Computers in Human Behaviour (2017), 450–454.

[42] Fiona Fui-Hoon Nah, Qing Zeng, Venkata Rajasekhar Telaprolu, Abhishek Padmanabhuni Ayyappa, and Brenda Eschenbrenner. 2014. Gamification of education: a review of literature. In International conference on hci in business. Springer, 401–409. https://doi.org/10.1007/978-3-319-07293-7_39

[43] Wilk Oliveira and Ig Ibert Bittencourt. 2019. Tailored Gamification to Educational Technologies. Springer Nature.

[44] Wilk Oliveira, Armando Toda, Paula Toledo, Lei Shi, Julita Vassileva, Ig Ibert Bittencourt, and Seiji Isotani. 2020. Does Tailoring Gamified Educational Systems Matter? The Impact on Students' FlowExperience. In Proceedings of the 53rd Hawaii International Conference on System Sciences. ScholarSpace, 1226–1235.

[45] Rita Orji. 2014. Exploring the Persuasiveness of Behavior Change Support Strategies and Possible Gender Differences.. In BCSS@PERSUASIVE. 41–57.

[46] Rita Orji, Lennart E Nacke, and Chrysanne Di Marco. 2017. Towards personality-driven persuasive health games and gamified systems. In Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems. 1015–1027.

[47] Rita Orji, Gustavo F Tondello, and Lennart E Nacke. 2018. Personalizing persuasive strategies in gameful systems to gamification user types. In Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems. ACM, 435. https://doi.org/10.1145/3173574.3174009

[48] Kiemute Oyibo, Rita Orji, and Julita Vassileva. 2017. The influence of culture in the effect of age and gender on social influence in persuasive technology. In Adjunct publication of the 25th conference on user modeling, adaptation and personalization. 47–52.

[49] Kiemute Oyibo, Rita Orji, and Julita Vassileva. 2017. Investigation of the Persuasiveness of Social Influence in Persuasive Technology and the Effect of Age and Gender.. In PPT@ PERSUASIVE. 32–44.

[50] Kiemute Oyibo, Rita Orji, and Julita Vassileva. 2017. Investigation of the social predictors of competitive behavior and the moderating effect of culture. In Adjunct publication of the 25th conference on user modeling, adaptation and personalization. 419–424.

[51] Luiz Rodrigues, Robson Bonidia, and Jacques Duílio Brancher. 2020. Procedural versus human level generation: Two sides of the same coin? International Journal of Human-Computer Studies 141 (2020), 102465. https://doi.org/10.1016/j.ijhcs.2020.102465

[52] Luiz Rodrigues, Wilk Oliveira, Armando Toda, Paula Palomino, and Seiji Isotani. 2019. Thinking Inside the Box: How to Tailor Gamified Educational Systems Based on Learning Activities Types. In Proceedings of the Brazilian Symposium of Computers on Education. SBC, 823–832. https://doi.org/10.5753/cbie.sbie.2019.823

[53] Fatemeh Roosta, Fattaneh Taghiyareh, and Maedeh Mosharraf. 2016. Personalization of gamification-elements in an e-learning environment based on learners' motivation. In 2016 8th International Symposium on Telecommunications (IST). IEEE, 637–642. https://doi.org/10.1109/ISTEL.2016.7881899

[54] Michael Sailer and Lisa Homner. 2020. The Gamification of Learning: a Meta-analysis. Educ Psychol Rev 32 (2020), 77–112.

[55] Isabelle Savard and Riichiro Mizoguchi. 2019. Context or culture: what is the difference? Research and Practice in Technology Enhanced Learning 14, 1 (2019), 1–12.

[56] Hallifax Stuart, Elise Lavoué, and Audrey Serna. 2020. To tailor or not to tailor gamification? An analysis of the impact of tailored game elements on learners' behaviours and motivation. In 21th International Conference on Artificial Intelligence in Education.

[57] Bahar Taspinar, Werner Schmidt, and Heidi Schuhbauer. 2016. Gamification in education: a board game approach to knowledge acquisition. Procedia Computer Science 99 (2016), 101–116.

[58] Armando M Toda, Ana CT Klock, Wilk Oliveira, Paula T Palomino, Luiz Rodrigues, Lei Shi, Ig Bittencourt, Isabela Gasparini, Seiji Isotani, and Alexandra I Cristea. 2019. Analysing gamification elements in educational environments using an existing Gamification taxonomy. Smart Learning Environments 6, 1 (2019), 16. https://doi.org/10.1186/s40561-019-0106-1

[59] Gustavo Fortes Tondello. 2019. Dynamic Personalization of Gameful Interactive Systems. Ph.D. Dissertation. University of Waterloo.

[60] Gustavo F Tondello, Alberto Mora, Andrzej Marczewski, and Lennart E Nacke. 2019. Empirical validation of the gamification user types hexad scale in English and Spanish. International Journal of Human-Computer Studies 127 (2019), 95–111.

[61] Gustavo F Tondello, Alberto Mora, and Lennart E Nacke. 2017. Elements of gameful design emerging from user preferences. In Proceedings of the Annual Symposium on Computer-Human Interaction in Play. ACM, 129–142. https://doi.org/10.1145/3116595.3116627

[62] Gustavo F Tondello, Rita Orji, and Lennart E Nacke. 2017. Recommender systems for personalized gamification. In Adjunct Publication of the 25th Conference on User Modeling, Adaptation and Personalization. ACM, 425–430. https://doi.org/10.1145/3099023.3099114

[63] Gustavo F Tondello, Rina RWehbe, Lisa Diamond, Marc Busch, Andrzej Marczewski, and Lennart E Nacke. 2016. The gamification user types hexad scale. In Proceedings of the 2016 annual symposium on computer-human interaction in play. ACM, 229–243. https://doi.org/10.1145/2967934.2968082

[64] Andrea C Tricco, Erin Lillie, Wasifa Zarin, Kelly K O'Brien, Heather Colquhoun, Danielle Levac, David Moher, Micah DJ Peters, Tanya Horsley, Laura Weeks, *et al.* 2018. PRISMA extension for scoping reviews (PRISMA-ScR): checklist and explanation. Annals of internal medicine 169, 7 (2018), 467–473.

ANNEX

B

JUST BEAT IT: EXPLORING THE INFLUENCES OF COMPETITION AND TASK-RELATED FACTORS IN GAMIFIED LEARNING ENVIRONMENTS

Just beat it: Exploring the influences of competition and task-related factors in gamified learning environments

Luiz Rodrigues[1], Armando M. Toda[1], Wilk Oliveira[1], Paula T. Palomino[1], Seiji Isotani[1]

[1]Institute of Mathematics and Computer Science – University of São Paulo (ICMC-USP)
400 Trabalhador São-carlense Avenue São Carlos, SP 13566-590 – Brazil.

{lalrodrigues, armando.toda, wilk.oliveira, paulatpalomino}@usp.br

sisotani@icmc.usp.br

Abstract. *Understanding how each game element in isolation affects learners' motivation and contextual factors' moderator effects is needed to improve gamified interventions. Thus, this paper explored the impact of one of the most used game elements - Competition - on motivation and whether task-related contextual factors (e.g., familiarity with the task's subject) moderate that impact. In a within-subject quasi-experimental design, graduate students from an Artificial Intelligence course created a reflexive intelligent agent for a console-based fight simulator in a non-gamified condition. Then, they improved their agents to compete against their peers' agents (gamified condition). Based on motivation levels measured in both conditions, we found that Competition was positive for students and that task-related contextual factors influenced that effect. Therefore, suggesting i) Competition alone can be positive for motivation and ii) contextual moderators should be considered in defining gamified designs.*

1. Introduction

Gamification is the use of game elements outside their original context [Deterding et al. 2011]. Whereas its overall effect is positive, some studies show uncertain and negative findings (e.g., demotivation and disengagement), with gamification designs often blamed for these undesired outcomes [Koivisto and Hamari 2019]. Although scholars have presented approaches to aid that process (e.g., [van Roy and Zaman 2017, Toda et al. 2019a]), two main problems arise from this context. First, as approaches often lead to designs featuring multiple game elements, it is unclear which game element is actually causing the outcomes (positive or negative), and to what extent each one is contributing [Mekler et al. 2017]. Second, which factors moderate gamification's success are not completely defined, demanding studies to better understand pre-determinants of gamification's effectiveness [Sailer and Homner 2019].

Furthermore, in the context of gamification, the Competition element[1] is one of the most used in gamified environments and it is recommended for males [Klock et al. 2020]. Motivation is often considered the psychological outcome sought by gamification [van Roy and Zaman 2017] and context has been discussed as a moderator of gamification's success [Liu et al. 2017], with activities/tasks being a central part of the context

[1]As discussed by [Toda et al. 2019b], we consider Competition a game element that might be represented by, for instance, a leaderboard or player-to-player conflict.

DOI: 10.5753/cbie.sbie.2020.461

[Savard and Mizoguchi 2019]. Hence, the purpose of this paper is to verify the effects of Competition and task-related factors on users' motivation. Given that gamification is often expected to affect user motivation, contextual data are likely to moderate that effect, and the aforementioned problems, we sought to answer (**RQ1**) *how does Competition affect learners' motivation?* and (**RQ2**) *how do task-related factors moderate Competition's impact on learners' motivation?*.

Compared to previous research, this study is different in two points. First, whereas most studies implement Competition as leaderboards (e.g., [Chan et al. 2018, Mekler et al. 2017]), we implemented it as a player-to-player (PvP) competition, which avoids having learners down in the leaderboard as the competition occurs amongst them. Second, studies evaluating Competition's impact often analyze moderators linked to users' characteristics (e.g., [Landers et al. 2019, Papadopoulos et al. 2015]), whilst we explore those related to the task, guided by literature research agenda [Hallifax et al. 2019, Rodrigues et al. 2019, Liu et al. 2017]. Thus, this study contributes by analyzing the impact of a single game element (i.e., Competition) on learners' motivation and investigating how task-related factors (i.e., those related to the gamified task) moderate that effect.

Moreover, we link gamification on practice to theory by exploiting [Pink 2011]'s theory when discussing the Competition's as well as the moderator's impacts and motivation. This author advocates intrinsic motivation is preferred over the extrinsic one. Intrinsic motivation relates to pleasure from the activity itself and extrinsic motivation relates to pleasure from what one receives in exchange for performing the activity [Pink 2011]. It might be split into external and identified regulations, the former relates to behaviors originated by interests in rewards and the latter relates to performing an activity due to its end (e.g., possible rewards) but assuming it as valuable and chosen by oneself. Amotivation relates to no intention in performing a task [Guay et al. 2000]. [Pink 2011] also proposes achieving intrinsic motivation is based on *autonomy* (i.e., freedom to do as desired), *mastery* (i.e., become better at something relevant), and *purpose* (i.e., a cause/reason).

2. Study

The goal of this study was to **analyze the impact of Competition on different motivation types and how task-related factors moderate this impact**. To achieve this goal, we performed a within-subject quasi-experimental study following the one factor with two treatments design. Treatments were a gamified condition, implemented through the Competition game element [Toda et al. 2019b], and a non-gamified condition. To compare conditions, we used a within-subject (paired) design that, according to [Wohlin et al. 2012], improves the experiment precision. Thus, there was no random assignment, characterizing a quasi-experiment. We also aimed to analyze it within a real learning setting (i.e., a real class), then, by convenience sampling, we recruited students from a graduate course where we were allowed to perform the study. Artificial Intelligence (AI) was the class' topic, in which students were introduced to Python programming language, AI itself, and intelligent agents types. Subjects were 15 Brazilian males with an average age of 31 (± 8.43) years. All of them agreed to participate in the study and were in accordance with the use of their (non)personal data for research ends.

In the study task, participants worked with a console-based fight simulator, which was developed specifically to be used as a learning resource in the course this quasi-experiment was performed. As the quasi-experiment was performed in an introductory AI class, its tasks concerned developing the simplest intelligent agent type, a reflexive agent, which is a kind of agent that chooses actions based on its current perception of the environment [Russell and Norvig 2010]. Subjects were provided with a random reflexive agent that they could use to test their agents' performance and with instructions on which actions their agents could use. Then, the quasi-experiment's tasks were accomplished in the learning resources developed for the course. At first (Task A), participants were asked to implement their agents by analyzing the environment and selecting which action to do. All other functionalities were provided (e.g., simulating the fights and updating the environment based on agents' actions). No gamification was deployed in Task A as we do not consider fighting to the random agent a competition because learners were not considering competing or in conflict with another person [Toda et al. 2019b]. Thereafter (Task B), participants were required to improve their agents compared to the version developed in Task A. As such task is unlikely to be motivating to the students, and code improvement is necessary in many cases, so improving the motivation for this task is valuable.

The condition manipulation was performed by deploying a PVP unplugged competition [Toda et al. 2020] in Task B. Choosing this game is recommended to our male sample [Klock et al. 2020]. Right after participants finished Task A and before they started Task B, subjects were warned that the result of the task would be competing with their peers (i.e., fighting against other subjects' agents). This inserts participants into a competitive environment as they readily internalize they are performing a task with a competitive end [Toda et al. 2019b]. Henceforth, adding an unplugged gamification design into the task by inciting a PVP competition, exposing all participants to conditions with and without gamification to allow us to evaluate whether the Competition game element impacts their motivation.

To measure motivation, we used the Situational Motivation Scale (SIMS) [Guay et al. 2000] because it captures subjects' motivation while performing a learning activity and assesses intrinsic, identified, external, and amotivation. Also, it is psychometrically validated and has been widely used and cited in the literature. Furthermore, unlike most related works, we evaluated task-dependent factors as possible moderators. In that sense, analyzing users' previous knowledge contributes to understanding the effect of context-related moderators as the activity is fundamental to the context [Savard and Mizoguchi 2019]. Accordingly, given that participants worked with AI in the Python programming language, we assessed their self-reported familiarity with these aspects, which were all approached in the class: AI, programming, and Python. These data were measured in five-point Likert-scales. Hence, allowing us to collect contextual data related to the task they would perform. Table 1 demonstrates the sample's percentages in each familiarity degree for each measure.

Table 1. Participants' contextual-factors (five-point Likert scale), shown as N(%).

Familiarity with	1	2	3	4	5
Programming	0 (0.00)	1 (0.07)	3 (0.20)	7 (0.47)	4 (0.27)
Python	2 (0.13)	9 (0.60)	2 (0.13)	1 (0.07)	1 (0.07)
AI	6 (0.40)	6 (0.40)	2 (0.13)	1 (0.07)	0 (0.00)

The study was conducted in six steps. **First**, learners self-grouped into groups of three to discuss the activity. We adopted this approach because the study concerns the last activity students performed at the course's first class (after roughly 6 hours of learning, without considering the intervals) and we wanted to foster collaborative learning (e.g., exchanging knowledge, opinions and/or strategies regarding the activity itself, the programming language, and the programming approaches they could adopt during the tasks), as well as get to know each other, which is facilitated by this kind of grouping intervention [Swaray 2012]. The remaining steps were accomplished by each student individually. **Second**, each participant completed the contextual info questionnaire. This was necessary to provide data to analyze the possible moderation effect of contextual, task-related factors on Competition's impact. **Third**, they had to perform Task A, which was the accomplishment of the first part (control condition) of the within-subject design[2]. **Fourth**, subjects completed the SIMS (i.e., pre-test, the first part of the paired comparison). **Fifth**, they had to perform Task B, the second and final part (experimental condition) of the within-subject design as we were analyzing two conditions only. This was the gamified task as participants performed in within the unplugged PVP competition previously introduced. **Sixth**, students completed the SIMS again (i.e., post-test, the second part of the paired comparison) right after finishing Task B. The reliability of participants' answers (Cronbach's alpha $\alpha \geq 0.7$) was adequate for all motivation types.

Lastly, for data analysis, we analyzed paired measures' difference to evaluate Competition's impact following the suggestion by [Wohlin et al. 2012]. This difference was calculated subtracting the pre- from the post-measure, which yields higher values for larger impacts, and vice-versa. Due to our reduced sample size, we chose not to perform inferential statistical tests as those would have no power to yield reliable results. Considering our quasi-experiment context (N = 15; paired), a large effect size (ES; ≥ 0.8) would be needed for a t-test to achieve the usual 0.8 power (Calculated using the *pwr* package from R) under the standard 0.05 significance level; similar for an ANOVA to assess moderations. However, gamification's effects mostly range from small to moderate [Sailer and Homner 2019], and the evaluation of moderator effects was exploratory. Therefore, we used descriptive measures (mean (M); standard deviation (SD); and ES) for data analysis, especially considering ES is sample size-independent, unlike inferential tests [Wohlin et al. 2012]. To calculate ES, we used Hedge's g, which is recommended over Cohen's d for small samples, such as in this study's case [Ellis 2010].

3. Results

Overall, subjects reported high levels of intrinsic and identified motivation (> 6), *indifferent* levels of external motivation ($\pm$ 4), and low levels of amotivation ($\pm$ 2), as shown in Table 2. Additionally, the impact of adding Competition to the tasks the participants performed was small (between $|0.14|$ and $|0.27|$) [Ellis 2010] for all motivation types. However, the *difference's* standard deviations in Table 2 suggest that either the motivation change was positive or negative depended on the subject. Hence, reinforcing that one gamification design for all (one-size-fits-all) is unlikely to be completely successful, demonstrating the need for understanding what factors moderate/pre-determine gamification effectiveness [Koivisto and Hamari 2019]. We approach this need by analyzing the

[2]Note that the subjects experienced the control condition first because once a subject experience a gamified condition, they cannot be readily *ungamified* [Thom et al. 2012].

impact of subjects' familiarity with aspects related to the activities' context, in accordance with literature suggestions [Hallifax et al. 2019, Liu et al. 2017].

Table 2. Descriptive statistics (SD: Standard Deviation) for each motivation type.

	Intrinsic		Identified		External		Amotivation	
	Mean	SD	Mean	SD	Mean	SD	Mean	SD
Pre	6.10	1.08	6.03	0.97	3.63	1.52	2.25	1.42
Post	6.30	0.70	6.28	0.80	3.85	1.37	2.03	1.55
Difference	0.20	1.39	0.25	1.37	0.22	2.24	-0.22	2.41
Effect size	0.21	-	0.27	-	0.15	-	-0.14	-

Figures 1 to 4 present participants' motivation changes for intrinsic, identified, external, and amotivation, respectively, grouped by their familiarity degree for the three moderators (programming, Python, and AI). Colors are used to represent specific degrees (e.g., yellow for two) along the X-axis, independent of the moderator, and boxes represent the distribution of motivation changes for participants within that group. The Y-axis represents the extent of the change. Next, as these analyses are exploratory, we discuss findings and rise some hypotheses on possible reasons for them based on the motivation type's definition and [Pink 2011]'s theory.

For intrinsic motivation (Figure 1), our findings suggest previous general programming familiarity had the highest impact. Users with less familiarity (< four) were demotivated by competition, whereas roughly all of those with higher familiarity reported increased intrinsic motivation. Similarly, those with the highest AI familiarity (three and four) presented higher intrinsic motivation gains than other subjects, although the small difference and the overlap with those that reported two. We speculate that participants' background supported them in feelings of mastery, therefore, increasing their desire to do so for no reason besides themselves. On the other hand, familiarity with Python on intrinsic motivation changes appears to have a U function. Possibly, those with some to moderate experience (two and three) felt the activity was not challenging enough (no purpose) whereas those with none (i.e., one) or more than moderate (four and five) were motivated to do so by themselves either to try to achieve or because already felt mastery.

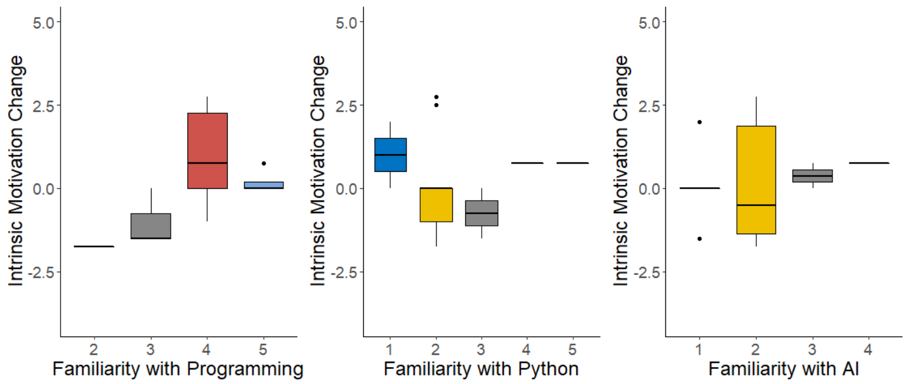

Figure 1. Intrinsic motivation changes per contextual factor.

For identified motivation (Figure 2), it is possible to note a roughly U function on Python experience as well. A possible reason is those around the middle not being interested in the outcomes of the competition due to their moderate familiarity (e.g., no purpose), while those with the smallest and the two highest familiarity scores were concerned with either improving or showing up their skills seeking for acknowledgment (i.e., external reward; mastery). In terms of AI familiarity, a weak positive association appears to exist from this factor to identified motivation gains, in which those that reported higher familiarity also reported higher gains. Regarding programming familiarity, the results show those with high familiarity (four and five) reported positive changes whereas the others' reports demonstrated a decrease. In the last two cases, it might be that learners with more familiarity assumed they had to perform better in order to show their skills (external reward), or that they felt more confidence and interest in achieving the competition's outcomes compared to the remaining subjects (external reward; mastery).

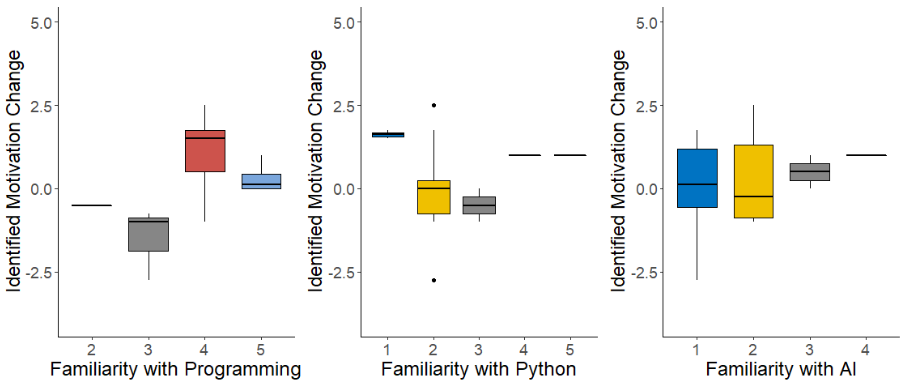

Figure 2. Identified motivation changes per contextual factor.

For external motivation (Figure 3), those with high levels (four or five) of familiarity in both Python and AI were highly demotivated. A hypothesis for that is those subjects were confident in their skills or were not afraid of possible punishments (no purpose). In addition, for AI familiarity, overall motivation practically did not change for those with moderate familiarity or less ($<$ four), which might be that the lack of this skill had no effect on learners' external regulations. For Python familiarity, from one to three it is possible to identify a positive association to motivation gain, a cue that as learners' familiarity approached a moderate point, their motivation also increased. Contrary, from programming familiarity two to four, a negative association is shown, which might be because as the more the participants were familiar with programming, the less they cared about external regulations (loss of purpose).

For amotivation (Figure 4), it appears that subjects with moderate or higher familiarity with AI were not affected by competition whereas the less their familiarity in this factor the more their amotivation decreased. This is similar to the results for Python familiarity, with exception to those in the middle point (i.e., three). Probably, Competition helped to foster interest in performing the activity for those that were not interested due to their lack of background on the previously mentioned factors (purpose; mastery). For the aforementioned exception, it might be that those subjects are not interested in competition itself (no purpose) or that their moderate familiarity led to discouragement (no mastery), therefore, demotivating them to perform the task. On the other hand, in terms of programming familiarity, those with high levels (four and five) presented decreased amotivation whereas the remaining became more amotivated with the competition, which

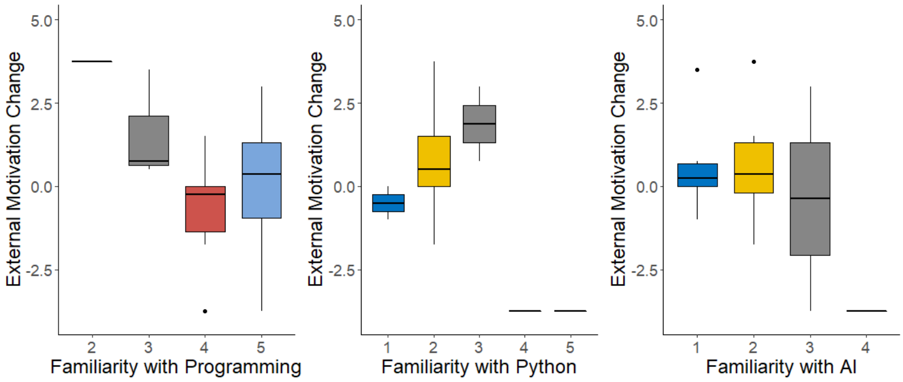

Figure 3. External motivation changes per contextual factor.

possibly emerged as those with more background were more confident whilst the others felt insecure (mastery).

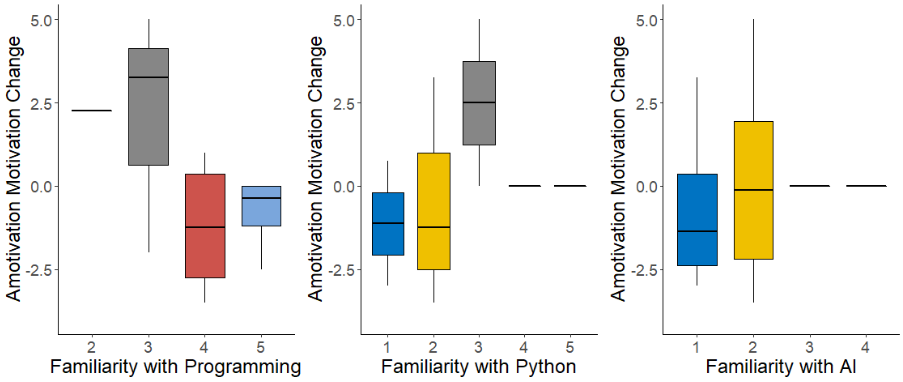

Figure 4. Amotivation motivation changes per contextual factor.

In summary, these findings suggest that Competition positively affected users' motivations, especially the intrinsic and identified ones (RQ1), and that the task-related factors influenced such effect, increasing and decreasing participants' motivation change depending on their affinity level (RQ2). Next, we further discuss these findings.

4. Discussion

According to our results, Competition has an overall small positive effect on learners' motivation, decreasing their amotivation whilst increasing intrinsic, identified, and external motivations (RQ1). This finding corroborates those of a recent meta-analysis of gamification on education, indicating gamification has overall small impacts on motivational outcomes [Sailer and Homner 2019]. Those are valuable findings for educators as the motivation types influenced the most (i.e., identified and intrinsic) are related to autonomous motivation, which is discussed as the ideal motivation type to education [van Roy and Zaman 2017].

Furthermore, although the average impact was positive, the effect on some learners was more than small (positive), for others were none, and for others was negative. These findings also corroborate indications that gamification designs should consider people's characteristics [Liu et al. 2017] as research has shown, in different contexts, the impact of individual characteristics [Rodrigues and Brancher 2019, Hallifax et al. 2019]. Hence, reinforcing the need for providing personalized gamified interventions to improve gamification effectiveness whilst preventing it from delivering undesired outcomes.

Also, our findings tackle a perspective of gamification studies that demand research and that has increasingly attracted researchers attention: the analysis of pre-determinants related to the context (operationalized as the activity in this study; RQ2) [Koivisto and Hamari 2019, Savard and Mizoguchi 2019]. We found the task-dependent factors we assessed moderated Competition's effect on learners' motivation in varied ways: AI familiarity appears to have no overall impact; more familiarity with general programming appears to slightly improve motivation, except that it decreased external motivation for those with high familiarity; and the familiarity with Python programming language showed a U function impact on autonomous motivation and, for external and amotivation, a positive effect up to moderate familiarity with a decrease thereafter.

These findings are valuable to inform both researchers and practitioners. They contribute to the knowledge on how a specific game element impacts different types of motivation, which is the main psychological aspect scholars argue gamification seeks to affect. Hence, these results can be used to inform the design of gamified interventions by suggesting that even the isolated use of competition can lead to motivation gains, as suggested by previous research [Landers et al. 2019]. Furthermore, the insights can be used to inform research on personalized gamification in terms of how Competition worked for our sample and the hypotheses we raised for context-related moderations, exploiting [Pink 2011]'s theory, can be analyzed in future studies.

4.1. Limitations and Threats to Validity

This study has some limitations and threats to validity that must be considered. Concerning the sample, our findings are based on a small, homogeneous one, which reduces findings generalization; however, it was necessary to improve the study by performing it in a real class due to the costs involved. Additionally, participants' gender might have affected gamification's effects as all participants were males and given that the Competition game element was selected accordingly. Nevertheless, Competition might not work even for males - as our findings suggest - and while participants of a single gender limit findings generalization, it reduces the possible confound of multiple genders. Concerning the intervention, participants were just introduced to the gamified (i.e., Competition) context, weakening our understanding about to what extent the improvements would last, despite it is not clear whether the novelty effect plays a role in gamified environments [Sailer and Homner 2019]. However, competition is something people are often exposed to in their daily lives, mitigating this limitation. Concerning the instruments, the gamification design was not defined based on any gamification framework because of the study goal. Rather than following a framework to choose the gamification design, we deliberately implemented an ad-hoc PVP competition [Toda et al. 2020] to identify its effects, which can inform choices of those following some framework based on the effects found on our study sample and context. Furthermore, although we selected a psychometric validated scale to measure motivation, its language (English) is different from that of the participants, which is mitigated because participants belong to a field study that often interacts with English and studied it in high school.

5. Concluding Remarks

This paper analyzed the impact of the Competition game element on the motivation of learners from a graduate AI class. Mainly, we found using Competition positively affected a specific population (which may provide some guidelines to educators, designers, and developers on how to use Competition to achieve better results) and indication that learners' familiarity to task's topic, which concerns contextual characteristics, is likely to moderate gamification's effectiveness. Hence, contributing to the scarce literature on analyzing the students' motivation based on isolated game elements, as well as how context-related factors moderate gamification's success, besides using a design that can be replicated in other studies to further analyze this and other game elements.

Acknowledgments

The authors would like to thank the funding provided by CNPq, CAPES, and FAPESP (Projects 2016/02765- 2; 2018/11180-3; 2018/15917-0; 2018/07688-1).

References

Chan, E., Nah, F. F.-H., Liu, Q., and Lu, Z. (2018). Effect of gamification on intrinsic motivation. In *International Conference on HCI in Business, Government, and Organizations*, pages 445–454. Springer.

Deterding, S., Dixon, D., Khaled, R., and Nacke, L. (2011). From game design elements to gamefulness: defining gamification. In *Proceedings of the 15th international academic MindTrek conference: Envisioning future media environments*, pages 9–15.

Ellis, P. D. (2010). *The essential guide to effect sizes: Statistical power, meta-analysis, and the interpretation of research results*. Cambridge University Press.

Guay, F., Vallerand, R. J., and Blanchard, C. (2000). On the assessment of situational intrinsic and extrinsic motivation: The situational motivation scale (sims). *Motivation and emotion*, 24(3):175–213.

Hallifax, S., Serna, A., Marty, J.-C., Lavoué, G., and Lavoué, E. (2019). Factors to consider for tailored gamification. In *Proceedings of the Annual Symposium on Computer-Human Interaction in Play*, CHI PLAY '19, page 559–572.

Klock, A. C. T., Gasparini, I., Pimenta, M. S., and Hamari, J. (2020). Tailored gamification: A review of literature. *International Journal of Human-Computer Studies*.

Koivisto, J. and Hamari, J. (2019). The rise of motivational information systems: A review of gamification research. *International Journal of Information Management*, 45:191–210.

Landers, R. N., Collmus, A. B., and Williams, H. (2019). The greatest battle is within ourselves: An experiment on the effects of competition alone on task performance. *International Journal of Human-Computer Studies*, 127:51–61.

Liu, D., Santhanam, R., and Webster, J. (2017). Toward meaningful engagement: A framework for design and research of gamified information systems. *MIS quarterly*, 41(4).

Mekler, E. D., Brühlmann, F., Tuch, A. N., and Opwis, K. (2017). Towards understanding the effects of individual gamification elements on intrinsic motivation and performance. *Computers in Human Behavior*, 71:525 – 534.

Papadopoulos, P. M., Lagkas, T., and Demetriadis, S. N. (2015). How revealing rankings affects student attitude and performance in a peer review learning environment. In *International Conference on Computer Supported Education*, pages 225–240. Springer.

Pink, D. H. (2011). *Drive: The surprising truth about what motivates us*. Penguin.

Rodrigues, L. and Brancher, J. D. (2019). Playing an educational game featuring procedural content generation: which attributes impact players' curiosity? *RENOTE*, 17(1).

Rodrigues, L., Oliveira, W., Toda, A., Palomino, P., and Isotani, S. (2019). Thinking inside the box: How to tailor gamified educational systems based on learning activities types. In *Brazilian Symposium on Computers in Education (Simpósio Brasileiro de Informática na Educação-SBIE)*, volume 30, page 823.

Russell, S. J. and Norvig, P. (2010). Intelligent agents. In *Artificial Intelligence: A Modern Approach*. Prentice Hall Press.

Sailer, M. and Homner, L. (2019). The gamification of learning: a meta-analysis. *Educational Psychology Review*.

Savard, I. and Mizoguchi, R. (2019). Context or culture: what is the difference? *Research and Practice in Technology Enhanced Learning*, 14(1):1–12.

Swaray, R. (2012). An evaluation of a group project designed to reduce free-riding and promote active learning. *Assessment & Evaluation in Higher Education*, 37(3).

Thom, J., Millen, D., and DiMicco, J. (2012). Removing gamification from an enterprise sns. In *Proceedings of the acm 2012 conference on computer supported cooperative work*, pages 1067–1070. ACM.

Toda, A. M., do Carmo, R. M., da Silva, A. P., Bittencourt, I. I., and Isotani, S. (2019a). An approach for planning and deploying gamification concepts with social networks within educational contexts. *International Journal of Information Management*, 46:294–303.

Toda, A. M., Klock, A. C., Oliveira, W., Palomino, P. T., Rodrigues, L., Shi, L., Bittencourt, I., Gasparini, I., Isotani, S., and Cristea, A. I. (2019b). Analysing gamification elements in educational environments using an existing gamification taxonomy. *Smart Learning Environments*, 6(1):16.

Toda, A. M., Palomino, P. T., Oliveira, W., Rodrigues, L., Klock, A. C., Gasparini, I., Cristea, A. I., and Isotani, S. (2020). How to gamify learning systems? an experience report using the design sprint method and a taxonomy for gamification elements in education. *Journal of educational technology and society.*, 22(3):1–14.

van Roy, R. and Zaman, B. (2017). Why gamification fails in education and how to make it successful: Introducing nine gamification heuristics based on self-determination theory. In *Serious Games and Edutainment Applications: Volume II*, pages 485–509.

Wohlin, C., Runeson, P., Höst, M., Ohlsson, M. C., Regnell, B., and Wesslén, A. (2012). *Experimentation in software engineering*. Springer Science & Business Media.

ANNEX

C

GAMIFICATION WORKS, BUT HOW AND TO WHOM? AN EXPERIMENTAL STUDY IN THE CONTEXT OF PROGRAMMING LESSONS

Gamification Works, but How and to Whom? An Experimental Study in the Context of Programming Lessons

Luiz Rodrigues
lalrodrigues@usp.com
University of São Paulo, Brazil

Armando M. Toda
armando.toda@usp.br
University of São Paulo, Brazil

Wilk Oliveira
wilk.oliveira@usp.br
University of São Paulo, Brazil

Paula T. Palomino
paulatpalomino@usp.br
University of São Paulo, Brazil

Anderson Paulo Avila-Santos
anderson.avila@usp.br
University of São Paulo, Brazil

Seiji Isotani
sisotani@icmc.usp.br
University of São Paulo, Brazil

ABSTRACT

Programming is a complex, not trivial to learn and teach task, which gamification can facilitate. However, how gamification affects learning and the influence of context-related aspects on that effect demand research to better understand how and to whom gamification enhances programming learning. Therefore, we conducted an experimental study analyzing how gamification worked and the role of context-related aspects in terms of intervention duration and learners' familiarity with programming (i.e., the task's topic). It was a six-week study with 19 undergraduate students from an *Algorithms* class that measured their learning gains, intrinsic motivation, and number of completed quizzes. Mainly, we found gamification affected learning via intrinsic motivation, effect that depended on intervention duration and learners' familiarity with programming. That is, intrinsic motivation strongly predicted learning gains and gamification's effect on intrinsic motivation changed over time, decreasing from positive to negative as learners had less familiarity with programming. Thus, showing gamification can positively impact programming learning by improving students' intrinsic motivation, although that effect changes over time depending on one's previous familiarity with programming.

CCS CONCEPTS

• **Human-centered computing → Empirical studies in HCI**; • **Applied computing → Interactive learning environments**.

KEYWORDS

gamification, learning, motivation, moderators, longitudinal

ACM Reference Format:
Luiz Rodrigues, Armando M. Toda, Wilk Oliveira, Paula T. Palomino, Anderson Paulo Avila-Santos, and Seiji Isotani. 2021. Gamification Works, but How and to Whom? An Experimental Study in the Context of Programming Lessons. In *Proceedings of the 52nd ACM Technical Symposium on Computer Science Education (SIGCSE '21), March 13–20, 2021, Virtual Event, USA.* ACM, New York, NY, USA, 7 pages. https://doi.org/10.1145/3408877.3432419

ACM ISBN 978-1-4503-8062-1/21/03…$15.00
https://doi.org/10.1145/3408877.3432419

1 INTRODUCTION

Learning to program is challenging. Students present difficulties in syntactic, conceptual, and strategic knowledge [28], and often lack motivation to learn, leading to low grades and high rates of dropout [20, 24]. To address motivational concerns aiming to improve learning, recent research started to explore gamified learning [2, 25] (i.e., adding game elements to change the learning process [19]), which is associated to overall positive effects on learning outcomes [34], including applications to Computer Science (CS) [6].

However, studies on CS applications often focus on gamification's impact on behavioral learning outcomes, such as student performance (e.g., [5, 8, 20, 22]), neglecting learners' motivation and context of use, despite those are inherently connected to learning [1, 36, 38, 42]. Thus, there is a need to understand how gamification affects programming learning while also considering motivation, as well as the role of context, corroborating recent calls for better understanding how gamification works and the context's role aiming to improve gamification's positive outcomes [15, 31, 35].

To advance this understanding, we conducted an experimental study. The experiment context is a Brazilian undergraduate *Algorithms* class focused on programming lessons. Students were randomly assigned to complete quizzes in one of two Moodle versions (gamified | non-gamified) during half semester. Then, we analyzed how gamification affected programming learning based on cognitive (learning gains), motivational (intrinsic motivation), and behavioral (number of tasks completed) learning outcomes [34]; the last two were weekly measured. Moreover, as context involves someone, with their mental state, performing an activity in a given environment for a given period (e.g. a learner, with their current knowledge, completing quizzes in a gamified system) [36], we analyzed its role based on intervention duration and students' familiarity with general course topics. Accordingly, the purpose of this paper was to answer *how does using gamification for half semester affect Brazilian undergraduates programming learning?*

We found gamification affected programming learning by influencing learners' intrinsic motivation. Additionally, we found intervention duration and familiarity with programming together moderated gamification's effect: at the intervention beginning, it was positive but then decreased for students with little familiarity with programming; whereas it started negative and then became positive for those with high familiarity with programming. Thus, we contribute with empirical evidence revealing through which construct gamification affected programming learning and when/to whom that effect was positive or negative, besides a theory-grounded gamification design likely to contribute to undergraduates' learning depending on their familiarity with programming and usage time.

2 BACKGROUND AND RELATED WORK

Often, students consider learning to program difficult and of low motivation and, thus, perform poorly [1, 24]. A possible explanation for this phenomenon is found in Self-determination theory (SDT) [3], a well-know theory often related to learning that considers three motivation types: intrinsic (i.e., a desire/interest that comes from within), external (i.e., desire/interest in the rewards/outcomes), and amotivation (i.e., no motivation). The literature suggests the first type is ideal for learning contexts [42], and that intrinsically motivated learners are more engaged and retain information better [3]. Accordingly, evidence that learners with higher intrinsic motivation levels achieve higher exam scores has been found (e.g., [11]). Hence, suggesting intrinsically motivated people will achieve better learning than amotivated and extrinsically ones.

Gamification might aid with the motivation issue as it is often concerned with improving it [43]. However, between gamification impacting motivation, there are the moderator factors: those that increase/decrease - or pre-determinate - gamification's effect, possibly even changing it from positive to negative [19]. Studies have investigated moderators related to users' profiles (e.g., [26, 29, 37]), whereas more attention to the intervention duration and context is needed [10, 14, 31, 35]. Here, we consider the context involves users performing some action (e.g., activity) in a given environment [36]. Accordingly, we might expect that users' familiarity with topics related to actions will affect how they perceive the environment.

Moreover, by affecting motivation, gamification can help in influencing learners' behaviors [23]. Studies have shown that completing quizzes (self-testing) also relates to higher exam scores (e.g., [4, 35]). This relationship relies on the *testing effect*, which has been supported in numerous experimental settings [33]. Thereby, demonstrating gamification's potential to indirectly affect learners' behavior is of value to their learning.

In light of this context, we review empirical research on the effects of gamification on learning outcomes within the context of programming concepts. For studies selection, we analyzed two recent secondary studies [2, 34] as they were published less than one year before our time of writing and map empirical gamification research found in a broad range of databases.

Hakulinen et al. [8] evaluated the effect of badges on behavioral learning outcomes. Based on data from 281 students from a *Data Structures and Algorithms* course (around eight weeks long), both positive (e.g., badges earned and time in the system) and null (e.g., completed exercises) results were found when log data from learners' interactions with the gamified educational system were compared to log data from those who interacted with the non-gamified version of the same system.

Krause et al. [17] compared the impacts of gamification and social gamification to a non-gamified condition in the context of a course for learning Python as a statistical analysis tool (four weeks long). Considering retention, in-system quizzes' performance (n=206), and a post-test score (n=101), they found gamification versions overcame the control condition in all three measures; social gamification overcame the simpler gamification version. No moderator effect of age neither sex was found.

Fotaris et al. [5] gamified a Python programming course (12 weeks long) using Kahoot! and Codeacademy. They longitudinally compared attendance, late arrivals, and number of material downloads. Also, they compared the students' academic performance in this course to that of a non-gamified version of the same course, offered in the previous semester. Overall results were positive in favor of gamification, but only descriptive analyses were performed.

Moreno and Pineda [22] analyzed the impact of using a gamified educational system featuring automatic code judging compared to traditional workshops. Participants (n=43) were split into two groups (one for each condition) and had their learning compared in terms of performance on programming tests (e.g., conditionals and loops). Findings show those who used the gamified system achieved higher scores compared to the remaining, suggesting the benefits of the system as no performance difference was found between groups in the pre-test. However, it is unclear whether the effect emerged from gamification or other system features because gamification was not the only difference between conditions.

Marin et al. [20] analyzed data from two semesters (n=817) of a C programming course (four weeks long) to assess gamification's impact on students' performances. Despite they measured learning performance from two exams, the first one was administered after the intervention began (middle semester), not characterizing a pre-test. Overall results are positive for gamification, but the decrease from exam one to exam two was similar in both semesters and the change from one to another was not considered.

Table 1 summarizes this paper and main related works' characteristics, allowing the identification of the gaps this study faces. First, not using pre-tests, which opens the possibility of not acknowledging when one group has, for instance, an initial motivation higher than the other, which might mislead conclusions. Additionally, using pre-tests has been acknowledged as a characteristic needed for gamification studies to present high methodological rigor [34]. To address this issue, our experiment employed pre-tests right before the intervention begin. Second, the lack of longitudinal studies, which disables the possibility of understanding gamification's effects over time. To tackle this lack, we studied learners' psychological states and behavior at each experiment's week.

Table 1: Related research characterization.

Ref.	ECG	PT	LA	IA	MA	ID	SC	TB	LO
[8]	Y	N	N	Y	N	8	Y	N	B
[17]	Y	N	N	Y	Y	4	Y	N	B,C
[5]	Y	N	Y	N	N	12	N	N	B
[22]	N	Y	N	Y	N	?	Y	N	B
[20]	Y	N	N	Y	N	16	N	N	B
This	Y	Y	Y	Y	Y	6	Y	Y	M, B, C

ECG = equivalent control group; PT = pre-test; LA = longitudinal analysis; IA = inferential analysis; MA = moderation analysis; ID = intervention duration in weeks; SC = same class; TB = theoretical background; LO = learning outcomes; Y = yes; N = no; ? = undefined; B = behavioral; C = cognitive; M = motivational.

Third, the little attention to moderators that, otherwise, would advance the understanding of when/to whom gamification is more or less suited. To expand this understanding, we analyzed the impact of possible moderators that have been suggested in the literature, such as intervention time and contextual characteristics [14, 30, 35]. Fourth, not grounding gamification designs on learning-related theories neither exploring motivational or cognitive learning outcomes, which would better explain the process through which gamification affects learning, rather than just showing if it does. We approached this need grounding the gamification design on SDT, one of the most used in gamification studies and considered relevant to learning [11, 42, 43], as well as evaluated cognitive, behavioral and motivational learning outcomes.

Given this context, we tested the following hypotheses to answer our research question: (**H1**) Intrinsic motivation and completing quizzes positively affect learning gains; (**H2**) Gamification improves intrinsic motivation, effect moderated by intervention duration and users' familiarity with class' topics; and (**H3**) The more the learners' intrinsic motivation, the more quizzes they complete.

3 EXPERIMENT

To achieve the goal of understanding *how using gamification for half semester affects Brazilian undergraduates programming learning*, this experiment was performed in an undergraduate Software Engineering course of a private institution in Londrina - Brazil, with approval of the university's ethical committee. We conducted the experiment from March to June, 2020, on the *Algorithms* discipline, which explores pseudo-language to introduce first-term students to programming. The discipline instructor is male, holds a MsC. degree in CS, and had taught for six years at that time. Topics taught during the experiment included conditionals, loops, and arrays (initialization and manipulation). Our participants provided informed consent and represent 68% of those initially enrolled in the discipline. Inclusion criteria was being enrolled in the discipline and completing pre- and post-tests. From the 28 possible participants (all males), 19 met the criteria: all males with an average age of 20.32 years (±3.64).

The materials related to this experiment are the educational system, experimental task, gamification design, measures, and moderators. The educational system used to accomplish the experiment's tasks was Moodle because it is the standard educational system in the university, enabling the intervention to be within the environment students and course instructor regularly use.

The task participants had to accomplish was completing quizzes, an optional task that added extra points in the course. We explored extra activities to reduce the influence on the original course program, and along with the instructor, defined quizzes would give extra points to encourage students to engage with the tasks. Each quiz featured from three to five items, and six new quizzes were provided each week. The rationale for six quizzes was that students could complete one quiz per non-class day, aiming not to overcharge them with extra work. To align quizzes with the course design, we followed the revision of the cognitive process dimensions of Bloom's Taxonomy, since it is a renowned structure to support the development of learning outcomes [16]. Then, at each week, quizzes complexity increased according to those dimensions. That is, quizzes' highest dimension was *remember* in the first week, *understand* in the second, and so on. Multiple knowledge dimensions were intentionally explored within each week.

For gamification design, we implemented gamification heuristics focused on SDT, aiming to affect intrinsic motivation. Next, we introduce the heuristics, taken verbatim from [40], and how we implemented them.

#1 Avoid obligatory uses: Completing quizzes was optional and students could solve the week quizzes in their preferred order.

#2 Provide a moderate amount of meaningful options: Quizzes were aligned to each week's class topic, based on the instructor's schedule, offering six of those per week (i.e., learners could do one per day, excluding the class day).

#3 Set challenging but manageable goals: The discipline instructor revised all quizzes' items and provided feedback for adapting them when necessary.

#4 Provide positive, competence-related feedback: We added weekly, unannounced badges that acknowledged students according to the cognitive dimension of the quiz.

#5 Facilitate social interaction: Two out of the six weekly quizzes were team/group activities. In those, students could see peers' answers after completing the quizzes, analyze and discuss the answers, and update their owns.

#6 When supporting a particular psychological need, wary to not thwart the other needs: We i) used Cooperation to support relationships feelings, rather than Competition, which could make users feel incompetent, and ii) offered unannounced badges to prevent, for instance, feelings of *needing* to receive it (e.g., anxiety).

#7 Align gamification with the goal of the activity in question: We transformed quizzes into missions that encouraged learners to complete the quizzes.

#8 Create a need-supporting context: Implemented by attending user needs, autonomy, competence, and relatedness, through heuristics #1, #4, and #5, respectively.

#9 Make the system flexible: Not implemented because personalizing/adapting gamification is a recent, open research field [9, 32].

Nevertheless, the implementation of many of those heuristics were available in the regular Moodle version as well. As not all aspects are directly related to adding game elements to change the learning process (i.e., gamifying learning), we intentionally allowed the non-gamified Moodle version to feature them. Compared to the regular version, the gamified one adds unannounced badges and a more gameful experience by presenting quizzes as graphic-enriched missions, similar to [41], and working groups as teams, resembling a game rather than a regular group learning activity. The missions and badges used can be seen at: shorturl.at/gkrDN.

As measures, we captured cognitive, behavioral, and motivational learning outcomes [34]. To measure cognitive learning outcomes (i.e., learning gains), we designed a test featuring 15 multiple choice items to assess *remembering* and *understanding* domain processes on three programming topics: conditionals, loops, and arrays (five items per topic). The test was revised by the discipline instructor, being considered a suitable formative evaluation instrument. The behavioral outcome was operationalized as the number of completed quizzes, following previous similar research (e.g., [35]), which Moodle automatically collected. Intrinsic motivation (motivational outcome) was measured with Portuguese version of the *interest/enjoyment* sub-scale of the Intrinsic Motivation Inventory as validated in [27].

As moderators, we captured intervention time and contextual characteristics. For intervention time, we considered the experiment week (*week*; 0 for the pre-test, 6 for the last week). For contextual characteristics, we captured learners' self-reports of their familiarity with topics related to the course in five-point Likert-scales: familiarity with algorithms (FAlg), C programming language (FC), programming (FProg), and pseudo-language (FPseu). These aspects concern the context because they are directly related to the action (activity) users performed in the experiment [36].

As experimental design, we employed a 2x6 mixed factorial design with random assignment to the between-subject independent variable Condition: control (i.e., used non-gamified Moodle; N = 10) and experimental (i.e., used gamified Moodle; N = 9). The within-subject independent variable concerned six repeated measures of motivational and behavioral data collected weekly. Moodle automatically collected the former, whereas the professor asked students to complete the measures of the latter during the weekly classes.

The data collection procedure followed three steps. First, pre-tests were administered to serve as baseline comparisons for cognitive and motivational learning outcomes. Then, the intervention started, which lasted for six weeks. During this phase, participants were offered a new set of extra activities related to the class' topics each week. Activities were the same for all participants, with the only difference being the Moodle version to be used. Lastly, after the sixth experiment week, the post-test was administered, generating the learning gain measure (post - pre). Completing the motivation scales and quizzes was optional. Consequently, some participants completed no quiz as well as the number of motivational measures completed varies per week. The number of completions of control and experimental groups is, respectively: W0, 7 and 7; W1, 6 and 6; W2, 3 and 6; W3, 7 and 7; W4, 6 and 3; W5, 6 AND 3; W6, 9 and 5. Completions reliability was good in all weeks ($\alpha > 0.8$) . Summarizing, we captured cognitive (learning gains; pre- and post-tests), behavioral (Moodle's logs; one per week), and motivational (self-reports; pre-test plus one per week) learning outcomes.

For data analysis, we explored regression methods. We used multiple linear regression to test **H1**, as this approach enables understanding whether and how various independent variables predict a dependent variable [7]. In **H1** case, intrinsic motivation and completed quizzes (independent) and learning gains (dependent) are the variables. Learning gains were measured based on the difference between post- and test-tests. Consequently, there is one measure per participant. However, this analysis' independent variables were repeatedly measured. Therefore, we aggregated them (average and sum), creating one measure of each variable per participant. Furthermore, we excluded outliers based on standard deviation ($|x| > 2 * SD$; two SD due to the small sample size, i.e., 19), after Shapiro-wilk tests suggested both independent variables follow a normal distribution, as regression analyses are sensitive to outliers [18]; a single item was excluded.

Testing **H2** involves dependent data (i.e., multiple answers from the same participants), which violates the independence assumption of classical regression methods [7]. Multilevel models are regression-based models that properly account for dependent data, and can be seen as a hierarchical system of regression equations (one for each level/group) [12]. This is achieved by allowing each group (e.g., of subjects) to have its own intercept and slope coefficients, which are often referred to as *random* coefficients. Additionally, these models contain *fixed* coefficients, which do not vary across groups. By doing so, multilevel methods model the groups' variance as well as find estimates applicable to the whole sample. Measuring random coefficients, however, requires sample sizes larger than that of this study. Therefore, we focus on analyzing fixed effects, which are reproducible properties of the overall data [21]. Moreover, compared to repeated-measures ANOVA, multilevel analysis has more power and handles data with varied numbers of answers per subject, besides accounting for dependencies [7]. Therefore, we test **H2** using multilevel analysis.

To test **H2**, we followed guidelines [7, 21] for properly identifying and defining multilevel models. The recommended approach to evaluate the relevance of a specific parameter (independent variable; e.g., Condition) is to test whether the parameter significantly increases model fit compared to the model without it. In the context of multilevel analyses, this is often accomplished through the Likelihood Ratio Test (LRT). Furthermore, multilevel models might be developed from bottom-up (starts simple and increasingly complexify) or top-down (starts complex and removes irrelevant parameters). The former requires multiple steps, whereas the latter allows creating a model with all parameters to be tested and, then, remove those that do not decrease model fit [7, 21].

We tested **H2** following the top-down approach to reduce the number of tests in model development. Accordingly, we defined a model that accounts for relationships between our dataset's three-level hierarchy: repeated measures (first; e.g., intrinsic motivation and experiment week), nested within students (second; e.g., FAlg) nested within a condition (third; control or experimental). That is, a model accounting for interactions between repeated measures-level and both student- and condition-level variables, as well as interactions from student- to condition-level variables. Centering variables is recommended for models with interactions [7], therefore, with did so to all independent variables before fitting the model.

To determine relevant variables, we removed each of the three-level interactions (e.g., week, FAlg and gamification) at a time, testing if some removal significantly changed model fit based on the LRT. We only tested removing the three-way interactions because terms involved in a significant interaction should be kept in the model even when the term itself is nonsignificant [7]. **H3** is similar to **H2** (involves dependent data), but no moderators were involved in this step. Therefore, the testing procedure was similar, but we only compared whether removing the single independent variable significantly decreased model fit. We adopted a more lenient alpha level (0.1) for all LRTs due to the exploratory testing of moderators of gamification's effects [12]. P-values were adjusted with the False Discovery Rate approach due to multiple comparisons [13].

4 RESULTS

H1 predicted intrinsic motivation and the number of completed quizzes would positively affect participants' learning gains. Overall regression results (N = 18; $R^2 = 0.76$; R^2-adj = 0.73; $F(2,15) = 23.61$; $p < 0.01$) and individual predictors (intrinsic motivation average: B = 1.24; SE = 0.24; $\beta = 0.66$; $p < 0.01$; sum of completed quizzes:

B = 0.07; SE = 0.02; β = 0.60; p < 0.01) were significant. Thus, suggesting strong positive effects of the predictors on learning gains, supporting **H1**.

H2 concerns gamification affecting intrinsic motivation while being moderated by context-related factors. Results from the LRTs (Table 2) demonstrate the only term to significantly affect model fit is the three-way interaction between gamification, week, and FAlg. Therefore, we followed literature recommendation [12] and fitted a new model adding only the variables and interactions involved in that significant term (Table 3). Figure 1 helps understanding the model. It shows how the intrinsic motivation (Y-axis) of those with (right) and without (left) gamification changed over time (X-axis) depending on their previous FAlg. The figure shows intrinsic motivation changed over time for all participants, and that gamification's impact was mixed, increasing from negative to positive inasmuch learners had more familiarity with algorithms at the beginning of the experiment. Therefore, suggesting that gamification affected intrinsic motivation and that this impact was moderated by intervention time and FAlg but not by other contextual factors analyzed. Thus, partially supporting **H2**.

H3 predicted intrinsic motivation would positively affect the number of completed quizzes. The LRT showed removing intrinsic motivation insignificantly changes the model fit (F(1, 67.425) = 2.01; p = 0.16), indicating this model is no better than an intercept-only one. Thus, we have no evidence intrinsic motivation affected the number of completed quizzes, failing to support **H3**.

In summary, our findings indicate that students who completed more quizzes and were more intrinsically motivated achieved higher learning gains (**H1**), that completing quizzes was unlikely driven by intrinsic motivation (**H3**), and that gamification's effects depended on intervention duration and learners' previous familiarity with algorithms (**H2**).

5 DISCUSSION

This section discusses our results from four perspectives. First, our research question. In terms of how gamification affected Brazilian undergraduates programming learning in a six-week period, we found that was accomplished by influencing their motivation. That is, gamification influenced learners' motivation (**H2**) that, in turn, positively predicted learning gains (**H1**). We also predicted gamification would indirectly affect learning via users' behaviors, which would be affected by intrinsic motivation (**H3**), however, we found no support for that. Furthermore, **H2** also predicted intervention duration and context-related factors would moderate gamification's effect on intrinsic motivation. Intervention duration and only one (FAlg) out of the four context-related factors were significant moderators. Besides changing the effect's strength, these moderators affected its direction: over time, gamification was positive when participants had at least some familiarity to algorithms but negative otherwise. Hence, the way gamification affected participants learning was through intrinsic motivation, positively or negatively affecting it over time according to learners' previous FAlg.

The second perspective is findings' relationship to previous work. Most studies applying gamification in programming education reported positive outcomes [17, 20, 22], with few null results [5, 8]. Differently, our results were mixed. A possible rationale is that most reviewed studies focused on analyzing behavioral outcomes, whereas we analyzed motivational and cognitive ones as well. Moreover, despite gamification's overall effect is positive, multiple factors moderate its success [34], likely leading to cases where results are null/negative [39]. As we analyzed moderators, we were able to

Table 2: Likelihood ratio tests assessing significant terms when modeling gamification's effect on intrinsic motivation while accounting for moderators. *gamified* dummy coded.

Terms	F (df_n, df_d)	p	p-adj
gamified:week:FAlg	7.669 (1, 59.920)	0.007	0.090
gamified:week:FC	3.236 (1, 64.686)	0.077	0.307
gamified:week:FPseu	2.161 (1, 63.285)	0.146	0.352
gamified:week:FPprog	0.001 (1, 64.439)	0.973	0.973
gamified:FAlg	2.603 (1, 15.450)	0.127	0.352
gamified:FC	0.778 (1, 17.923)	0.389	0.673
gamified:FPseu	0.779 (1, 13.652)	0.393	0.673
gamified:FProg	0.010 (1, 15.638)	0.923	0.973
week:FAlg	0.316 (1, 59.810)	0.576	0.739
week:FC	0.285 (1, 63.364)	0.595	0.739
week:FPseu	3.873 (1, 60.173)	0.054	0.307
week:FProg	0.255 (1, 59.716)	0.616	0.739

F = familiarity to; Alg = algorithms; Prog = programming; Pseu = Pseudo-language; C = C programming language.

Table 3: Multilevel model of gamification's effect on intrinsic motivation controlling for intervention duration (week) and learners' previous familiarity with algorithms (FAlg; Likert-scale). *gamified* dummy coded; * p < 0.1.

Coefficient	Est. (SE)	Coefficient	Est. (SE)
Intercept	5.64 (0.90)	gamified:week	-0.32 (0.21)
gamified	0.41 (1.24)	gamified:FAlg	-0.26 (0.47)
week	0.06 (0.15)	week:FAlg	-0.04 (0.06)
FAlg	-0.02 (0.35)	gamified:week:FAlg	0.17 (0.08)*

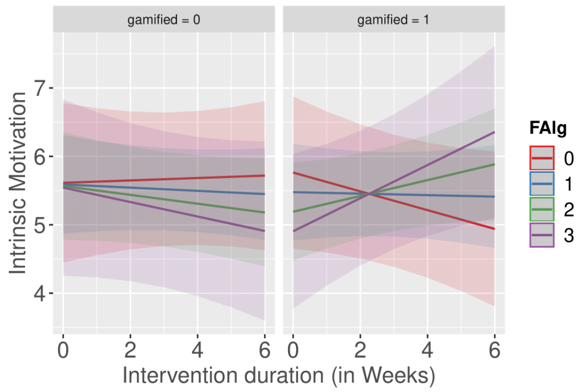

Figure 1: Effects of gamification on intrinsic motivation (Y-axis), moderated by intervention duration (X-axis) and previous familiarity with algorithms (FAlg; Likert-scale).

understand which factors affected gamification's impact, whereas a single related study performed a similar analysis [17]. Based on this context, our findings corroborate the overall gamification literature by achieving mixed outcomes [15], provide evidence for research claiming the moderator effect of context and intervention duration [10, 14, 31], and suggest the predominance of positive reports might be due to not considering moderators' impact.

The third perspective concerns our findings' implications to programming teaching. From testing our hypotheses, we found evidence towards two main directions, which poses two implications for programming learning. First, we found intrinsic motivation and the number of competed quizzes positively predicted learning gains (**H1**), further grounding previous discussions (e.g., [33, 42]) with evidence in the context of programming learning (conditionals, loops, and arrays). Therefore, the implication is instructors should seek to improve this psychological state of learners, as well as having them self-tested (e.g., completing quizzes), as this likely enhances their learning. Second, we found intervention duration moderated the impact of gamification, which was positive for those with previous familiarity to the task's topic (another moderator) but null or negative for those with no previous familiarity (**H2**). Therefore, the implication is that the decision to gamify an educational system must be made with caution, considering not only users' demographics and profiles (c.f. [23]), but also for how long it will be used as well as users' previous familiarity with the system's topics.

Lastly, we discuss three implications from our findings to future research. The first concerns results from **H1**. Despite those corroborate previous research [11, 35], further research are still needed to ground whether intrinsic motivation and the testing effect (e.g., completing quizzes) hold within the context of programming learning. The second concerns results from **H2**. The fact that the same gamification is unlikely to work for all users has been recently discussed, calling for the need of tailored gamification [40]. Within this context, recent studies have called for considering aspects related to the context and learning activities when tailoring gamification [9, 10, 30, 32, 38]. Our findings' implication to this vein is that learners' previous familiarity with the learning activity, which relates to the context, moderates how gamification impacts their intrinsic motivation. This implies research on tailored gamification should investigate these factors as tailoring gamification to such familiarity might be crucial to improve its effectiveness. The third implication relates to intrinsic motivation not driving users' behavior, unlike our expectations (**H3**). In this research, completing quizzes worth extra points (external rewards), then, students possibly were motivated to complete them due to extrinsic motivation [3]. The implication for future research, therefore, is that studies should seek to motivate learners participation through intrinsic rather than extrinsic approaches.

5.1 Limitations and Threats to Validity

First, our sample is restricted in size (19), limiting our findings' generalization. We opted for this approach to perform the study within a real class, which is costly and hard to perform with large samples. Additionally, the sample concerns a single class, which might lead to groups' contamination (information from one group leaking to another). We chose to study a single class to increase the study internal validity, guaranteeing all participants would learn from the same instructor and lessons, increasing the chances that differences are due to gamification. Second, there were missing data (38%) in the motivational outcomes, possibly because completing the scales did not worth extra points, unlike completing quizzes. Multilevel analysis handle such missings on the dependent variable, but in independent variables the common approach is deletion [12]. Hence, threatening our conclusion validity only with regards to **H3**. Moreover, the limited sample size also impacted the data analyses (e.g., low statistical power, possibly inaccurate estimates). To address this, we focused the analyses on fixed effects, as they can be estimated with smaller samples than random effects [12]. Third, there is the learning gains measuring. We opted for measuring this construct through pre- and post-tests, as recommended in the literature [34], in which we used the same test in both occasions. As there was a 42 days interval between the tests, we believe threats related to memorizing test's items were mitigated. Additionally, the test was validated by the discipline instructor, guaranteeing it was aligned to and measured the topics approached during the experiment. Fourth, there was no control over how/where/when participants completed the experiment's task. This approach increases external validity, as this freedom is similar to when students are given homework or optional tasks. However, as participants are likely to have distinct routines and livings, which might have affected our results. Fifth, the intervention only lasted for half semester due to restrictions from the university and the effort needed by the discipline instructor. Although our results suggest how gamification's impact would change in a longer intervention, this can only be ensured with more research. Lastly, during the intervention, the class changed from face-to-face to online due to covid-19 quarantine. The between-subject design mitigated this threat as participants continued using the same condition regardless of the change.

6 CONCLUSIONS

Students often lack motivation to learn to program, jeopardizing their learning. Gamification can aid with this issue, but the understanding of how it contributes to programming learning, and which aspects moderate that contribution, is scarce. Thus, we conducted a longitudinal experiment with Brazilian undergraduate students, comparing the motivation, behavior, and learning gains of those interacting with gamified quizzes to those engaged with non-gamified ones. Mainly, we found that gamification contributed to students' programming learning via intrinsic motivation and that this effect changed as intervention duration time increased, decreasing from positive to negative inasmuch learners had less familiarity with programming.

In summary, our main contributions are i) empirical evidence revealing through which construct gamification affected programming learning, ii) when/to whom that effect was positive or negative, and iii) a theory-grounded gamification design likely to improve the learning of undergraduates with previous familiarity to programming after a six-week use. As future works, we recommend conducting similar experiments with different samples to ground our findings, developing/testing other gamification designs aiming to mitigate cases of negative effects, and advancing the understanding of moderators of gamification's success.

ACKNOWLEDGMENTS

This research was partially funded by CNPq, CAPES, and FAPESP (Projects 2016/02765-2; 2018/11180-3; 2018/15917-0; 2018/07688-1).

REFERENCES

[1] Raad A Alturki et al. 2016. Measuring and improving student performance in an introductory programming course. *Informatics in Education-An International Journal* 15, 2 (2016), 183–204.
[2] Shurui Bai, Khe Foon Hew, and Biyun Huang. 2020. Is gamification "bullshit"? Evidence from a meta-analysis and synthesis of qualitative data in educational contexts. *Educational Research Review* (2020), 100322.
[3] Edward L. Deci and Richard M. Ryan. 2000. The "What" and "Why" of Goal Pursuits: Human Needs and the Self-Determination of Behavior. *Psychological Inquiry* 11, 4 (2000), 227–268. https://doi.org/10.1207/S15327965PLI1104_01 arXiv:https://doi.org/10.1207/S15327965PLI1104_01
[4] Paul Denny, Fiona McDonald, Ruth Empson, Philip Kelly, and Andrew Petersen. 2018. Empirical Support for a Causal Relationship Between Gamification and Learning Outcomes. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems* (Montreal QC, Canada) *(CHI '18)*. Association for Computing Machinery, New York, NY, USA, 1–13. https://doi.org/10.1145/3173574.3173885
[5] Panagiotis Fotaris, Theodoros Mastoras, Richard Leinfellner, and Yasmine Rosunally. 2016. Climbing up the Leaderboard: An Empirical Study of Applying Gamification Techniques to a Computer Programming Class. *Electronic Journal of e-learning* 14, 2 (2016), 94–110.
[6] MN Gari, GS Walia, and AD Radermacher. 2018. Gamification in computer science education: A systematic literature review. In *American Society for Engineering Education*.
[7] Andrew Gelman and Jennifer Hill. 2006. *Data analysis using regression and multilevel/hierarchical models*. Cambridge university press.
[8] Lasse Hakulinen, Tapio Auvinen, and Ari Korhonen. 2015. The Effect of Achievement Badges on Students' Behavior: An Empirical Study in a University-Level Computer Science Course. *International Journal of Emerging Technologies in Learning* 10, 1 (2015).
[9] Stuart Hallifax, Audrey Serna, Jean-Charles Marty, and Élise Lavoué. 2019. Adaptive Gamification in Education: A Literature Review of Current Trends and Developments. In *European Conference on Technology Enhanced Learning*. 294–307.
[10] Stuart Hallifax, Audrey Serna, Jean-Charles Marty, Guillaume Lavoué, and Elise Lavoué. 2019. Factors to Consider for Tailored Gamification. In *Proceedings of the Annual Symposium on Computer-Human Interaction in Play* (Barcelona, Spain) *(CHI PLAY '19)*. 559–572. https://doi.org/10.1145/3311350.3347167
[11] Michael D. Hanus and Jesse Fox. 2015. Assessing the effects of gamification in the classroom: A longitudinal study on intrinsic motivation, social comparison, satisfaction, effort, and academic performance. *Computers & Education* 80 (2015), 152 – 161. https://doi.org/10.1016/j.compedu.2014.08.019
[12] Joop J Hox, Mirjam Moerbeek, and Rens Van de Schoot. 2010. *Multilevel analysis: Techniques and applications*. Routledge.
[13] Mohieddin Jafari and Naser Ansari-Pour. 2019. Why, when and how to adjust your P values? *Cell Journal (Yakhteh)* 20, 4 (2019), 604.
[14] Jonna Koivisto and Juho Hamari. 2014. Demographic differences in perceived benefits from gamification. *Computers in Human Behavior* 35 (2014), 179 – 188. https://doi.org/10.1016/j.chb.2014.03.007
[15] Jonna Koivisto and Juho Hamari. 2019. The rise of motivational information systems: A review of gamification research. *International Journal of Information Management* 45 (2019), 191 – 210. https://doi.org/10.1016/j.ijinfomgt.2018.10.013
[16] David R Krathwohl. 2002. A revision of Bloom's taxonomy: An overview. *Theory into practice* 41, 4 (2002), 212–218. https://doi.org/10.1207/s15430421tip4104_2
[17] Markus Krause, Marc Mogalle, Henning Pohl, and Joseph Jay Williams. 2015. A Playful Game Changer: Fostering Student Retention in Online Education with Social Gamification. In *Proceedings of the Second (2015) ACM Conference on Learning @ Scale* (Vancouver, BC, Canada) *(L@S '15)*. 95–102.
[18] Max Kuhn and Kjell Johnson. 2013. *Applied predictive modeling*. Vol. 26. Springer.
[19] Richard N Landers, Elena M Auer, Andrew B Collmus, and Michael B Armstrong. 2018. Gamification science, its history and future: Definitions and a research agenda. *Simulation & Gaming* 49, 3 (2018), 315–337.
[20] B. Marín, J. Frez, J. Cruz-Lemus, and M. Genero. 2018. An Empirical Investigation on the Benefits of Gamification in Programming Courses. *ACM Trans. Comput. Educ.* 19, 1, Article 4 (Nov. 2018), 22 pages. https://doi.org/10.1145/3231709
[21] Daniel Mirman. 2016. *Growth curve analysis and visualization using R*. CRC press.
[22] Julian MORENO and Andres F PINEDA. 2018. Competitive programming and gamification as strategy to engage students in computer science courses. *Revista ESPACIOS* 39, 35 (2018).
[23] Benedikt Morschheuser, Lobna Hassan, Karl Werder, and Juho Hamari. 2018. How to design gamification? A method for engineering gamified software. *Information and Software Technology* 95 (2018), 219–237.
[24] IT Chan Mow. 2008. Issues and difficulties in teaching novice computer programming. In *Innovative techniques in instruction technology, e-learning, e-assessment, and education*. Springer, 199–204.
[25] Paula Palomino, Armando Toda, Wilk Oliveira, Luiz Rodrigues, and Seiji Isotani. 2020. From the Lack of Engagement to Motivation: Gamification Strategies to Enhance Users Learning Experiences. In *2020 19th Brazilian Symposium on Computer Games and Digital Entertainment (SBGames) - GranDGamesBR Forum*.
[26] Paula T Palomino, Armando M Toda, Wilk Oliveira, Luiz Rodrigues, and Seiji Isotani. 2019. Gamification journey: A Novel approach for classifying gamer types for gamified educational systems. *2019 18th Brazilian Symposium on Computer Games and Digital Entertainment (SBGames)* (2019), 165–173.
[27] Laís Zagatti Pedro. 2016. *Uso de gamificação em ambientes virtuais de aprendizagem para reduzir o problema da externalização de comportamentos indesejáveis*. Ph.D. Dissertation. Universidade de São Paulo.
[28] Yizhou Qian and James Lehman. 2017. Students' Misconceptions and Other Difficulties in Introductory Programming: A Literature Review. *ACM Trans. Comput. Educ.* 18, 1, Article 1 (Oct. 2017), 24 pages.
[29] Luiz Rodrigues and Jacques Duílio Brancher. 2019. Playing an Educational Game Featuring Procedural Content Generation: Which Attributes Impact Players' Curiosity? *Journal New Technologies on Education (Revista Novas Tecnologias na Educação - RENOTE)* (2019).
[30] Luiz Rodrigues, Wilk Oliveira, Armando Toda, Paula Palomino, and Seiji Isotani. 2019. Thinking Inside the Box: How to Tailor Gamified Educational Systems Based on Learning Activities Types. In *Proceedings of the Brazilian Symposium on Computers in Education (Simpósio Brasileiro de Informática na Educação - SBIE)*.
[31] Luiz Rodrigues, Armando Toda, Wilk Oliveira, Paula Palomino, and Seiji Isotani. 2020. Just beat it: Exploring the influences of competition and task-related factors in gamified learning environments. In *Proceedings of the Brazilian Symposium on Computers in Education (Simpósio Brasileiro de Informática na Educação-SBIE)*, Vol. 31.
[32] Luiz Rodrigues, Armando M. Toda, Paula T. Palomino, Wilk Oliveira, and Seiji Isotani. 2020. Personalized gamification: A literature review of outcomes, experiments, and approaches. In *Proceedings of the 8th International Conference on Technological Ecosystems for Enhancing Multiculturality (TEEM 2020) (Salamanca, Spain, October 21-23, 2020)*, F. J. García-Peñalvo (Ed.).
[33] Christopher A Rowland. 2014. The effect of testing versus restudy on retention: a meta-analytic review of the testing effect. *Psychological Bulletin* 140, 6 (2014), 1432.
[34] Michael Sailer and Lisa Homner. 2019. The Gamification of Learning: a Meta-analysis. *Educational Psychology Review* (15 Aug 2019).
[35] Diana R. Sanchez, Markus Langer, and Rupinder Kaur. 2020. Gamification in the classroom: Examining the impact of gamified quizzes on student learning. *Computers & Education* 144 (2020), 103666.
[36] Isabelle Savard and Riichiro Mizoguchi. 2019. Context or culture: what is the difference? *Research and Practice in Technology Enhanced Learning* 14, 1 (2019).
[37] Armando Toda, Wilk Oliveira, Lei Shi, Ig Ibert Bittencourt, Seiji Isotani, and Alexandra I. Cristea. 2019. Planning gamification strategies based on user characteristics and DM : a gender-based case study. In *Proceedings of the 12th International Conference on Educational Data Mining.*, Michel Desmarais, Collin F. Lynch, Agathe Merceron, and Roger Nkambou (Eds.). Educational Data Mining 2019, Montréal, Canada, 438–443. http://dro.dur.ac.uk/28609/
[38] Armando Toda, Filipe Dwan Pereira, Ana Carolina Tomé Klock, Luiz Rodrigues, Paula Palomino, Wilk Oliveira, and Elaine Oliveira. 2020. For whom should we gamify? Insights on the users intentions and context towards gamification in education. In *Proceedings of the Brazilian Symposium on Computers in Education (Simpósio Brasileiro de Informática na Educação-SBIE)*, Vol. 31.
[39] Armando M. Toda, Pedro H. D. Valle, and Seiji Isotani. 2018. The Dark Side of Gamification: An Overview of Negative Effects of Gamification in Education. In *Higher Education for All. From Challenges to Novel Technology-Enhanced Solutions*, Alexandra Ioana Cristea, Ig Ibert Bittencourt, and Fernanda Lima (Eds.). Springer International Publishing, Cham, 143–156.
[40] Rob van Roy and Bieke Zaman. 2017. Why gamification fails in education and how to make it successful: introducing nine gamification heuristics based on self-determination theory. In *Serious Games and edutainment applications*. Springer, 485–509.
[41] Rob Van Roy and Bieke Zaman. 2018. Need-supporting gamification in education: An assessment of motivational effects over time. *Computers & Education* 127 (2018), 283–297.
[42] Maarten Vansteenkiste, Eline Sierens, Bart Soenens, Koen Luyckx, and Willy Lens. 2009. Motivational profiles from a self-determination perspective: The quality of motivation matters. *Journal of educational psychology* 101, 3 (2009), 671.
[43] Zamzami Zainuddin, Samuel Kai Wah Chu, Muhammad Shujahat, and Corinne Jacqueline Perera. 2020. The impact of gamification on learning and instruction: A systematic review of empirical evidence. *Educational Research Review* (2020), 100326.

ANNEX

D

ARE THEY LEARNING OR PLAYING? MODERATOR CONDITIONS OF GAMIFICATION'S SUCCESS IN PROGRAMMING CLASSROOMS

Are they learning or playing? Moderator conditions of gamification's success in programming classrooms

LUIZ RODRIGUES, University of SÃčo Paulo, Brazil
FILIPE PEREIRA, Federal University of Roraima, Brazil and Durham University, United Kingdom
ARMANDO TODA, University of SÃčo Paulo, Brazil and Durham University, United Kingdom
PAULA PALOMINO, University of SÃčo Paulo, Brazil and University of Waterloo, Canada
WILK OLIVEIRA, University of SÃčo Paulo, Brazil and Tampere University, Finland
MARCELA PESSOA, Amazonas State University, Brazil, and Federal University of Amazonas, Brazil
LEANDRO CARVALHO, Federal University of Amazonas, Brazil
DAVID OLIVEIRA, Federal University of Amazonas, Brazil
ELAINE OLIVEIRA, Federal University of Amazonas, Brazil
ALEXANDRA CRISTEA, Durham University, United Kingdom
SEIJI ISOTANI, University of SÃčo Paulo, Brazil

Students face several difficulties in introductory programming courses (CS1), often leading to high dropout rates, student demotivation, and lack of interest. The literature has indicated that the adequate use of gamification might improve learning in several domains, including CS1. However, the understanding of which (and how) factors influence gamification's success, especially for CS1 education, is lacking. Thus, there is a clear need to shed light on *pre-determinants of gamification's impact*. To tackle this gap, we investigate how user and contextual factors influence gamificationâĂŹs effect on CS1 students through a quasi-experimental retrospective study ($N = 399$), based on a between-subject design (conditions: gamified or non-gamified) in terms of final grade (academic achievement) and the number of programming assignments completed in an educational system (i.e., how much they practised). Then, we evaluate whether and how user and contextual characteristics (such as age, gender, major, programming experience, working situation, internet access, and computer access/sharing) moderate that effect. Our findings indicate that gamification amplified to some extent the impact of practising. Overall, students practising in the gamified version presented higher academic achievement than those practising the same amount in the non-gamified version. Intriguingly, those in the gamified version that practised much more extensively than the average showed lower academic achievements than those who practised comparable amounts in the non-gamified version. Furthermore, our results reveal gender as the only statistically significant moderator of gamification's effect: in our data, it was positive for females, but nonsignificant for males. These findings suggest which (and how) personal and contextual factors moderate gamification's

Authors' addresses: Luiz Rodrigues, lalrodrigues@usp.br, University of SÃčo Paulo, Brazil; Filipe Pereira, filipe.dwan@ufrr.br, Federal University of Roraima, Brazil and Durham University, United Kingdom; Armando Toda, armando.toda@usp.com, University of SÃčo Paulo, Brazil and Durham University, United Kingdom; Paula Palomino, paulatpalomino@usp.br, University of SÃčo Paulo, Brazil and University of Waterloo, Canada; Wilk Oliveira, wilk.oliveira@usp.br, University of SÃčo Paulo, Brazil and Tampere University, Finland; Marcela Pessoa, msppessoa@uea.edu.br, Amazonas State University, Brazil, and Federal University of Amazonas, Brazil; Leandro Carvalho, galvao@icomp.ufam.edu.br, Federal University of Amazonas, Brazil; David Oliveira, david@icomp.ufam.edu.br, Federal University of Amazonas, Brazil; Elaine Oliveira, elaine@icomp.ufam.edu.br, Federal University of Amazonas, Brazil; Alexandra Cristea, alexandra.i.cristea@durham.ac.uk, Durham University, United Kingdom; Seiji Isotani, sisotani@icmc.usp.br, University of SÃčo Paulo, Brazil.

1946-6226/2022/1-ART1 $15.00
https://doi.org/10.1145/3485732

effects, indicate the need to further understand and examine context's role, and show gamification must be cautiously designed to prevent students from playing instead of learning.

CCS Concepts: • **General and reference → Empirical studies**; • **Social and professional topics → Computing education**.

Additional Key Words and Phrases: Gamified Learning, Testing Effect, Moderation, Gaming the system, Context

1 INTRODUCTION

Learning to code is challenging and demands significant effort from the learners [75]. Research on introductory programming (i.e., CS1 [33]) has been conducted for several decades and, up to date, findings report high failure and dropout rates in these courses [8–10, 45]. Furthermore, CS1 is compulsory not only for computing-related majors but for STEM[1] courses as well [22, 69]. For students of the latter majors, the problem is enhanced, as students often lack affinity with programming and even fail to see its value for their professional lives [21, 87]. As such, students' lack of interest and effort negatively contribute to their achievement, given that learning to code requires significant amounts of practice [38, 75, 103].

Recently, there is increasing evidence that the need for motivating students to practice programming might be addressed with gamification: adding game elements into non-gaming contexts [17]. It has been widely used in the educational domain [52], with empirical results reporting overall positive outcomes in, for instance, academic performance [101]. Reasons for such positive effects include allowing goal-setting, providing performance feedback and recognition, and fostering enthusiasm [4]. To that end, it is important to design gamification to drive the expected motivational and behavioural outcomes [96], as well as enhance learning in the educational context [46]. Otherwise, the game elements might jeopardise students' learning, such as in cases when too much engagement with the system leads to behaviours such as gaming the system, distraction, or lack of utility [5, 86, 94].

1.1 Problem

Gamification's success is assumed to depend on multiple factors, such as who will interact with it (i.e., its users) and the context[2] in which it will be used [32, 50, 90]. Advancing the understanding regarding such moderators is important, to shed light on which aspects predetermine gamification's success, as well as how each one acts (i.e., maximising or minimising it) [43]. However, the precise factors that moderate gamification's effectiveness are not well known [85], especially in the context of CS1, where empirical studies assessing gamification's impact often lack moderator analyses [24, 53, 58]. Therefore, despite some of the existing literature arguing that gamification depends on factors related to the user (e.g., gender and age) and context (e.g., the environment or circumstance) [4, 31, 77], there is a gap in the understanding of which factors moderate gamification's success when applied to CS1 learning contexts. Thus, we tackle this gap by answering the following research question:

- **RQ:** *How do user and contextual factors influence the effect of gamification on the academic achievement of CS1 students?*

Given the current literature on gamification applied to CS1 education, we expand it, by evaluating *how* gamification improves CS1 students' academic achievement, assuming this happens by influencing their behaviours, along with an analysis of which contextual and demographic factors moderate that effect. Additionally, for the gamification design, we used fictional, social, and challenge-based game elements [93], whereas previous similar studies only focused on the last two kinds (e.g., [19, 44, 53]). Featuring fictional game elements is valuable, as a

[1]Science, Technology, Engineering, and Mathematics
[2]Resources, methods, people's mental representations, environment, and circumstance involved in an activity (see Section 1.3.3).

recent meta-analysis found those to maximise gamification's impact on learning outcomes [85]. Thus, the overall contribution of this study is *to offer empirical evidence revealing which user and contextual characteristics moderate the impact of a gamification design featuring fictional game elements on CS1 students' academic achievement, as well as how those moderators act, that is, maximising or minimising that impact.*

1.2 Literature Review

Most empirical research applying gamification to programming learning focuses on *whether* gamification had a positive effect on learning or not, compared to no gamification. In that context, [30] added badges to an eight-week-long Data Structures and Algorithms course, which resulted in a positive effect on studentsâĂŹ time-on-system, but not on the number of completed exercises, when compared to the condition with no badges. Similarly, [23] used Kahoot! and Codeacademy to gamify a 12-week-long programming course. Based on descriptive analyses, they found positive results, compared to the course's previous run that was not gamified, mainly in terms of final grades and attendance. [53] used the UDPiler to gamify a four-week-long C programming course. By analysing data collected for two semesters, they found positive results on learning performance from gamification usage. [19] used the OneUp platform to gamify a Data Structures (semester-long) course. Their findings suggested gamification was positive in terms of the number of challenge attempts and students' final grades. [57] gamified QueryCompetition by adding points and leaderboards and compared it to a nongamified version in terms of students' performance, motivation, and user experience. Based on a five-week intervention featuring pre- and post-tests, they found positive results favouring gamification, especially in performance. These studies support the claim that gamification has the potential to positively influence learner behaviour. However, they do not contribute to understanding what factors moderate the impact of gamification.

In contrast, few studies analyse moderators of the impact of gamified programming learning. For instance, [29] analysed the role of two factors: achievement goal orientation and motivation towards badges. They found a relationship among those factors, indicating that students with different goal orientations have distinct motivations towards badges. However, their findings suggested that those factors did not moderate students' behaviour in the gamified system. Two points of this study that must be noted are the intervention duration (half-semester) and the use of a single game element (badges). In another research, [44] evaluated whether learners' gender (male or female) and major (computer science or psychology) moderated gamification's impact on student retention, quiz accuracy, and test performance. They found no significant interactions between conditions (e.g., gamified and non-gamified) and moderators, suggesting neither gender nor major moderated gamification's effect.

Similarly, [64] assessed the role of three possible moderators: gender, major, and gaming experience. Again, their empirical findings suggested that the gamification's impact was not moderated by any of the three factors analysed. [2] studied the role of two moderators - group size and time - on students' performance and satisfaction. Based on a 16-week data collection involving 229 participants, they found both factors moderating the gamification's effect, alone and together, on both outcomes. [79] also evaluated the influence of two moderators on gamification's effect on intrinsic motivation: usage time and previous affinity to the content to be learned. They found that both moderators affected the gamification's impact only when considered together. Two limitations of [44], [64], and [79] must be acknowledged, though: the limited sample sizes - 71, 102, and 19, respectively - and intervention duration - four, four, and six weeks, respectively. On the other hand, [2] contributed to understanding a moderator's effect, but did not use fictional game elements and focused on a moderator related to the gamification design. Hence, while it advances the understanding of gamification design, it does not provide evidence on how contextual and situational factors predeterminate the effectiveness of gamification.

We showed thus that most previous studies failed to analyse moderators of gamification's success and that, those that did, are limited in terms of sample size, intervention duration, moderator's nature (i.e., user or context-related), and/or gamification design. In this study, we tackle these limitations, by presenting a 15-week empirical study analysing moderators of gamification's success, based on a wide sample of 399 learners and a well-planned gamification design. Furthermore, all of the reviewed studies explored social (e.g., collaboration) and challenge-based (e.g., challenges and rewards) gamification. Besides game elements similar to those, the gamification design employed in this study also features fictional elements [93]. Meta-analytic evidence indicates that this kind of game element is likely to improve gamification's effectiveness [85]. Hence, we expand the literature, by evaluating moderators of gamification's effectiveness based on a different design.

Lastly, studies performing moderator analysis [2, 44, 64] indicate gamificationâĂŹs impact does not depend on user characteristics (e.g., gender and goal orientation) and contextual factors (e.g., student major). These results contradict empirical findings from non-programming contexts. For instance, [67] found that gamification only worked for boys; that is, a moderator effect of gender. In contrast, [72] found that gamification influences women and older people more. Similarly, results from [47] suggest gamification only works for people with good attitudes towards it. Accordingly, it has been noted that studies must control contextual characteristics, such as the educational level of the learners, when analysing the gamification's effect [43]. Empirical findings suggest the relevance of other contextual factors, such as the moderator role of the geographic location [4, 81] and the motivational improvements from considering the learning task when designing gamification [78]. Some literature reviews also discuss contextual factors in general terms, such as usage domain (e.g., [31, 32]). However, because of the broadness of those factors, they are likely products of more specific ones [49]. Then, the contradiction might be attributed, for instance, to not finding and studying the most relevant moderators. Thus, this demonstrates the need for empirical studies investigating other moderators, to reveal which factors affect gamificationâĂŹs success [43, 85]. Table 1 contrasts the present study to those reviewed in this section, summarising the main points discussed here.

Table 1. Summary of related work on gamification applied to computing education.

Study, year	Moderator analysis?	Uses game fiction?	Intervention duration	Sample size
[30], 2015			8 weeks	281
[23], 2016			12 weeks	106
[53], 2018			16 weeks	817
[19], 2019			16 weeks	27
[57], 2020			5 weeks	139
[29], 2014	X		8 weeks	278
[44], 2015	X		8 weeks	71
[64], 2019	X		4 weeks	102
[2], 2020	X		16 weeks	229
[79], 2021	X		6 weeks	19
Our study	X	X	15 weeks	399

1.3 Hypothesis

As our RQ concerns understanding moderators of gamification's effect on academic achievement, we first need to test for such effects. For that, we rely on the Theory of Gamified learning [46], which is considered a framework suitable to understand how gamification acts according to recent research [85]. That framework advocates that for gamification to affect outcomes such as academic achievement, it affects user behaviour. Therefore, to answer our RQ, we first needed to consider the behavioural source of its effect, which we hypothesise to be *practising* (H2). Similarly, we need to ensure that practising is working as expected (H1). Lastly, we needed to test the user and contextual moderators (H3) to answer our RQ. Based on that, we discuss the relevant literature that supports our research model next.

1.3.1 Testing effect. The claim that learning to code requires practising can be supported by the theory of the testing effect. Also referred to as test-enhanced learning, it is concerned with the fact that long-term memory is often improved when learners dedicate some of their studying time to retrieve the information they expect to be remembered [27]. For instance, that might be achieved by completing quizzes or, in a more elaborated way, by retrieving programming information presented in lectures, to elaborate when performing problem-solving activities. Overall results for the testing effect are positive (e.g., [54, 66]), as the literature shows that learners who study and are tested (like in a school test) present higher long-term knowledge retention compared to those that studied but were not tested [83]. Furthermore, in a meta-analysis comparing restudy (i.e., reading again) to testing, Rowland [84] found the testing effect on recalling tasks (e.g., short answers) is much larger. Despite the fact that this effect is commonly studied in simple tasks, such as quizzes and short answers, the literature supports the effectiveness of test-enhanced learning for contents that are highly related to others (e.g., you need to know conditionals to understand loops) as well [40].

As programming presents learners with difficulties in aspects such as natural language, syntax, and abstraction [73], the need for practising can be related to improving long-term memory about these aspects, as well as receiving feedback on their programs, that is, testing themselves concerning their ability to code. Additionally, it has been claimed that students learn to code by doing (e.g., [99]). Accordingly, empirical evidence supports that need. For instance, [79] shows that the more students engaged with quizzes, the higher were their learning gains. Similarly, [70] demonstrated that the more students practised, the higher were their performances. Nevertheless, such effects are less known for complex materials, showing the need for empirical research to further examine it [74]. Consequently, we might expect that the more students practice, or test themselves, the more they will be successful in programming. Thus, our first hypothesis H1 is:

H1: Practising positively affects academic achievement in CS1.

1.3.2 Gamified Learning. Gamification has been widely explored within the educational domain, based on beliefs that it can improve, for instance, learners' engagement, motivation, and learning [18, 52, 65]. Consequently, several empirical studies assessing its effectiveness emerged, some of which have been recently summarised in secondary studies. Sailer and Homner [85] presents a meta-analysis of gamification's impact on three kinds of learning outcomes, in which they found it has, overall, small positive effects for motivational (e.g., intrinsic motivation), behavioural (e.g., performance), and cognitive (e.g., conceptual knowledge) learning outcomes. Results of the meta-analysis by [35] corroborate those findings, demonstrating a positive impact from gamification on cognitive learning outcomes. In another recent meta-analysis, Bai et al. [4] focused on gamification's effect on academic performance, finding it has a small-to-moderate positive influence.

Additionally, Bai et al. [4] summarised reasons for students enjoying gamification. These include inciting enthusiasm, providing performance feedback and means to be recognised (e.g., badges), and goal-setting. While enthusiasm is likely to motivate people to use the gamified system, goal-setting is valuable to drive performance [97], providing feedback is highly important for learning programming [68, 75, 103], and recognition is likely to incite feelings of self-efficacy and fulfil competence needs, aspects that are also positively related to academic achievement and performance [36, 71, 91, 102]. This demonstrates the importance of defining gamification designs that support the desired behaviours/outcomes (e.g., enhancing learning by fulfilling competence needs) and minimises side effects (e.g., gaming the system rather than using it to study). Hence, the overall literature provides evidence of the potential of well-designed gamification to contribute to learning, demonstrating it can affect cognition, motivation, and behaviour and indicating reasons for those effects to incite self-efficacy and fulfil competence needs as well as goal-setting.

Thereby, we might expect using gamification will enhance programming practice, consequently improving academic achievement, and formulate hypothesis H2 as:

H2: Gamification enhances the testing effect and, consequently, academic achievement.

1.3.3 Moderators of gamification's success. Although gamification has overall positive impacts on learning outcomes, there are cases in which results are null or negative [37, 94]. A recurrent justification for those outcomes is the quality of the gamification design [15, 51]. According to many discussions, what leads to such low-quality gamification designs is the lack of consideration of user and contextual characteristics (e.g., [18, 32, 60, 89, 90]). That is, many of such characteristics moderate whether gamification will be effective for a given user in a given context, hence, when those are not taken into account, it is likely that effectiveness will not be achieved for some users [48].

According to those discussions, studies have shown empirical support for the influence of users' characteristics regarding their experiences, perceptions, and preferences. For instance, research has demonstrated that people with different gender, age, socio-economic conditions, behavioural profiles, and personality traits have different preferences, perceptions, and experiences, even when doing the same task [12, 29, 49, 56, 62, 63, 80, 86, 92, 95]. Moreover, context-related aspects are acknowledged as relevant moderators of user experience as well [4, 32]. However, studies often understand context differently and define it in general terms (e.g., based on usage domain/aim [31, 43]). In contrast, in the scope of this paper, we interpret the context to involve resources (e.g., gamified system) and methods involving human activity (e.g., task to be done in a gamified system) [100]. Further, context might be seen as internal and external: the former relates to usersâĂŹ mental representations (e.g., user characteristics that influence the learning experience) while the latter relates to the environment/circumstance (e.g., where the activity takes place) [88]. Thus, by relying on context-specific definitions (i.e., from Human-Computer Interaction and Educational Technology literature), we can study and understand contextual factors in a fine-grained, comprehensive way.

However, empirical analyses of the role of contextual aspects are rare, although recent research highlights their importance and lack of studies in this regard [31, 41, 43, 50, 77]. For instance, research has explored majorâĂŹs moderator effect [44, 64], which fits within the external context, but found no significant effects. On the other hand, [79] found initial evidence on the moderator effect of a factor related to the internal context: oneâĂŹs previous affinity to the content to be learned, when considered together with intervention duration. Hence, whereas the literature acknowledges there are multiple factors likely to moderate gamification's effectiveness, evidence from related work is limited and suggests we need to explore new factors that could explain the role of context on the effects of gamification. Thus, highlighting the need for studies to confirm, discover, and better understand the role of those factors [36, 85]. Moreover, understanding whether these factors' moderator effect will be positive or negative, as well as its magnitude, is even more uncertain. Therefore, in a more exploratory hypothesis, we expect that user and contextual characteristics will moderate the effects of gamification, as predicted by **H3**, with no assumption on the direction (positive or negative) and the magnitude of these moderators:

H3: Gamification's effects will be moderated by a) user and b) contextual characteristics.

The research model illustrating this study's hypotheses is shown in Figure 1. It shows the testing effect prediction (**H1**), the assumption that gamification will enhance academic achievement by improving the testing effect (**H2**), and the expectation that user and contextual characteristics will moderate the gamification's impact (**H3**).

2 METHOD

This is a retrospective study, as we examine data captured in the past that was made available for analysis by Federal University of Amazonas (UFAM in Portuguese) from Brazil.

Collab - retrospective study]Research Model - Simplified.pdf

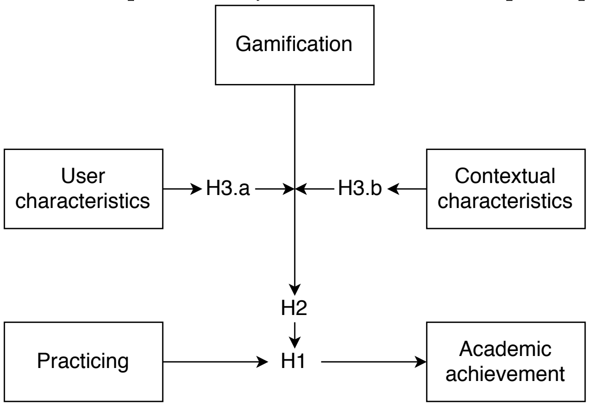

Fig. 1. Study research model on Gamification in CS1, hypothesising that i) academic achievement is positively affected by practice ii) and by gamification, iii) effect that is moderated by user and contextual characteristics.

2.1 Inclusion and Exclusion

As this is a retrospective study, we prepared the dataset (N = 1309) in four steps. First, we removed participants (325) that did not provide consent to participating in the research. Second, we removed participants (30) that completed the characterisation survey with unrealistic values (e.g., age of 4). Third, we removed participants (198) that dropped out due to two reasons. First, to ensure analysing data from subjects that participated in the whole semester, which is necessary because gamificationâĂŹs effect might decline with time [4, 85]. Therefore, analysing it based on long usage periods is imperative to achieve reliable findings. Second, our dependent variable (see Section 2.3) depends on several assignments completed from semesterâĂŹs week two to week 14. Importantly, those assignmentsâĂŹ weights increase progressively. Therefore, we had to remove those who dropped out because their final gradesâĂŹ measures would be misleading. Lastly, we removed data from majors in which the number of subjects in any condition (control or experimental) was five or less (357)[3].

2.2 Participant Characteristics

Our analysis concerns data from 399 CS1 students of seven majors of the UFAM. Majors are Materials Engineering, Electrical Engineering, Mechanical Engineering, Statistics, Physics, Mathematics, and Applied Mathematics. Although different, all majors followed the same methodological plan and pedagogical materials, and all activities offered for participants to practice were selected from the same database. The participants are 399 learners (64.2% males, 35.8% females) with an average age of 22.1 years (±3.6) who self-reported whether they had i) previous experience with any programming language (yes: 37.3%; no: 62.7%), ii) worked/interned before (yes: 20.1%; no: 79.9%), iii) had internet at home (yes: 82%; no: 18%), iv) a computer at home (yes: 87.7%; no: 12.3%), and whether they shared their home computer with someone else (yes: 27.3%; no: 72.7%).

Table 2 presents a comparison between control and experimental groups regarding their categorical demographic characteristics. Two significant differences were found: control group was more experienced (p-adj < 0.05) and worked/interned before the degree less (Worked; p-adj < 0.01) than the experimental one. Also, the experimental group (M = 21.84; ± 3.87) was younger than the control one (M = 22.78; ± 2.80), W = 23073; p < 0.001;

[3]The number of students removed in this step is large, because many majors were highly unbalanced among conditions (i.e., many students in one condition, very few in the other) because we analysed majors in **H3**.

CI = 1.000-2.000. While comparing groups with different characteristics might affect the results, we handled this limitation by inserting all these variables as covariates during the data analysis process. Then, if some difference was to be found due to a demographic characteristic (e.g., experience) rather than condition (i.e., gamification), the analysis would reveal it.

Table 2. Comparison of demographic data from study groups (i.e., control - Ctr, no gamification; experimental - Exp, gamification). Data represented as percentages; p-values adjusted using the False Discovery Rate approach [39]; all comparisons' degree of freedom was one.

	Gender	Has PC?	Shares PC?	Int?	Exp?	Worked?
	Mal/Fem.	No/Yes	No/Yes	No/Yes	No/Yes	No/Yes
Ctr	56/44	12/88	70/30	18/82	47/53	73/23
Exp	58/42	13/87	74/26	18/82	69/31	83/17
χ^2	0.927	0.000	0.311	0.000	18.439	5.222
P-val	0.336	1.000	0.577	1.000	0.000	0.030
P-adj	0.672	1.000	0.865	1.000	0.001	0.089

Has/Shares PC = whether the participant has/shares a PC at home; Int = whether the participant has internet at home; Exp = whether the participant has experience with any programming language; Worked = whether the participants worked/internet before the degree.

2.3 Measures and Covariates

We analysed two measures that concern a whole semester: academic achievement and practising to code. The former is the study dependent variable, measured as a student's final grade (Equation 1).

$$FG_s = \frac{(Ex_{t1} + Ex_{t2}).1 + (Ex_{t1} + Ex_{t2}).2 + (Ex_{t1} + Ex_{t2}).3 + \frac{PA_{t1}+\ldots+PA_{t7}}{7}}{16} \tag{1}$$

A final grade (*FG*) is calculated based on the student's (*s*) scores on the programming assignment (*PA*) and their exams' marks (at each two weeks, the students were required to take an exam − Ex − about the same topic of the PA.). As the content of programming is cumulative, the weights of the seven exams increased over the topics. The students' groups (control or experimental) did not affect how their final grades were calculated.

The latter − practising to code − acted as a proxy to the former, measuring the extent to which learners practised programming. We operationalised it based on how many times they submitted their codes during the semester, that is, the number of all attempts that a student made in the educational system throughout the whole semester (sum attempts hereafter).

The characterisation questionnaire captured the moderators analysed in this study, which were selected by convenience, due to this study's retrospective nature. Moderators consider user characteristics as well as data related to the internal and external context. Table 3 presents each moderator, along with a brief description that indicates its possible values, an alias that will be used hereafter, and the category it fits in. We consider age and gender as user characteristics because they are basic user information. Because the circumstance and environment wherein participants completed the activities differ depending on their major, we see it as part of the external context. The others are classified as internal context, as they are characteristics that influence users' learning experiences, such as previous experience, currently working/interning, or having internet access.

Note that despite covariates being selected by convenience, selecting them addresses literature limitations. On one hand, the frameworks recommended in the literature, which explain how gamification works, do not define

Table 3. Possible moderators of gamification's success analysed.

Description	Alias	Category
Student's age (numeric)	Age	User characteristic
Whether one is male (1) or female (0)	Male*	User characteristic
Whether one has a PC at home (1) or not (0)	PC	Internal context
Whether one share a PC at home (1) or not (0)	SharesPC	Internal context
Whether one has internet at home (1) or not (0)	Internet	Internal context
Whether one has previous experience with any programming language (1) or not (0)	Exp	Internal context
Whether one has worked/interned (1) or not (0)	Worked	Internal context
Which major one is enrolled at	Major	External context

* Because this information was dichotomous (male or female), we dummy coded it as 1 for Male and 0 for Female to facilitate the interpretation of the regression coefficient.

what are the possible moderators of its effects [46, 48], possibly because that question remains open [43, 85]. On the other hand, evidence from empirical studies is unclear in terms of what those moderators are, especially in terms of contextual factors (see Section 1.2). Therefore, while the literature often indicates moderators, what are these factors remain undefined. Thus, we approach it through an exploratory perspective, based on new factors.

2.4 Data Collection

First, all students completed a characterisation questionnaire that captured demographics and internal context-related information presented in Table 3. This was accomplished in the first week of the semester. *Second*, students were offered several programming assignments that they could complete to practice their programming skills. Completing these assignments was optional and had a very small impact (about 6 %) on the final grade of both groups. There were PA available during the whole term, that is, for 15 weeks. In 2016, the system did not feature gamification yet, whereas it was present in that system in 2017 and 2018. Therefore, students from 2016 feature the control group and students from 2017 and 2018 belong to the experimental group. *Third*, students had to complete a programming exam that worth part of their final grade every two weeks, starting in the term's second week. Thus, their final grade (the measure of academic achievement) was based on scores from exams collected throughout the whole semester.

2.5 Instrumentation

We used to instruments for data collections: an educational system and programming assignments implemented within the system.

2.5.1 Educational System. All participants used the CodeBench[4] system, which is a home-made online judge created by one of the authors. Through this system, instructors/monitors select problems to create assignment lists for programming classes. The system features an embedded Integrated Development Environment - IDE - where students develop solutions and submit them at the same place. When the learner submits a solution for a given problem (i.e., attempts), the system provides instantaneous feedback on whether the solution is correct, partially correct, or incorrect. Such automatic assessment system is both convenient for instructors due to the reduction of workload related to correcting students' attempts and for learners, as they receive instantaneous feedback about the correctness of their solutions.

[4]https://codebench.icomp.ufam.edu.br/

2.5.2 Programming Assignments. The task participants solved are programming assignments (PA) offered for them to practice programming concepts/techniques introduced in classes. In total they were required to solve seven PA, each one related to one of the programming topics (sequentially taught): (t1) sequential, (t2) conditional (if-then-else), (t3) nested conditionals, (t4) repetition by condition (using loops with *while*), (t5) vectors and strings, (t6) repetition by counting (using loops with *for*), (t7) matrices. Biweekly, students learned one of the seven topics and could solve the PA of the current topic.

2.6 Masking

No masking took place in this study because we collected data in a natural context.

2.7 Conditions and Design

The study follows a quasi-experimental between-subject design to evaluate whether gamification improves students' academic achievement by influencing their behaviours, as well as which factors moderate that effect. The quasi-experiment is characterised due to the lack of random assignment, a common feature for maintaining a natural setting [13]. Instead of random assignment, condition assignment was done based on the year a student was enrolled at a CS1 class of the majors considered in this study.

This study concerns data from five semesters from three different years (2016 to 2018). Data from the first year represent our control condition (control group; N = 118), in which subjects used a non-gamified educational system. Data from the subsequent years represent our experimental condition (experimental group; N = 281), in which subjects used the same system but with gamification. Hence, characterising the between-subject design as participants interacted with a single condition.

While students of the control group used the standard, non-gamified version of the system, those in the experimental group used it with gamification. The gamification design was planned according to the proposal of [98], which combined aspects of the ADDIE (Analyse, Design, Develop, Implement, and Evaluate) approach [11] with the Instructional Systematic Design model [20]. ADDIE is a product development concept used in educational environments to build student-centered learning [11]. Similarly, the Instructional Systematic Design model [20] is an instructional design process based on learning theories and research, and practical experience. According to those, the gamification design resulted in the following:

- *Goal*: Motivating students to solve the PA's problems.
- *Media type*: Digital and online.
- *Context*: Programming classes or when appropriate to the learner.
- *Interaction*: Single user.
- *Narrative*: A medieval fantasy world where characters (students from the same class) must face a monster (Chimera) to free their lands from domination.
- *Description*: The student chooses their avatar among several options available. When they solve a problem, they progress in the map towards the Chimera, winning strength points and weapons. The greater the strength and the better the avatar's weapon, the more hit points the student can take from the monster when facing it.
- *Results*: A percentage of the class must reach the end of the map to find and kill the Chimera. Although the interaction is individual, this collective objective aims to minimise undesired competition among students. Killing the Chimera leads to the âĂIJwinning stateâĂİ.
- *Feedback*: While using the gamified online judge, students receive additional feedback about their performance according to their characters' positions, weapons and strength.

Next, we further describe the system's gamified version, relating its main aspects to the game elements of a recent taxonomy of game elements for education [93]. In the gamified version, deployed since 2017, the students

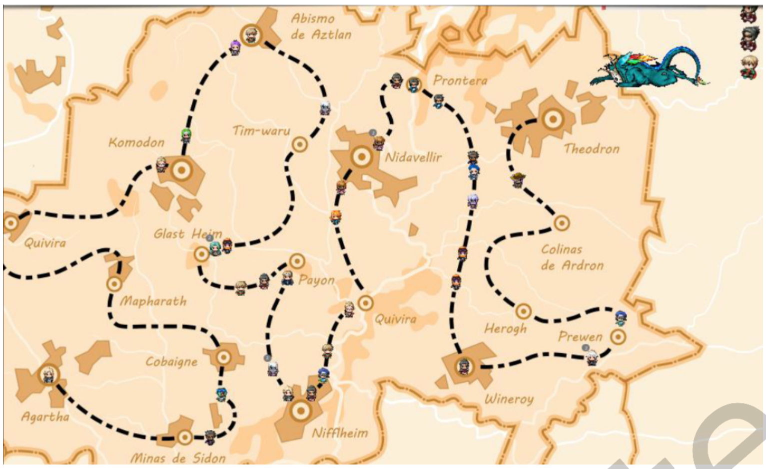

Fig. 2. Map in which learners' avatars advance within the story provided by the gamification design as they complete assignments.

see themselves as one of the characters in a fictional world, where they can walk through maps, overcome an enemy, investigate stories and explore environments (Storytelling and Narrative). The gamification does not influence the content of the course, nor completing its assignments is mandatory for students. However, when the students correctly solve the problems from the assignment lists created by the instructors/monitors, they receive rewards (Acknowledgement) that allow them to advance within the story (Progression), unlock new items and new interactions with the environment (Novelty). Thus, the reward is given to those who are more dedicated to solving the programming problems. Furthermore, the students can compete to each other (Competition) based on the rewards received after completing assignments. Nevertheless, they can only finish the fictional world's story if a large proportion of the class completes the assignments. That is, they have to work together (Cooperation). Summarising, the gamification design is a mix of immersive-, social-, and challenge-based gamification: it is mostly concerned with providing an immersive experience through fictional game elements (i.e., narrative and storytelling), while also presenting performance feedback (e.g., points and progression). Figure 2 shows the map in which learners explore the story.

2.8 Analytic Strategy

Because our data follows a hierarchical structure (e.g., students grouped by their majors), we used multilevel regression for data analysis. Multilevel regression is designed for statistically analysing, at the same time, multiple variables from different levels of a hierarchical structure, as well as taking into account their dependencies [34]. As we need to simultaneously analyse the relationship between variables of distinct hierarchical levels (e.g., users' data, level 1, and their classes, level 2) to achieve our goal, multilevel regression is adequate for testing our hypotheses and answering our research question.

To account for such group differences, multilevel models allow each one to have its intercept and regression coefficients, which are known as *random coefficients*. In the case of this study, students are grouped by majors, thus, each major will have an intercept. Additionally, we want to evaluate whether gamification's effect differs depending on the student's major (i.e., do majors moderate gamification's effect?). Therefore, our model will allow the gamification coefficient to vary across majors to estimate these possible differences. Complementary, multilevel models also estimate coefficients that apply to the overall model, which are known as *fixed coefficients*.

Fixed coefficients estimate a predictor's overall effect, whereas random coefficients capture how much a group differs from the sample's average. As our hypotheses assume a single effect depends on the grouping (i.e., gamification's influence depends on students' majors), the remaining variables we analyse are all considered as fixed coefficients only as including many random coefficients substantially increase model complexity and sample size needed [14].

For the analysis procedure, we followed a set of literature recommendations concerning data preparation and model development. For data preparation, we applied three transformations. First, we transformed the continuous variables using the squared root transformation to guarantee model validity as, without these transformations, some multilevel regression assumptions were being violated (e.g., heteroscedasticity) [25]. Second, we scaled these square-root-transformed continuous variables, which is recommended for models featuring interactions. Scaling is important to guarantee variables from an interaction are in similar scales, as well as for coefficients interpretation [25]. Third, we transformed nominal variables (e.g., gamified or not, is male or not) into factors. Scaling was not applied to non-numerical variables because they are all dummy coded (e.g., gamified or not, has previous experience or not), thereby, their values and interactions are meaningful without scaling.

For model development, we followed a top-down approach [34]. That is, starting with a *full* model (i.e., all interactions between all variables examined) and iteratively removing predictors that do not affect the model fit. In the context of this study, this was accomplished by starting with a model that can be represented as:

$$finalgrade = gamified \times sumattempts \times (usercharacteristics + internalcontextdata)$$

where the equation aims to model a student's final grade (the academic achievement measure) based on interactions between the condition (gamified or not; dummy coded), the measure of practice (*sumattempts*), and moderators under analysis (e.g., gender and previous experience with some programming language). Thus, this model allows us to analyse all influences and moderations predicted by our hypotheses. Note that the measure of external context (i.e., students' major) is the grouping factor, considered as a random coefficient. Then, to identify which predictors to remove, we used the Likelihood Ratio Test (LRT). The LRT is recommended as it calculates a predictor's impact based on the change in model fit when it is removed, indicating whether it is significant or not [25]. Hence, we iteratively removed all predictors that insignificantly affect model fit, starting from the three-way interactions (interactions between three predictors), then the two-way, and, lastly, the single terms; that is, from the most complex ones to the simplest.

Given the exploratory nature of investigating these moderations, we adopted a 10% alpha level, following literature suggestions [28, 34]. Furthermore, due to the multiple comparisons in this procedure, we adjusted p values using the False Discovery Rate approach as is has been recommended over the Bonferroni approach [39]. All analyses were conducted using R[5], R studio[6], and the *lme4* package [7].

3 RESULTS

Table 4 presents descriptive statistics of the main measures analysed in this study: students' final grade (academic achievement) and how much they practised programming (sumattempts). The table describes these values in the raw form as well as after the transformations applied for data analysis for the reader's reference when interpreting our results. Next, this section presents the development process for modelling students' academic achievement based on the predictors previously introduced. Then, we present how the multilevel model developed answers our hypotheses, as well as additional insights it reveals.

[5]https://www.r-project.org/
[6]https://rstudio.com/

Table 4. Measures' descriptive statistics.

	Academic Achievement			Sum of attempts		
Statistic	Raw	Root squared	S+S	Raw	Root squared	S+S
Mean	5.40	2.15	0	394.7	18.21	0
SD	3.22	0.87	1	349.70	7.95	1
Q1	2.39	1.55	-0.70	189.00	13.75	-0.56
Q3	8.20	2.86	0.82	507.50	22.53	0.54

S+S = Root squared then Scaled.

3.1 Modelling Students' Academic Achievement

Following the top-down approach [34], we fitted a full model as previously defined. We found no three-way interaction significantly affected model fit (p-adj > 0.1), as shown in Table 5. Therefore, we removed predictors corresponding to those interactions and moved to the next step: testing whether any two-way interaction affects the model fit (Table 6). Results show five two-way interactions significantly affect model fit: gamification and sumattempts, gamification and Male, sumattempts and Age, sumattempts and Male, and sumattempts and Internet. Therefore, we kept these interactions in the model. Then, we tested whether the predictors not involved in any significant interaction affect the model fit alone (Table 7), finding that Worked was the only non-significant predictor. Thus, we removed it and fitted the final model, which is summarised in Table 8.

Table 5. Likelihood ratio tests assessing three-way interactions in modelling students' academic achievement.

Predictors	F(num_df, den_df)	p-val	p-adj
gamified:sumattempts:Age	0.076(1, 380.38)	0.783	0.818
gamified:sumattempts:Male	0.053(1, 382.61)	0.818	0.818
gamified:sumattempts:PC	3.393(1, 326.67)	0.066*	0.465
gamified:sumattempts:SharesPC	0.856(1, 367.27)	0.355	0.718
gamified:sumattempts:Internet	0.680(1, 339.46)	0.410	0.718
gamified:sumattempts:Exp	0.380(1, 392.20)	0.538	0.753
gamified:sumattempts:Worked	0.876(1, 371.97)	0.350	0.718

* $p < 0.1$

Furthermore, to test whether the external context moderator (students' major) affects the model fit, we performed an LRT test comparing the full model shown in Table 8 to a version without the random coefficient that allows gamification's effect to be moderated by a student's major. This test yielded non-significant results (LRT(2) = 0.02; p = 0.989), indicating the random effect does not affect model fit. We highlight that adding or removing a random coefficient does not change a model's estimates, however, removing it increases the chances of finding false-positive fixed coefficients (i.e., inflating type I errors) [34]. Therefore, we left the random effect in our final model, although it does not significantly improve model fit. In the last step of this process, we tested the final model validity concerning aspects such as normality of random effects and heteroscedasticity; none was violated. The normality of residuals was not met, but due to our sample size, this is unlikely to affect model validity [55]. Assumptions' testing and a summary of all models developed in this process are available in the supplementary material.

Table 6. Likelihood ratio tests assessing two-way interactions in modelling students' academic achievement.

Predictors	F(num_df, den_df)	p-val	p-adj
gamified:sumattempts	6.623(1, 396.95)	0.010***	0.052**
gamified:Age	2.522(1, 385.17)	0.113	0.242
gamified:Male	10.476(1, 396.56)	0.001***	0.010**
gamified:PC	0.189(1, 392.07)	0.664	0.824
gamified:SharesPC	1.146(1, 391.44)	0.285	0.428
gamified:Internet	0.632(1, 391.69)	0.427	0.582
gamified:Exp	0.001(1, 371.16)	0.979	0.979
gamified:Worked	0.021(1, 389.76)	0.884	0.947
sumattempts:Age	20.543(1, 393.96)	0.000***	0.000***
sumattempts:Male	5.719(1, 396.58)	0.017**	0.065*
sumattempts:PC	2.703(1, 393.00)	0.101	0.242
sumattempts:SharesPC	1.673(1, 396.74)	0.197	0.369
sumattempts:Internet	4.683(1, 392.47)	0.031**	0.093*
sumattempts:Exp	1.256(1, 394.37)	0.263	0.428
sumattempts:Worked	0.135(1, 393.81)	0.714	0.824

* $p < 0.1$; ** $p < 0.05$; *** $p < 0.01$

Table 7. Likelihood ratio tests assessing single predictors in modelling students' academic achievement in the presence of significant interactions.

Predictors	F(num_df, den_df)	p-val	p-adj
PC	4.459(1, 392.58)	0.035**	0.052*
SharesPC	4.211(1, 392.23)	0.041**	0.052*
Exp	8.093(1, 393.18)	0.005***	0.011**
Worked	0.121(1, 392.14)	0.728	0.728
gamified:sumattempts	8.005(1, 395.10)	0.005***	0.011**
gamified:Male	9.756(1, 395.28)	0.002***	0.009***
sumattempts:Age	20.510(1, 393.89)	0.000***	0.000***
sumattempts:Male	5.276(1, 396.28)	0.022**	0.040**
sumattempts:Internet	3.390(1, 392.65)	0.066*	0.075*

* $p < 0.1$; ** $p < 0.05$; *** $p < 0.01$

3.2 Study Hypotheses

From our final model (Table 8), we can discuss our hypotheses as well as answer our research question. **H1** predicted that practising to code would positively influence students' academic achievement. According to the model, the extent to which students practised (sumattempts) has a positive, highly significant effect on academic achievement, suggesting the more students practised, the higher was their academic achievement. Therefore, supporting **H1**.

H2 predicted gamification would positively influence learners' academic achievement, by maximising the testing effect. Our model suggests gamification (gamified) has a positive, highly significant effect on academic achievement. However, its interaction with how much students practised (sumattempts) is negative and highly significant. This indicates gamification had a positive effect on academic achievement, as expected, but the more

Table 8. Multilevel model predicting students' academic achievement.

Predictors	Est.(SE)	CI	p-val
Fixed Effects			
(Intercept)	-0.55 (0.17)	-0.84 âĂŞ -0.27	0.001***
gamified	0.62 (0.14)	0.38 âĂŞ 0.86	0.000***
Male	0.40 (0.14)	0.17 âĂŞ 0.64	0.004***
sumattempts	0.58 (0.12)	0.38 âĂŞ 0.77	0.000***
Age	0.04 (0.04)	-0.02 âĂŞ 0.11	0.276
Internet	0.50 (0.15)	0.25 âĂŞ 0.76	0.001***
PC	-0.39 (0.18)	-0.69 âĂŞ -0.09	0.035**
SharesPC	-0.17 (0.08)	-0.31 âĂŞ -0.03	0.041**
Exp	0.22 (0.08)	0.09 âĂŞ 0.34	0.005***
gamified * sumattempts	-0.28 (0.10)	-0.44 âĂŞ -0.12	0.005***
gamified * Male	-0.54 (0.17)	-0.82 âĂŞ -0.25	0.002***
sumattempts * Age	0.17 (0.04)	0.11 âĂŞ 0.23	0.000***
sumattempts * Male	0.19 (0.08)	0.05 âĂŞ 0.33	0.022**
sumattempts * Internet	0.15 (0.08)	0.02 âĂŞ 0.28	0.065*
Random Effects			
	Var(SD)		
Residual	0.50 (0.71)		
Major	0.04 (0.19)		
gamified * Major	0.00 (0.01)		
Model fit			
Marginal R^2	0.46		
Conditional R^2	0.50		

* p < 0.1; ** p < 0.05; *** p < 0.01

students practised, the more this effect decreased (see Figure 3), unlike per our expectations. Thus, this only partially supports **H2**.

H3 predicted that gamification's effects would depend on user and contextual characteristics. The only significant moderator of gamification's effect was learners' gender (Figure 4). The moderator effect of the remaining user and contextual characteristics were non-significant. For data from the internal context and the user, this was found during the model development process as those interactions were found to not affect the model fit. For the external context moderator that we evaluated, this was shown by testing the random coefficient effect, but can also be seen by the small variances in the random part shown in Table 8, as well as the negligible improvement of the Conditional R^2 (4%; Conditional R^2 = 0.50), which considers the random and the fixed model parts, compared to the marginal R^2 (46%), which only considers the fixed part. Thus, only partially supporting **H3**.

3.3 Additional Findings

Moreover, our model revealed additional insights that do not directly relate to this study's hypotheses and research questions. Despite our focus on gamification's effect, we found insights concerning direct moderators of academic achievement as well as factors that moderate the impact of practising. Surprisingly, the developed model indicates having a PC at home has a negative, highly significant impact on academic achievement. The model also

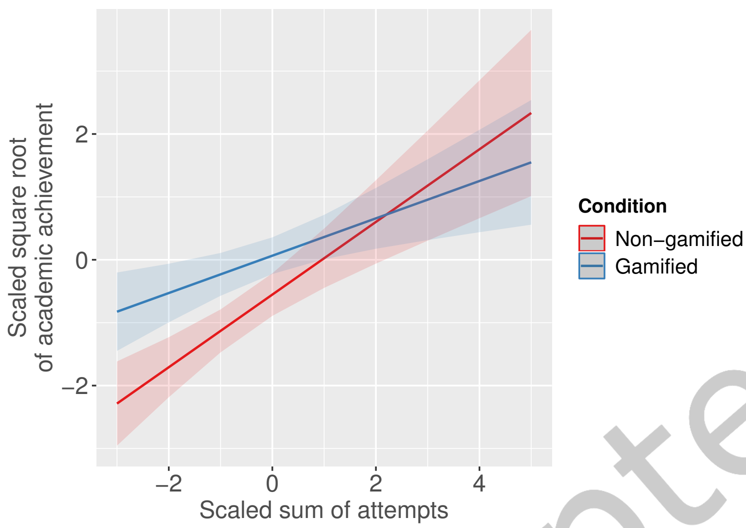

Fig. 3. Predicted effects of gamification on the impact that practising to code (sum of attempts) has on academic achievement.

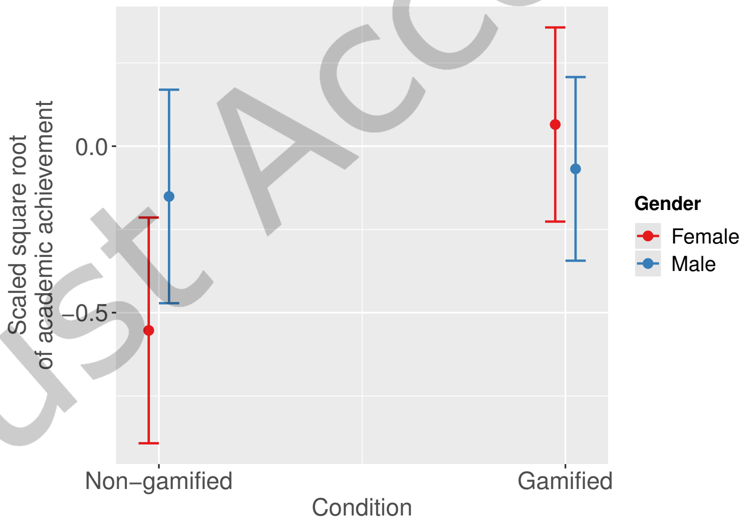

Fig. 4. Gamification's effects on academic achievement as moderated by students' gender.

indicates that sharing the PC at home has a negative, significant effect on academic achievement. On the other hand, the model indicates previous experience with any programming language and having internet have positive, highly significant impacts on academic achievement. The model also indicates males presented higher academic achievements than females. Additionally, the final model indicates the testing effect (i.e., practising to code to improve learning via self-testing) is moderated by students' age, gender, and whether they have internet at home. The indications are that i) compared to younger students, older ones accomplish higher academic achievement as they practice more; ii) the more males practice, the higher is the difference in their academic achievement compared to females, and iii) the more one with internet practices, the larger is their academic achievement compared to those with no internet.

3.4 Summary

From our findings, the answer to *how user and contextual factors influence gamificationâĂŹs effect on the academic achievement of CS1 students* is that gender significantly moderated that influence, positively for females and null for males, whereas user age and contextual factors (i.e., having a PC at home, sharing a PC at home, having internet, having previous experience with some programming language, having already worked/interned, and major enrolled at) had null (non-significant) influences on that effect. Furthermore, we found that for those who practised more than two standards deviations above the average (i.e., 1094 attempts), gamification's effect changed from positive to negative. Our results also revealed additional insights concerning moderators of academic achievement as well as of the testing effect.

4 DISCUSSION

Our results support the testing effect in the study context and show that students in the gamified version yielded higher academic achievement. The difference was moderated by participants' gender. We noted that some learners self-tested to a point in which their academic achievement was below that of those who practised less. In contrast to this study, previous work on gamification applied to CS1 education rarely performed moderator analysis and results were often based on small usage periods and relatively small samples. Differently, we analysed data from 399 learners, evaluated gamification's effects after they used it for a whole semester (15 weeks), and analysed the moderator role of eight factors, including user (e.g., age and gender) and contextual (major) information. Thus, our contribution is revealing user and contextual characteristics that moderate gamificationâĂŹs impact on CS1 studentsâĂŹ academic achievement, as well as whether those maximise or minimise that impact, based on empirical evidence build upon a sample of substantial size and long-term usage of the intervention. Next, we discuss our results related to relevant literature.

First, our findings provide empirical support for the testing effect [84] in the context of programming learning. The results demonstrate that the more the students practised programming, the higher their academic achievements were at the end of the semester. On one hand, previous research has explored and demonstrated the testing effect's benefits for programming learning, but either with experienced programmers or in a non-gamified environment (e.g., [75, 103]). On the other hand, studies have also shown evidence on the testing effect's positive impact on learning from using gamified systems (e.g., [16, 79, 86]) but in other contexts or based on small samples. Whereas our findings corroborate the results of those studies, we studied the testing effect based on data from beginner programmers (CS1 learners) using a gamified educational system. Therefore, we contribute with support for the testing effect within the context of CS1, providing evidence that practising and submitting programming assignments is positively related to higher academic achievement.

Second, our findings demonstrate gamification's contribution to students' academic achievement was positive. The analyses suggested gamification had a positive impact on learners' academic achievement. This finding corroborates the general gamification literature [43] as well as gamified learning effects, in which meta-analytic results indicate that gamified interventions lead to small positive learning outcomes, compared to non-gamified conditions [85]. This positive result is aligned to the positive outcomes of gamified programming learning as well (e.g., [23, 53]; Section 1.2). It should be noted, however, that empirical studies' methodological rigour has been a limitation of the field, which raises questions on findings' validity [18, 32, 90]. Differently, our findings are based on an examination featuring a control group, a substantial sample size and a long usage period, besides controlling for covariates in data analysis. Therefore, by corroborating the overall literature, our findings contribute robust empirical evidence on gamification's benefits to learning. Thus, we expand the literature both in terms of gamification's impact, with evidence on its effectiveness in the context of programming learning based on data from a significant sample size who used gamification for an entire semester, as well as to computing education, supporting the potential from gamification to support programming learning.

We also found that gamification's impact was influenced by two factors: the amount of practice and gender. Concerning the relationship between gamification and the testing effect, our analysis revealed that gamification's contribution was mixed. The results for this analysis indicate gamification improved the testing effect; however, such improvement not only vanished but became negative for students who practised substantially more than the average (see Figure 3). We interpret this finding from two perspectives. One possible explanation might lie behind the support gamification provides to learning. In a meta-synthesis of gamification literature, Bai et al. [4] found learners enjoy gamification because it provides performance feedback and means to be recognised (e.g., badges), as well as goal-setting. Feedback is important for programming learning [75, 103], being recognised is likely to fulfil competence needs, which also contributes to meaningful learning experiences, as well as goal-setting [91, 97]. Similarly, the game elements might have contributed to learners' feelings of self-efficacy, which is positively related to learning performance as well [36, 102].

On the other hand, Bai et al. [4] found gamification might cause anxiety or jealousy. Those feelings might lead students to do whatever it takes to not feel this way, such as substantially interacting with the system aiming to receive rewards and climb up the leaderboard, without paying attention to the learning task. For instance, students might start submitting several attempts, in a "desperate" effort to solve the question, wherein most submissions are likely to be partially or completely wrong. Specifically, the system we used does not check whether a new submission differs from previous ones. Then, in the gamified version, students might engage in behaviours such as adding small incremental changes (or even simple resubmissions) due to the anxiety to climb the leaderboard and/or their jealousy of those at the top [4], instead of putting the adequate effort to understand and correctly solve the question. Differently, in the version without gamification, students would have no motivation for engaging in such behaviours. Such potentially desperate, reward-driven behaviours indicate students were just gaming the system: seeking game-like rewards instead of properly using it to practice and learn [5]. Hence, explaining why some of our participants ended up yielding lower academic achievement than those who submitted fewer or the same number of attempts in the non-gamified version. A similar outcome was reported in Ghaban and Hendley [26], wherein learners of the gamified version dropped out less, but showed worse learning gains. Thus, we expand the literature with insights about gaming the system behaviour in such context, empirically demonstrating that, although gamification is of value, it might lead to outcomes opposed to the expected, which also contributes to the literature by responding to the need of analysing gamification's negative effects [43].

Concerning gender, our analysis revealed that the academic achievement of male learners was almost the same, regardless of using the gamified system or not, while showing that female learners' were positively affected by gamification (see Figure 4). This finding corroborates claims that user characteristics affect their experiences with gamified systems (e.g., [18, 60]), as well as has been shown in the context of games (e.g., [61, 76]). On the other hand, this finding is contrary to studies that found no moderator effect of gender [44, 72, 86]. A reason for that contradiction might be the gamification design. Research acknowledges that the gamification design is determinant for users' experience with gamified systems [51, 59]. Whereas we used a mix of immersion, challenge, and social gamification in this study, previous research [44, 72, 86] mostly focused on challenge-based designs. As different designs were used, it might be that one of them was equally good, bad, or null for all learners, whereas the other was different for females and males. Hence, explaining the fact that gender was a significant moderator in one case but not in others.

Despite gender being the only factor we found to moderate gamification's impact, we analysed another seven aspects, related to both the user and the context. Those are based on arguments that for gamification to yield positive outcomes, it needs to be designed according to the user and the context (e.g., [43, 59, 93]). Therefore, one might expect that user and contextual characteristics will moderate gamification's effect. There are studies exploring these claims, showing whether gamification usage is successful or not depends on aspects such as age [44], major [72], attitudes towards game-based learning [3], goal orientation [29], performance [1], and the user and context in general [18, 31]. Nevertheless, recent research still highlights the need for studying moderators, with calls to advance the understanding of pre-determinants and occasions in which learners take advantage of gamification [43, 49, 85]. Therefore, this study contributes by responding to such calls, analysing moderators not only related to user demographics (e.g., gender and age) but also exploring those related to internal (e.g., previous experience with programming) and external (e.g., the major a student is enrolled at) context. Nevertheless, our findings are mostly contradictory to the overall discussion in related work, as ours indicate the impact of gamification only depended on learners' gender. In contrast, what we found mostly corroborates moderator analysis in the context of programming learning, except for the moderator role of gender (see Section 1.2). Thus, reinforcing the need for advancing the understanding of moderators of gamification's success, especially in varied contexts.

Furthermore, we also found some factors moderated the testing effect as well as students' academic achievement. Despite the focus of this study is on gamification and its moderators, our analyses revealed interesting findings that contribute with evidence on factors likely to affect programming learning. On one hand, we found some characteristics related to internal context [88] moderated academic achievement. As expected, users with previous experience in some programming languages achieved higher academic achievement than those with no previous experience. Differently and surprisingly, students with at least one computer at home presented lower academic achievement compared to those with no computer. Similarly, having to share the home computer negatively affects academic achievement whereas the academic achievement of learners with internet at home was higher than that of learners without it. On the other hand, we also found three factors moderated the testing effect (i.e., age, gender, and internet), in which all of those maximised it. These findings do not directly relate to the objective of this study, they rather emerged as a consequence of our data analysis process. Then, we briefly presented and interpreted them so that the interested reader can understand what our analysis revealed. Nevertheless, we believe these findings are of value for those interested in moderators of learning as well as of the testing effect. Those provide insights on the characteristics of learners that are more likely to take more or less advantage from the testing effect, as well as factors that play a role in CS1 students' learning. Thus, opening directions for future research to investigate these insights.

4.1 Implications

We highlight five main implications of our findings. First, we have shown that students who practice more are likely to yield higher academic achievement at the end of the semester. Instructors can explore this finding by providing their students with opportunities to practice programming as much as possible, which can be accomplished by making several programming assignments available during the course. Second, we demonstrated gamification can enhance as well as mitigate the testing effect. On one hand, this suggests that instructors can rely on gamified systems to improve the effectiveness of the testing effect. On the other hand, this implies that gamification must be designed cautiously, aiming to prevent behaviours such as gaming the system or feelings of jealousy and anxiety, as for some users gamification might end up diminishing the testing effect. Third, our finding concerning students who gamed the system contributes to the design of online judges. The tool our participants used did not impose restrictions on the extent to which one submission should differ from previous ones, nor decreased gamification outcomes upon repeated attempts. Hence, designers of online judges might consider imposing such restrictions, to avoid the submission of (almost) identical solutions, aiming to push students to work on improving their code instead of just trying the correction system until getting it right. Fourth, we found gamification's impact was different depending on whether users were males or females. This further supports the claim that one size does not fit all [60], indicating the need for developing and providing gamification designs tailored to specific audiences. Fifth, we found no support for the moderator effect of various user and contextual characteristics, which is contrary to the overall discussion from previous studies (e.g., [31, 81, 93]) and our initial expectations. Hence, pointing that more research is needed to understand what moderates gamification's success.

4.2 Limitations

This study has some limitations that must be considered when interpreting its findings. First, as we explored data from three years, we could not guarantee majors had the same instructor in both the control and experimental condition. This was mitigated with all other aspects being the same (e.g., class program, handbooks, exam structure). Nevertheless, this is likely to not affect our findings as a recent secondary study suggests gamification's effect is the same regardless of instructors being different or not [4].

Second, because this is a retrospective study, we could not use random assignment, and groups significantly differed in some demographic characteristics. Considering covariates (e.g., age, gender, previous experience) helps to handle this limitation as the analysis would reveal if some difference is due to the covariate itself and not due to conditions. Additionally, the meta-analysis by Sailer and Homner [85] found randomisation did not affect gamification's impact on behavioural outcomes, suggesting our second limitation is unlikely to threaten our results. Also, most covariates are self-reported and, in some cases, based on binary answers, which respectively inserts subjectivity and limits the information they provide. While that limits statistical results, the successful use of similar data in prior research (e.g., [29, 64, 79]) suggests it represents a mitigated threat to the findings.

Third, we did not conduct a pre-test, which opens the possibility for students from one condition to have more previous knowledge than those in the other. Consequently, this could lead to misleading conclusions regarding a condition's contribution to students' academic achievement. To mitigate this threat, we analysed covariates related to possible prior knowledge (i.e., previous experience and having worked/interned before), hence, accounting for those possible differences in the data analysis. Meta-analytic evidence further supports the reduced role of this limitation (the lack of a pre-test) in our findings: it demonstrates gamification's effect on cognitive outcomes (e.g., academic achievement) does not significantly change, depending on whether only post-test or pre- and post-tests were used [85].

Fourth, the number of participants in each condition was unbalanced ($N_{control}$ = 118; $N_{experimental}$ = 281). This limitation emerged because data from a single year (2016) composed the control group, whereas data from two years (2017 and 2018) composed the experimental group. The reason is that gamification was deployed to the system used for data collection in 2017. To cope with this limitation, we performed statistical analysis to shed light on whether differences were not by mere chance. Nevertheless, the number of participants per group in our study is above the average of participants per study in similar research (e.g., 107, 95, and 75 according to Bai et al. [4], Sailer and Homner [85], and Koivisto and Hamari [42], respectively), which demonstrates our analysis is based on a representative sample size (N = 399) in terms of gamification studies for both study groups.

Lastly, we note our sample consists of STEM, not computing-related majors, such as Computer Science and Software Engineering. While studying STEM-related majors is important, because their students face difficulties with CS1 often [21, 87], we cannot ensure that our findings will hold with computing-related majors. However, we note that the effects did not vary from one major to another within our sample, so it is reasonable to expect similar outcomes for other majors.

4.3 Future Work

As future work, we suggest some lines of research. First, we call for more research to understand when gamification does not work. We found mixed effects, in which we discussed those in terms of gaming the system, possibly motivated by participants' feeling anxious or jealous. Further experiments to understand when and why such effects happen would contribute to the design of more effective gamified systems, consequently, contributing to students' learning, as has been suggested recently [43]. Towards preventing these undesired behaviours, game elements exploring game fiction and socialisation [93] are promising [85], as they do not rely on external motivators. Thus, we encourage future studies to evaluate the impacts of game fiction and socialisation on CS1 learning.

We also call for research revealing factors that moderate gamification's effectiveness. In this study, we found evidence gender was one of such moderators, while our results indicated other factors (e.g., age, previous experience, and student's major) were not. Such indications add to the literature, confirming the need to determine which are those factors, as well as understanding when and in which occasion gamification works [6, 85]. Accordingly, experiments to identify which are those moderators would contribute to the understanding of factors to consider when designing gamified interventions.

Based on that need, we call for research on designing more tailored gamified interventions. This also emerges from our finding that the gamification design we used was effective for females but null for males, highlighting the need for developing gamification designs that are tailored to the target sample [35]. That is, once we are aware there are males and females in the target population, gamification should be planned accordingly, such as in personalised gamification [82]. Consequently, this calls for research to understand how to gamify an educational environment to a target population while considering all of its relevant characteristics [41], including the users, the task, and the context [31, 50, 77].

5 CONCLUSIONS

Through a quasi-experimental retrospective study (N = 399), we showed that gamification positively affected CS1 students' academic achievement, by enhancing the testing effect. However, the results also suggested that some learners were gaming the system rather than studying, which led to negative learning outcomes. This leads us to conclude that gamification can contribute to CS1 education, by enhancing behaviours valuable to learning (e.g., self-testing). However, it should be planned and deployed with caution, to prevent inciting undesired behaviours. Furthermore, we examined the moderator role of several user and contextual characteristics, finding only gender as a significant moderator. Surprisingly, this contradicts previous research advocating that gamification's success depends on multiple user and contextual characteristics, which was not supported by our analyses. Nevertheless, more research is needed to better understand and ground which factors moderate the impact of gamification, given that, currently, there are many theoretical discussions and few empirical examinations in this direction, especially those concerning the role of context, as in this study.

6 ACKNOWLEDGMENTS

The authors are thankful to the financial support of the following agencies: Brazilian National Council for Scientific and Technological Development (CNPq) - processes 141859/2019-9, 308458/2020-6, and 308513/2020-7; Coordination for the Improvement of Higher Education Personnel (CAPES) - Finance Code 001; and SÃčo Paulo State Research Support Foundation (FAPESP) - processes 2018/15917-0, 2018/07688-1, and 2013/07375-0. Additionally, this research was carried out within the scope of the Samsung-UFAM Project for Education and Research (SUPER), according to Article 48 of Decree nÂž 6.008/2006 (SUFRAMA), and partially funded by Samsung Electronics of Amazonia Ltda., under the terms of Federal Law nÂž 8.387/1991, through agreements 001/2020 and 003/2019, signed with Federal University of Amazonas and FAEPI, Brazil.

REFERENCES

[1] Samuel Abramovich, Christian Schunn, and Ross Mitsuo Higashi. 2013. Are badges useful in education?: It depends upon the type of badge and expertise of learner. *Educational Technology Research and Development* 61, 2 (2013), 217–232.

[2] Adnan Ahmad, Furkh Zeshan, Muhammad Salman Khan, Rutab Marriam, Amjad Ali, and Alia Samreen. 2020. The impact of gamification on learning outcomes of computer science majors. *ACM Transactions on Computing Education (TOCE)* 20, 2 (2020), 1–25.

[3] Michael B. Armstrong and Richard N. Landers. 2017. An Evaluation of Gamified Training: Using Narrative to Improve Reactions and Learning. *Simulation & Gaming* 48, 4 (2017), 513–538. https://doi.org/10.1177/1046878117703749 arXiv:https://doi.org/10.1177/1046878117703749

[4] Shurui Bai, Khe Foon Hew, and Biyun Huang. 2020. Is gamification âĂĲbullshitâĂİ? Evidence from a meta-analysis and synthesis of qualitative data in educational contexts. *Educational Research Review* (2020), 100322.

[5] Ryan SJd Baker, Albert T Corbett, Ido Roll, and Kenneth R Koedinger. 2008. Developing a generalizable detector of when students game the system. *User Modeling and User-Adapted Interaction* 18, 3 (2008), 287–314.

[6] Gonçalo Baptista and Tiago Oliveira. 2019. Gamification and serious games: A literature meta-analysis and integrative model. *Computers in Human Behavior* 92 (2019), 306–315.

[7] Douglas Bates, Martin Mächler, Ben Bolker, and Steve Walker. 2014. Fitting linear mixed-effects models using lme4. *arXiv preprint arXiv:1406.5823* (2014).

[8] Brett A Becker and Keith Quille. 2019. 50 years of cs1 at sigcse: A review of the evolution of introductory programming education research. In *Proceedings of the 50th acm technical symposium on computer science education*. 338–344.

[9] Jens Bennedsen and Michael E Caspersen. 2019. Failure rates in introductory programming: 12 years later. *ACM Inroads* 10, 2 (2019), 30–36.

[10] Paulo Blikstein. 2011. Using learning analytics to assess students' behavior in open-ended programming tasks. In *Proceedings of the 1st international conference on learning analytics and knowledge*. 110–116.

[11] Robert Maribe Branch. 2009. *Instructional design: The ADDIE approach*. Vol. 722. Springer Science & Business Media.

[12] Patrick Buckley and Elaine Doyle. 2017. Individualising gamification: An investigation of the impact of learning styles and personality traits on the efficacy of gamification using a prediction market. *Computers & Education* 106 (2017), 43–55.

[13] John W Creswell and J David Creswell. 2017. *Research design: Qualitative, quantitative, and mixed methods approaches*. Sage publications.

[14] Jan De Leeuw, Erik Meijer, and Harvey Goldstein. 2008. *Handbook of multilevel analysis*. Springer.

[15] Luis De-Marcos, Adrián Domínguez, Joseba Saenz-de Navarrete, and Carmen Pagés. 2014. An empirical study comparing gamification and social networking on e-learning. *Computers & education* 75 (2014), 82–91.

[16] Paul Denny, Fiona McDonald, Ruth Empson, Philip Kelly, and Andrew Petersen. 2018. Empirical Support for a Causal Relationship Between Gamification and Learning Outcomes. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems* (Montreal QC, Canada) *(CHI âĂŹ18)*. Association for Computing Machinery, New York, NY, USA, 1âĂŞ13. https://doi.org/10.1145/3173574.3173885

[17] Sebastian Deterding, Dan Dixon, Rilla Khaled, and Lennart Nacke. 2011. From game design elements to gamefulness: defining gamification. In *Proceedings of the 15th international academic MindTrek conference: Envisioning future media environments*. ACM, 9–15.

[18] Christo Dichev and Darina Dicheva. 2017. Gamifying education: what is known, what is believed and what remains uncertain: a critical review. *International journal of educational technology in higher education* 14, 1 (2017), 9.

[19] Darina Dicheva, Keith Irwin, and Christo Dichev. 2019. OneUp: Engaging Students in a Gamified Data Structures Course. In *Proceedings of the 50th ACM Technical Symposium on Computer Science Education*. 386–392.

[20] Walter Dick, Lou Carey, and James O Carey. 2005. The systematic design of instruction. (2005).

[21] Leovy Echeverría, Ruth Cobos, Liliana Machuca, and Ivan Claros. 2017. Using collaborative learning scenarios to teach programming to non-CS majors. *Computer applications in engineering education* 25, 5 (2017), 719–731.

[22] Samuel C Fonseca, Filipe Dwan Pereira, Elaine HT Oliveira, David BF Oliveira, Leandro SG Carvalho, and Alexandra I Cristea. 2020. Automatic Subject-based Contextualisation of Programming Assignment Lists. EDM.

[23] Panagiotis Fotaris, Theodoros Mastoras, Richard Leinfellner, and Yasmine Rosunally. 2016. Climbing up the Leaderboard: An Empirical Study of Applying Gamification Techniques to a Computer Programming Class. *Electronic Journal of e-learning* 14, 2 (2016), 94–110.

[24] MN Gari, GS Walia, and AD Radermacher. 2018. Gamification in computer science education: A systematic literature review. In *American Society for Engineering Education*.

[25] Andrew Gelman and Jennifer Hill. 2006. *Data analysis using regression and multilevel/hierarchical models*. Cambridge university press.

[26] Wad Ghaban and Robert Hendley. 2020. Can We Predict the Best Gamification Elements for a User Based on Their Personal Attributes?. In *International Conference on Human-Computer Interaction*. Springer, 58–75.

[27] E Bruce Goldstein. 2014. *Cognitive psychology: Connecting mind, research and everyday experience*. Nelson Education.

[28] Joseph F Hair Jr, G Tomas M Hult, Christian Ringle, and Marko Sarstedt. 2016. *A primer on partial least squares structural equation modeling (PLS-SEM)*. Sage publications.

[29] Lasse Hakulinen and Tapio Auvinen. 2014. The effect of gamification on students with different achievement goal orientations. In *2014 international conference on teaching and learning in computing and engineering*. IEEE, 9–16.

[30] Lasse Hakulinen, Tapio Auvinen, and Ari Korhonen. 2015. The Effect of Achievement Badges on Students' Behavior: An Empirical Study in a University-Level Computer Science Course. *International Journal of Emerging Technologies in Learning* 10, 1 (2015).

[31] Stuart Hallifax, Audrey Serna, Jean-Charles Marty, Guillaume Lavoué, and Elise Lavoué. 2019. Factors to Consider for Tailored Gamification. In *Proceedings of the Annual Symposium on Computer-Human Interaction in Play* (Barcelona, Spain) *(CHI PLAY âĂŹ19)*. Association for Computing Machinery, New York, NY, USA, 559âĂŞ572. https://doi.org/10.1145/3311350.3347167

[32] Juho Hamari, Jonna Koivisto, and Harri Sarsa. 2014. Does gamification work?–a literature review of empirical studies on gamification. In *2014 47th Hawaii international conference on system sciences*. Ieee, 3025–3034.

[33] Matthew Hertz. 2010. What do" CS1" and" CS2" mean? Investigating differences in the early courses. In *Proceedings of the 41st ACM technical symposium on Computer science education*. 199–203.

[34] Joop J Hox, Mirjam Moerbeek, and Rens Van de Schoot. 2010. *Multilevel analysis: Techniques and applications*. Routledge.

[35] Rui Huang, Albert D Ritzhaupt, Max Sommer, Jiawen Zhu, Anita Stephen, Natercia Valle, John Hampton, and Jingwei Li. 2020. The impact of gamification in educational settings on student learning outcomes: a meta-analysis. *Educational Technology Research and Development* (2020), 1–27.

[36] Xiaoxia Huang and Richard E Mayer. 2019. Adding self-efficacy features to an online statistics lesson. *Journal of Educational Computing Research* 57, 4 (2019), 1003–1037.

[37] Sami Hyrynsalmi, Jouni Smed, and Kai Kimppa. 2017. The Dark Side of Gamification: How We Should Stop Worrying and Study also the Negative Impacts of Bringing Game Design Elements to Everywhere.. In *GamiFIN*. 96–104.

[38] Petri Ihantola, Arto Vihavainen, Alireza Ahadi, Matthew Butler, Jürgen Börstler, Stephen H Edwards, Essi Isohanni, Ari Korhonen, Andrew Petersen, Kelly Rivers, et al. 2015. Educational data mining and learning analytics in programming: Literature review and case studies. In *Proceedings of the 2015 ITiCSE on Working Group Reports*. 41–63.

[39] Mohieddin Jafari and Naser Ansari-Pour. 2019. Why, when and how to adjust your P values? *Cell Journal (Yakhteh)* 20, 4 (2019), 604.

[40] Jeffrey D Karpicke and William R Aue. 2015. The testing effect is alive and well with complex materials. *Educational Psychology Review* 27, 2 (2015), 317–326.

[41] Ana Carolina Tomé Klock, Isabela Gasparini, Marcelo Soares Pimenta, and Juho Hamari. 2020. Tailored gamification: A review of literature. *International Journal of Human-Computer Studies* (2020), 102495.

[42] Jonna Koivisto and Juho Hamari. 2014. Demographic differences in perceived benefits from gamification. *Computers in Human Behavior* 35 (2014), 179 – 188. https://doi.org/10.1016/j.chb.2014.03.007

[43] Jonna Koivisto and Juho Hamari. 2019. The rise of motivational information systems: A review of gamification research. *International Journal of Information Management* 45 (2019), 191 – 210. https://doi.org/10.1016/j.ijinfomgt.2018.10.013

[44] Markus Krause, Marc Mogalle, Henning Pohl, and Joseph Jay Williams. 2015. A Playful Game Changer: Fostering Student Retention in Online Education with Social Gamification. In *Proceedings of the Second (2015) ACM Conference on Learning @ Scale* (Vancouver, BC, Canada) *(L@S âĂŹ15)*. Association for Computing Machinery, New York, NY, USA, 95âĂŞ102. https://doi.org/10.1145/2724660.2724665

[45] Carmen Lacave, Ana I Molina, and José A Cruz-Lemus. 2018. Learning Analytics to identify dropout factors of Computer Science studies through Bayesian networks. *Behaviour & Information Technology* 37, 10-11 (2018), 993–1007.

[46] Richard N Landers. 2014. Developing a theory of gamified learning: Linking serious games and gamification of learning. *Simulation & gaming* 45, 6 (2014), 752–768.

[47] Richard N Landers and Michael B Armstrong. 2017. Enhancing instructional outcomes with gamification: An empirical test of the Technology-Enhanced Training Effectiveness Model. *Computers in human behavior* 71 (2017), 499–507.

[48] Richard N Landers, Elena M Auer, Andrew B Collmus, and Michael B Armstrong. 2018. Gamification science, its history and future: Definitions and a research agenda. *Simulation & Gaming* 49, 3 (2018), 315–337.

[49] Richard N Landers, Gustavo F Tondello, Dennis L Kappen, Andrew B Collmus, Elisa D Mekler, and Lennart E Nacke. 2019. Defining gameful experience as a psychological state caused by gameplay: Replacing the term âĂŸGamefulnessâĂŹ with three distinct constructs. *International Journal of Human-Computer Studies* 127 (2019), 81–94.

[50] De Liu, Radhika Santhanam, and Jane Webster. 2017. Toward Meaningful Engagement: A Framework for Design and Research of Gamified Information Systems. *MIS quarterly* 41, 4 (2017), 1011–1034.

[51] Kevin Loughrey and Daire Broin. 2018. Are We Having Fun Yet? Misapplying Motivation to Gamification. In *2018 IEEE Games, Entertainment, Media Conference (GEM)*. IEEE, 1–9.

[52] Jenni Majuri, Jonna Koivisto, and Juho Hamari. 2018. Gamification of education and learning: A review of empirical literature. In *Proceedings of the 2nd International GamiFIN Conference, GamiFIN 2018*. CEUR-WS.

[53] B. Marín, J. Frez, J. Cruz-Lemus, and M. Genero. 2018. An Empirical Investigation on the Benefits of Gamification in Programming Courses. *ACM Trans. Comput. Educ.* 19, 1, Article 4 (Nov. 2018), 22 pages. https://doi.org/10.1145/3231709

[54] KB McDermott, S Kang, and HL Roediger III. 2005. Test format and its modulation of the testing effect. In *biennial meeting of the Society for Applied Research in Memory and Cognition, Wellington, New Zealand*.

[55] Douglas C Montgomery, Elizabeth A Peck, and G Geoffrey Vining. 2012. *Introduction to linear regression analysis.* Vol. 821. John Wiley & Sons.

[56] Alberto Mora, Gustavo F Tondello, Laura Calvet, Carina González, Joan Arnedo-Moreno, and Lennart E Nacke. 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.* ACM, 1.

[57] Miguel Ehécatl Morales-Trujillo and Gabriel Alberto García-Mireles. 2020. Gamification and SQL: An Empirical Study on Student Performance in a Database Course. *ACM Transactions on Computing Education (TOCE)* 21, 1 (2020), 1–29.

[58] Julian MORENO and Andres F PINEDA. 2018. Competitive programming and gamification as strategy to engage students in computer science courses. *Revista ESPACIOS* 39, 35 (2018).

[59] Benedikt Morschheuser, Lobna Hassan, Karl Werder, and Juho Hamari. 2018. How to design gamification? A method for engineering gamified software. *Information and Software Technology* 95 (2018), 219–237. https://doi.org/10.1016/j.infsof.2017.10.015

[60] Lennart E Nacke and Christoph Sebastian Deterding. 2017. The maturing of gamification research. *Computers in Human Behaviour* (2017), 450–454.

[61] Rita Orji, Regan L Mandryk, and Julita Vassileva. 2017. Improving the efficacy of games for change using personalization models. *ACM Transactions on Computer-Human Interaction (TOCHI)* 24, 5 (2017), 1–22.

[62] Rita Orji, Kiemute Oyibo, and Gustavo F. Tondello. 2017. A Comparison of System-Controlled and User-Controlled Personalization Approaches. In *Adjunct Publication of the 25th Conference on User Modeling, Adaptation and Personalization* (Bratislava, Slovakia) *(UMAP '17)*. ACM, New York, NY, USA, 413–418. https://doi.org/10.1145/3099023.3099116

[63] Rita Orji, Gustavo F Tondello, and Lennart E Nacke. 2018. Personalizing persuasive strategies in gameful systems to gamification user types. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems.* ACM, 435. https://doi.org/10.1145/3173574.3174009

[64] Margarita Ortiz-Rojas, Katherine Chiluiza, and Martin Valcke. 2019. Gamification through leaderboards: An empirical study in engineering education. *Computer Applications in Engineering Education* 27, 4 (2019), 777–788.

[65] Paula T Palomino, Armando M Toda, Luiz Rodrigues, Wilk Oliveira, and Seiji Isotani. 2020. From the Lack of Engagement to Motivation: Gamification Strategies to Enhance Users Learning Experiences. In *Brazilian Symposium on Computer Games and Digital Entertainment.* 1127–1130.

[66] Harold Pashler, Nicholas J Cepeda, John T Wixted, and Doug Rohrer. 2005. When does feedback facilitate learning of words? *Journal of Experimental Psychology: Learning, Memory, and Cognition* 31, 1 (2005), 3.

[67] Lais Z Pedro, Aparecida MZ Lopes, Bruno G Prates, Julita Vassileva, and Seiji Isotani. 2015. Does gamification work for boys and girls? An exploratory study with a virtual learning environment. In *Proceedings of the 30th annual ACM symposium on applied computing.* 214–219.

[68] Filipe Pereira, Elaine Oliveira, David Fernandes, Leandro Silva Galvão de Carvalho, and Hermino Junior. 2019. Otimização e automação da predição precoce do desempenho de alunos que utilizam juízes online: uma abordagem com algoritmo genético. In *Brazilian Symposium on Computers in Education (Simpósio Brasileiro de Informática na Educação-SBIE)*, Vol. 30. 1451.

[69] Filipe D Pereira, Elaine HT Oliveira, David BF Oliveira, Alexandra I Cristea, Leandro SG Carvalho, Samuel C Fonseca, Armando Toda, and Seiji Isotani. 2020. Using learning analytics in the Amazonas: understanding studentsâĂŹ behaviour in introductory programming. *British Journal of Educational Technology* (2020).

[70] Filipe D Pereira, Armando Toda, Elaine HT Oliveira, Alexandra I Cristea, Seiji Isotani, Dion Laranjeira, Adriano Almeida, and Jonas Mendonça. 2020. Can we use gamification to predict studentsâĂŹ performance? A case study supported by an online judge. In *International Conference on Intelligent Tutoring Systems.* Springer, 259–269.

[71] Daniel H Pink. 2011. *Drive: The surprising truth about what motivates us.* Penguin.

[72] Ana Isabel Polo-Peña, Dolores María Frías-Jamilena, and María Lina Fernández-Ruano. 2020. Influence of gamification on perceived self-efficacy: gender and age moderator effect. *International Journal of Sports Marketing and Sponsorship* (2020).

[73] Yizhou Qian and James Lehman. 2017. StudentsâĂŹ Misconceptions and Other Difficulties in Introductory Programming: A Literature Review. *ACM Trans. Comput. Educ.* 18, 1, Article 1 (Oct. 2017), 24 pages. https://doi.org/10.1145/3077618

[74] Katherine A Rawson. 2015. The status of the testing effect for complex materials: Still a winner. *Educational Psychology Review* 27, 2 (2015), 327–331.

[75] Anthony V Robins. 2019. Novice programmers and introductory programming. *The Cambridge Handbook of Computing Education Research, Cambridge Handbooks in Psychology* (2019), 327–376.

[76] Luiz Rodrigues, Robson Bonidia, and Jacques Brancher. 2020. Procedural versus human level generation: Two sides of the same coin? *International Journal of Human-Computer Studies* 141 (2020), 102465.

[77] Luiz Rodrigues, Wilk Oliveira, Armando Toda, Paula Palomino, and Seiji Isotani. 2019. Thinking Inside the Box: How to Tailor Gamified Educational Systems Based on Learning Activities Types. In *Proceedings of the Brazilian Symposium on Computers in Education (SimpÃşsio Brasileiro de InformÃątica na EducaÃğÃčo - SBIE).*

[78] Luiz Rodrigues, Paula T. Palomino, Armando M. Toda, Ana C. T. Klock, Wilk Oliveira, Anderson P. Avila-Santos, Isabela Gasparini, , and Seiji Isotani. 2021. Personalization Improves Gamification: Evidence from a Mixed-methods Study. *Proc. ACM Hum.-Comput. Interact.* 5, CHI PLAY, Article 287 (sep 2021), 25 pages. https://doi.org/10.1145/3474714

[79] Luiz Rodrigues, Armando M Toda, Wilk Oliveira, Paula T Palomino, Anderson Paulo Avila-Santos, and Seiji Isotani. 2021. Gamification Works, but How and to Whom? An Experimental Study in the Context of Programming Lessons. In *Proceedings of the 52nd ACM Technical Symposium on Computer Science Education*. 184–190.

[80] Luiz Rodrigues, Armando M Toda, Wilk Oliveira, Paula T Palomino, and Seiji Isotani. 2020. Just beat it: Exploring the influences of competition and task-related factors in gamified learning environments. In *Anais do XXXI Simpósio Brasileiro de Informática na Educação*. SBC, 461–470.

[81] Luiz Rodrigues, Armando M. Toda, Wilk Oliveira, Paula T. Palomino, Julita Vassileva, and Seiji Isotani. 2021. Automating Gamification Personalization: To the User and Beyond. arXiv:2101.05718 [cs.HC]

[82] Luiz Rodrigues, Armando M Toda, Paula T Palomino, Wilk Oliveira, and Seiji Isotani. 2020. Personalized gamification: A literature review of outcomes, experiments, and approaches. In *Eighth International Conference on Technological Ecosystems for Enhancing Multiculturality*. 699–706.

[83] Henry L Roediger-III and Jeffrey D Karpicke. 2006. Test-enhanced learning: Taking memory tests improves long-term retention. *Psychological science* 17, 3 (2006), 249–255.

[84] Christopher A Rowland. 2014. The effect of testing versus restudy on retention: a meta-analytic review of the testing effect. *Psychological Bulletin* 140, 6 (2014), 1432.

[85] Michael Sailer and Lisa Homner. 2019. The Gamification of Learning: a Meta-analysis. *Educational Psychology Review* (15 Aug 2019). https://doi.org/10.1007/s10648-019-09498-w

[86] Diana R. Sanchez, Markus Langer, and Rupinder Kaur. 2020. Gamification in the classroom: Examining the impact of gamified quizzes on student learning. *Computers & Education* 144 (2020), 103666. https://doi.org/10.1016/j.compedu.2019.103666

[87] Bianca L Santana and Roberto A Bittencourt. 2018. Increasing motivation of cs1 non-majors through an approach contextualized by games and media. In *2018 IEEE Frontiers in Education Conference (FIE)*. IEEE, 1–9.

[88] Isabelle Savard and Riichiro Mizoguchi. 2019. Context or culture: what is the difference? *Research and Practice in Technology Enhanced Learning* 14, 1 (2019), 1–12.

[89] Sofia Schöbel and Matthias Söllner. 2016. How to Gamify Information Systems - Adapting Gamification to Individual User Preferences. In *24th European Conference on Information Systems*.

[90] Katie Seaborn and Deborah I Fels. 2015. Gamification in theory and action: A survey. *International Journal of human-computer studies* 74 (2015), 14–31.

[91] Geneviève Taylor, Tomas Jungert, Geneviève A Mageau, Kaspar Schattke, Helena Dedic, Steven Rosenfield, and Richard Koestner. 2014. A self-determination theory approach to predicting school achievement over time: The unique role of intrinsic motivation. *Contemporary Educational Psychology* 39, 4 (2014), 342–358.

[92] Armando Toda, Filipe Dwan Pereira, Ana Carolina Tomé Klock, Luiz Rodrigues, Paula Palomino, Wilk Oliveira, Elaine Harada Teixeira Oliveira, Isabela Gasparini, Alexandra Ioana Cristea, and Seiji Isotani. 2020. For whom should we gamify? Insights on the users intentions and context towards gamification in education. In *Anais do XXXI Simpósio Brasileiro de Informática na Educação*. SBC, 471–480.

[93] Armando M Toda, Ana CT Klock, Wilk Oliveira, Paula T Palomino, Luiz Rodrigues, Lei Shi, Ig Bittencourt, Isabela Gasparini, Seiji Isotani, and Alexandra I Cristea. 2019. Analysing gamification elements in educational environments using an existing Gamification taxonomy. *Smart Learning Environments* 6, 1 (2019), 16.

[94] Armando M. Toda, Pedro H. D. Valle, and Seiji Isotani. 2018. The Dark Side of Gamification: An Overview of Negative Effects of Gamification in Education. In *Higher Education for All. From Challenges to Novel Technology-Enhanced Solutions*, Alexandra Ioana Cristea, Ig Ibert Bittencourt, and Fernanda Lima (Eds.). Springer International Publishing, Cham, 143–156.

[95] Gustavo F Tondello, Alberto Mora, and Lennart E Nacke. 2017. Elements of gameful design emerging from user preferences. In *Proceedings of the Annual Symposium on Computer-Human Interaction in Play*. ACM, 129–142.

[96] Gustavo F Tondello and Lennart E Nacke. 2020. Validation of User Preferences and Effects of Personalized Gamification on Task Performance. *Frontiers in Computer Science* 2 (2020), 29.

[97] Gustavo F Tondello, Hardy Premsukh, and Lennart Nacke. 2018. A theory of gamification principles through goal-setting theory. Hawaii International Conference on System Sciences.

[98] Christiane Gresse von Wangenheim and Aldo von Wangenheim. 2012. Ensinando computação com jogos. *Bookess Editora, Florianópolis, SC, Brasil* (2012).

[99] Szymon Wasik, Maciej Antczak, Jan Badura, Artur Laskowski, and Tomasz Sternal. 2018. A survey on online judge systems and their applications. *ACM Computing Surveys (CSUR)* 51, 1 (2018), 1–34.

[100] Robert Wellington. 2015. Context to Culture for Gamification HCI Requirements: Familiarity and Enculturement. In *Gamification in Education and Business*. Springer, 151–163.

[101] Zamzami Zainuddin, Samuel Kai Wah Chu, Muhammad Shujahat, and Corinne Jacqueline Perera. 2020. The impact of gamification on learning and instruction: A systematic review of empirical evidence. *Educational Research Review* (2020), 100326.

[102] Amy L Zeldin, Shari L Britner, and Frank Pajares. 2008. A comparative study of the self-efficacy beliefs of successful men and women in mathematics, science, and technology careers. *Journal of Research in Science Teaching: The Official Journal of the National Association for Research in Science Teaching* 45, 9 (2008), 1036–1058.

[103] Wayne Xin Zhao, Wenhui Zhang, Yulan He, Xing Xie, and Ji-Rong Wen. 2018. Automatically learning topics and difficulty levels of problems in online judge systems. *ACM Transactions on Information Systems (TOIS)* 36, 3 (2018), 1–33.

ANNEX

E

GAMIFICATION SUFFERS FROM THE NOVELTY EFFECT BUT BENEFITS FROM THE FAMILIARIZATION EFFECT: FINDINGS FROM A LONGITUDINAL STUDY

International Journal of Educational Technology in Higher Education

RESEARCH ARTICLE

Open Access

Gamification suffers from the novelty effect but benefits from the familiarization effect: Findings from a longitudinal study

Luiz Rodrigues[1*], Filipe D. Pereira[2,3], Armando M. Toda[1,3], Paula T. Palomino[1,4], Marcela Pessoa[5,6], Leandro Silva Galvão Carvalho[6], David Fernandes[6], Elaine H. T. Oliveira[6], Alexandra I. Cristea[3] and Seiji Isotani[1]

*Correspondence: lalrodrigues@usp.br
[1] Institute of Mathematics and Computer Science, University of São Paulo, Avenida Trabalhador São Carlense, 400 - Centro, São Carlos, SP 13566-590, Brazil
Full list of author information is available at the end of the article

Abstract

There are many claims that gamification (i.e., using game elements outside games) impact decreases over time (i.e., the novelty effect). Most studies analyzing this effect focused on extrinsic game elements, while fictional and collaborative competition have been recently recommended. Additionally, to the best of our knowledge, no long-term research has been carried out with STEM learners from introductory programming courses (CS1), a context that demands encouraging practice and mitigating motivation throughout the semester. Therefore, the main goal of this work is to better understand how the impact of a gamification design, featuring fictional and competitive-collaborative elements, changes over a 14-week period of time, when applied to CS1 courses taken by STEM students (N = 756). In an ecological setting, we followed a 2x7 quasi-experimental design, where Brazilian STEM students completed assignments in either a gamified or non-gamified version of the same system, which provided the measures (number of attempts, usage time, and system access) to assess user behavior at seven points in time. Results indicate changes in gamification's impact that appear to follow a U-shaped pattern. Supporting the novelty effect, the gamification's effect started to decrease after four weeks, decrease that lasted between two to six weeks. Interestingly, the gamification's impact shifted to an uptrend between six and 10 weeks after the start of the intervention, partially recovering its contribution naturally. Thus, we found empirical evidence supporting that gamification likely suffers from the novelty effect, but also benefits from the familiarization effect, which contributes to an overall positive impact on students. These findings may provide some guidelines to inform practitioners about how long the initial contributions of gamification last, and how long they take to recover after some reduction in benefits. It can also help researchers to realize when to apply/evaluate interventions that use gamification by taking into consideration the novelty effect and, thereby, better understand the real impact of gamification on students' behavior in the long run.

Keywords: Gameful, Education, Learning, Quasi-experiment Repeated-measures

Introduction

Game elements have been widely added outside games, approach known as gamification, with the goal of affecting user behavior Deterding et al. (2011); Helmefalk (2019). The educational domain is where gamification has been researched the most Koivisto and Hamari (2019), which is in line with the view that it is an educational innovation that can enhance learning experiences Palomino et al. (2020); Hernández et al. (2021). Accordingly, empirical evidence supports its positive effect on, for instance, behavioral (e.g., class attendance) and cognitive (e.g., test performance) learning outcomes Briffa et al. (2020); Huang et al. (2020). However, researchers often discuss that gamification suffers from the *novelty effect* Hamari et al. (2014); Nacke and Deterding (2017). That is, it leads to positive outcomes when introduced and, as its novelty goes, so does the positive effect Clark (1983).

Studies have provided empirical evidence supporting the novelty effect in gamified learning. Those studies show that, for instance, users' perceived benefits from a gamified service decreased as the time using that service increased Koivisto and Hamari (2014) and that there are cases when gamification's impact on student behavior and motivation changed after an initial positive effect Sanchez et al. (2020); Rodrigues et al. (2021). Such findings question whether gamification can lead to long-lasting benefits. Meta-analytic evidence, on the other hand, suggest the moderator effect of intervention duration on behavioral outcomes is nonsignificant Sailer and Homner (2019). Accordingly, researchers have called for longitudinal studies tracking gamification's impact over long time-periods Bai et al. (2020); Alsawaier (2018); Kyewski and Krämer (2018).

This article responds to literature calls by presenting a longitudinal analysis of gamification's effects on behavioral outcomes over the course of a full semester. Specifically, we analyze data of Brazilian undergraduate students from 14 STEM[1] courses, which were enrolled at the Introductory Programming (CS1) subject. Gamification is especially valuable in that context because, oftentimes, CS1 is not motivating for STEM students, which, along with the complexity and substantial practice needed to learn the topic, leads to high drop-out and failure rates Santana and Bittencourt (2018); Fonseca et al. (2020); Pereira et al. (2020). However, while gamification might be a way to encourage positive learning behaviors for CS1, STEM learners, there is a lack of empirical evidence on gamification's long-term effects in that context. We address this gap by answering *How does gamification's effects on CS1, STEM students' behavioral outcomes change over time?*

Despite past research has studied similar questions Hanus and Fox (2015); Sanchez et al. (2020); Tsay et al. (2020), this one differs in two main perspectives, according to our best knowledge. First, research context: The literature indicates gamification's effect varies depending on the context Hamari et al. (2014); Liu et al. (2017); Toda et al. (2020), and a single research analyzed how gamification's effect changes over time when applied to CS1 students Rodrigues et al. (2021). Second, the gamification design. Meta-analyses indicate that gamified systems including *collaborative competition* and *fictional* game elements improve gamification's effect Sailer and Homner (2019); Huang et al. (2020);

[1] Science, Technology, Engineering, and Mathematics.

Toda et al. (2019), but only Putz et al. (2020) implemented both of those and Mavletova (2015) implemented one: fictional elements. However, those studies are limited to interventions lasting from a day to six weeks, while past research has called for longer analyses Bai et al. (2020); Alsawaier (2018); Nacke and Deterding (2017). This study address both of these perspectives, thus, contributing *new empirical evidence revealing how the effect of a gamification design featuring both fictional and competitive-collaborative elements changes, over a 14-week period, when applied to CS1, STEM students.*

Next, we review related work in terms of longitudinal studies assessing gamification's impact over time, discussing the main points in which this article adds to the literature ("Related work" Section). Then, we describe our study ("Method" Section) and present its results ("Results" Section). Subsequently, we discuss our findings as well as their implications and limitations ("Discussion" Section). Lastly, we draw our conclusions in Conclusions" Section.

Related work

Research has provided empirical evidence that users' perceptions about gamification usage change over time. Koivisto and Hamari found that the more users used a gamified app, the less the perceived benefit they reported was Koivisto and Hamari (2014). Similarly, Van Roy and Zaman found learners' motivations regarding a gamified course decreased throughout the semester, with indication that it then increased towards the end of the semester Van Roy and Zaman (2018). However, these studies are limited by not featuring a baseline comparison group. That is, a condition with no gamification. Hence, they do not allow the identification of whether the gamification's impact (the difference between conditions) changed overtime. This study addresses this need by comparing data from a gamified setting to a control condition with no gamification. Given that context, the remainder of this section reviews longitudinal studies focused on changes of gamification's effects over time. That is, empirical research comparing a gamified condition to a non-gamified setting (to identify gamification's effects) based on measures captured two or more times.

Some research analyzed changes on gamification's effect over time within classrooms. Hanus and Fox implemented gamification (using Badges, Coins, and a Leaderboard) to assess its effects on Communications students Hanus and Fox (2015). Measures were captured three times at every four weeks, for 14 weeks, revealing that the gamification's impact was nonsignificant at the first time point and negative for the subsequent ones for motivation, satisfaction, and empowerment. Grangeia et al. evaluated gamification's impact with final-year medical students, implemented using Levels, Medals, and Encouraging messages Grangeia et al. (2019). The authors evaluated log-data on a daily basis, longitudinally measuring student behavior and interaction with the technology used during a whole course (four weeks). Their results suggest a positive impact of gamification throughout the four weeks. Sanchez et al. added Progress bars, a Wager option, and Encouraging messages to quizzes offered in classrooms for psychology-related majors Sanchez et al. (2020). Students' performances in three tests were analyzed, which were captured from varied intervals (i.e., spacing varied between 3 and 5 weeks). Results indicated gamification was positive for the first test, but nonsignificant for the remaining ones. Tsay et al. explored gamification effects with undergraduate learners in the context

Table 1 Summary of related works and this study

Source	GF	CC	Context	#M	MS	ID
Mavletova (2015)	X		Online survey completion	2	2m	<1d
Hanus and Fox (2015)			Communication classroom	3	4w	16w
Mitchell et al. (2017)			Adults physical activities	3	1w	4w
Grangeia et al. (2019)			Medical students	28	1d	4w
Sanchez et al. (2020)			Psychology-related classrooms	3	3-5w	13w
Tsay et al. (2020)			Personal and Professional Development classroom	24	1w	24w
Putz et al. (2020)	X	X	Workshops on sustainable supply chain management	2	2w	<1d
Rodrigues et al. (2021)			CS1, Software Engineering classrooms	6	1w	6w
This study	X	X	CS1, STEM classrooms	7	2w	14w

GM Game fiction, *CC* Competitive-collaborative, *#M* Number of measures, *MS* Measures spacing, *ID* Intervention duration, *m* months, *w* weeks, *d* day

of a Personal and Professional Development module Tsay et al. (2020). Based on log data from two terms (24 weeks), the authors analyzed gamification's (Quests, Progression, Badges, and Leaderboard) impact weekly, finding that it decreased over time in the first term, but not in the second, results they attributed to changing the gamification design. Rodrigues et al. applied gamification (Badges, Objectives, and Collaboration) with CS1 undergraduates majoring in Software engineering Rodrigues et al. (2021). Measuring their motivation weekly, during six weeks, they found mixed effects that changed over time: at first, gamification was positive and negative for participants with low and high familiarity with coding, respectively. After the six weeks, the effects were the opposite.

Others performed similar research in domains apart from classrooms. Mitchell et al. experimented with a gamified app—featuring Variable difficulty levels, Player choice, and Dynamic feedback—in the context of physically active adults Mitchell et al. (2017). Authors measured participants' motivational and behavioral engagement three times over the course of four weeks, with one week of spacing between them. Unlike most research, findings indicated a positive gamification effect over time, although the decrease in the effect's magnitude. Mavletova conducted a study in the context of children completing surveys, analyzing gamification impact in two waves that had a two-month spacing Mavletova (2015). Results from using Narrative (a form of game fiction), Rules, Challenges, and Rewards were only positive in the first wave and, for some measures (e.g., survey test-retest reliability), negative in the second. Putz et al. analyzed gamification's impact on knowledge retention from workshops on sustainable supply chain management for two time points: 20 minutes and 14 days after the end of the intervention, which lasted around six hours Putz et al. (2020). The game elements used in the workshops were Time constraint, Storytelling (a form of game fiction), Rewards, Leaderboard, Collaborative Competition, Clear goals, Immediate feedback, Avatar, Points. Their findings suggest that gamification was positive for short-term knowledge retention, but roughly ineffective at the second time-point.

Summary

Compared to related work, this study advances the literature based on two main perspectives, summarized in Table 1. The first concerns the context of the study. The literature argues that gamification's success depends on the context Hamari et al. (2014); Liu et al. (2017); Toda et al. (2020). Thus, performing similar studies in different contexts is important to advance the understanding to new contexts Seaborn and Fels (2015). This study differs from the available body of research by analyzing changes of gamification's effect over time, in the context of CS1. While one study also worked with CS1 learners Rodrigues et al. (2021), they were not STEM students and the experiment was limited to 19 participants studied for six weeks. Differently, this study features 756 participants from STEM courses that were studied over a 14-week period.

The second perspective concerns the gamification design. The gamification design employed in most studies explored extrinsic, challenge-based game elements, such as rewards and leaderboards Toda et al. (2019). However, meta-analytic evidence demonstrated that adding fictional game elements (e.g., Narrative and Storytelling Palomino et al. (2019)), as well as competitive-collaborative social interactions, enhanced the effect of gamification on behavioral learning outcomes Sailer and Homner (2019). Despite that, a single study, in which the intervention lasted less than a day and measures were taken only two times, explored both kinds of elements Putz et al. (2020). Differently, this research explores both fictional and competitive-collaborative elements in a 14-week intervention, analyzing data from seven time-points. Thereby, our study is contributing to the understanding of how the impact of gamification design changes over a semester.

Based on those premises, this article's novelty relies in analyzing how the gamification effect change over time, in a context (CS1, STEM students) it has not been previously explored in the long-run (14 weeks), with a large sample (756) and using both fictional and competitive-collaborative game elements.

Method

This study responds to calls for longitudinal gamification studies Nacke and Deterding (2017); Bai et al. (2020). Accordingly, the objective is to answer the following research question:

> *How does gamification's effects on CS1, STEM students' behavioral outcomes change over time?*

Therefore, we report results stemming from a quasi-experimental study, conducted in an ecological setting, from 2016 to 2018. The study was conducted in the context of CS1 courses of 14 STEM undergraduation programs from Federal University of Amazonas (UFAM). The experimental tasks concern activities that were already included in the course schedule, which allowed studying learners in an ecological setting. The quasi-experimental characteristic emerges because random assignment was not possible, a common feature of studies performed in ecological settings Creswell and Creswell (2017). Instead, the assignment of conditions was based on the year in which a student enrolled in a discipline, within the context of the study. This assignment procedure was necessary, because the system used for data collection was developed and deployed in 2016, without the innovation of gamification and, only since 2017, its gamification design was deployed. Hence, students from 2016 were assigned to the control condition ($N = 138$), whereas those from 2017 and 2018 were assigned to the experimental condition ($N = 659$).

Design

This study follows a 2x7 design. The first factor is *gamification*, in which levels are *yes* and *no*. The second factor is *time*, wherein levels range from one to seven. Time one refers to data captured until the end of the second week of the intervention, and subsequent times refer to data captured within two-week intervals, until week 14. Based on this design, the first factor enables the identification of the gamification impact, whilst the second provides data for analysing changes of the gamification effect over time.

Participants

Seven-hundred and fifty-six ($N = 756$[2]) Brazilian students participated in this research. Their average age was 22.2 years old (Standard Deviation, SD: 4.2); 62.3% were males and 37.7% were females. All of them were students of the UFAM and were enrolled in a CS1 course of one of the following STEM classes: Computer Science (2%), Materials Engineering (9%), Oil and Gas Engineering (9%), Production Engineering (9%), Electrical Engineering (13%), Mechanical Engineering (9%), Chemical Engineering (10%), Statistics (7%), Bachelor in Physics (7%), Licensure in Physics (6%), Bachelor in Mathematics (5%), Licensure in Mathematics (11%), and Applied Mathematics (3%). As this study was conducted in an ecological setting, all students enrolled in any of those courses were eligible as participants. However, we only considered those 756 learners, because of two reasons. First, they properly completed the form providing informed consent to having their data used in scientific research and basic demographic information (e.g., age and gender). That was accomplished at the beginning of the first class of the term, before students started using the system. Hence, our approach complied to ethical standards and allowed us to understand characteristics of our study sample. Second, students completed the semester without failing by attendance. This is desirable as our goal is evaluating how the effect of gamification changes over time. That led to the removal of 19.7% and 24.2% of the students, who failed by attendance in the non-gamified and gamified groups, respectively.

Materials

Educational system

According to the ecological setting of the study, data was collected through Codebench[3]. It has been developed at UFAM, Brazil, and since 2016 has been used for supporting teaching the CS1 subject of all STEM classes of the university. Figure 1 shows a snapshot of the system's main usage page, taken from Galvão et al. (2016) with copyright holder's authorization.

Codebench is an online judge, which has been used to help both instructors and students. For instructors, the system contributes by providing a list of programming problems they can use to create assignment lists, to use as activities in their classes. For students, the system offers an IDE (Integrated Development Environment) where they can develop solutions for programming problems, suggested by instructors. Furthermore, the system provides automatic feedback about solution correctness, which is

[2] Forty-one participants were in both groups due to the ecological setting of the experiment.

[3] https://codebench.icomp.ufam.edu.br/

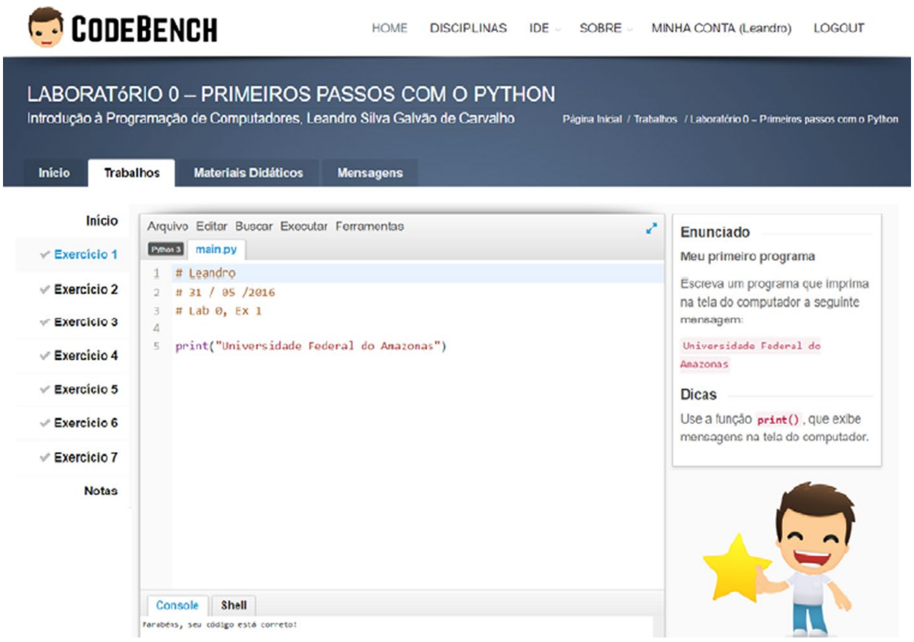

Fig. 1 Snapshot of the system participants used to complete this study's task. That image was taken from Galvão et al. (2016) with copyright holder's authorization

possible by testing each solution of a given problem against a set of hidden test cases. Hence, it helps students with instantaneous feedback, and instructors, by reducing the workload of correcting learner solutions.

Task

Given the nature of the system used for data collection, the task analyzed in this study concerns programming assignments students solved in Codebench. These assignments were offered by instructors, mainly for students to practice concepts and techniques introduced in classes, although solving them was only worth a small proportion of their final grade (around 6%). Hence, when analyzing gamification impact over time, we consider data from students solving non-compulsory programming assignments, which had a small impact on their final grade and were suggested by instructors. Moreover, we highlight that although the participants were learners from different STEM classes, all programming problems were selected from the same database, regardless of the class, and the same pedagogical materials and methodological plan were followed in the CS1 subject for all STEM courses. Thereby, this is mitigating confounding effects, which could emerge from having participants from different courses.

Experimental conditions

According to the need for first identifying gamification impact and, then, how it changes over time, this study features two experimental conditions. The first is the *control* condition, in which participants used Codebench without gamification. The second condition is the *experimental* one, in which participants used the same system, but with game elements; that is, they used the system's gamified version.

The gamified version was developed following the proposal by von Wangenheim and von Wangenheim (2012), integrating the Instructional Systematic Design model Dick et al. (2005) and the instrumental design ADDIE Branch (2009). The gamification goal is to motivate students to complete programming assignments, either during face-to-face or online classes, when more suitable to the learners, but always using Codebench. To achieve that goal, the gamification allows the students to see themselves as characters from a medieval fantasy world (narrative). The students can customize these characters (i.e., their avatar) and, if they receive a high score in a series of assignments, their avatar gets a badge that leads to a more powerful weapon. Furthermore, the gamification design also enables students to join their classmates (collaboration) to fight a monster, aiming to free the fantasy world from it. For this goal, there is a map that the students explore, as they solve programming assignments.

Note that as Codebench can be used by different subjects (i.e., not only CS1), the gamification design is not tied to any specific topic. Therefore, the integration between programming assignments and gamification happens once a student completes any assignment. When they submit a solution to the system, if the solution is correct, the student receives an in-game reward. That reward is a card selected through a draw, which can bring strength points and weapons to the student's avatar. The greater the strength, the better the student's weapon becomes, and the more hit points they take from the monster. Consequently, they have more chances of remaining in the best positions on the leaderboard, which is helpful, because the top three students receive a special reward. Nevertheless, a single student cannot defeat the monster alone, as this is a collective mission. Thus, a percentage of the class must reach the end of the map, to be able to beat the monster and, finally, win the challenge together (cooperation).

Given this context, we might summarize the two main elements of the gamification design as follows. First, the narrative aspect immerses students into a fictional story. To advance in that story, students have to complete real programming assignments—hence, connecting the learning activities to the gamified design. Second, the students have to collaborate to finish the gamified story, while they also compete with each other during that process, towards staying at the top of the leaderboard and gaining rewards to improve and customize their avatars. Thus, the competitive-collaborative aspect emerges, because students must work together; at the same time, they have to outperform others, to receive more rewards.

Measures

We analyze gamification impact based on three behavioral outcomes, which were all collected through data logs from Codebench. Those are:

- **Attempts:** The number of attempts that a students made. That is, how many times they submitted their solution to the online judge for checking the correctness of the code. Consequently, this measure accounts for both correct and incorrect submissions.
- **IDE Usage:** The time (in minutes) that a student used the IDE integrated in the online judge. Note that inactive usage does not count in this metric. Hence, we can see it as

a measure of the time a student spent working towards improving their programming skills; that is, practicing to code.

- **System Access:** How many times a student accessed the online judge. Thereby, it accounts for each time a student logged into the system. Thus, we can see it as a measure of students' intentions to practice, consult learning materials, and interact with the gamification.

Those measures were selected considering that practicing is prominent to learn to code Estey and Coady (2016); Pereira et al. (2021). We can see **attempts** as a measure of relevant learning behavior based on the testing effect: the idea that the more one is tested, the more one learns Rowland (2014); Rodrigues et al. (2021); Sanchez et al. (2020). Specifically, empirical evidence supports that the simple act of trying to complete a test is valuable for learning Goldstein (2014); Ahadi et al. (2016). Therefore, we might expect that the more attempts a student has, even if those are failed attempts, the better for their learning. For similar reasons, we might expect that **IDE usage** will benefit learning as well. For instance, empirical evidence demonstrates that time-on-task is positively related to learning outcomes Landers and Landers (2014); Pereira et al. (2020), which further supports the positive role spending time on the IDE has on learning. Lastly, **system access** reflects valuable learning behaviors, as it captures students' interest in practicing (i.e., using the IDE to submit a solution—attempt) and interacting with the gamification.

Procedure

According to the study setting, participants followed a straight-forward procedure. First, they provided informed consent for their data to be used in scientific research, and completed a brief demographic questionnaire, with information such as gender and age. Then, students were provided with assignment lists they could complete in Codebench. Note that although optional, completing these lists was worth 6% of the students' final grade. By completing the programming task-lists on the online judge, data used for analysis was generated, as the online judge automatically captured the participants' interactions, storing it onto log data. At the end of the semester, these logs were grouped, to represent students' interactions at each of the seven points in time analyzed in this study.

Data analysis process

Given the design of this study, we use a Robust ANOVA based on trimmed means Wilcox (2017) for data analysis. We chose this method because it behaves better than the traditional ANOVA for real world situations, while maintaining the flexibility of the traditional approach, as the literature discusses and recommends Cairns (2019). The rationale is that robust ANOVA using trimmed means maintains the ability to compare changes in location (e.g., how much one is bigger than the other), unlike rank-based tests (e.g., Kruskal-Wallis), while still providing results robust to violations of assumptions, such as non-normality and homogeneity of sphericity Keselman et al. (2000); Kowalchuk et al. (2003) (for a comprehensive discussion, see Cairns (2019)). Hence, we use the

Robust ANOVA[4] to test whether there is an interaction between the two factors (gamification and time). This can indicate how the effect of gamification changes over time.

In cases of significant interactions, we perform post-hoc tests using the Yuen-Welch test Yuen (1974), following literature recommendations Cairns (2019). Here, we limit the post-hoc analysis to testing the difference between conditions (gamified and non-gamified) for each time-point, aiming to identify the magnitude of the difference, as well as finding how they change (by how much they increase/decrease) over time. Accordingly, difference magnitudes are estimated using the explanatory measure of effect size introduced by Wilcox and Tian Wilcox and Tian (2011), which is a robust measure, aligned to the Yuen-Welch test. Effect sizes of 0.1, 0.3, and 0.5 are interpreted as small, moderate, and large, respectively. The alpha level is set to 0.05 for all tests and p-values are corrected using the false-discovery-rate approach, which has been recommended over the more common Bonferroni alternative Jafari and Ansari-Pour (2019). Inferential analyses were conducted using R R Core Team (2020), R studio RStudio Team (2019), and the *WRS2* package Mair and Wilcox (2018).

Results

First, this section presents preliminary analyses due to the study design. Then, it introduces analyses and results answering our research question.

Preliminary analyses

Due to the lack of random assignment, we compare the demographic characteristics of participants for each experimental condition. In terms of gender, 38% and 62% of the participants were females and males, respectively, for both experimental conditions. In terms of age, those in the control condition had an average age of 23.3 years (SD: 4.0), whereas those in the experimental setting were, on average, 22 years old (SD: 4.2). While there was no difference in gender, the difference in age might confound the results. Then, we further examined whether age is a predictor for any of the behavioral measures.

- For attempts, linear regression results were statistically significant, $F(1, 7782) = 18.42$, $p < 0.001$, $R^2 = 0.002$, adj-$R^2 = 0.002$, and age was a significant predictor, $t = -4.29$, $p < 0.001$, $B = -0.58$, $\beta = -0.05$.
- For IDE usage, linear regression results were not statistically significant, $F(1, 7782) = 2.48$, $p = 0.116$, $R^2 <$ textless 0.001, adj-$R^2 < 0.001$.
- For system access, linear regression results were statistically significant, $F(1, 7782) = 12.25$, $p < 0.001$, $R^2 = 0.002$, adj-$R^2 = 0.001$, and age was a significant predictor, $t = -3.5$, $p < 0.001$, $B = -0.38$, $\beta = -0.04$.

These findings indicate that, although statistically significant for two of the three behavioral measures, age's practical effect is negligible, based on the proportion of variance explained (all R^2 <1%) and standardized coefficients (all $\beta \leq 0.05$) Kotrlik and Williams (2003). Therefore, these results mitigate the threat of age confounding our results as the

[4] Note that we rely on a robust alternative because it handles such violations when they are present, as well as provides similar results to the standard approach, otherwise Cairns (2019); Wilcox (2017). Additionally, we do not test for assumption violations (e.g., normality) as the reliability of such tests has been critiqued Cairns (2019).

Table 2 Descriptive statistics for behavioral outcomes of participants from control (CTR; non-gamified) and experimental (EXP; gamified) conditions for seven time-points (TP) of the semester, as well as overall (1-7). Data represented as Mean±Standard Deviation

	Attempts		IDE usage		System access	
TP	CTR	EXP	CTR	EXP	CTR	EXP
1-7	40±51	47±48	80±94	119±118	29±24	64±39
1	62±53	85±52	112±64	185±103	19±19	48±27
2	35±33	57±57	56±42	126±82	22±12	52±30
3	68±73	61±53	134±134	185±159	33±16	60±32
4	42±59	36±36	89±113	108±112	42±23	66±36
5	29±37	32±35	76±90	95±107	25±26	67±39
6	22±26	28±31	48±68	74±93	30±27	73±44
7	24±35	29±37	47±81	60±83	35±31	79±47

difference in groups' overall age is unlikely to be the source of relevant differences in student behavior.

Does gamification's effect change over time?

Table 2 presents an overview of the behavioral measures, overall (1-7) and for each time-point, for both experimental conditions. The table shows that participants in the experimental condition (gamified): attempted problems, used the IDE, and accessed the system 17.5%, 48.6%, and 220% more, respectively, than those with the non-gamified setting. These magnitude differences are small ($g = 0.144$), moderate ($g = 0.341$), and large ($g = 0.948$) according to Hedge's g[5]. Similarly, Fig. 2 presents boxplots that demonstrate the distribution of each behavioral measure for each group in each session. Next, we present results from hypothesis tests, to provide statistical evidence on how the impact of gamification changes over time.

First, we analyzed student attempts. A robust two-way ANOVA on 20% trimmed means was used to evaluate the effects of *gamification* and *time* on student attempts. A statistically significant interaction between the effects of gamification and time was found, $F(6, 410.3948) = 4.5677$, $p < 0.001$. The main effects of gamification, $F(1, 429.5147) = 41.2818$, $p < 0.001$, and time, $F(6, 409.9150) = 73.1199$, $p < 0.001$, were statistically significant as well. Next, we conducted post-hoc analyses using the Yuen-Welch tests to assess differences of gamification impact over the seven time points. Table 3 presents the results from comparing the attempts of participants of both experimental conditions. It shows results from the Yuen-Welch test, comparing groups and estimates of the difference in location among groups. The results reveal moderate to large statistically significant differences in the number of attempts in time points one and two, insignificant differences in time-points three to five, and small to moderate statistically significant differences in time-points six and seven.

Second, we analyzed students' IDE usage. We repeated the previous analysis, but with students' IDE usage as dependent variable. A statistically significant interaction between

[5] Calculated using www.polyu.edu.hk/mm/effectsizefaqs/calculator/calculator.html

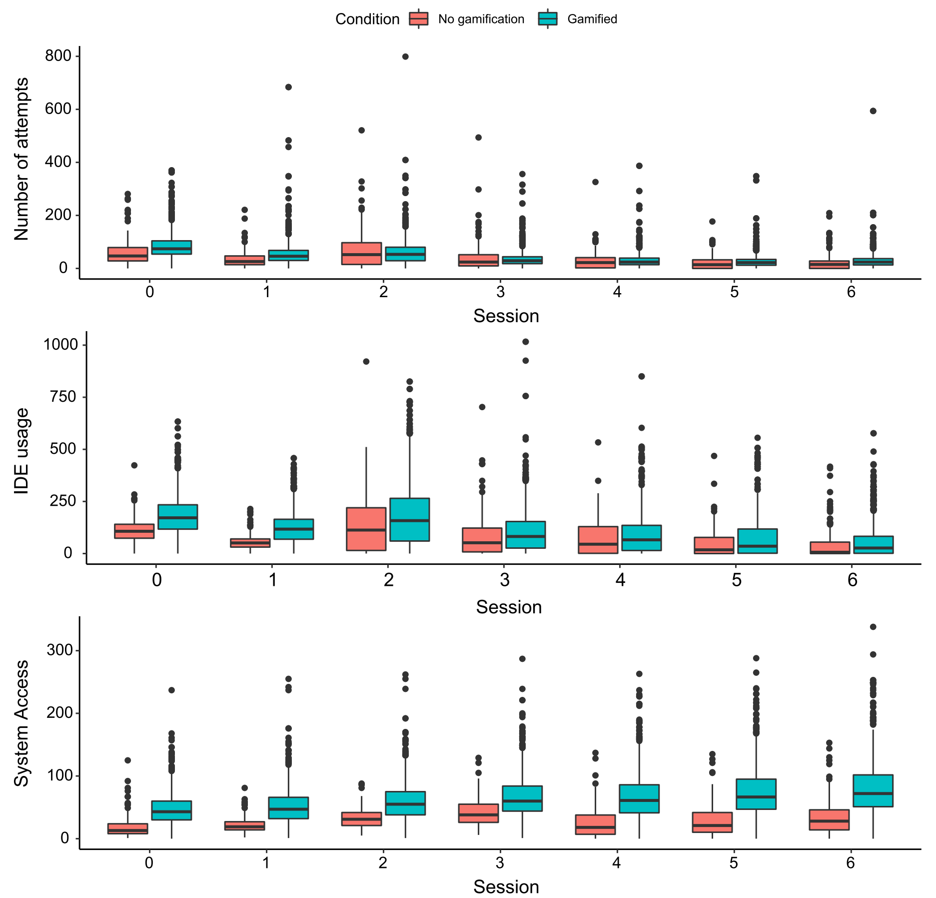

Fig. 2 Boxplots demonstrating the distribution of each behavioral measure of each condition for each session

Table 3 Comparison of experimental conditions (control, non-gamified; and experimental; gamified) in terms of participant attempts for each point in time (TP)

	Yuen-Welch test				Effect size	
TP	t(df)	p	p-adj	95% CI	ES	95% CI
1	7.872 (121.11)	0.000	0.000	− 33.123 − 19.811	0.448	0.305 0.596
2	7.781 (130.84)	0.000	0.000	− 22.894 − 13.613	0.443	0.296 0.584
3	0.023 (97.45)	0.982	0.982	− 10.969 10.717	0.033	0.000 0.194
4	0.717 (95.90)	0.475	0.555	− 8.500 3.991	0.061	0.000 0.249
5	1.468 (102.10)	0.145	0.203	− 8.285 1.236	0.100	0.000 0.260
6	3.523 (103.19)	0.001	0.001	− 11.192 − 3.130	0.240	0.070 0.378
7	5.407 (113.15)	0.000	0.000	− 13.916 − 6.453	0.331	0.178 0.467

the effects of gamification and time was found, $F(6, 412.7856) = 7.5375$, $p < 0.001$. The main effects of gamification, $F(1, 440.7169) = 91.5288$, $p < 0.001$, and time, $F(6, 412.3179) = 93.1663$, $p < 0.001$, were statistically significant as well. Next, we conducted post-hoc analyses using the Yuen-Welch tests, to assess differences in gamification impact over the seven time-points. Table 4 presents the results from comparing the IDE

Table 4 Comparison of experimental conditions (control, non-gamified; and experimental; gamified) in terms of participants' IDE usage in each time-point (TP)

	Yuen-Welch test				Effect size	
TP	t(df)	p	p-adj	95% CI	ES	95% CI
1	12.132 (186.58)	0.000	0.000	− 78.997 − 56.898	0.599	0.469 0.719
2	18.060 (282.93)	0.000	0.000	− 73.875 − 59.354	0.717	0.594 0.812
3	3.347 (120.61)	0.001	0.002	− 72.908 − 18.713	0.213	0.064 0.367
4	3.692 (126.42)	0.000	0.001	− 44.746 − 13.517	0.214	0.067 0.368
5	1.935 (113.97)	0.056	0.056	− 34.214 0.405	0.123	0.000 0.288
6	3.233 (154.58)	0.002	0.002	− 31.532 − 7.611	0.199	0.038 0.346
7	3.227 (132.07)	0.002	0.002	− 25.468 − 6.109	0.199	0.037 0.354

Table 5 Comparison of experimental conditions (control, non-gamified; and experimental; gamified) in terms of participants' system access for each time-point (TP)

	Yuen-Welch test				Effect size	
TP	t(df)	p	p-adj	95% CI	ES	95% CI
1	22.458 (199.67)	0.000	0.000	− 32.564 − 27.307	0.794	0.721 0.882
2	22.252 (303.48)	0.000	0.000	− 30.655 − 25.673	0.822	0.737 0.899
3	15.019 (193.11)	0.000	0.000	− 27.913 − 21.433	0.708	0.582 0.806
4	10.696 (152.51)	0.000	0.000	− 26.744 − 18.404	0.561	0.440 0.678
5	18.783 (152.56)	0.000	0.000	− 47.941 − 38.816	0.783	0.680 0.863
6	18.640 (163.20)	0.000	0.000	− 49.943 − 40.375	0.780	0.670 0.860
7	16.877 (153.56)	0.000	0.000	− 50.217 − 39.692	0.749	0.629 0.832

usage of participants of both experimental conditions. It shows results from the Yuen-Welch test, comparing groups and estimates of the difference in location among groups. The results reveal large, statistically significant differences in IDE usage for the first two time-points and small, statistically significant differences for all the remaining ones, with the exception of point five, in which the difference is only marginally significant.

Third, we analyzed system access, with the approach used for the other measures. A statistically significant interaction between the effects of gamification and time was found, $F(6, 422.4725) = 7.4256$, $p < 0.001$. The main effects of gamification, $F(1, 840.7586) = 672.3375$, $p < 0.001$, and time, $F(6, 422.6884) = 30.5048$, $p < 0.001$, were statistically significant, as well. Next, we conducted post-hoc analyses using the Yuen-Welch tests, to assess differences of gamification impact over the seven points in time. Table 5 presents the results from comparing the system access of participants of both experimental conditions. It shows results from the Yuen-Welch test, comparing groups and estimates of the differences among groups. The results reveal large, statistically significant differences in IDE usage for all time points, despite the decrease in time-point four, which was readily recovered in time-point five.

How does the effect of gamification change over time?

The way how the effect of gamification changes over time, based on the effect sizes estimated in previous analyses, is shown in Fig. 3. In this figure, time-points are

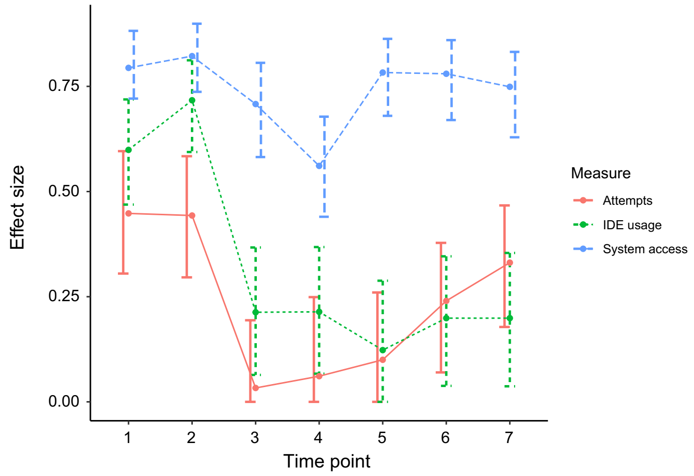

Fig. 3 Gamification's effect (compared to no gamification) over time, based on effect sizes estimated from differences in behavioral outcomes. Error bars represent the 95% confidence intervals estimated from effect sizes

shown on the x-axis and the magnitude of the effect size is shown on the y-axis, with each measure represented by a different line type and color and presenting error bars, based on the confidence intervals estimated in the analyses presented previously. The figure demonstrates that, in terms of attempts, the effect of gamification substantially decreased after the second time-point but, thereafter, progressively increased, until time-point seven. For IDE usage, the figure demonstrates that the effect of gamification increased from time one to time two but, subsequently, incrementally decreased, until time-point five and, in the end, suffered a small increase. In addition, the figure shows that the effect of gamification on system access also increased from time point one to time-point two, then decreased until time-point four, subsequently shifting to an increase in time-point five and, lastly, roughly decreasing, until time-point seven.

Overall, these results suggest that, over time, the effect of gamification on behavioral learning outcomes change, following, to some extent, a U-curve. We must note, however, that this conclusion is the product of an exploratory analysis. That is, while we expected that the effect of gamification would vary over time, we had no assumption on the shape of such a variation, prior to the data analysis. Accordingly, we intentionally chose to first analyze variations in gamification effect based on effect sizes estimated during post-hoc testing. Then, we analyzed such results by inspecting plots featuring confidence intervals. Importantly, our approach differs from the standard, quantitative analyses that mainly rely on p-values (e.g., a polynomial regression). Our justification is that the literature discourages using such a quantitative approach in the context of exploratory studies Vornhagen et al. (2020); Cairns (2019); Dragicevic (2016). Therefore, according to the exploratory nature of our analysis, we followed a more qualitative, visual interpretation, based on confidence intervals, as the above literature recommends. Thus, summarizing such results, we have i) empirical evidence that the gamification effect does change over time and ii) a model (i.e., the U-shaped function) of how such variation occurs during the course of a 14-week period, which must be further explored and validated in future research.

Discussion

Based on our results, this article's novelty lies in providing the following contributions, which have not been addressed in prior research, to the best of our knowledge:

- Empirical evidence showing the long-term effectiveness of an innovative gamification design, which features fictional and competitive-collaborative game elements, in improving positive learning behaviors;
- Empirical evidence revealing how the effect of gamification decreases (novelty effect) and naturally increases, with no intervention (familiarization effect) over time;
- A heuristic on a 'safe period' in which educators can use gamification before its effect starts to decrease;
- A heuristic on how long gamification studies should last to ensure their findings are roughly stabilized.

To further support these contributions, the remainder of this section discusses our findings, their implications, this article's limitations, and future research recommendations.

Findings

First, our findings revealed that the effect of gamification effect decreased after four weeks, for all behavioral measures, although the magnitude of that change differed between measures. THe effect of gamification either was maintained from the first time-point, or it increased between time-points one and two. However, it decreased from moderate-to-large to negligible, large to small, and had it size diminished for attempts, IDE usage, and system access, respectively. These findings support the novelty effect as, during the first four weeks, the effects were positive, after which they decreased. These findings corroborate most previous research, showing gamification impact decreasing over a period of time (e.g., Mitchell et al. (2017); Putz et al. (2020); Sanchez et al. (2020)). Compared to these, our study expands the literature, by providing evidence that the novelty effect holds in a different context, as well as for a distinct gamification design. In addition, this study expands the literature by demonstrating the extent to which the novelty effect influences the gamification impact, in terms of changes on effect size magnitudes. We also show that changes in magnitudes differ among similar measures, similar to what Mavletova (2015) found in the context of online surveys, yet extended to our context and gamification design. On the other hand, results by Grangeia et al. Grangeia et al. (2019) suggested that gamification impact was positive throughout the whole intervention. Whereas we show that this impact decreased after four weeks, their intervention lasted only four weeks. Therefore, considering the design we employed features elements likely to enhance gamification potential Sailer and Homner (2019), their results might be due to the limited intervention duration, or the different gamification design.

Second, our results indicate that after the decrease, the gamification effect showed a trend of increasingly positive impacts. While the decrease first appeared from time-point two to three, it promptly ended for attempts, but lasted until time-points four and five for system access and IDE usage, respectively. After the end of the decline, our results show an increase in gamification impact for all measures, which either continued to increase (attempts) or achieved a rough plateau (IDE usage and system access). These findings support the *familiarization effect*, in which users need some time to fully familiarize themselves with gamification. This effect was previously discussed by Van Roy and Zaman Van Roy and Zaman (2018), when they found that learner motivation appeared to follow a polynomial behavior over time. Similarly, Tsay et al. Tsay et al. (2020) noted that, over time, the impact of gamification decreased in the first term, but not in the second one, which also suggests a familiarization effect. However, neither of these studies performed such analyses by comparing data from the gamification intervention to a control, non-gamified group. Therefore, our contribution here is providing empirical support for the familiarization effect, based on a quasi-experimental study, comparing data from gamified and non-gamified interventions, besides extending it to the context of Brazilian students from CS1 classes from STEM courses, as well as to our gamification design. Additionally, it is worth noting other long-term studies (e.g., Hanus and Fox (2015); Sanchez et al. (2020)) are limited in the number of time-points considered (three) and data analysis method (linear modeling), which prevented them from finding such a non-linear pattern. Thus, our findings suggest that the decrease in gamification impact, likely due to the novelty effect, might happen because learners need some time to familiarize themselves with it.

Third, our findings revealed that, overall, the gamification effect was positive, and, importantly, that it was not negative at any point in time. Considering the whole intervention period, all behavioral measures of participants in the gamified condition outperformed those in the control setting. Considering each time-point separately, effect sizes estimated in the analyses corroborate that positive impact, with the exception of a few points in which gamification impact was nonsignificant. This finding is aligned to the overall gamification literature, which shows most studies report positive with null results Koivisto and Hamari (2019). Additionally, these findings are aligned to the gamified learning literature Sailer and Homner (2019), revealing average small-to-moderate positive impacts on behavioral learning outcomes. The literature has demonstrated cases in which gamification applied to education is associated with negative outcomes Toda et al. (2018); Hyrynsalmi et al. (2017), which is harmful to students' learning experience and reinforces the need to carefully designing gamification, to prevent such cases. A possible reason for avoiding these pitfalls might be the inclusion of fictional and competitive-collaborative elements, which have been suggested as positive moderators of gamification's impact on behavioral outcomes Sailer and Homner (2019). In this sense, this study contributes by scrutinizing the effects from the employed gamification design over several time periods, finding that besides being positive on average, it was not harmful to learners at any point in time.

Finally, these findings additionally demonstrate the effectiveness of adding an innovative design to the educational process of teaching programming. On one hand, gamification has been seen as an educational innovation, with the potential to improve learning experiences Hernández et al. (2021); Palomino et al. (2020). However, most research on gamification has been limited to standard designs, featuring elements such as points, badges, and leaderboards Bai et al. (2020); Koivisto and Hamari (2014);

prior gamification research also lacked long-term studies Nacke and Deterding (2017); Alsawaier (2018); Kyewski and Krämer (2018). On the other hand, unlike most prior research, this paper analyzed a gamification design featuring fictional game elements Toda et al. (2019). Additionally, that design also features competitive-collaborative elements, while most research is often limited to leaderboard-based competition Mustafa and Karimi (2021). Such distinction is important, because, despite being rarely used, such fictional and competitive-collaborative game elements are considered maximizers of gamification's effects Sailer and Homner (2019). Therefore, Codebench provides to students an innovative educational design, not only in terms of being gamified, but also because its gamification design itself is innovative, compared to the prior works. Thus, this article contributes with evidence demonstrating that adding an innovative educational design to programming lessons, implemented through gamification featuring both fictional and competitive-collaborative game elements, is able to improve learning behaviors in the long run.

Nevertheless, in addition to the gamification design, this study differs from previous research in terms of the context. Compared to Mavlova Mavletova (2015), Mitchell et al. Mitchell et al. (2017), and Putz et al. Putz et al. (2020), contexts differ, because we focused on the learning domain and they did not. Compared to Hanus and Fox Hanus and Fox (2015), Grangeia et al. Grangeia et al. (2019), Sanchez et al. Sanchez et al. (2020), and Tsay et al. Tsay et al. (2020), contexts differ because neither of them involved CS1. Consequently, our study context differs from theirs in terms of the task: we gamified programming assignments, whereas coding is not part of the classes involved in their research (e.g., psychology, medicine, and communication). Compared to Rodrigues et al. Rodrigues et al. (2021), who also gamified CS1 activities, most of those activities were not coding assignments. Additionally, their participants were Software Engineering students, a course highly linked to coding. In our study, most participants were students of STEM courses, such as Electrical Engineering and Mathematics. That difference is key, as such STEM students often struggle to see the value of coding for their future occupations Fonseca et al. (2020); Pereira et al. (2020). Therefore, we can summarize that our study context differs from those of prior research in terms of domain, learning task, and major. Note that acknowledging such differences is important because research suggests that one's experience with gamification might change, depending on the task Liu et al. (2017); and that different designs should be adopted depending on the learning activity Rodrigues et al. (2019); Hallifax et al. (2019). Consequently, the difference in contexts might have affected our findings when compared to similar studies. Thus, we also contribute to the discussions on the role of contextual aspects (e.g., domain, task type, and major) in gamification, raising the question of how those factors affect gamification effectiveness.

Implications

Our findings have major implications for higher education. That is, *they provide consistent support for the effectiveness of an increasingly popular approach, which has, nevertheless, been rarely analyzed in the long term: gamification*. From our findings, we have empirical evidence that gamifying programming assignments has an overall positive effect on students' submissions (attempts), IDE usage, and system access. Such behaviors are valuable for learning, as empirical evidence shows that the more one solves questions and spends time on a task, the better their learning is Rowland (2014); Rodrigues et al. (2021); Sanchez et al. (2020). Importantly, our study demonstrates that the innovative design was able to improve positive learning behaviors throughout the course of a whole semester. Thus, our findings' educational implication is that if practitioners deploy a gamification design similar to ours, they are likely to be adopting an innovative educational design that will improve their students' learning.

From that main contribution, we have derived five additional implications that contribute to educational practice and gamification research. First, *despite the fact that the gamification effect is likely to decrease over time, it is unlikely to start decreasing before four weeks.* Beyond supporting the *novelty effect*, our findings open the debate in terms of *when it starts to act*. Our results suggest that the effect od gamification only started to decrease after four weeks of use. Despite several discussions in the literature that gamification effect suffers from the novelty effect, little has been explored and discussed in terms of when the decrease starts. Therefore, this finding has both practical and theoretical implications, informing practitioners on a *safe period* in which gamification can be used without losing its power, as well as providing researchers with a threshold that can be assessed in further research, to ground the extent to which the novelty effect starts acting.

Second, *whereas the novelty effect is often present, the familiarization effect seems to naturally address it.* Our findings demonstrate that the impact of gamification on all measures started to decrease at some point in time. However, we also showed that such impact then shifted back to an increasing trend, without any intervention (e.g., changing the gamification design). This behavior, which resembles a U-shaped curve, has been called the *familiarization effect*: after some (familiarization) time, gamification's effect is enhanced. Additionally, it should be noted that such effect appeared after the downtrend of the novelty effect. Providing empirical support for the familiarization has practical implications as, despite the fact that the impact of gamification might decrease after a period of time, a recovery is likely to happen, after users familiarize themselves with the game elements. Furthermore, empirically supporting the familiarization effect has theoretical implications. From our findings, the familiarization effect tackles the drawback from the novelty effect within a few weeks (two to six) after the end of the downtrend. To our best knowledge, however, such analysis, based on a comparison to a control group, has not been performed before in previous research. Hence, our findings call for future long-term, longitudinal gamification studies, to better understand the familiarization effect.

Third, *gamification likely suffers from the novelty effect, but benefits from the familiarization effect, which contributes to an overall positive impact.* Our findings corroborate gamification literature on three perspectives. They demonstrate that i) using gamification positively impacted on learner behavior, ii) gamification impact suffered from the novelty effect, and iii) some time after the end of the novelty effect, gamification's effect was enhanced via the familiarization effect. Therefore, from a practical point of view, despite a gamified intervention seemingly loosing power or failing to work after some time, it is likely to gain power again in the future, although the effect might not go back to its initial values. From a theoretical point of view, more research is needed to understand the magnitude of both the novelty and the familiarization effects, as well as for shedding light into how long the uptrend of the familiarization effect lasts.

Fourth, *studies lasting less than 12 weeks are likely to be insufficient for truly revealing gamification's impact*. Our findings indicate that the novelty effect starts to act only after four weeks. In addition, our findings show that the final uptrend of the familiarization might start only after 10 or 12 weeks. Therefore, studies lasting less than this period of time might reveal unreliable findings, as they might fail to consider such changes that happen during learners' experiences with a gamified intervention. Nevertheless, it might be that even longer studies are necessary, given that, for attempts, the increase continued until the end of the intervention (14 weeks). These findings have practical and theoretical implications as well, informing researchers on the duration of the experiments they will conduct, and practitioners on how long they should use gamification before assessing its impacts.

Lastly, *our article has implications on how researchers can deploy similar studies in other contexts, learning activities, countries, and for other student types, further underpinning our findings*. As our evidence is limited to the context of STEM students completing programming assignments, our results should be seen in this context. Then, as the literature will benefit from understanding how the gamification effect varies in other settings, future research needs to generalize our study design to that end. For instance, one could use gamification to motivate completing different learning activities (e.g., multiple-choice quizzes) or following other behaviors (e.g., class attendance). In exploring other behaviors/learning activities, researchers can develop similar studies outside the STEM context. Thus, we contribute to future research with directions on how to ground our findings based on other samples/populations.

Limitations

This study generated several limitations that must be acknowledged when interpreting its findings. First, participants were not randomly assigned to experimental conditions, because the gamified version of the system used in this research was only available after 2016. To mitigate that, we compared participants in terms of demographic characteristics and, through preliminary analyses, showed the roughly one-year difference among groups is unlikely to confound our results. Additionally, meta-analytic evidence suggests the lack of random assignment does not affect the effect of gamification on behavioral learning outcomes Sailer and Homner (2019), further mitigating the impact of this limitation.

Second, the disciplines involved in this study did not have the same instructors throughout the three years of data collection. This limitation especially emerges because UFAM has a policy that professors take turns with the disciplines they teach. Despite that, pedagogical plans, learning materials, and the problems within the tasks participants completed were similar. In addition, meta-analytic evidence suggests that experimental conditions having different instructors does not affect gamification impact Bai et al. (2020), further mitigating this limitation's impact.

Third, group sizes were highly unbalanced. From the perspective of sample sizes, this limitation is handled by the number of participants for each condition (≥ 138) being above the average of the total sample of gamification studies, as reported by secondary studies (see Koivisto and Hamari (2019) and Sailer and Homner (2019), for instance). From the perspective of conclusion validity, we addressed this limitation, by answering our research question using robust statistical analyses that, among other advantages, handle groups with different sizes Wilcox (2017).

Fourth, our discussions regarding the shape (change) of the effect of gamification over time are based on visual inference. Although we selected robust statistical methods to guarantee conclusion validity, we did not perform, for instance, growth curve analysis. We chose this approach to enable comparing gamification effect at each time-point, as well as determining the effect's magnitude, rather than understanding the effect's curve. Consequently, we provided evidence on each time-point's difference and inferred the curve shape from visual analysis, upon that evidence. While that is a limitation, we opted for it instead of more conclusive approaches (such as regression analysis based on p-values), due to our exploratory analysis goal, as the literature recommends Vornhagen et al. (2020); Cairns (2019); Dragicevic (2016).

Fifth, regarding our study design, as this was a quasi-experimental study conducted over three academic years, its internal validity is likely affected. Indeed, gamification studies are often criticized, due to lack of methodological rigor Hamari et al. (2014); Dichev and Dicheva (2017). However, we made this choice in exchange of having a higher external validity, as it was conducted using a real system, within the context of real classrooms. Another aspect that could be discussed is the lack of a pre-test, in which pre-existent differences could be the source of differences rather than the experimental manipulation, especially given the lack of random assignment. Nevertheless, it must be noted that we analyzed gamification effect in terms of student behavior. Consequently, this is mitigating the lack of a pre-test, which would have to be based on learners' intentions (e.g., to access the system) rather than on actual behavior. Additional positive points are that this study features a control group and that the dependent variables reflect participants' behavior.

Lastly, in terms of study context, this research concerned STEM students completing programming assignments in either a gamified or non-gamified setting. That is, the study is limited to a single learning activity type. Additionally, we only deployed a single gamification design throughout the data collection period. In contrast, research suggests that one's experience with gamification might change depending on the task Liu et al. (2017) and that different designs should be adopted, depending on the learning activity Rodrigues et al. (2019); Hallifax et al. (2019). Therefore, we cannot rule out the possibility that we observed an effect due to the combination of the gamification design and the task type (i.e., programming assignments). However, given that the task type was invariant in our study, the effect observed could only be attributed to the gamification. Nevertheless, we opted for a single-factor study (i.e., only manipulating designs—gamified or not—within a single task) to maximize the findings' interval validity, following literature recommendations Cairns (2019).

Future work

Based on our results and their implications, we call for further similar research (i.e., long-term, longitudinal gamification studies featuring a control group) in different contexts, based on other measures, to further ground our findings as well as determine whether they generalize to new contexts. Future research is also needed to confirm for how long gamification positively acts, until the novelty effect starts to act—a four-week period according to this study. Also, the literature demands such longitudinal research to ground the familiarization effect. We found that it takes between six and 10 weeks to start the uptrend. However, the lack of longitudinal studies and appropriate data analysis methods lead to little evidence on whether this period differs in other contexts. Additionally, future research should seek for further evidence on how much of the initial effect of gamification the familiarization can recover, for how long its uptrend lasts, and, mainly, ground whether there is a point in which the gamification impact actually achieves a plateau. In doing so, researchers could explore mixed-methods approaches. Specifically, qualitative data would allow understanding what maintains user motivation and persistence over time, based on their subjective experiences. Then, researchers could triangulate quantitative and qualitative data to advance the overall understanding of gamification's effect over time.

Furthermore, once researchers are aware that these changes in gamification effect happen over time, interventions to mitigate them should be sought for. Although the familiarization effect appears to naturally address the novelty effect's negative impact, it seems not to recover gamification's initial benefit. Nevertheless, the period in which the gamification impact decreases should be tackled, to maximize its contributions. To that end, a promising research direction is tailored gamification, which can be accomplished through personalization or adaptation Klock et al. (2020); Rodrigues et al. (2020). In this approach, game elements are tailored based on user and/or contextual information, with the goal of enhancing gamification potential (e.g., Rodrigues et al. (2021); Lopez and Tucker (2021)). The rationale is that the same gamification design is unlikely to work for all users (i.e., one size does not fit all) because people have different preferences, perceptions, and experiences, even under the same conditions Van Roy and Zaman (2018); Rodrigues et al. (2020). Accordingly, in future studies, tailoring can be triggered once a user logs into the system (i.e., personalization), providing game elements that better suit the user, whilst expecting to mitigate the magnitude of the novelty effect, due to this choice, or during system usage (i.e., adaptation), changing the gamification design when its effect starts to decrease Tondello (2019). As personalized gamification is a recent research field Klock et al. (2020); Tondello et al. (2017), long-term experimental studies assessing its effects compared to general gamification approaches would be beneficial.

Conclusions

Gamification is an educational innovation that has been widely used in several domains. Researchers often advocate that gamification's impact is positive due to its novelty and that it consequently vanishes as the novelty passes (i.e., the novelty effect). However, little is known about how the gamification effect changes over time, due to the lack of longitudinal studies in this field. Accordingly, recent literature has called for research to address this gap. This study responded to such calls with a 14-week longitudinal study, in which we analyzed the impact of gamification on students' behavioral outcomes at seven points in time. Specifically, our findings are threefold. First, they confirm the effectiveness of the gamification design we proposed to innovate the educational process of teaching programming. Second, concerning the novelty effect, we found that novelty: i) starts to act after four weeks of intervention, ii) lasts for two to six weeks, and iii) diminishes a moderate effect to null in the worst case. Third, concerning the familiarization effect, we found that i) its positive trend starts between six and ten weeks after the start of the intervention and that ii) it seems to naturally tackle the negative effects of novelty expiration, although it did not restore the positive effect to the level of the initial contribution of the gamification.

These findings have several implications. Mainly, they reveal that innovating the educational practice with a gamification strategy similar to the one we used is likely to improve students' behavioral learning outcomes, even in the long-run. Additional implications are: informing practitioners on how long it takes for the novelty effect to start acting; revealing that although gamification's impact might decrease over time, it will likely recover after a familiarization period; and making the case that assessing gamification effect in less than 12 weeks is likely to yield unreliable results, due to the joint effects of novelty and familiarization. Furthermore, our findings have raised concerns that demand further research, such as confirming for how long the novelty effect decreases gamification impact and for how long (and to what extent) the familiarization effect recovers that effectiveness lost. Hence, our main contribution is advancing the understanding of how gamification impacts students' behavior over time, by providing empirical evidence on how the novelty effect acts, as well as supporting the familiarization effect. Ultimately, our conclusion is that gamification's impact on students' behaviors suffers from the novelty effect, but benefits from the familiarization effect, a counter-balance that is likely responsible for the overall positive effect gamification often demonstrates in maximizing behavioral learning outcomes.

Abbreviations
STEM: Science, Technology, Engineering, and Mathematics; CS1: Introductory Programming; IDE: IntegratedDevelopment Environment; ANOVA: Analysis of Variance; SD: Standard Deviation.

Acknowledgements
The authors are thankful for all organizations that provided funding to make this research possible.

Authors' contributions
LR: conception, design, data analysis, data interpretation, drafting the work; FP: conception, design, data analysis, data interpretation, drafting the work; AT: conception, design, drafting the work; PP: data interpretation, revised the work; MP: data acquisition, software creation, revised the work; LO: data acquisition, revised the work; DF: data acquisition, software creation, revised the work; EO: data acquisition, revised the work; AC: revised the work; SI: revised the work. All authors read and approved the final manuscript.

Funding
This research was partially funded by the following organizations: Brazilian National Council for Scientific and Technological Development (CNPq)—processes 141859/2019-9, 163932/2020-4, 308458/2020-6, and 308513/2020-7; Coordination for the Improvement of Higher Education Personnel (CAPES)—Finance Code 001; and São Paulo State Research Support Foundation (FAPESP)—processes 2018/15917- 0 and 2013/07375-0. Additionally, this research was carried out within the scope of the Samsung-UFAM Project for Education and Research (SUPER), according to Article 48 of Decree n° 6.008/2006 (SUFRAMA), and partially funded by Samsung Electronics of Amazonia Ltda., under the terms of Federal Law n° 8.387/1991, through agreements 001/2020 and 003/2019, signed with Federal University of Amazonas and FAEPI, Brazil.

Availability of data and materials
The datasets used and/or analyzed during the current study are available from the corresponding author on reasonable request

Declarations

Ethics approval and consent to participate
All participants provided consent allowing collection and usage of their data, in an anonymized way, for scientific purposes.

Consent for publication

All authors and institutions involved in this research are aware of and agree with the publication of this document.

Competing interests

The authors declare that they have no competing interests.

Author details

[1]Institute of Mathematics and Computer Science, University of São Paulo, Avenida Trabalhador São Carlense, 400 - Centro, São Carlos, SP 13566-590, Brazil. [2]Department of Computer Science, Federal University of Roraima, Boa Vista, Brazil. [3]Durham University, Durham, UK. [4]HCI Games Group, University of Waterloo, Stratford, Canada. [5]Amazonas State University, Manaus, Brazil. [6]Institute of Computing, Federal University of Amazonas, Manaus, Brazil.

Received: 2 September 2021 Accepted: 21 December 2021
Published online: 15 February 2022

References

Ahadi, A., Lister, R., Vihavainen, A. (2016). On the number of attempts students made on some online programming exercises during semester and their subsequent performance on final exam questions. In: Proceedings of the 2016 ACM Conference on Innovation and Technology in Computer Science Education, pp. 218–223

Alsawaier, R. S. (2018). The effect of gamification on motivation and engagement. *The International Journal of Information and Learning Technology*. https://doi.org/10.1108/IJILT-02-2017-0009.

Bai, S., Hew, K. F., & Huang, B. (2020). Does gamification improve student learning outcome? evidence from a meta-analysis and synthesis of qualitative data in educational contexts. *Educational Research Review, 30,* 100322.

Branch, R. M. (2009). *Instructional design: The ADDIE approach* (Vol. 722). Berlin: Springer.

Briffa, M., Jaftha, N., Loreto, G., Pinto, F. C. M., Chircop, T., & Hill, C. (2020). Improved students' performance within gamified learning environment: A meta-analysis study. *International Journal of Education and Research, 8*(1), 223–244.

Cairns, P. (2019). *Doing better statistics in human-computer interaction*. Cambridge: Cambridge University Press.

Clark, R. E. (1983). Reconsidering research on learning from media. *Review of educational research, 53*(4), 445–459.

Creswell, J. W., & Creswell, J. D. (2017). *Research design: qualitative, quantitative, and mixed methods approaches*. New York: Sage publications.

Deterding, S., Dixon, D., Khaled, R., Nacke, L. (2011). From game design elements to gamefulness: defining gamification. In: Proceedings of the 15th International Academic MindTrek Conference: Envisioning Future Media Environments, pp. 9–15 . ACM

Dichev, C., & Dicheva, D. (2017). Gamifying education: What is known, what is believed and what remains uncertain: a critical review. *International Journal of Educational Technology in Higher Education, 14*(1), 9.

Dick, W., Carey, L., Carey, J. O. (2005). The systematic design of instruction.

Dragicevic, P. (2016). Fair statistical communication in hci. In: Modern Statistical Methods for HCI, pp. 291–330. Springer

Estey, A., Coady, Y. (2016). Can interaction patterns with supplemental study tools predict outcomes in CS1? Proceedings of the 2016 ACM Conference on Innovation and Technology in Computer Science Education—ITiCSE '16, 236–241

Fonseca, S. C., Pereira, F. D., Oliveira, E. H., Oliveira, D. B., Carvalho, L. S., Cristea, A. I. (2020). Automatic subject-based contextualisation of programming assignment lists. In *Proceedings of The 13th International Conference on Educational Data Mining (EDM 2020)*.

Galvão, L., Fernandes, D., Gadelha, B.(2016). Juiz online como ferramenta de apoio a uma metodologia de ensino híbrido em programação. In: Brazilian Symposium on Computers in Education (Simpósio Brasileiro de Informática na Educação-SBIE), vol.27, p. 140

Goldstein, E. B. (2014). *Cognitive Psychology: Connecting Mind*. Nelson Education: Research and Everyday Experience.

Grangeia, T.d.A.G., de Jorge, B., Cecílio-Fernandes, D., Tio, R.A., de Carvalho-Filho, M.A. (2019). Learn+ fun! social media and gamification sum up to foster a community of practice during an emergency medicine rotation. *Health Professions Education, 5*(4), 321–335.

Hallifax, S., Serna, A., Marty, J.-C., & Lavoué, É. (2019). Adaptive gamification in education: A literature review of current trends and developments. In M. Scheffel, J. Broisin, V. Pammer-Schindler, A. Ioannou, & J. Schneider (Eds.), *Transforming learning with meaningful technologies* (pp. 294–307). Cham: Springer.

Hamari, J., Koivisto, J., Sarsa, H. (2014). Does gamification work?–a literature review of empirical studies on gamification. In: 2014 47th Hawaii International Conference on System Sciences, pp. 3025–3034 . Ieee

Hanus, M. D., & Fox, J. (2015). Assessing the effects of gamification in the classroom: A longitudinal study on intrinsic motivation, social comparison, satisfaction, effort, and academic performance. *Computers& Education, 80,* 152–161. https://doi.org/10.1016/j.compedu.2014.08.019.

Helmefalk, M. (2019). An interdisciplinary perspective on gamification: Mechanics, psychological mediators and outcomes. *International Journal of Serious Games, 6*(1), 3–26.

Hernández, M. I. O., Lezama, R. M., Gómez, S. M.(2021). Work-in-progress: The road to learning, using gamification. In: 2021 IEEE Global Engineering Education Conference (EDUCON), pp. 1393–1397 . IEEE

Huang, R., Ritzhaupt, A. D., Sommer, M., Zhu, J., Stephen, A., Valle, N., et al. (2020). The impact of gamification in educational settings on student learning outcomes: A meta-analysis. *Educational Technology Research and Development*. https://doi.org/10.1007/s11423-020-09807-z.

Hyrynsalmi, S., Smed, J., Kimppa, K. (2017). The dark side of gamification: How we should stop worrying and study also the negative impacts of bringing game design elements to everywhere. In: GamiFIN, pp. 96–104

Jafari, M., & Ansari-Pour, N. (2019). Why, when and how to adjust your p values? *Cell Journal (Yakhteh), 20*(4), 604.

Keselman, H., Algina, J., Wilcox, R. R., & Kowa, R. K. (2000). Testing repeated measures hypotheses when covariance matrices are heterogeneous: Revisiting the robustness of the welch-james test again. *Educational and Psychological Measurement, 60*(6), 925–938.
Klock, A. C. T., Gasparini, I., Pimenta, M. S., & Hamari, J. (2020). Tailored gamification: A review of literature. *International Journal of Human-Computer Studies*. https://doi.org/10.1016/j.ijhcs.2020.102495.
Koivisto, J., & Hamari, J. (2014). Demographic differences in perceived benefits from gamification. *Computers in Human Behavior, 35,* 179–188. https://doi.org/10.1016/j.chb.2014.03.007.
Koivisto, J., & Hamari, J. (2019). The rise of motivational information systems: A review of gamification research. *International Journal of Information Management, 45,* 191–210. https://doi.org/10.1016/j.ijinfomgt.2018.10.013.
Kotrlik, J., & Williams, H. (2003). The incorporation of effect size in information technology, learning, information technology, learning, and performance research and performance research. *Information Technology, Learning, and Performance Journal, 21*(1), 1.
Kowalchuk, R. K., Keselman, H., & Algina, J. (2003). Repeated measures interaction test with aligned ranks. *Multivariate Behavioral Research, 38*(4), 433–461.
Kyewski, E., & Krämer, N. C. (2018). To gamify or not to gamify? an experimental field study of the influence of badges on motivation, activity, and performance in an online learning course. *Computers& Education, 118,* 25–37.
Landers, R. N., & Landers, A. K. (2014). An empirical test of the theory of gamified learning: The effect of leaderboards on time-on-task and academic performance. *Simulation& Gaming, 45*(6), 769–785.
Liu, D., Santhanam, R., & Webster, J. (2017). Toward meaningful engagement: A framework for design and research of gamified information systems. *MIS quarterly, 41*(4), 1011–1034.
Lopez, C. E., Tucker, C. S. (2021). Adaptive gamification and its impact on performance. In: International Conference on Human-Computer Interaction, pp. 327–341 . Springer
Mair, P., Wilcox, R. (2018). Robust statistical methods using wrs2. The WRS2 Package.
Mavletova, A. (2015). A gamification effect in longitudinal web surveys among children and adolescents. *International Journal of Market Research, 57*(3), 413–438.
Mitchell, R., Schuster, L., & Drennan, J. (2017). Understanding how gamification influences behaviour in social marketing. *Australasian Marketing Journal (AMJ), 25*(1), 12–19.
Mustafa, A. S., Karimi, K. (2021). Enhancing gamified online learning user experience (ux): A systematic literature review of recent trends. Human-Computer Interaction and Beyond-Part I, 74–99
Nacke, L. E., & Deterding, C. S. (2017). The maturing of gamification research. *Computers in Human Behaviour*. https://doi.org/10.1016/j.chb.2016.11.062.
Palomino, P.T., Toda, A.M., Oliveira, W., Cristea, A.I., Isotani, S. (2019). Narrative for gamification in education: why should you care? In: 2019 IEEE 19th International Conference on Advanced Learning Technologies (ICALT), vol. 2161, pp. 97–99 . IEEE
Palomino, P., Toda, A., Rodrigues, L., Oliveira, W., Isotani, S.(2020). From the lack of engagement to motivation: Gamification strategies to enhance users learning experiences. In: 2020 19th Brazilian Symposium on Computer Games and Digital Entertainment (SBGames)-GranDGames BR Forum, pp. 1127–1130
Pereira, F. D., Fonseca, S. C., Oliveira, E. H., Cristea, A. I., Bellhäuser, H., Rodrigues, L., Oliveira, D. B., Isotani, S., Carvalho, L. S. (2021). Explaining individual and collective programming students' behaviour by interpreting a black-box predictive model. IEEE Access
Pereira, F. D., Oliveira, E. H., Oliveira, D. B., Cristea, A. I., Carvalho, L. S., Fonseca, S. C., Toda, A., Isotani, S. (2020). Using learning analytics in the amazonas: Understanding students' behaviour in introductory programming. *British Journal of Educational Technology*. https://doi.org/10.1111/bjet.12953.
Pereira, F. D., Toda, A., Oliveira, E. H., Cristea, A. I., Isotani, S., Laranjeira, D., Almeida, A., Mendonça, J. (2020). Can we use gamification to predict students' performance? a case study supported by an online judge. In: International Conference on Intelligent Tutoring Systems, pp. 259–269. Springer
Putz, L.-M., Hofbauer, F., & Treiblmaier, H. (2020). Can gamification help to improve education? Findings from a longitudinal study. *Computers in Human Behavior*. https://doi.org/10.1016/j.chb.2020.106392.
R Core Team. (2020). R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria. R Foundation for Statistical Computing. Retrieved from https://www.R-project.org/
Rodrigues, L., Oliveira, W., Toda, A., Palomino, P., Isotani, S. (2019). Thinking inside the box: How to tailor gamified educational systems based on learning activities types. In: Proceedings of the Brazilian Symposium on Computers in Education (Simpósio Brasileiro de Informática na Educação—SBIE)
Rodrigues, L., Palomino, P.T., Toda, A.M., Klock, A.C., Oliveira, W., Avila-Santos, A.P., Gasparini, I., Isotani, S. (2021). Personalization improves gamification: Evidence from a mixed-methods study. Proceedings of the ACM on Human-Computer Interaction 5(CHI PLAY), 1–25
Rodrigues, L., Toda, A. M., Oliveira, W., Palomino, P. T., Avila-Santos, A. P., Isotani, S. (2021). Gamification works, but how and to whom? an experimental study in the context of programming lessons. In: Proceedings of the 52nd ACM Technical Symposium on Computer Science Education, pp. 184–190
Rodrigues, L., Toda, A. M., Oliveira, W., Palomino, P. T., Isotani, S. (2020). Just beat it: Exploring the influences of competition and task-related factors in gamified learning environments. In: Anais do XXXI Simpósio Brasileiro de Informática na Educação, pp. 461–470 . SBC
Rodrigues, L., Toda, A. M., Palomino, P. T., Oliveira, W., Isotani, S. (2020). Personalized gamification: A literature review of outcomes, experiments, and approaches. In: Eighth International Conference on Technological Ecosystems for Enhancing Multiculturality, pp. 699–706
Rowland, C. A. (2014). The effect of testing versus restudy on retention: A meta-analytic review of the testing effect. *Psychological Bulletin, 140*(6), 1432.
RStudio Team. (2019). RStudio: Integrated Development Environment for R. RStudio, Inc., Boston, MA . RStudio, Inc. Retrieved from http://www.rstudio.com/
Sailer, M., & Homner, L. (2019). The gamification of learning: A meta-analysis. *Educational Psychology Review*. https://doi.org/10.1007/s10648-019-09498-w.

Sanchez, D. R., Langer, M., & Kaur, R. (2020). Gamification in the classroom: Examining the impact of gamified quizzes on student learning. *Computers& Education, 144,* 103666. https://doi.org/10.1016/j.compedu.2019.103666.
Santana, B. L., Bittencourt, R. A. (2018). Increasing motivation of cs1 non-majors through an approach contextualized by games and media. In: 2018 IEEE Frontiers in Education Conference (FIE), pp. 1–9 . IEEE
Seaborn, K., & Fels, D. I. (2015). Gamification in theory and action: A survey. *International Journal of human-computer studies, 74,* 14–31.
Toda, A. M., Klock, A. C., Oliveira, W., Palomino, P. T., Rodrigues, L., Shi, L., Bittencourt, I., Gasparini, I., Isotani, S., Cristea, A. I. (2019). Analysing gamification elements in educational environments using an existing gamification taxonomy. *Smart Learning Environments, 6*(1), 16.
Toda, A., Pereira, F. D., Klock, A. C. T., Rodrigues, L., Palomino, P., Oliveira, W., Oliveira, E. H. T., Gasparini, I., Cristea, A. I., Isotani, S. (2020). For whom should we gamify? insights on the users intentions and context towards gamification in education. In: Anais do XXXI Simpósio Brasileiro de Informática na Educação, pp. 471–480 . SBC
Toda, A. M., Valle, P. H. D., & Isotani, S. (2018). The dark side of gamification: An overview of negative effects of gamification in education. In A. I. Cristea, I. I. Bittencourt, & F. Lima (Eds.), *Higher education for all. From challenges to novel technology-enhanced solutions* (pp. 143–156). Cham: Springer.
Tondello, G. F. (2019). Dynamic personalization of gameful interactive systems. PhD thesis, University of Waterloo
Tondello, G. F., Orji, R., Nacke, L. E. (2017). Recommender systems for personalized gamification. In: Adjunct Publication of the 25th Conference on User Modeling, Adaptation and Personalization, pp. 425–430 . https://doi.org/10.1145/3099023.3099114.ACM
Tsay, C.H.-H., Kofinas, A. K., Trivedi, S. K., & Yang, Y. (2020). Overcoming the novelty effect in online gamified learning systems: An empirical evaluation of student engagement and performance. *Journal of Computer Assisted Learning, 36*(2), 128–146.
Van Roy, R., & Zaman, B. (2018). Need-supporting gamification in education: An assessment of motivational effects over time. *Computers& Education, 127,* 283–297.
von Wangenheim, C. G., & von Wangenheim, A. (2012). *Ensinando computação com jogos*. Florianópolis, SC, Brasil: Bookess Editora.
Vornhagen, J. B., Tyack, A., Mekler, E. D. (2020). Statistical significance testing at chi play: Challenges and opportunities for more transparency. In: Proceedings of the Annual Symposium on Computer-Human Interaction in Play, pp. 4–18
Wilcox, R. (2017). *Introduction to robust estimation and hypothesis testing* (4th ed.). Amsterdam: Elsevier.
Wilcox, R. R., & Tian, T. S. (2011). Measuring effect size: A robust heteroscedastic approach for two or more groups. *Journal of Applied Statistics, 38*(7), 1359–1368.
Yuen, K. K. (1974). The two-sample trimmed T for unequal population variances. *Biometrika, 61*(1), 165–170.

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

ANNEX

F

THINKING INSIDE THE BOX: HOW TO TAILOR GAMIFIED EDUCATIONAL SYSTEMS BASED ON LEARNING ACTIVITIES TYPES

Thinking Inside the Box: How to Tailor Gamified Educational Systems Based on Learning Activities Types

Luiz Rodrigues[1], Wilk Oliveira[1], Armando M. Toda[1], Paula T. Palomino[1], Seiji Isotani[1]

[1]Institute of Mathematics and Computer Science -- University of São Paulo 400 Av. Trabalhador São-carlense Avenue São Carlos, SP 13566-590 -- Brazil

{lalrodrigues, wilk.oliveira, armando.toda, paulatpalomino}@usp.br, sisotani@icmc.usp.br

Abstract. *Selecting gamification elements suitable for specific players (personalization) has been sought to improve the impacts of Gamified Educational Systems (GES). However, the lack of context might be a factor on the inconsistent results of those approaches. To address this lack, we introduce a method for personalizing GES based on learning activities types. The assumption is that selecting gamification elements for specific types of learning activities has the potential to improve GES impact on users by considering the context of each activity and, thus, contributing to their learning process. We describe how to apply our approach, how it differs from user-based methods, as well as discuss three cases of application and challenges yet to be tackled.*

1. Introduction

Gamification, the use of game elements in non-gaming contexts [Deterding *et al*. 2011], has been widely applied on educational systems [Borges *et al*. 2014]. Whereas many different studies have demonstrated positive results towards learners' motivation and engagement [Klock *et al*. 2018], others have shown a negative impact on learners [Toda *et al*. 2018]. Aiming to achieve the desired outcomes as well as to mitigate its negative effects, researchers have advocated towards personalizing Gamified Educational Systems (GES) (*e.g.,* [Borges *et al*. 2017; Santos *et al*. 2018a; Santos *et al*. 2018b; Lopez and Tucker 2018]). Personalization is a persuasive strategy that provides users with personalized content, commonly based on their personal characteristics [Oinas-Kukkonen and Harjumaa 2008]. Hence, personalizing a GES is often accomplished through selecting which gamification elements (*e.g.* points, leaderboard or badges) will be provided to the users based on their specific characteristics.

As users' demographic data might impact their opinions [Rodrigues and Brancher 2019a], many personalized GES are developed thinking in collecting users' demographic profile and giving them a reasonable group of game elements [Santos *et al*. 2018a; Lavoue *et al*. 2018; Monterrat *et al.* 2017]. Although, recent studies on personalization demonstrate that these tailored game elements might not impact positively on aspects related to the students' learning process [Santos et al. 2018a; Monterrat *et al.* 2017]. This allow us to assume that thinking from outside the box may not be the most suitable way to personalize those GES. Based on this premise, we propose a novel approach to use gamification by thinking from inside the box. In this approach, we aim at using the learning activities types (LAT; see, *e.g.* [Toetenel and Rienties 2016; Krathwohl 2002]) present on the educational system as the basis for generating the game

DOI: 10.5753/cbie.sbie.2019.823

elements that will be presented to the students. For instance, our approach would provide elements in line with each type of learning activity (*e.g.,* attending to information [Toetenel and Rienties 2016] or understanding a subject [Krathwohl 2002]) instead of elements users with specific characteristics (*e.g.*, males or females) prefer in a general picture. A specific case would be to add time pressure (gamification element) on an activity that intents to make learners effectively summarize (activity type) an essay, rather than inserting a set of gamification elements that, for example, female users (characteristic) prefer.

Thereby, to shed light on how to do so, this paper expands towards addressing the following research question: *how to tailor GES design based on LATs?*. The rationale is that different gamification elements having different functions/impacts [Toda *et al.* 2019; Sailer *et al.* 2017] as well as different types of learning activities also do [Toetenel and Rienties 2016; Chang 2016; Krathwohl 2002]. Thus, selecting the most appropriate gamification elements for each of these activity types based on its specific aims can be a valuable approach. However, the literature lacks evidence of how each gamification element alone, given a specific context, impacts users. To aid in this context, besides introducing our personalization approach, we provide a discussion concerning the reasons of using one element or another in three types of learning activities, based on their educational goals.

Following, this paper presents and discusses studies related to our presented approach in Section 2. Next, the approach presented in this paper is introduced (Section 3). Then, Section 4 describes this paper's case study settings and Section 5 discusses these. Thereafter, Section 6 steps to be tackled as well as challenges from our approach. Lastly, Section 7 provides our final remarks.

2. Related Works

Although personalizing GES aims to improve systems' impact on users [Oinas-Kukkonen and Harjumaa 2008], how those personalized systems affect users, compared to non-personalized systems, is yet uncertain [Santos *et al.* 2018a]. Roosta *et al.* (2016) and Lavoue *et al*. (2018) used adaptation to personalize GES. While the former found no significant impact, the latter found a positive effect on time spent on the system, despite users of the non-personalized version experienced higher motivation levels compared to those of the personalized version. Similarly, whilst the study of Monterrat *et al.* (2017) showed that interacting with the personalized version led to improvements on motivation and performance of learners, compared to those which interacted with a non-personalized version, Santos et al. (2018a) demonstrated that both personalized and non-personalized versions of a GES led to statistically insignificant differences in terms of users' flow state. Therefore, studies concerning the personalization's impact on users presented mixed results, showcasing the need for further research.

How each gamification element impacts users might be related to those findings. Different elements have distinct effects on users and, depending on the context in which those are applied, as well as on the gamification design (set of elements), distinct outcomes are likely to be found [Toda *et al.* 2019; Mekler *et al.* 2017; Sailer *et al.* 2017]. For instance, points alone might not affect users academic performance in terms of solving math problems [Attali and Arielli-Attali 2015], whereas along to badges and leaderboards, those might affect users psychological needs [Sailer *et al.* 2017]. Therefore, despite users of specific characteristics preferring different gamification elements [Santos et al. 2018a], those leading to different outcomes, according to both context and others elements applied together [Toda *et al.* 2019; Mekler *et al.* 2017; Sailer *et al.* 2017; Attali and Arielli-Attali 2015], might be related to the mixed findings of personalization studies.

In summary, literature presents gaps in terms of the effects of personalizing GES [Santos et al. 2018a; Lavoue *et al*. 2018; Monterrat *et al.* 2017] and in the sense that further studies are required to ground how different gamification designs, as well as gamification elements alone, impact learners [Mekler *et al.* 2017; Sailer *et al.* 2017; Attali and Arielli-Attali 2015]. Consequently, these gaps difficulties the selection of which element (or set of these) is the most suitable to use on each LAT.

3. Inside-out Personalization

This section aims at demonstrating the rationale behind our proposal, starting with the development process. Thereafter, the section highlights how it differs from the interventions that have been explored by the scientific community and, last, it provides a thoughtful explanation of what is necessary and how the "inside-out" approach can be employed in practice.

Figure 1 illustrates the development process of our approach. Firstly, a literature review was conducted in order to find studies concerning the use of personalization strategies on GES. Second, we sought to identify what attributes were used to drive the personalization strategy. Third, as highlighted in Section 2, we identified that users data were the most used through the analysis of the related works. Based on this context, along with results indicating that distinct gamification elements might lead to different outcomes, as previously mentioned, our fourth step was to develop our proposal. We defined it to personalize, specifically, each type of learning activity, with the goals of *i)* considering that each activity type has different aims and, therefore, will benefit from a specific gamification design, *ii)* exploring the outcomes that each gamification element can lead to, and *iii)* not relying on users data.

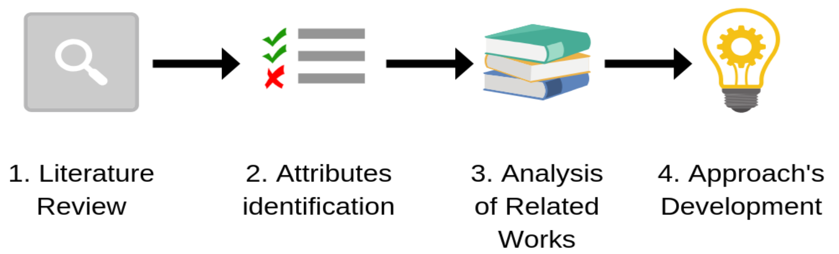

Figure 1. "Inside-out" approach development process.

To showcase the key difference from our LAT-based approach compared to the user-based approach that has been explored in the literature [Santos et al. 2018a; Lavoue *et al*. 2018; Monterrat *et al.* 2017; Roosta *et al.* 2016], Figure 2 illustrates the personalization processes of those two. It shows the entities that we consider to be inside (*i.e.,* educational system and personalization strategy) and outside (*i.e.*, users) the system's context. Also, the figure demonstrates that the key difference in these processes is the kind of data the personalization strategy receives as input. Thereby, while the "outside-in" relies on user data (*e.g.*, players' demographics and gamer types), which came from outside the system's context according to the figure, the "inside-out" considers LAT data that, on the other hand, came from the system itself, thus from inside its context.

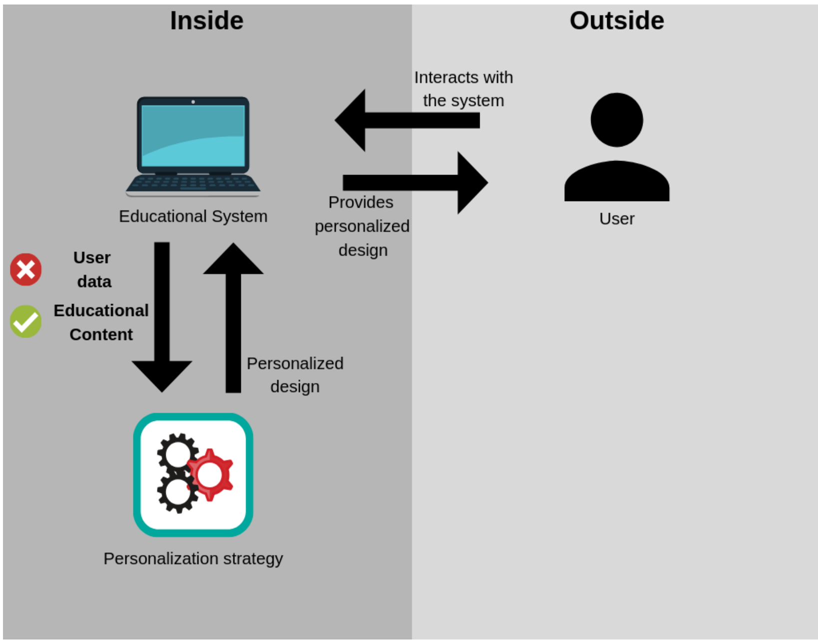

Figure 2. Personalization processes according to both our proposed approach (right) and those of the literature (left).

Given the definition of the "inside-out" approach, how to perform it is highly inspired on studies that employed the user-based methods. However, instead of seeking to collect, for instance, players' gamer type in order to stipulate the best design for each of them, tailoring a GES through the "inside-out" perspective depends on each specific part of the system. Thus, one should note that those personalization approaches differ in terms of *how* to personalize, not in terms of the process itself. We consider that different parts of the educational system will provide the user with different LATs. Hence, to accomplish this we must map the best gamification elements to feature the design of each of those types a GES features because, depending on the LAT, it is better to use one gamification element or another (or a specific set of them) considering that each element might lead the user to a different necessity/emotion/behavior [Toda *et al.* 2019; Sailer *et al.* 2017], as well as each learning activity has its specific requirements and goals [Toetenel and Rienties 2016; Chang 2016].

To compare those different forms of personalization in a practical example, suppose that one must personalize the gamification design of *watching a video lesson* in an educational system, which might be seen as an *attending to information* activity type [Toetenel and Rienties 2016]. Additionally, suppose that the aforementioned mapping indicates that the best design for this activity is to use *points, badges, and leaderboards*. Thereby, through the "inside-out" approach, those three elements would feature the system's design with the goal of improving this activity's execution. On the other hand, in the case of the GES used in both Lavoue *et al.* (2018) and Santos et al. (2018a), the gamification elements available would be defined considering users' preference, based on their gamer types or, in the case of Monterrat *et al.* (2017), according to users preferred mechanics. However, in neither of those, the gamification design is based on LATs. Therefore, it might be the case that the best design for a player type, in general, or the most preferred mechanic, is not the most suitable design for this activity, as the findings of the studies that adopted those approaches suggest.

Hence, when using the “inside-out” tailoring method, every LAT in a GES will contain a gamification design that is personalized specifically to it, based on these activities types, aiming to tailor the system as a whole through tailoring each of its parts. Consequently, this is expected to improve how the system impacts its users. We believe that adopting this approach might aid to guide players towards *(i)* being more focused on the educational tasks they have to perform, *(ii)* feeling more motivated to achieve better results, and *(iii)* being more engaged with the system itself and, thus spending more time in a learning context. We believe so because the gamification elements available will be in line with the context of each educational activity and, consequently, it is expected to efficiently influence the learners. In sum, by means of successfully accomplishing these aims, our proposed approach has the potential to contribute to users’ learning, besides contributing to designers by providing insights into how to create gamification designs to different LATs.

4. Case Study Settings

To address our research question, this paper presents examples of how to tailor three LATs through our “inside-out” method with the goal of showcasing how it can be employed on different occasions. The types of the activities addressed in the examples are: *(i)* attending to information (*e.g.* reading, watching or listening about a specific subject in a web browser), *(ii)* applying learning in a simulated setting (*e.g.* exploring, experimenting or simulating a problem), and *(iii)* assessment (*e.g.* writing, presenting, reporting, demonstrating or criticizing a subject with peers) [Toetenel and Rienties 2016].

When discussing how to tailor each of those, we rely on commonly used gamification elements to inspire and facilitate for teachers/researchers to apply our proposal. We adopted elements described in the taxonomy of Toda et al. (2019), which is focused on gamification applied on education and was validated by 19 experts. The authors described as well as indicated how 21 game elements that can be used to gamify educational systems impact users behavior (*i.e.*, engagement or motivation). Those are: Acknowledgement, chance, competition, cooperation, economy, imposed choice, level, narrative, novelty, objectives, point, progression, puzzles, rarity, renovation, reputation, sensation, social pressure, stats, storytelling, and time pressure. For a description of those we refer the reader to Toda et al. (2019).

5. Inside-out Tailoring of GES

First, consider students have to watch a video lesson in an educational system, which aims at providing the learners with explanations regarding a specific subject (*e.g.* balanced search trees) [Chang 2016]. Hence, an activity type in which learners have to **attend to**/**understand** information [Toetenel and Rienties 2016; Krathwohl 2002]. If the system featured the user-based approach, the following would be necessary. At first, users have to provide some demographic data (*e.g.*, genre and age) or to complete a questionnaire that would, *e.g.* allow the identification of their gamer type to the system (*e.g.*, *Daredevil*). Thereafter, the personalization strategy would update the system according to each user and, then, present the video player with the personalized gamification design. For instance, a user classified as *Daredevil* would be provided with Levels, Acknowledgement, and Competition [Santos et al. 2018a]. Similarly, when users of other gamer types accessed the system, other gamification elements would be available accordingly. If the system featured the "inside-out" personalization approach, users would not be required to complete a questionnaire neither to provide any demographic information because this strategy relies on data from inside the box (system), thus preventing the dependence of users' data. Hence, when the user access the video player screen, the system recognizes the type of the activity (*attending to information*) and the personalization strategy would select the gamification elements (*e.g.*, Progression) to be available in the screen accordingly. Figure 3 illustrates both personalization processes described here, showcasing their differences within this instance.

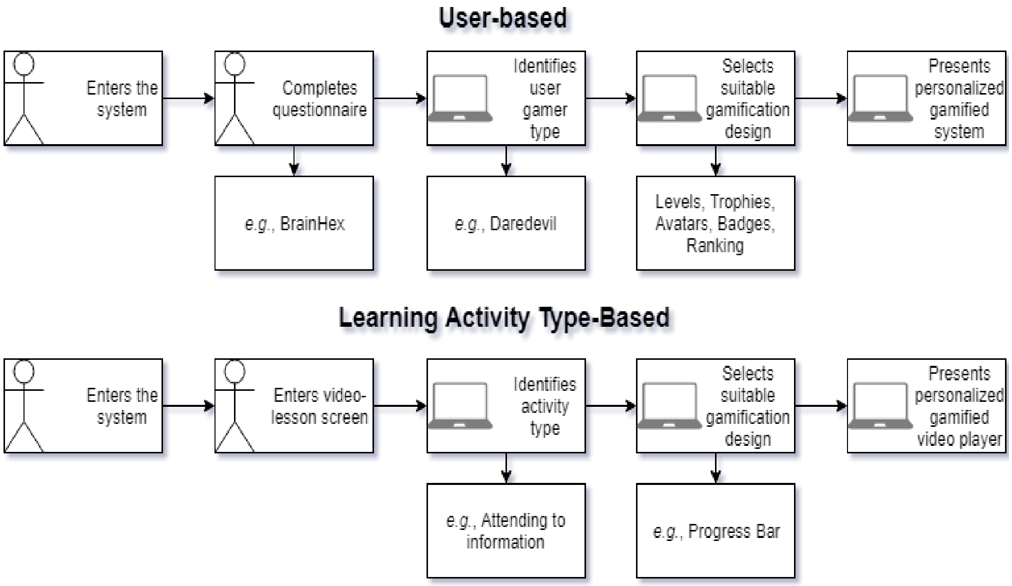

Figure 3. Tailoring the gamification design through user- (top) and LAT-based approaches (bottom) for the attending to information activity type instantiated as a watching video lesson.

As seen in Figure 3, whereas the "inside-out" personalization approach specifically selects the gamification design based on the type of the learning activity, the user-based one only considers the user gamer type. Following this example, while the gamification elements *Daredevil* users prefer might work for some types of activities, those might not for other types. Hence, focusing on users preferences might fail to help them when performing one learning activity or another. Differently, the "inside-out" approach aims to present gamification designs that match the most the goal of the learning activity based on its type. For instance, we argue that a gamification design featuring a Progression (*e.g.,* Progress-bar) would be interesting in this example. As the users have to attend to information through watching the whole video lesson, the Progress-bar would reflect the proportion of the video already watched, and it would be completely fulfilled only when the video was completely watched. Thereby, requiring the user to attend to all information presented in the video (activity's objective) in order to complete the Progress-bar and, consequently, achieving a milestone (*i.e.* Objective).

Second, assume students have to work on a scrum-development-simulation wherein each one of them plays a specific role (*e.g.* scrum master or developers). This activity aims to foster interpersonal skills and human endeavor [Chang 2016]. Hence, this is an activity in which users have to **apply** learning in a simulated setting [Toetenel and

Rienties 2016; Krathwohl 2002]. The personalization process for both approaches (user- and LAT-based) is fundamentally the same as the first example. In fact, the user-based method will yield the same design for *Daredevil* users, as well as those will be the same for users of other gamer types in both examples. Contrary, despite the "inside-out" personalization method will follow the same process as in the previous example, it will yield a design personalized to this type of learning activity. For instance, a design featuring a Narrative, which will emerge the learners into the simulation.

Third, consider students need to criticize colleagues' essays in a grammar class. This activity aids in providing and receiving critics and using these to improve their own works [Chang 2016]. Thereby, this is a type of **evaluation/assessment** learning activity [Toetenel and Rienties 2016; Krathwohl 2002]. In the same vein, the user-based approach would provide the same design for users of a specific gamer type, whilst the LAT-based method will present a design focused on this LAT. Therefore, in summary, the introduced method will systematically seek to present gamification designs that improve the execution of the learning activity, based on its type, whereas the user-based one employs the same design regardless of the LAT in order to meet users' preferences.

We highlight that, in the user-based method, different users will be provided with gamification designs based on different sets of gamification elements, based on their personal characteristics; however, regardless of the learning activity they are performing, those designs will still be based on the same set of gamification elements. Contrary, in the "inside-out" approach, different users will be provided with gamification designs also based on the same set of gamification elements, however, these will be selected based on the LATs. That is, users of different characteristics will receive the same designs, however, different LATs will offer different sets of gamification elements. Table 1 summarizes these differences.

Table 1. Design differences between the user-based and the "inside-out" methods.

Approach/Occasion	**Different users**	**Different LATs**
User-based	Different designs	Same design
"Inside-out"	Same design	Different designs

Moreover, from one side, the user-based approach relies on a number of theoretical studies that suggests different users have different preferences (c.f. [Santos et al. 2018a; Monterrat *et al. 2017*]). The problem is that gamification leads to different outcomes, according to both the context it is applied as well as how it is designed [Mekler *et al.* 2017; Sailer *et al.* 2017; Attali and Arielli-Attali 2015]. From the other side, we are introducing an approach that relies on the fact that different gamification elements will lead to different outcomes. Hence, based on literature's findings, assuming that when the right element is applied in the right context (*i.e.* LAT), the potential for positive results is enhanced. In spite of that, the little evidence of how each gamification element influences users [Mekler *et al.* 2017] limit our ability to claim our assumption is correct. Thus, the next section discusses what is necessary to address and validate our hypothesis.

6. Discussion

Based on this context, there are three next steps we believe to be of utmost importance to advance the understanding of how to personalize GES through the "inside-out" approach. We believe that it is necessary to develop a mapping between gamification elements and educational contents. To advance on how the design of a specific educational content should be, we must know which is the best set of gamification elements to feature it. As proposed here, the "inside-out" approach is a generic method and, in order to make it suitable for any GES, we recommend this mapping to be system-independent. Thereby, developers and designers will be able to rely on the mapping to tailor the educational contents of any system. Moreover, this mapping could be used to guide the development of a mechanism able to automatically create the designs of GES, reducing human efforts by facilitating the development of those.

Empirically/experimental validating the "inside-out" personalization approach is the second step that we recommend to be tackled. By doing so, it will be possible to identify whether using this approach is relevant to users' learning, experience, concentration, and other aspects of relevance to educational systems. Supported by the empirical evidence of the most appropriate way to design GES based on their learning activities, researchers can conduct experiments to validate whether those generic evidence actually leads to a positive effect on users, compared to, *e.g.* generic designs or user-based personalization approaches, through different case studies. These experiments' results will provide insights to practitioners concerning whether it is interesting for them to use the presented approach on their systems and classes, for example.

Furthermore, there are some limitations and further challenges which are likely to be faced on applications of our proposed approach. One might note that the "inside-out" approach might represent a threat to some users. In the same way some users prefer specific gamification elements, there are a set of these they dislike [Santos et al. 2018a], which is motivated by the fact that different users behave differently [Rodrigues and Brancher 2019b] and correspond better to different designs [Denden et al. 2018]. Thereby, despite using a gamification element suitable to the LAT is valuable from the perspective of education, it might exert a negative influence on users that dislike elements of that design. Nevertheless, although the gamification design proposed for the aforementioned examples featured a single element, there exist other options that might be adopted for the same type of learning activity as well as users of a specific gamer type prefer multiple gamification elements.

Hence, the set of gamification elements featuring a specific activity type could seek to avoid those that specific users dislike, besides selecting the elements based on the LAT. Thus, yielding a personalization approach based on both users and LATs. However, seeking to mitigate this based on user data (*e.g.,* users' gamer types or demographics), is uncertain, considering the literature's evidence [Santos et al. 2018a; Lavoue et al. 2018]. In spite of that, more advanced and user-specific data-driven techniques (*e.g.*, data mining/machine learning) could be explored to tackle that limitation as the potential of those have been suggested in literature [Meder *et al.* 2017].

7. Final Remarks

This paper introduced a GES personalization approach, presenting examples of how to apply it, discussing steps to both implement and validate it, and challenges and suggestions to mitigate these. This paper's main contribution is the introduction of the approach, which personalizes each system's learning activity specifically and does not consider users data, besides being generic (*i.e.* might be employed on any GES). After the initial mapping provided in this paper, we are in the process of designing the experiment to both refine it and validate its impacts on learners.

Acknowledgment

We would like to thank the funding provided by CNPq, FAPESP (Projects 2018/15917-0; 2016/02765-2; 2018/11180-3; 2018/07688-1), and CAPES.

References

Attali, Y., & Arieli-Attali, M. (2015). Gamification in assessment: Do points affect test performance?. Computers & Education, 83, 57-63.

Borges, S., Durelli, V., Reis, H., et al. (2017). Selecting Effective Influence Principles for Tailoring Gamification-Based Strategies to Player Roles. n. Cbie, p. 857.

Chang, V. (2016). Review and discussion: E-learning for academia and industry. *International Journal of Information Management*, v. 36, n. 3, p. 476–485.

De Sousa Borges, S., Durelli, V. H. S., Reis, H. M. and Isotani, S. (2014). A systematic mapping on gamification applied to education. *Proceedings of the 29th Annual ACM Symposium on Applied Computing - SAC '14*, n. Icmc, p. 216–222.

Denden, M., Tlili, A., Essalmi, F. and Jemni, M. (2018). Does personality affect students' perceived preferences for game elements in gamified learning environments? *Proceedings - IEEE 18th International Conference on Advanced Learning Technologies, ICALT 2018*, n. July, p. 111–115.

Deterding, S., Dixon, D., Khaled, R. and Nacke, L. (2011). From Game Design Elements to Gamefulness: Defining "Gamification." *Proceedings of the 2011 annual conference extended abstracts on Human factors in computing systems - CHI EA '11*, p. 2425.

Krathwohl, D. R. (2002). A revision of Bloom's taxonomy: An overview. Theory into practice, 41(4), 212-218.

Klock, A. C. T., Ogawa, A. N., Gasparini, I. and Pimenta, M. S. (2018). Does gamification matter ? *Proceedings of the 33rd Annual ACM Symposium on Applied Computing*, p. 2006–2012.

Lavoue, E., Monterrat, B., Desmarais, M. and George, S. (2018). Adaptive Gamification for Learning Environment. *IEEE Transactions on Learning Technologies*, p. 1–13.

Lopez, C. and Tucker, C. (2018). Towards Personalized Adaptive Gamification: A Machine Learning Model for Predicting Performance. *IEEE Transactions on Games*, v. PP, n. c, p. 1–1.

Meder, M., Plumbaum, T. and Albayrak, S. (2017). A primer on data-driven gamification design. In Proceedings of the Data-Driven Gamification Design Workshop.

Mekler, E. D., Brühlmann, F., Tuch, A. N., & Opwis, K. (2017). Towards understanding the effects of individual gamification elements on intrinsic motivation and performance. Computers in Human Behavior, 71, 525-534.

Monterrat, B., Lavoué, E., & George, S. (2017). Adaptation of gaming features for motivating learners. Simulation & Gaming, 48(5), 625-656.

Nah, F. F.-H., Zeng, Q., Rajasekhar Telaprolu, V., Padmanabhuni Ayyappa, A. and Eschenbrenner, B. (2014). Gamification of Education: A Review of Literature. *Lncs*, v. 8527, p. 401–409.

Oinas-Kukkonen, H. and Harjumaa, M. (2008). A systematic framework for designing and evaluating persuasive systems. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, v. 5033 LNCS, p. 164–176.

Rodrigues, L. and Brancher J. (2019a). Playing an Educational Game Featuring Procedural Content Generation: Which Attributes Impact Players' Curiosity?. Revista Novas Tecnologias, RENOTE, 17(1).

Rodrigues, L. A. L. and Brancher, J. D. (2019b). Improving Players’ Profiles Clustering from Game Data Through Feature Extraction. *Brazilian Symposium on Games and Digital Entertainment, SBGAMES*, v. 2018–November, p. 177–186.

Roosta, F., Taghiyareh, F., & Mosharraf, M. (2016, September). Personalization of gamification-elements in an e-learning environment based on learners' motivation. In 2016 8th International Symposium on Telecommunications (IST) (pp. 637-642). IEEE.

Sailer, M., Hense, J. U., Mayr, S. K., & Mandl, H. (2017). How gamification motivates: An experimental study of the effects of specific game design elements on psychological need satisfaction. Computers in Human Behavior, 69, 371-380.

Santos, W. O., Bittencourt, I. I., & Vassileva, J. (2018a). Design of tailored gamified educational systems based on gamer types. Proceedings of Workshops of Brazilian Conference on Computers in Education (Vol. 7, No. 1, p. 42).

Santos, W. O., Toda, A., Isotani, S., & Bittencourt, I. I (2018b). Does Gamified Educational Systems Change Students’ Learning Behaviors? A Case Study with Postgraduate Students. RENOTE, 16(2).

Toda, A. M., Oliveira, W., Klock, A. C., et al. (2019). A Taxonomy of Game Elements for Gamification in Educational Contexts : Proposal and Evaluation. In *19th International Conference on Advanced Learning Technologies*.

Toda, A. M., Valle, P. H. D. and Isotani, S. (2018). *The Dark Side of Gamification: An Overview of Negative Effects of Gamification in Education*. Springer International Publishing. v. 832

Toetenel, L. and Rienties, B. (2016). Analysing 157 learning designs using learning analytic approaches as a means to evaluate the impact of pedagogical decision making. *British Journal of Educational Technology*, v. 47, n. 5, p. 981–992.

ANNEX

G

AUTOMATING GAMIFICATION PERSONALIZATION TO THE USER AND BEYOND

Automating Gamification Personalization: To the User and Beyond

Luiz Rodrigues, Armando M. Toda, Wilk Oliveira, Paula T. Palomino, Julita Vassileva, Seiji Isotani

***Abstract*—Personalized gamification explores user models to tailor gamified designs to mitigate cases wherein the one-size-fits-all approach ineffectively improves learning outcomes. The tailoring process should simultaneously consider user and contextual characteristics (e.g., activity to be done and geographic location), which leads to several occasions to tailor. Consequently, tools for automating gamification personalization are needed. The problems that emerge are that which of those characteristics are relevant and how to do such tailoring are open questions, and that the required automating tools are lacking. We tackled these problems in two steps. First, we conducted an exploratory study, collecting participants' opinions on the game elements they consider the most useful for different learning activity types (LAT) via survey. Then, we modeled opinions through Conditional Decision Trees to address the aforementioned tailoring process. Second, as a product from the first step, we implemented a recommender system that suggests personalized gamification designs (which game elements to use), addressing the problem of automating gamification personalization. Our findings i) present empirical evidence that LAT, geographic locations, and other user characteristics affect users' preferences, ii) enable defining gamification designs tailored to user and contextual features simultaneously, and iii) provide technological aid for those interested in designing personalized gamification. The main implications are that demographics, game-related characteristics, geographic location, and LAT to be done, as well as the interaction between different kinds of information (user and contextual characteristics), should be considered in defining gamification designs and that personalizing gamification designs can be improved with aid from our recommender system.**

***Index Terms*—Gamified Learning; Personalization; Educational System; Recommender Systems; Context-aware.**

I. Introduction

To improve learning technologies ability to engage and motivate users, practitioners and researchers have used gamification: the use of game elements in non-gaming contexts [1], [2]. Overall results from these applications are positive, showing improvements in learning outcomes such as academic achievement, conceptual and application-oriented knowledge, and motivation to learn [3]. However, there are situations in which gamification is ineffective in impacting learning outcomes, or even negative [4]. Often, those happen due to poorly designed gamification [5], such as assuming that the same choices will work for all users, the one-size-fits-all approach [6]. To overcome such failures, researchers started to investigate personalized gamification [7].

Personalized gamification concerns exploring knowledge about the users to enable providers (e.g., instructors or the system itself) to offer game elements tailored to those users [8]. For instance, a case would be a system changing from game elements set A to game elements set B when users are females because the latter is tailored to these users. The premise for personalizing gamification emerged from discussions that people with different demographic characteristics and cultural background have distinct preferences [9], behaviors [10], and are motivated differently [11]. Consequently, those might experience and respond to the same conditions in distinct ways [12], [13]. The common practice for gamification is selecting which game elements to add to the system from a list of available elements [14], [15]. Accordingly, researchers invested in providing recommendations indicating which game elements suit better users of different groups to provide personalized gamification, predominantly based on their preferences (e.g., [16], [17]).

Despite personalized gamification is commonly build upon user preference, it is mainly personalized to users' profiles [18], [19]. However, the application context is relevant for gamification's success as well [14], [20], and gamification designs should be aligned to it [6]. Furthermore, multiple factors (e.g., users' demographics [21]–[23] and the system's context [5]) moderate users' experience, either positive or negatively. Although, tailoring approaches often consider a single one, reflecting current gaps in the field of personalized gamification [7], [18], [24]: the fact that i) personalization models should consider more than users' characteristics, such as encompassing the learning activities and geographic locations, and that ii) personalization methods should consider multiple aspects simultaneously, as well as their interactions.

To address these gaps, we sought to understand how to tailor gamified systems to the education domain by considering the learning activity at hand, the user's characteristics, and the geographic location simultaneously, as well as the interactions between all aspects taken into account. To achieve that goal, we performed an exploratory, survey-based research to capture users' preferences, a methodology that has been widely accepted and adopted by related research, as personalization is often based on user preference [15], [25]. As this process is concerned with understanding which aspects (i.e., among learning activity at hand, user's characteristics, and the geographic location) affect user preference, as well as the most suitable game elements for each aspects combination, we sought to answer research question 1 (**RQ1**): *Does users'*

L. Rodrigues, A. M. Toda, W. Oliveira, P. T. Palomino, and S. Isotani are with the Laboratory of Applied Computing to Education and Advanced Social Technology, Institute of Mathematics and Computer Science, University of São Paulo, São Carlos, Brazil. E-mail: {lalrodrigues, armando.toda, wilk.oliveira, paulatpalomino}@usp.br, sisotani@icmc.usp.br. J. Vassileva is with the Multi-User Adaptive Distributed Mobile and Ubiquitous Computing (MADMUC) Lab, Department of Computer Science, University of Saskatchewan, Saskatoon, Canada. E-mail: jiv@cs.usask.ca.

preferences differ depending on (a) their characteristics, (b) geographic location, and (c) the type of the learning activity to be performed?; and **RQ2**: *What is the most useful game elements set, from users' preferences, according to their characteristics, geographic location, and Learning Activity Type (LAT)*[1]*?*

RQ1 informs RQ2 as it reveals which aspects should be considered when defining the most useful game elements set. That is, the combination of game elements (e.g., points, badges, and leaderboards) users prefer the most, an interpretation based on personalization being commonly built upon user preference [7]. Consequently, the challenge that emerges is that interactions from multiple characteristics lead to several combinations. For instance, five binary characteristics would lead to 25 combinations (i.e., 25 recommendations); the number of recommendations for five three-valued characteristics would exponentially increase. Thereby, providing a way to automate such recommendations becomes imperative, which corroborates another challenge of personalized gamification: automating the personalization process [24]. Thus, our **RQ3**: *How to automate gamification personalization?* To answer RQ3, we implement a Recommender System (RS) for personalized gamification [26] based on RQ1's and RQ2's answers. Our RS informs the most useful set of game elements, according to users' preferences, given an input of user's characteristics and their geographic location along with the LAT to be performed. Hence, it enables automating gamification personalization to multiple factors.

Thus, our contributions are threefold. First, evidence from users' preferences that can be used to inform researchers and practitioners on how to tailor Gamified Educational Systems (GES) to LAT, geographic location, and user characteristics. Second, an RS to automate gamification personalization, which performs recommendations by considering multiple aspects simultaneously (i.e., user characteristics, geographic location, and LAT), enabling the implementation of gamification designs more aligned to their preferences. Third, demonstrating which user characteristics impacted their preferences, along with the degree of each one's influence; thus, one might decide which user characteristic to prioritize, take into account, and/or pay more attention as, for instance, moderators of gamification's effectiveness.

II. Literature Review

This section provides background information on the topics covered by this article, reasons about the literature to justify research choices, and highlights the contribution our study provides to existing literature compared to similar works.

A. Game elements

There are many definitions and categorizations of game elements. In the scope of this article, we consider game elements similar to the definition adopted by [15]: the building blocks impacting users' experience with the system, which are characteristic to gameful systems [2], following the vocabulary used more often by similar research [18].

Given the numerous game elements available, it has been common practice for each study to self-select which set of those elements to use. Based on a literature review, [15] presented 59 general elements. In [20], the authors reviewed the literature to select 12 common game elements, without considering any content game element [1] due to the generic nature of their research. In both studies, game elements were selected with no consideration for the domain application, according to their purposes. Differently, [16] explored an element set created from gamification on education literature [27], which is composed by eight options.

Given that our research focuses on a specific domain, education, this article differs from [15], [20] by exploring a taxonomy [28] containing the most common game elements (N = 21) from GES. This taxonomy was created through a rigorous, systematic process, and was validated by 19 experts in the field of gamification and games. Differently, [16] relied on a simpler, reduced game elements set, which was created based on a literature review. Furthermore, by selecting an expert-validated taxonomy, we ensure the game elements available are well defined, avoid using elements with the same purpose but different names, and prevent possible bias from the selection process. Additionally, the selected taxonomy also provides guidance on how the elements are expected to affect users [29], another advantage to those using it [15].

B. Personalized Gamification

Given personalization's importance to information systems, it should be deployed to enhance these systems' relevance to users [6]. Within the scope of gamified systems, a common practice to achieve personalization has been to tailor the gamification design (set of game elements) to specific user's characteristics [7]. In other words, gamified systems have been personalized by performing static adaptations on the game elements it features, based on pre-defined characteristics (i.e., behavioral profile), to tailor the gamification designs [18].

Recent literature reviews [7], [18] found that information used to drive personalization are, predominantly, users' player/gamer types (e.g., HEXAD user types) [19], followed by personality [30]. Nevertheless, it has been shown that other user characteristics, such as gaming habits (e.g., weekly playing time) [31] and gender [32], also impact their preference, as well as the relationship between user demographics (i.e., age and gender) and player types suggest the impact of those aspects [19]. Despite that, these aspects have been rarely explored in methods for tailoring gamification designs in education [33]. This research addresses this need by introducing an approach that exploits demographic and gaming habits as information used to drive the gamified designs' tailoring.

Furthermore, the user is not the only factor to be considered when defining gamification designs. A factor that has been often discussed as relevant for gamification effectiveness [3], [14], [34], which is rarely considered by tailoring methods, is the application context (e.g., geographic location). Specifically in the context of educational systems, an aspect researchers

[1] In the scope of this study, a LAT is defined based on its main expected outcome (see Section II for further details)

have recently argued as relevant, and recommended to consider when tailoring gamified systems, is the learning activity [18], [35]. This is related to the recommendation that gamified designs should match the task [6] and, given that tasks of educational systems are almost ever learning activities, personalizing the gamified designs to these activities should be accomplished.

Despite that, to the best of our knowledge, there are only two approaches for personalizing gamified designs based on learning activities [17], [25]. In [17], learning activities are considered based on their main expected objective, similar to this article. In [25], the learning activities are activities from Moodle (e.g., forum and quizzes). Hence, while recommendations from [17] can be extended to any learning activity (linked to their objective), those from [25] are limited to a specific set of Moodle activities. In addition, both works consider one user characteristic, personality trait and player type, respectively. Thus, they provide valuable contributions in terms of exploring learning activities, as well as presenting recommendations that consider the interaction between those and a user characteristic (e.g., player type X, learning activity Y).

However, these studies fall into the category of methods that rely on the most often researched user characteristic, a single user characteristic is considered in each one, and the guideline from [25] cannot be generalized to any learning activity. Therefore, the main advances of this article compared to those works are: i) considering multiple user characteristics rarely explored simultaneously, ii) taking the context into account via learning activities and users' geographic location, and iii) providing recommendations that consider the interaction between all of those aspects that are relevant for users.

C. Learning Tasks

To generally describe a task, one might rely on its desirable outcomes, behavioral requirements, and/or complexity [6], [36]. Similarly, from the human-computer interaction perspective, a task refers to the activities required to achieve a specific goal [37]. Consequently, given the context of our study, a learning task refers to a set of activities that aim at some educational outcome. From this definition, it is possible to note that numerous tasks might be found in GES, which makes it infeasible to develop a specific personalization approach for each one. An alternative to that limitation is categorizing the activities, which can substantially reduce their quantity; consequently, enabling the recommendation of gamification designs to each category.

To overcome the numerous learning tasks and categorize them, we opted to rely on the revision of Bloom's taxonomy of educational objectives [38]. This approach contributes to the learning process by matching the educational activities' gamification designs to a cognitive taxonomy [17]. Although there are other options available, the revision of Bloom's taxonomy is a widely cited, well-accepted taxonomy, similar to its original version [39]. It acts as a framework that can be used to classify what is expected from an educational activity (outcome), as well as its complexity [38]. The revised version is composed of two dimensions: knowledge (concerned with what is to be learned; e.g., the subject of matter) and cognitive process (concerned with actions associated with learning; e.g., how to learn) [38].

In the scope of this research, we consider the second dimension, similar to related work [17]. By categorizing learning activities based on the cognitive domain of such a taxonomy, we avoid having the gamification focused on the activity itself (e.g., completing a quiz or answering a forum) and allow it to be aligned with the activity's expected learning outcome. Thereby, addressing the recommendation that gamification should match the task [6]. Moreover, as many GES feature tasks of varied subjects, the second dimension choice makes the approach subject-independent, focusing the gamification designs' tailoring on the activities' particular objectives while allowing it to be used regardless of the system's educational topic.

The structure of the cognitive process dimension is split into six categories: remember, understand, apply, analyze, evaluate, and create. Here, we consider each dimension a different LAT, wherein their complexity increases following the order in which they were introduced (i.e., remember is the less complex and create is the most complex). Hereafter, we refer to those as LAT1 to LAT6, also following the introduced order. Furthermore, although an activity might fit in more than one LAT, our approach considers every activity will have a predominant, main objective to be achieved. Hence, the personalization process should be based on that main goal. It is worth noting that those LAT might be split again, however, we opted to work with the high-level abstraction given that the similarities within these sub-categories might be even higher. Thus, this article contributes a proposal that is based on the six high-level types of cognitive processes established in [38], that aids in tailoring gamification designs to different LAT, according to their predominant goal.

D. Recommender Systems for Personalized Gamification

An RS can be seen as a technique, or software tool, able to recommend items to users [40]. Such systems are especially valuable for cases in which several options are available, alleviating the burden of human selection by providing recommendations, often based on what other people recommend. Common applications of such systems are e-commerce, movies, and music. Recently, the use of RS has been suggested for personalized gamification [26], which corroborates to our research in terms of, for instance, reducing the burden of selecting the most suitable game elements for several combinations of user characteristics, geographic location, and LAT. Next, we provide a brief overview of RS for personalized gamification following the framework by [26].

RS have three main elements: inputs, outputs, and process. Inputs concerns all the aspects that are received by the RS to be taken into account before doing the recommendations. There are four main types of input: user profile (e.g., demographics, personality, behavioral profile), items (e.g., game elements), transactions (e.g., the relationship between users and items; using or preferring a game element), and context (e.g., geographic location, activities to be done). Outputs are ratings related to the choices that the RS made from the input received. For instance, if items are game elements, the output would be the rating of each one. The process is the core part of the RS, concerning the method through which it will perform the recommendations. There also are four main recommendation methods. Content-based recommenders are based on knowledge of the application, such as data log, or empirical and theoretical information. The collaborative filtering method exclusively depends on data collected implicitly or explicitly from interactions with a system. Context-aware recommenders are those that explore information of the context to make their choices. Lastly, hybrid recommenders aim at using two or more of the previous approaches together.

Generally, there is a lack of technological support for gamifying educational environments [41]. Accordingly, the literature on personalized gamification lacks concrete RS implementations, demonstrated by recent literature reviews finding only four studies that relate to RS or other forms of automating gamification personalization [24]. Among those, one is the framework proposal itself [26], whereas the remaining are theoretical/conceptual models with no concrete implementations available for third-parties use [12], [42], [43]. Differently, we present and provide an RS for personalizing gamification, which was built upon findings from the study of this article. Hence, we advance the literature with a free, hybrid RS as it uses both contextual as well as empirical information from users' preferences.

E. Summary

Table I summarizes and demonstrates the points in which this study differs from related works based on our previous discussion. As shown, most studies focus on user characteristics, few consider the task to be done, and none but this one takes into account geographic locations. Additionally, the few works that consider information from the user and the task provide recommendations based on two factors (one from each kind). On the other hand, our approach was developed considering nine aspects, of which eight were found to be significant (see Section IV) and, therefore, are considered in the product from our research (see Section IV-D). This final product is another key difference. Whereas previous research only provides conceptual/visual guidelines, this study contributes with technological aid for the design of personalized gamified systems. This also differs from research on recommender systems for gamification [12], [42], [43] as those provide no concrete implementations from their proposals.

TABLE I
SUMMARY COMPARISON OF RELATED WORKS.

	Recommends game elements based on				
Study	User	Task	GL	N Factors	Product
[15]	X				Conceptual
[20]	X				Conceptual
[16]	X				Conceptual
[19]	X				Conceptual
[31]	X				Conceptual
[32]	X				Conceptual
[25]	X	X		2	Conceptual
[17]	X	X		2	Conceptual
This	X	X	X	8	Technology

GL = Geographic location.

III. STUDY

The goal of this research is to understand how to tailor GES to LAT, geographic location, and users' characteristic. To achieve that goal, we performed a survey-based research asking participants to indicate their preferred game elements for each LAT. Up to date, this methodology is the most used by similar works [7], [24] and has been widely accepted given the number of related research following it [15], [25]. Therefore, we considered it the most adequate approach to adopt. This study also follows an exploratory approach, which aims to understand possible relations between the observable variables to create possible research guidance [44]. Based on that, this section presents an overview of this study development process, as well as further describes the material and methods followed.

A. Overview

In developing this research, three factors had to be defined: what domain, how to interpret the tasks, and which user characteristic to consider. **First**, we opted for the education domain, which is the one gamification research has focused the most [14] and, both positive [3] and negative [4] outcomes have been found, showing the need for further research. **Second**, given the domain, users will perform learning activities when using the gamified systems. As one might create numerous of those activities, our approach considers activities types based on the revised Bloom's taxonomy [38], an established, well-accepted taxonomy within the educational context. **Third**, we chose to focus on users' demographic characteristics and gaming habits and preferences, deepening into aspects that have been discussed as relevant factors [21], [31] but received less attention from the academic community compared to the most used ones [7], [24].

Then, to achieve the desired understanding, we developed Conditional Decision Trees (CDT) [45], which takes into account the interactions between all input variables to provide recommendations on the most suitable game elements given an input set. During data collection, we operationalized gamification designs as the top three game elements participants prefer the most, provided game elements (N = 21) extracted from an expert-validated taxonomy [28], and operationalized LAT as the six cognitive process types defined in [38]. Note that we chose this top-three design to match the number of elements of the most used gamification design (PBL - points, badges, and leaderboards) [46] because the number of game elements might affect gamification's effect [47].

B. Procedure

The following five steps were performed to develop our approach for tailoring gamified designs to LAT and users.

1) **Survey development**: defining the survey design and sections and the game elements and LAT to consider;

2) **Data collection**: disclosing the survey online, through Amazon's Mechanical Turk[2] (MTurk), to collect participants opinions.
3) **Data analysis**: running analyses to identify which characteristics impact users' preferences.
4) **Users preferences analysis**: investigating our findings to identify how to tailor educational systems' gamification designs to users, geographic location, and LAT.
5) **RS design**: developing a free, ready-to-use resource, based on our findings, to aid those who want to tailor their educational systems' gamified designs.

C. Survey

The survey was developed online[3] and can be viewed in the appendix. Its design was defined in four steps. First, two researchers brainstormed and developed an initial version. Second, three other researchers revised it and provided feedback on how to improve it. Following, the survey was improved accordingly and, lastly, we ran a pilot study with 50 participants.

The final version has four sections: consent form, demographics, gaming background, and preferences. In the **consent form**, all respondents were informed to be participating in a research and agreed all information provided would be used to research ends only. The **demographics** and **gaming background** captured participants' gender, age, living country, highest level of education, and MTurk identifier to avoid repeated completions, and for how many years the participants researched/worked with gamification (0 for those who did not), how much time (in hours) they spend with games per week, and their preferred game genre and playing setting, respectively. Lastly, in the **preferences** section, participants ranked the top three game elements they prefer the most when performing each of the six LAT. To aid users, this section described each game element along with examples.

The 21 game elements available were: Acknowledgment; Chance; Competition; Cooperation; Economy; Imposed Choice; Level; Narrative; Novelty; Objectives; Point; Progression; Puzzles; Rarity; Renovation; Reputation; Sensation; Social Pressure; Stats; Storytelling; Time Pressure. Further descriptions of these elements can be seen in [28]. The LAT are those introduced in Section II (remember, understand, apply, analyze, evaluate, and create). For further information about each one, see [38]. Thus, the last section had seven items - one for each LAT - and a repeated item to assess participant attention/consistency (see next section).

Each of those seven items had three sub-items, allowing the participant to select the rank-one, -two, and -three game elements, in which the 21 game elements were possible answers. Nevertheless, the same game element could not be selected twice within the same item. That is, each participant's top three should be composed of three different game elements. A sample question was *Indicate the three gamification elements you consider will help you the most when performing an activity you need to REMEMBER something (e.g., remember what the '+' symbol means in arithmetic operations).*, whereas other items of the same section differed only in the LAT (e.g., understand instead of remember) and the example at the end of the item. All items had basic mathematical examples due to the generality of the topic.

Additionally, we highlight that this top-three survey design was adopted due to the number of game elements (21) and LAT (six), which would lead to a questionnaire with 126 items if subjects should, similar to related work [15], [16], provide a rating for each gamification element through a Likert-scale. That is, participants would answer to 21 items six times; one time per LAT. Thus, we opted for one item per LAT, featuring three options each, to reduce effort, tiredness, and time spent in completing the survey, aiming to improve answers' reliability. Lastly, note that the survey sections' order was fixed (the same as previously introduced) but, within each section, the items' order was randomized.

D. Data Collection and Filtering

We recruited participants through crowdsourcing (MTurk). We made this choice to increase our sample size, similarly to related research [15], [20], an approach that has been recommended in the literature [48], [49] to improve external reliability [50]. No participant restriction was enforced to avoid selection biases and everyone who completed the survey received a fixed remuneration.

Nevertheless, similar studies [15], [20] have employed additional items to survey's long sections to assess whether participants are paying attention and providing consistent answers. Then, based on those specific items' answers, researchers filter participants according to some assertion threshold (e.g., discarding those who failed in more than one item [20]). In this study, we adopted a similar approach. On the *preferences* section, we added a repeated question for one LAT, which allowed us to assess whether the participant was consistent their answer (i.e., did they select the same top-three game elements in both items?). Participants' remuneration was not conditional to consistently answering, neither participants were warned about the repeated item, aiming to improve the reliability assessment.

Following related work, we adopted a tolerance for inconsistent completions. Hence, we removed all participants that provided consistent answers in less than two out of the three game elements. For instance, one selected Acknowledgment, Chance, and Competition and, then, in the repeated question, selected Acknowledgment, Cooperation, and Economy. This participant would be discarded by selecting two different game elements for the same question. In total, 1018 individuals have completed the survey, from which 657 answers were discarded based on our criteria. Thus, the final dataset contains 361 consistent answers. The description of these reliable, valid answers is shown in Table II.

Overall, our sample is composed of adults (51.5% males, 47.4% females, and 1.1% others) with 32 years on average (± 11) and undergraduate or higher degrees (65.4%). Hence, we might expect our sample to feature responsible people with good educational background. Furthermore, despite the

[2]https://www.mturk.com/
[3]Online survey: http://bit.ly/2JWxwqs

large majority never researched gamification (91%), there is an interesting variation in their preferred playing setting (Singleplayer: 59%; Multiplayer: 41%) as well as game genre (20% for the most preferred genre: Role Playing Game), with an overall playing time of 12 hours (±13) per week. Thereby, we might expect participants to be familiar with games and their elements.

TABLE II
DATASET DESCRIPTION.

Value	N(%)	Value	N(%)
Gender		*Preferred playing setting*	
Female	186 (0.515)	Singleplayer	214(0.593)
Male	171 (0.474)	Multiplayer	147(0.407)
Other Gender	4 (0.011)	*Researched gamification*	
Country		No	329 (0.911)
United States	259 (0.717)	Yes	32 (0.089)
India	22 (0.061)	*Age*	
United Kingdom	20 (0.055)	Mean	32.615
Canada	18 (0.050)	SD	11.299
Brazil	16 (0.044)	Min.	18.000
Others	26 (0.072)	25%	24.000
Highest education level		50%	29.000
Undergraduate	161 (0.446)	75%	39.000
High School	81 (0.224)	Max.	75.000
MsC	63 (0.175)	*Weekly playing time (hours)*	
Technical education	30 (0.083)	Mean	12.874
Other Education	14 (0.039)	SD	13.782
Ph.D	12 (0.033)	Min.	0.000
Preferred game genre		25%	4.000
Role Playing Game	75 (0.208)	50%	10.000
Adventure	61 (0.169)	75%	20.000
Action	60 (0.166)	Max.	112.000
Strategy	50 (0.139)		
Other Genre	115(0.319)		

Others, shown as: Country (Count) = Italy (6), Germany (5), Spain (3), Australia (2), Netherlands (1), Albania (1), France (1), Ireland (1), Poland (1), Turkey (1), Austria (1), Nigeria (1), Belize (1), Jamaica (1).

E. Data Analysis

For data analyses, we decided to work with decision trees, algorithms that determine an output based on the interaction between elements from an input set [51]. Besides handling interactions, which is key for our objective, a decision tree provides other three positive points that led us to choose it. First, it allows visualizing the rules followed to determine the output. Therefore, we can comprehensively discuss and understand how game elements are selected, given an input set (user data and LAT). Second, it demonstrates which aspects are more or less important, as the main ones are in the tree's top, and vice-versa. It also ignores unnecessary inputs, excluding from the tree those that do not contribute. That is, it works as a feature selection method. Hence, providing insights on which aspects influence users' preferences, as well as which are most influencing ones from those we studied. Third, the algorithm itself determines how each characteristic will be split (e.g., should age be split in 18-28, 29-39 or 18-23, 24-29, 30-39?), removing human bias that are likely to be inserted in this process.

Nevertheless, decision trees might be implemented in varied ways. The standard classification approach is optimized towards predicting a single output class (i.e., one of the output's values) given an input set [52]. Then, for our aim of creating an RS (RQ3), this approach would yield limited performance, in terms of the RS's ratings, because of the focus on a single, definitive output. Additionally, to cope with our survey design (top-three selections), we would need to create tree decision trees. An alternative is multi-label classification [53], wherein the output has multiple values (e.g., A-B instead of just A or just B). While this alternative would lead to a single, multi-label decision tree, it also has limitation concerning the RS ratings. Besides, it would limit our answer to RQ1 because we would not be able to distinguish how trees' rules change as the importance of users' preferences change from first to second to third selection. Moreover, there are algorithms that learn to rank [54], wherein the input is a list of items that can be ordered (i.e, they have a rating that is used as the sorting criterion). However, because those algorithms do not indicate by how much rank-one and rank-two differ, they would limit our answer to RQ3 as well. Additionally, the need of creating three trees would remain.

Apart from those, there are CDT, an alternative based on a statistical approach [45]. Consequently, CTD allow working with frequency tables as outputs, such as in a Chi-square test. That is, a distribution-based output. Hence, differently from standard singe output classification, the algorithm is optimized to such output format that gives a rate for each game element. Consequently, the output has ratings for our RS (RQ3), while we can also identify the most (and least) useful game elements (RQ2) based on highest (and lowest) values. Additionally, because CDT are tree-based algorithms, we can still identify which factors are relevant and how choices are made (RQ1) from their visual and feature selection properties. Therefore, we use CTD because they i) are designed to work with distribution-based outputs, ii) conduct internal feature selection, and iii) provide visual interpretations of their rules.

Furthermore, as CDT follow a statistical approach, they optimize the model evaluation and validation process [45]. In creating a model, for each factor (i.e., a feature of the input set), the algorithm analyses whether splitting it has a significant impact on the output (i.e., p $\leq$ alpha). For instance, an example would be checking whether including gender as a factor would change the distribution of participants' preferences for each game element. For our analysis, because the output is a distribution (i.e., a frequency table of game elements' selections), CTD use a test like a Chi-square. Hence, the algorithm itself performs the feature selection process, based on whether a factor significantly changes the output. Note that the algorithm only creates a node after analyzing all factors and selecting the most discriminant one (based on statistical significance). Then, the algorithm recursively repeats that procedure until no more significant splits are found. Consequently, the higher the node in a tree, the higher its importance.

Importantly, the algorithm defines how to split a feature (e.g., United State vs all others or United States and Canada vs all others). Therefore, CTD allow an statistical, in-deep analysis of which input factors are associated with the output, as well as reveal each factor's importance [45]. Additionally, notice that all splits are statistically significant. Hence, them as well as the factors included in the trees are expected to be variables with roles that generalize from the sample to the population, according to the Null Hypothesis Significance Testing Framework [55]. Thereby, that statistical approach acts as an embedded method of model validation, similar to cross-validation. Consequently, the model is evaluation depends on the extent to which the splits are statistically significant. While that approach differs from standard classification metrics (e.g., accuracy and precision), it is aligned to our method, especially because we used distributions as the CDT's outputs. Hence, metrics such as accuracy and precision would not properly evaluate our models.

Thus, according to our goal, data captured via our survey (Table II) was entered as input to generate three CDT. Each tree's output was users' preferences for either their first, second, or third selected game elements. Accordingly, each tree predicts the distribution of users' preferred game elements. Note that the dataset described in Table II is in the *wide format*. That is, one row per participant, and one column for their preference on each LAT (one column for LAT1, one for column LAT2, and so on). Then, to generate trees able to distinct users' preferences from one LAT to another, we converted the dataset to the *long format*; that is, six rows per participant, a new column indicating the LAT each row corresponds to, and a single column indicating the preferred game element from each user for each LAT. We highlight that, although this increases the size of the dataset inputted to the CDT, the characteristics' distribution remains the same.

IV. RESULTS

First, this section explores the CDT generated from our data, discussing the answers for our research questions in light of insights gained from them.

A. Conditional Decision Trees Overview

We built our first CDT - CDT1 - from participants' number one choice. That is, the game element they prefer the most for each LAT. Similarly, our second and third CDT - CDT2 and CDT3, respectively - concern the game element participants selected as the second- and third-preferred ones for each LAT. CDT1 is shown in Figure 1, where circles represent decision nodes and rectangles are leaf ones. Decision nodes function as if/else statements. For instance, the first node tests if, for a given input, the preferred game genre is equal to adventure, other genres, role playing game, or strategy (left), or equal to action (right). Based on the answer, it is decided whether one should follow to the left or right path of the tree. This procedure is iteratively repeated for each decision node until reaching a leaf node. Leaf nodes indicate the tree's output, which are the game elements' ratings (for simplicity, Figure 1 shows the game element(s) with the highest rating). Hence, for someone whose preferred the game genre is action and lives in the Netherlands or Spain, CDT1 recommends Objectives.

Note, however, that the tree in Figure 1 is a simplified version compared to the original tree generated from the R package *party*. That version has two main differences. First, it shows p values for each decision node, demonstrating they are significant splits. Second, their leaf nodes present bar plots, demonstrating to each game element's rating. Such information can be used to recommend the most preferred game element (i.e., highest rating) or to provide ratings on the most likely preferred ones (i.e., output all elements' ratings). Considering this context, we highlight Figure 1 only presents the most preferred game element due to the limited space, as the full image would not be readable within the article template. For similar reasons, CDT2 and CDT3 are not shown in the article. Nevertheless, the full images, with barpots for all leaf nodes, from all CDT we created, are available in the appendix.

B. RQ1: Characteristics that Impact User Preferences

RQ1 concerns finding which aspects, among user characteristics, geographic location, and LAT, impact users' preferences for game elements. Therefore, we analyze which of those appear in our CDT to identify the ones that influence participants' choices.

CDT1 used six of the nine (eight from Table II plus LAT) inputs: preferred game genre, LAT, gender, country, experience researching gamification, and education, which appeared in the tree in this order. Thus, for participants number one choice, those are the characteristics that impacted their preferences, with preferred game genre and education being the most and the less influencing ones. *CDT2* also used six out of the nine inputs: country, LAT, preferred game genre, gender, preferred playing setting, and weekly playing time, with the same order of relevance as presented here. Thus, for participants number two choice, those are the six characteristics that impacted their preferences, with country and weekly playing time being the most and the less influencing ones. *CDT3* used five of the nine inputs: country, preferred game genre, experience researching gamification, LAT, and education, which appeared in the tree in this order. Thus, these are the characteristics that influenced participants' preferences for their third choice, in which country was the most relevant one, as opposed to education and LAT that were both the less relevant ones.

Based on these findings, we answer **RQ1** with evidence that factors impacting users' preferences are country, LAT, preferred game genre and playing setting, gender, experience researching gamification, weekly playing time, and education. Additionally, we also found the order of importance of these characteristics for each of the three selections. This finding is summarized in Table III, which demonstrates the highest level of the tree where each characteristic appears (because one might appear multiple times and at different levels). Consequently - as the higher the level, the more the importance - allowing us to identify each one's importance.

C. RQ2: Most useful Game Elements Sets from User' Preferences

RQ2 concerns identifying the most useful game elements, given users' characteristics, the LAT they will perform, and their geographic location, according to participants' preferences. From the three CDT we generated, note that CDT1, CDT2, and CDT3 have 17, 16, and 15 terminal nodes, respectively. This means that, together, all trees provide recommendations for 48 combinations of the input set. Consequently,

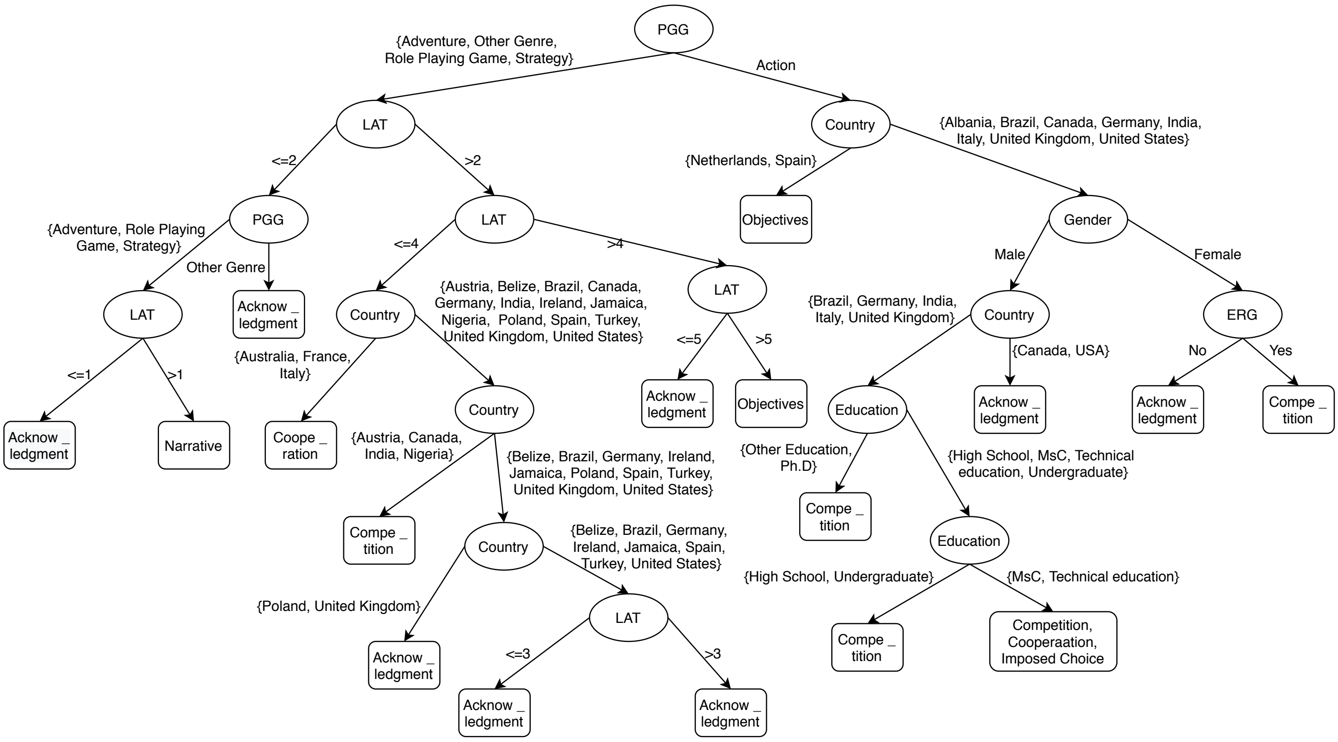

Fig. 1. Conditional decision tree for participants most preferred game element. Codes refer to preferred game genre (PGG), learning activity type (LAT), and experience researching gamification (ERG). Note that in cases of ties, leaf nodes present all game elements tied in alphabetical order.

TABLE III
LEVEL IN WHICH EACH CHARACTERISTIC APPEARED IN THE CDT OF EACH USERS' CHOICES.

Choice	Cnt	LAT	PGG	PPS	G	ERG	WPT	Edu
First	4	2	1		3	4		5
Second	1	2	2	5	5		6	
Third	1	4	2			2		4

Cnt = country; LAT = learning activity type; PGG = preferred game genre; PPS = preferred playing setting; G = gender; ERS = experience researching gamification; WPT = weekly playing time; Edu = education.

presenting a complete description of the recommendation for each of these combinations is unfeasible. Nevertheless, we demonstrate recommendations for specific cases to illustrate the most useful game elements for some cases, according to our findings.

First, consider the simple case wherein one wants to personalize gamification to LAT only, without considering any user characteristic. To illustrate that case, we split our dataset in six: each one containing only rows of one LAT. Then, we predict the output from each of our CDT using each sub-dataset. The results (Table IV) show there are cases (e.g., first, second, and third rows) in which the same element is recommended as second and third preferred. Although one participant could not select the same element for both cases, this corroborates the fact that the most select game element as second and third, considering the overall sample, was Objectives. Accordingly, our CDT recommend the same element as the second and third choices. With that in mind, our findings suggest that the most useful game elements set, considering LAT and no user characteristic, for LAT1 is Acknowledgment and Objectives, for LAT2 is Narrative and Objectives, and so on.

TABLE IV
RECOMMENDATIONS FOR PERSONALIZING GAMIFICATION TO LAT ONLY, WITHOUT CONSIDERING ANY USER CHARACTERISTIC, BASED ON OUR DATASET.

LAT	First	Second	Third
1	Acknowledgment	Objectives	Objectives
2	Narrative	Objectives	Objectives
3	Acknowledgment	Objectives	Objectives
4	Acknowledgment	Objectives	Acknowledgment
5	Acknowledgment	Level	Point
6	Objectives	Objectives	Progression

Among the main contributions from our approach, is its ability to handle multiple characteristics simultaneously, as well as the interaction between these characteristics. Therefore, we exemplify cases of personalizing gamification for a learning activity wherein students need to remember (LAT1) some content from long-term memory and then perform a second activity in which they need to evaluate (LAT5) others' opinions. Additionally, let us compare the most useful game elements set for Brazilian and Americans performing such activities. For simplicity, assume all students are males, never researched gamification, High School degree is their highest education level, play similar amounts of time per week (10 hours), and prefer the same game genre and playing setting: action and singleplayer, respectively[4]. In this context, the

[4]This fixed combination was selected arbitrarily, aiming to simplify the illustration. Other characteristics were not mentioned as they were found not to influence user preferences (see Table III)

recommendations are likely to vary due to changes in LAT, as well as geographic location (country), as all other relevant characteristics are the same. Finally, Table V demonstrates the recommendations for those cases.

TABLE V
RECOMMENDATIONS, DEPENDING ON LAT AND COUNTRY, FOR AN ARBITRARILY SELECTED SAMPLE: MALES, WHO NEVER RESEARCHED GAMIFICATION AND HAVE HIGH SCHOOL DEGREE AS THEIR HIGHEST EDUCATION LEVEL, PLAY 10 HOURS PER WEEK, AND PREFER PLAYING ACTION GAMES ALONE.

Combination	First	Second	Third
LAT1 - USA	Acknowledgment	Competition	Competition
LAT1 - Brazil	Competition	Competition	Time pressure
LAT5 - USA	Acknowledgment	Level	Point
LAT5 - Brazil	Competition	Level	Point

Our results (Table V) suggest the most useful game element set for LAT1 for Brazilians is Acknowledgment and Competition, whereas that for Americans is Competition and Time Pressure. For LAT5, the recommendation for Brazilians is Acknowledgment, Level, and Point, while for Americans the difference is Competition rather than Acknowledgment. Hence, highlighting the impact of contextual factors on users' preference, which differed depending on the LAT they were expecting to perform, as well as their geographic location.

In summary, we demonstrated which are the most useful game element set for specific combinations of user and contextual characteristics. In doing so, we selected the elements with the highest ratings according to our CDT. We did not show the recommendations for all combinations due to space restrictions. However, our CTD can be analyzed in their completeness - incuding each element's ratings - in the appendix for finding the recommendations.

D. RQ3: Recommender System to automate personalization

To provide technological aid that helps automating gamification personalization, consequently coping with the complexity of determining recommendations from visual inspection of CDT, we converted our tree CDT into an RS (**RQ3**). This system encapsulates all trees and simplifies the task of determining which game elements to use given a user, a LAT, and a geographic location. In summary, the RS's algorithm is as follows:

// 1. receives external information to create the trees's input
$a \leftarrow$ *user's preferred game genre*
$b \leftarrow$ *user's preferred playing setting*
$c \leftarrow$ *user's weekly playing time*
$d \leftarrow$ *user's gender*
$e \leftarrow$ *user's highest educational level*
$f \leftarrow$ *whether the user researched gamification before*
$g \leftarrow$ *user's living country*
$h \leftarrow$ *LAT to be gamified*
// 2. creates the trees's input
$input \leftarrow [a, b, c, d, e, f, g, h]$
// 3. uses the trees to get recommendations based on the input
$rec\_cdt1 \leftarrow CDT1(input)$
$rec\_cdt2 \leftarrow CDT2(input)$
$rec\_cdt3 \leftarrow CDT3(input)$
// 4. creates the output merging the recommendations from all trees
$output \leftarrow [rec\_cdt1, rec\_cdt2, rec\_cdt3]$

Specifically, the algorithm's first block receives external information in the same format as users passed to our survey (see Section III). Second, the algorithm merges such information to create the CDT's input. Third, the algorithm passes that same input to each of the CTDs, which return a list with a rating for each game element (values might be zero). Lastly, the algorithm merges the rating lists from all trees into a single matrix-like output similar to Table VI. Next, we describe the characteristics of our RS and briefly present technical concerns on how we converted our CDT into a free, easy-to-use system able to automate the personalization of GES.

We characterize our RS according to the framework for RS for personalized gamification introduced in [26]. Our RS considers six user inputs. Those are their preferred game genre and playing setting, weekly playing time, gender, highest education level, and whether the user researched gamification before. In addition, the user's living country, as well as the LAT that will be gamified, must also be entered, inputs related to the context [56]. Items are the game elements users could choose in the survey, the 21 game elements from the taxonomy proposed and validated in [28]. Lastly, the transactions concern users' preferred game elements (i.e., user with characteristics X, from geographic location Y, prefers element Z for LAT W), which is defined according to our findings.

The method adopted for output selection characterizes our RS as a hybrid recommender [26]. The input involves two contextual characteristics, geographic location and the LAT to be performed. Accordingly, the method would be characterized as a context-aware recommender. However, the selection process also relies on empirical information from our findings, which concerns a content-based recommender. Thus, our RS is a hybrid recommender due to exploring the characteristics of two methods. Lastly, our RS outputs are the ratings for game elements, for each of the top-three recommendations, defined according to the percentual of each game element's selection for that input. Consequently, the highest percentual reflects a recommendation's accuracy, given that the element with the highest rating is recommended. For instance, if we consider the case shown in Table VI, the Acknowledgment game element rating would be roughly 0.25 for the first selection. Accordingly, the recommendation accuracy is 0.25 because 25% of the observations within those criteria selected Acknowledgment.

Table VI demonstrates an output of our RS for people who live in the United States, have no experience in researching gamification, the preferred game genre is RPG, and will complete an analyzing learning activity. It demonstrates a full output of the RS with the ratings for all game elements when considered as first-, second-, and third-preference. For participants' number one preference, the game element with the highest rating is Acknowledgment (0.246), followed by Novelty (0.116). For participants number two choice, the highest rating is for Level (0.234), followed by Novelty (0.107).

For their third preferred element, Point holds the highest rating (0.125), followed by Novelty (0.125). When using the RS for other input sets, similar outputs will be given, likely with different ratings for each game element. Hence, based on outputs as that shown in Table VI, one can assess which elements are more likely to be the preferred ones for a given situation and define their gamification design accordingly.

TABLE VI
RATINGS OF OUR RS FOR PEOPLE WHO LIVE IN THE UNITED STATES, HAS NO EXPERIENCE IN RESEARCHING GAMIFICATION, PREFERRED GAME GENRE IS RPG, AND WILL COMPLETE AN ANALYZING LEARNING ACTIVITY.

Game element	First	Second	Third
Acknowledgment	0.246	0.069	0.086
Chance	0.043	0.032	0.036
Competition	0.050	0.060	0.056
Cooperation	0.076	0.095	0.076
Economy	0.040	0.041	0.040
Imposed Choice	0.050	0.091	0.076
Level	0.106	0.123	0.046
Narrative	0.003	0.009	0.010
Novelty	0.116	0.107	0.112
Objectives	0.023	0.079	0.063
Point	0.027	0.057	0.125
Progression	0.053	0.038	0.030
Puzzles	0.007	0.006	0.013
Rarity	0.013	0.022	0.017
Renovation	0.013	0.022	0.050
Reputation	0.003	0.019	0.033
Sensation	0.010	0.003	0.023
Social pressure	0.076	0.069	0.069
Stats	0.017	0.035	0.017
Storytelling	0.027	0.022	0.023
Time Pressure	0.000	0.000	0.000

Aiming to improve the usability of our RS, we reimplemented the CDT generated through the R package *party* [45] in Javascript. Although one could access the R objects, or try to make some external connection to R code from, for instance, a web browser, this process could be laborious and discouraging. On the other hand, Javascript can be easily run in most web browsers, as well as be easily plugged-in into a web site. Furthermore, as decision trees can be represented through a set of if/else statements, the conversion from R objects to Javascript does not require handling complex programming technical challenges. This is another advantage because the procedure of transforming our CDT into a Javascript plugin can be replicated to any other programming language.

Our RS is freely available (see the appendix), and there are two main use cases in which we believe it can be explored. First, the main case of automating gamification personalization, in which other systems use it as an external resource/tool. In this case, a gamified system can explore our RS as a plug-in that is consulted to find which game elements should be available for some occasion. To this end, the system would call the plug-in, passing the needed inputs as parameters to receive the ratings of each game element. Then, the system could, for instance, turn on those elements with the highest ratings. This procedure could be iteratively repeated, when the type of the learning activity to be performed changed, for instance. Thus, the RS would aid the system in performing dynamic adaptations [18] of its gamification design according to the user's characteristics and geographic locations as well as the tasks performed. The second case is using our RS as a standalone tool to provide recommendations for one interested in, for instance, personalizing an unplugged gamified environment [57] or to manually define their system gamification.

V. Discussion

Based on participants' preferences captured though a survey, our findings provided evidence that users' preferences differ depending on their characteristics, geographic location, and the LAT to be performed (RQ1). Also, we were able to develop an RS that recommends the preferred gamification design for a LAT to be performed by a user with some specific characteristic in a defined geographic location (RQ2). The main contribution of this research is, therefore, providing a free RS for personalized gamification, built upon a state-of-the-art approach, that aids in automating the tailoring of gamification designs by suggesting which game elements to use (RQ3). This RS is based on three aspects of personalization: domain, user, and task [6], implemented as the educational domain, demographics and gaming characteristics, and LAT and geographic location, respectively. Additionally, we revealed which context and user characteristics impact their preferences, and which of those are more or less relevant, contributing to expanding and grounding knowledge from previous studies (e.g., [15], [31]).

Concerning the results on users' characteristics impacting their preferences, our findings are aligned with the literature. Previous studies have shown that, for instance, demographics [19], [32] and attributes related to users' gaming habits [31] affect user preference. We corroborate those by providing more empirical evidence that users with different characteristics have different preferences, as well as presenting which of those are more important than others. For instance, we found simple user attributes, such as gender and having researched gamification, are less relevant than gaming-related characteristics (see Table III), which is in line with previous literature suggestions [58]. Furthermore, it also has been discussed that the task to be performed influences the perceptions of gamified systems' users [6]. Following that and within the educational context, suggestions to consider learning activities within the tailoring process of educational systems have emerged [18], [35]. Our findings are aligned with those theories as well, showing that users' preferences differ depending on the LAT they expect to perform (see Figure 1 and Section IV-C). Additionally, we found geographic location to be another relevant factor, a finding consistent with recent literature suggestions [24].

Concerning the results on users' preferred gamification design for each LAT given their characteristics, we expand the literature by i) providing recommendations applicable to any task (by considering its main objective - type) and ii) exploring less studied user characteristics (i.e., demographics and gaming-related) as well as taking into account their geographic location. On one hand, besides not guiding on how to tailor to LAT and geographic location, other personalization approaches often rely on user profiles [7]. However, as shown by our findings (see Table III), demographics and gaming-related characteristics are relevant as well. On the other hand, despite the recent calls for considering learning activities when personalizing [18], [35], available approaches considering such aspects are yet limited, mainly due to considering only two characteristics (one for from user and the learning activity) [17], [25]. Although, if multiple aspects are relevant, they all should be considered, as well as their interaction [6], [22]. Our research contributes to these concerns, guiding how to personalize gamification to users (i.e., demographics and game-related) and contextual (i.e., LAT and geographic location) aspects simultaneously.

Moreover, this article advances the literature by providing an RS for personalized gamification. In [26], a framework for such RS has been proposed, however, the literature still lacks concrete implementations of these systems. On the other hand, recent research has highlighted the need for research to aid in the automation of gamification personalization [24]. This article contributes to this vein by introducing a free RS for personalized gamification that can be both plugged-in gamified systems to automate their personalization process, as well as independently used as a guide for defining personalized gamification designs. As this system is built upon the findings from this article, it implements a state-of-the-art personalization approach, which addresses a couple of literature challenges, namely the need for considering contextual factors along with user information, as well as the interaction between all relevant characteristics (see Section II). Nevertheless, note that our sample size limits our RS, as well as the unbalance in various characteristics such as participants' countries. Therefore, while we validated our models based on standard statistical practices, their recommendations must be interpreted with caution, always analyzing ratings and having their limitations in mind.

A. Implications

There are five main implications of our findings. First, demographics and gaming-related characteristics are moderators of user preference that should be prioritized differently. We showed that these characteristics do affect user preference but that each one's importance differs from one to another. Additionally, those exploring gamification effectiveness might rely on our results to define which data to capture from their samples to further assess whether these characteristics also play a role in other aspects (e.g., motivation or learning from interacting with GES).

Second, personalization approaches should be expanded beyond the user. We have shown that the game elements people prefer when expecting to perform one LAT differ from what they prefer when expecting to perform another; similarly for users who live in different countries. These findings' implication is empirical evidence that rather than just thinking on what users generally prefer, aspects of the task that will be performed and the user's geographic location should be taken into account, supporting recent literature arguments [6], [18], [24], [35].

Third, the interaction between relevant characteristics cannot be ignored. Our results demonstrated that the game elements preferred the most are likely to change when a single characteristic (e.g., country) changes. For example, we demonstrated that the recommended game elements for the same LAT will differ for Brazilian and American users, even if all other characteristics are the same. Thus, confirming the need for tailoring gamification designs not only to the user but also to the context [18], [35] as well as considering the interaction between different aspects [6], [22]. Hence, the implication is that only one side of the whole is likely not to work in full potential.

Fourth, when designing GES, two people might prefer the same game elements, but with different priorities. When surveying participants, we asked them to rank the top three game elements that would help them the most in learning activities of a specific type. Hence, gathering data able to inform not only which game elements are the most preferred on each occasion, but also the importance order of the selected elements. Thus, we imply that when relying on our findings to design GES, one should define the emphasis each game element will receive based on users' selection order (see Section IV-C) because despite different individuals might prefer the same game elements set, they might prefer those with different priorities.

Lastly, putting together our findings and analyses, one can use our RS (see the appendix) to automate gamified systems' personalization process as well as be informed on how to tailor gamification designs of educational systems. Practitioners can exploit our RS to define their systems' gamification designs, as well as researchers can apply its recommendations on their studies to assess the effectiveness of users' preferred designs. To aid those interested in using our RS, we have made it freely available for use and briefly discussed how it can be either incorporated into an existing system as well as using it as a guide. Thus, to the best of our knowledge, these findings pose a direct implication to the design and development of GES as it offers the first technological for personalization of gamification. Nevertheless, one should not disregard the RS's limitations, which are consequences from our sample size and characteristics and must be taken into account when using it.

B. Limitations

This section discusses our study limitations. Concerning the survey: It presented a description for each LAT and each game element to avoid misinterpretations and, consequently, guarantee answers reliability. However, this likely increased the complexity of understanding it as well as the time required to complete the survey, possibly contributing to tiring the participants throughout the process. To address this limitation, we adopted the rank-based design, which reduced the number of items, and added an attention question, which allowed us to discard inconsistent answers.

Concerning the sample: Some participants' attributes were highly unbalanced. For instance, 71% of the participants are from the United States, and many countries have less than 10 participants. Consequently, the external validity of recommendations for groups with a small presence in the dataset (e.g., those from countries other than the United States) is substantially affected. We addressed that limitation using CDT, which decide how to split countries (e.g., having recommendations for each country or for a group of countries; see Figure 1) based on statistical significance, which handles the different subsample sizes to some extent. Thus, despite sample size limitations do affect our findings' external validity, our recommendations were built to control for those.

Concerning recommendations' effectiveness: This was an exploratory, preference-based research, following a methodology commonly adopted by related research. Consequently, as in previous research, we cannot ensure that personalizing to users' preferences will be effective. However, given the number of game elements to be considered (21) as well as LAT (six), thousands of combinations would have to be tested in user studies, which is unfeasible. Our survey-based study addresses this limitation by presenting a valuable first step in suggesting which game elements to use for specific conditions and providing guidance for future studies to test our preference-based recommendations. Furthermore, whereas our survey-based recommendations have to be empirically tested, such recommendations would only be possible after data collection. Therefore, our RS helps to address the cold-starting problem [26] while it can be enhanced with real usage data in the future.

Nevertheless, readers must consider that the unbalanced features from our sample limits the generalization from our recommendations. To address that, we used CTD, which handle such variations through its statistical approach. Consequently, that led to the limitation of validating and evaluating our models based on statistical rather than standard machine learning approaches. Hence, while we used a well-established approach according to our goal, we call for future research to extend our contribution based on larger, more heterogeneous samples.

Concerning the RS's input: Although we selected the revision of Bloom's taxonomy due to its relevance within the education context, the lack of a systematic selection process also limits our findings in terms of how our study interprets LAT. Also, as our recommendations are based on averages, it might be that it will not work for some users. Lastly, although our RS is a ready-to-use resource, it is a plug-in in its initial version that can be further enhanced to improve, for instance, its compatibility with other systems, as well as its presentation for independent use.

VI. Final Remarks

Personalization emerged as an alternative to improve gamification effectiveness. Most studies in this field exploit user profiles to tailor the gamified designs. Hence, they ignore the fact that, besides the user, tasks and domain play a significant role in gamification's success. Additionally, studies often do not consider the interaction between multiple relevant characteristics, neither offer concrete resources to help in automating gamification personalization. To address these gaps, this article introduced a preference-based RS that suggests game elements tailored to the user (demographics and game-related characteristics) and the context (LAT - tasks - and geographic location), focused on the educational domain. This RS considers the interaction between its inputs and is freely available for anyone to use it.

Thus, our contributions are twofold. First, we provided practitioners with a ready-to-use resource able to guide them on how to design GES that are tailored to users' characteristics, as well as geographic location, according to the tasks they will perform. Second, we expanded the literature on how to tailor gamification designs to any learning activity (based on its type) by presenting recommendations that might be empirically tested in future research, providing empirical evidence on which demographics and game-related user characteristics impact their preferences, as well as which one is more important than the others, and supporting literature suggestions by showing that LAT and geographic location do affect user preference.

As future studies, we mainly recommend validating the effectiveness of our RS recommendations (e.g., ability to improve user motivation, flow, academic performance, or learning gains), compared to one-size-fits-all and other personalization methods, to identify whether personalizing to users' preferences will positively impact them as expected. Another line of future research is expanding our RS, especially because of our sample restrictions due to some unbalanced attributes, with more heterogeneous samples. Additional improvements are transforming it into a service to mitigate compatibility problems as well as the need for manually adding the code to the project. Additionally, future studies might tackle the limitation of not assessing the match between all game elements and all LAT from our methodology, which might be accomplished in steps (e.g., assessing one LAT per experiment) to cope with the complexity of testing all at once.

Appendix A

Appendixes are available at: shorturl.at/aguQT.

Acknowledgment

The authors would like to thank the funding provided by CNPq (141859/2019-9; 308458/2020-6), CAPES (Finance code - 001), and FAPESP (Projects: 2018/07688-1; 2018/15917-0; 2016/02765-2; 2018/11180-3; 2013/07375-0).

References

[1] K. M. Kapp, *The gamification of learning and instruction*. Wiley San Francisco, 2012.
[2] S. Deterding, D. Dixon, R. Khaled, and L. Nacke, "From game design elements to gamefulness: Defining "gamification"," in *Proceedings of the 15th International Academic MindTrek Conference: Envisioning Future Media Environments*, ser. MindTrek '11, Tampere, Finland, Sep. 28-30, 2011, p. 9–15. [Online]. Available: 10.1145/2181037.2181040
[3] M. Sailer and L. Homner, "The gamification of learning: a meta-analysis," *Educational Psychology Review*, Aug 2019. [Online]. Available: https://doi.org/10.1007/s10648-019-09498-w
[4] A. M. Toda, P. H. D. Valle, and S. Isotani, "The dark side of gamification: An overview of negative effects of gamification in education," in *Higher Education for All. From Challenges to Novel Technology-Enhanced Solutions*, A. I. Cristea, I. I. Bittencourt, and F. Lima, Eds. Cham: Springer International Publishing, 2018, pp. 143–156. [Online]. Available: 10.1007/978-3-319-97934-2_9
[5] B. Morschheuser, L. Hassan, K. Werder, and J. Hamari, "How to design gamification? a method for engineering gamified software," *Information and Software Technology*, vol. 95, pp. 219–237, Mar 2018. [Online]. Available: https://doi.org/10.1016/j.infsof.2017.10.015

[6] D. Liu, R. Santhanam, and J. Webster, "Toward meaningful engagement: A framework for design and research of gamified information systems," *MIS quarterly*, vol. 41, no. 4, pp. 1011–1034, Jun 2017. [Online]. Available: 10.25300/MISQ/2017/41.4.01

[7] L. Rodrigues, A. M. Toda, P. T. Palomino, W. Oliveira, and S. Isotani, "Personalized gamification: A literature review of outcomes, experiments, and approaches," in *Eighth International Conference on Technological Ecosystems for Enhancing Multiculturality*, ser. TEEM'20, Salamanca, Spain, Oct. 21-23, 2020, p. 699–706. [Online]. Available: https://doi.org/10.1145/3434780.3436665

[8] G. F. Tondello, "Dynamic personalization of gameful interactive systems," Ph.D. dissertation, University of Waterloo, 2019.

[9] N. Yee, "Gaming motivations align with personality traits," jan 2016, accessed in January 2020. [Online]. Available: quanticfoundry.com/2016/01/05/personality-correlates/

[10] L. Rodrigues and J. D. Brancher, "Improving players' profiles clustering from game data through feature extraction," in *Proceedings of SBGames 2018 - Computing Track*. Foz do Iguaçu, Brazil: IEEE, Oct. 29-01, 2018, pp. 177–186. [Online]. Available: 10.1109/SBGAMES.2018.00029

[11] R. Orji, J. Vassileva, and R. L. Mandryk, "Modeling the efficacy of persuasive strategies for different gamer types in serious games for health," *User Modeling and User-Adapted Interaction*, vol. 24, no. 5, pp. 453–498, Jul. 7-11, 2014. [Online]. Available: 10.1007/s11257-014-9149-8

[12] A. Knutas, R. Van Roy, T. Hynninen, M. Granato, J. Kasurinen, and J. Ikonen, "A process for designing algorithm-based personalized gamification," *Multimedia Tools and Applications*, vol. 78, no. 10, pp. 13 593–13 612, Dec 2019. [Online]. Available: 10.1007/s11042-018-6913-5

[13] L. Rodrigues, A. Toda, W. Oliveira, P. Palomino, and S. Isotani, "Just beat it: Exploring the influences of competition and task-related factors in gamified learning environments," in *Proceedings of the Brazilian Symposium of Computers on Education*, Porto Alegre, Brasil, Nov. 24-28, 2020, pp. 461–470. [Online]. Available: 10.5753/cbie.sbie.2020.461

[14] J. Koivisto and J. Hamari, "The rise of motivational information systems: A review of gamification research," *International Journal of Information Management*, vol. 45, pp. 191–210, Apr 2019. [Online]. Available: 10.1016/j.ijinfomgt.2018.10.013

[15] G. F. Tondello, A. Mora, and L. E. Nacke, "Elements of gameful design emerging from user preferences," in *Proceedings of the Annual Symposium on Computer-Human Interaction in Play*, Amsterdam, The Netherlands, Oct. 15-18, 2017, pp. 129–142. [Online]. Available: 10.1145/3116595.3116627

[16] W. Oliveira and I. I. Bittencourt, *Tailored Gamification to Educational Technologies*. Springer Nature, 2019.

[17] J. Baldeón, I. Rodríguez, and A. Puig, "Lega: A learner-centered gamification design framework," in *Proceedings of the XVII International Conference on Human Computer Interaction*, Salamanca, Spain, Sep. 13-16, 2016, pp. 45:1–45:8. [Online]. Available: http://doi.acm.org/10.1145/2998626.2998673

[18] S. Hallifax, A. Serna, J.-C. Marty, and É. Lavoué, "Adaptive gamification in education: A literature review of current trends and developments," in *Transforming Learning with Meaningful Technologies*, M. Scheffel, J. Broisin, V. Pammer-Schindler, A. Ioannou, and J. Schneider, Eds. Cham: Springer International Publishing, 2019, pp. 294–307. [Online]. Available: 10.1007/978-3-030-29736-7_22

[19] G. F. Tondello, R. R. Wehbe, L. Diamond, M. Busch, A. Marczewski, and L. E. Nacke, "The gamification user types hexad scale," in *Proceedings of the 2016 annual symposium on computer-human interaction in play*, Austin, Texas, USA, Oct. 16-19, 2016, pp. 229–243. [Online]. Available: 10.1145/2967934.2968082

[20] S. Hallifax, A. Serna, J.-C. Marty, G. Lavoué, and E. Lavoué, "Factors to consider for tailored gamification," in *Proceedings of the Annual Symposium on Computer-Human Interaction in Play*, ser. CHI PLAY '19, Barcelona, Spain, Oct. 22-25, 2019, p. 559–572. [Online]. Available: https://doi.org/10.1145/3311350.3347167

[21] A. Mora, G. F. Tondello, L. Calvet, C. González, J. Arnedo-Moreno, and L. E. Nacke, "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*, Donostia Gipuzkoa, Spain, Jun. 25-28, 2019, p. 1. [Online]. Available: 10.1145/3335595.3335625

[22] P. Buckley and E. Doyle, "Individualising gamification: An investigation of the impact of learning styles and personality traits on the efficacy of gamification using a prediction market," *Computers & Education*, vol. 106, pp. 43–55, Mar 2017. [Online]. Available: 10.1016/j.compedu.2016.11.009

[23] S. Borges, V. Durelli, H. Reis, I. I. Bittencourt, R. Mizoguchi, and S. Isotani, "Selecting effective influence principles for tailoring gamification-based strategies to player roles," in *Brazilian Symposium on Computers in Education*, vol. 28, Recife, Brazil, Oct. 30-02 2017, pp. 857–866. [Online]. Available: 10.5753/cbie.sbie.2017.857

[24] A. C. T. Klock, I. Gasparini, M. S. Pimenta, and J. Hamari, "Tailored gamification: A review of literature," *International Journal of Human-Computer Studies*, vol. 144, p. 102495, Dec 2020. [Online]. Available: https://doi.org/10.1016/j.ijhcs.2020.102495

[25] K. Bovermann and T. J. Bastiaens, "Towards a motivational design? connecting gamification user types and online learning activities," *Research and Practice in Technology Enhanced Learning*, vol. 15, no. 1, pp. 1–18, Jan 2020. [Online]. Available: 10.1186/s41039-019-0121-4

[26] G. F. Tondello, R. Orji, and L. E. Nacke, "Recommender systems for personalized gamification," in *Adjunct Publication of the 25th Conference on User Modeling, Adaptation and Personalization*, Bratislava, Slovakia, Jul. 09-12 2017, pp. 425–430. [Online]. Available: 10.1145/3099023.3099114

[27] F. F.-H. Nah, Q. Zeng, V. R. Telaprolu, A. P. Ayyappa, and B. Eschenbrenner, "Gamification of education: a review of literature," in *International conference on hci in business*, Heraklion, Greece, Jun 22-27 2014, pp. 401–409. [Online]. Available: 10.1007/978-3-319-07293-7_39

[28] A. M. Toda, W. Oliveira, A. C. Klock, P. T. Palomino, M. Pimenta, I. Gasparini, L. Shi, I. Bittencourt, S. Isotani, and A. I. Cristea, "A taxonomy of game elements for gamification in educational contexts: Proposal and evaluation," in *2019 IEEE 19th International Conference on Advanced Learning Technologies (ICALT)*, vol. 2161-377X, Maceió, Brazil, Jul. 15-18 2019, pp. 84–88. [Online]. Available: 10.1109/ICALT.2019.00028

[29] A. M. Toda, A. C. Klock, W. Oliveira, P. T. Palomino, L. Rodrigues, L. Shi, I. Bittencourt, I. Gasparini, S. Isotani, and A. I. Cristea, "Analysing gamification elements in educational environments using an existing gamification taxonomy," *Smart Learning Environments*, vol. 6, no. 1, p. 16, Dec 2019. [Online]. Available: 10.1186/s40561-019-0106-1

[30] R. R. McCrae and O. P. John, "An introduction to the five-factor model and its applications," *Journal of personality*, vol. 60, no. 2, pp. 175–215, Jun 1992. [Online]. Available: 10.1111/j.1467-6494.1992.tb00970.x.

[31] M. Denden, A. Tlili, F. Essalmi, and M. Jemni, "An investigation of the factors affecting the perception of gamification and game elements," in *2017 6th International Conference on Information and Communication Technology and Accessibility (ICTA)*, Muscat, Sultanate of Oman, Dec. 19-21 2017, pp. 1–6. [Online]. Available: 10.1109/ICTA.2017.8336019

[32] A. Toda, W. Oliveira, L. Shi, I. Bittencourt, S. Isotani, and A. Cristea, "Planning gamification strategies based on user characteristics and dm: A gender-based case study," in *12th International Conference on Educational Data Mining*, Montreal, Canada, Jul. 02-05, 2019, pp. 438–443.

[33] A. M. Toda, R. M. do Carmo, A. P. da Silva, I. I. Bittencourt, and S. Isotani, "An approach for planning and deploying gamification concepts with social networks within educational contexts," *International Journal of Information Management*, vol. 46, pp. 294 – 303, Jun 2019. [Online]. Available: https://doi.org/10.1016/j.ijinfomgt.2018.10.001

[34] S. Deterding, "The lens of intrinsic skill atoms: A method for gameful design," *Human–Computer Interaction*, vol. 30, no. 3-4, pp. 294–335, May 2015. [Online]. Available: 10.1080/07370024.2014.993471

[35] L. Rodrigues, W. Oliveira, A. Toda, P. Palomino, and S. Isotani, "Thinking inside the box: How to tailor gamified educational systems based on learning activities types," in *Proceedings of the Brazilian Symposium of Computers on Education*, Brasilia, Brazil, Nov. 11-14 2019, pp. 823–832. [Online]. Available: 10.5753/cbie.sbie.2019.823

[36] R. E. Wood, "Task complexity: Definition of the construct," *Organizational Behavior and Human Decision Processes*, vol. 37, no. 1, pp. 60–82, Feb 1986. [Online]. Available: https://doi.org/10.1016/0749-5978(86)90044-0

[37] D. Diaper and N. Stanton, *The handbook of task analysis for human-computer interaction*. CRC Press, 2003.

[38] D. R. Krathwohl, "A revision of bloom's taxonomy: An overview," *Theory into practice*, vol. 41, no. 4, pp. 212–218, Jun 2002. [Online]. Available: 10.1207/s15430421tip4104_2

[39] B. S. Bloom, *Taxonomy of educational objectives: The classification of educational goals*. New York: Longman, 1956.

[40] F. Ricci, L. Rokach, and B. Shapira, “Introduction to recommender systems handbook,” in *Recommender Systems Handbook*, F. Ricci, L. Rokach, B. Shapira, and P. B. Kantor, Eds. Boston, MA: Springer US, 2011, pp. 1–35. [Online]. Available: 10.1007/978-0-387-85820-3_1

[41] D. Dicheva, K. Irwin, and C. Dichev, “Motivational factors in educational gamification,” in *2018 IEEE 18th International Conference on Advanced Learning Technologies (ICALT)*, Mumbai, India, Jul. 09-13, 2018, pp. 408–410. [Online]. Available: 10.1109/ICALT.2018.00102

[42] Y. Xu and Y. Tang, “Based on action-personality data mining, research of gamification emission reduction mechanism and intelligent personalized action recommendation model,” in *Cross-Cultural Design Methods, Practice and Impact*, P. Rau, Ed. Cham: Springer International Publishing, 2015, pp. 241–252. [Online]. Available: 10.1007/978-3-319-20907-4_22

[43] B. Monterrat, É. Lavoué, and S. George, “Toward an adaptive gamification system for learning environments,” in *Computer Supported Education*, S. Zvacek, M. T. Restivo, J. Uhomoibhi, and M. Helfert, Eds. Cham: Springer International Publishing, 2015, pp. 115–129. [Online]. Available: 10.1007/978-3-319-25768-6_8

[44] J. Lazar, J. H. Feng, and H. Hochheiser, *Research methods in human-computer interaction*. Morgan Kaufmann, 2017.

[45] T. Hothorn, K. Hornik, and A. Zeileis, “Unbiased recursive partitioning: A conditional inference framework,” *Journal of Computational and Graphical Statistics*, vol. 15, no. 3, pp. 651–674, Jan 2006. [Online]. Available: https://doi.org/10.1198/106186006X133933

[46] S. Bai, K. F. Hew, and B. Huang, “Does gamification improve student learning outcome? evidence from a meta-analysis and synthesis of qualitative data in educational contexts,” *Educational Research Review*, vol. 30, p. 100322, Jun 2020. [Online]. Available: https://doi.org/10.1016/j.edurev.2020.100322

[47] R. N. Landers, K. N. Bauer, R. C. Callan, and M. B. Armstrong, *Psychological Theory and the Gamification of Learning*. Cham: Springer International Publishing, 2015, pp. 165–186. [Online]. Available: https://doi.org/10.1007/978-3-319-10208-5_9

[48] K. Casler, L. Bickel, and E. Hackett, “Separate but equal? a comparison of participants and data gathered via amazon’s mturk, social media, and face-to-face behavioral testing,” *Computers in human behavior*, vol. 29, no. 6, pp. 2156–2160, Nov 2013. [Online]. Available: 10.1016/j.chb.2013.05.009

[49] R. N. Landers and T. S. Behrend, “An inconvenient truth: Arbitrary distinctions between organizational, mechanical turk, and other convenience samples,” *Industrial and Organizational Psychology*, vol. 8, no. 2, pp. 142–164, Mar 2015. [Online]. Available: 10.1017/iop.2015.13

[50] C. Wohlin, P. Runeson, M. Hst, M. C. Ohlsson, B. Regnell, and A. Wessln, *Experimentation in Software Engineering*. Springer Publishing Company, Incorporated, 2012.

[51] S. Safavian and D. Landgrebe, “A survey of decision tree classifier methodology,” *IEEE Transactions on Systems, Man, and Cybernetics*, vol. 21, no. 3, pp. 660–674, May 1991. [Online]. Available: 10.1109/21.97458

[52] J. Mingers, “Expert systems—rule induction with statistical data,” *Journal of the Operational Research Society*, vol. 38, no. 1, pp. 39–47, Dec 1987. [Online]. Available: 10.1057/jors.1987.5

[53] G. Tsoumakas and I. Katakis, “Multi-label classification: An overview,” *International Journal of Data Warehousing and Mining (IJDWM)*, vol. 3, no. 3, pp. 1–13, Jul 2007. [Online]. Available: 10.4018/jdwm.2007070101

[54] X. Amatriain and J. Basilico, “Past, present, and future of recommender systems: An industry perspective,” in *Proceedings of the 10th ACM Conference on Recommender Systems*, Boston, Massachusetts, USA, Sep. 15-19, 2016, p. 211–214. [Online]. Available: 10.1145/2959100.2959144

[55] D. J. Sheskin, *Handbook of parametric and nonparametric statistical procedures*. Chapman and Hall/CRC, 2003.

[56] M. Khemaja and F. Buendia, “Building context-aware gamified apps by using ontologies as unified representation and reasoning-based models,” in *Serious Games and Edutainment Applications*. Springer, 2017, pp. 675–702. [Online]. Available: 10.1007/978-3-319-51645-5_29

[57] A. M. Toda, P. T. Palomino, W. Oliveira, L. Rodrigues, A. C. Klock, I. Gasparini, A. I. Cristea, and S. Isotani, “How to gamify learning systems? an experience report using the design sprint method and a taxonomy for gamification elements in education,” *Journal of Educational Technology & Society*, vol. 22, no. 3, pp. 47–60, Dec 2019.

[58] R. N. Landers, G. F. Tondello, D. L. Kappen, A. B. Collmus, E. D. Mekler, and L. E. Nacke, “Defining gameful experience as a psychological state caused by gameplay: Replacing the term ‘gamefulness’ with three distinct constructs,” *International Journal of Human-Computer Studies*, vol. 127, pp. 81–94, Jul 2019. [Online]. Available: 10.1016/j.ijhcs.2018.08.003

Luiz Rodrigues Luiz Rodrigues is a Ph.D. candidate at the University of São Paulo. His research focuses on understanding and improving gamification effectiveness within the context of educational systems. His main research interests are gamification, personalization, user modeling, and procedural content generation.

Armando M. Toda Armando Toda is, currently, a PhD. candidate at University of Sao Paulo working with gamification design in educational contexts. Other main areas of research are concerned with: data mining, serious games, evaluation and instrument design.

Wilk Oliveira Wilk Oliveira is a Ph.D. candidate at the University of São Paulo and MSc. in Computer Science for the Federal University of Alagoas with exchange programs at the University of Saskatchewan. He was a guest lecturer at the Tiradentes University Center. His main topics include Artificial Intelligence in Education, Human-computer Interaction, and Gamification.

Paula T. Palomino Paula Palomino is a PhD. candidate at University of São Paulo, researching and developing a gamification content-based framework for educational purposes. Other areas of study include Games, HCI and Cyberculture.

Julita Vassileva Julita Vassileva is Professor at the Department of Computer Science, University of Saskatchewan. Julita does research in Social Computing, Persuasive Technologies, Personalization, User Modeling, AI in Education, Peer-to-Peer systems, Trust and Reputation Mechanisms, and Human-computer Interaction. She is currently working on Personalized Persuasive Systems, Learning Data-Driven Persuasion, and Distributed Ledgers for Storing and Sharing Personal Data.

Seiji Isotani Seiji Isotani is a full professor at the University of São Paulo (USP). His research topics include Artificial Intelligence in Education, Intelligent Tutoring Systems, Computer-Supported Collaborative Learning (CSCL), Gamification, Ontology Engineering and Linked Open Data. He has published more than a hundred scientific papers, has more than 3000 citations and, according to Google Scholar, is among the 15 top researchers in gamification

ANNEX

H

PERSONALIZATION IMPROVES GAMIFICATION: EVIDENCE FROM A MIXED-METHODS STUDY

Personalization Improves Gamification: Evidence from a Mixed-methods Study

LUIZ RODRIGUES, University of São Paulo, Brazil
PAULA T. PALOMINO, University of São Paulo, Brazil and University of Waterloo, Canada
ARMANDO M. TODA, University of São Paulo, Brazil
ANA C. T. KLOCK, Federal University of Rio Grande do Sul, Brazil and Tampere University, Finland
WILK OLIVEIRA, University of São Paulo, Brazil and Tampere University, Finland
ANDERSON P. AVILA-SANTOS, University of São Paulo, Brazil
ISABELA GASPARINI, Santa Catarina State University, Brazil
SEIJI ISOTANI, University of São Paulo, Brazil

Personalization of gamification is an alternative to overcome the shortcomings of the one-size-fits-all approach, but the few empirical studies analyzing its effects do not provide conclusive results. While many user and contextual information affect gamified experiences, prior personalized gamification research focused on a single user characteristic/dimension. Therefore, we hypothesize if a multidimensional approach for personalized gamification, considering multiple (user and contextual) information, can improve user motivation when compared to the traditional implementation of gamification. In this paper, we test that hypothesis through a mixed-methods sequential explanatory study. First, 26 participants completed two assessments using one of the two gamification designs and self-reported their motivations through the Situational Motivation Scale. Then, we conducted semi-structured interviews to understand learners' subjective experiences during these assessments. As result, the students using the personalized design were more motivated than those using the one-size-fits-all approach regarding intrinsic motivation and identified regulation. Furthermore, we found the personalized design featured game elements suitable to users' preferences, being perceived as motivating and need-supporting. Thus, informing i) practitioners on the use of a strategy for personalizing gamified educational systems that is likely to improve students' motivations, compared to OSFA gamification, and ii) researchers on the potential of multidimensional personalization to improve single-dimension strategies. For transparency, dataset and analysis procedures are available at https://osf.io/grzhp/.

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

Additional Key Words and Phrases: Gamification in education; gameful; tailoring; adaptation; self-determination theory

Authors' addresses: Luiz Rodrigues, lalrodrigues@usp.br, University of São Paulo, Av. Trab. São Carlense, 400, São Carlos, São Paulo, Brazil, 13566-590; Paula T. Palomino, paulatpalomino@usp.br, University of São Paulo, Av. Trab. São Carlense, 400, São Carlos, São Paulo, Brazil, 13566-590, University of Waterloo, Canada; Armando M. Toda, armando.toda@usp.br, University of São Paulo, Av. Trab. São Carlense, 400, São Carlos, São Paulo, Brazil, 13566-590; Ana C. T. Klock, actklock@inf.ufrgs.br, Federal University of Rio Grande do Sul, Brazil, Tampere University, Finland; Wilk Oliveira, wilk.oliveira@usp.br, University of São Paulo, Av. Trab. São Carlense, 400, São Carlos, São Paulo, Brazil, 13566-590, Tampere University, Finland; Anderson P. Avila-Santos, anderson.avila@usp.br, University of São Paulo, Av. Trab. São Carlense, 400, São Carlos, São Paulo, Brazil, 13566-590; Isabela Gasparini, isabela.gasparini@udesc.br, Santa Catarina State University, Brazil; Seiji Isotani, sisotani@icmc.usp.br, University of São Paulo, Av. Trab. São Carlense, 400, São Carlos, São Paulo, Brazil, 13566-590.

https://doi.org/10.1145/3474714

ACM Reference Format:
Luiz Rodrigues, Paula T. Palomino, Armando M. Toda, Ana C. T. Klock, Wilk Oliveira, Anderson P. Avila-Santos, Isabela Gasparini, and Seiji Isotani. 2021. Personalization Improves Gamification: Evidence from a Mixed-methods Study. *Proc. ACM Hum.-Comput. Interact.* 5, CHI PLAY, Article 287 (September 2021), 25 pages. https://doi.org/10.1145/3474714

1 INTRODUCTION

Motivation is important for positive learning experiences and academic success, since it predicts learning gains and enhances performance [27, 57]; while its lack is associated with lower performance and high drop-out rates [55, 78]. Moreover, students' lack of motivation is among the major issues faced by instructors [41, 53, 68]. Gamification can aid with that issue as it promotes game-like experiences aiming to improve users' motivations [18, 30]. Overall, gamification has a positive impact on psychological learning outcomes (e.g., intrinsic motivation), compared to non-gamified interventions [62]. However, there are cases in which it is associated with the opposed: undermined intrinsic motivation [71]. Many scholars attribute such cases to bad designs [44, 69], often arguing that providing the same game elements to everyone (i.e., the one-size-fits-all - OSFA - approach) is unlikely to make gamification work in its full potential since each user has different expectations, needs, and preferences [31, 49, 73]. Thus, the need for an approach to avoid the limitations of OSFA gamification, which might harm students' learning experiences while expecting to improve them.

Tailoring gamification might be the approach to mitigate OSFA gamification's drawbacks. It provides different game elements to different users/contexts through *customization* (i.e., users freely choose the game elements they prefer the most) or *personalization*[1] (i.e., the system or designer adapts the game elements according to each user/context) [72]. While there is some evidence customization improves OSFA gamification in terms of, for instance, user performance, customization requires the users/learners to indicate their game elements preferences before using the gamified system [42, 74]. Differently, personalization of gamification does not have such requirement because it is accomplished through predefined rules applied by the system/designer [72]. However, despite much research on personalization of gamification has been published, most empirical studies compare it to random or counter-tailored gamification [60]. Practitioners will not design gamified systems based on such approaches because of the uncertain effect of the former and the probable negative impact of the latter. Therefore, such comparisons do not provide evidence that advances our understanding of whether personalization is beneficial in practice. In contrast, a few studies compared personalized to standard, OSFA gamification empirically (e.g., [46, 48]), but those are unclear on whether the former improves the latter, especially in the educational domain.

A possible reason for the inconclusive findings is that such research (e.g., [46, 48]) applied personalization strategies that consider a single user dimension for the personalization. Whereas, a dual personalization approach resulted in positive motivational outcomes in [67], providing a valuable contribution on the potential of considering more than one personalization criterion. From such context, we expect multidimensional personalization[2] to approximate the positive effects of customization (e.g., [42, 66]) because users are not reduced to a single dimension; while avoiding the burden of asking their game element preferences [2, 54]. However, the only study (to our best knowledge) applying multidimensional personalization of gamification with users did not compare it to the OSFA approach [67]. Thus, the need for empirical research to advance our understanding of the potential of personalized gamification in educational contexts, compared to what is likely to be used in practice (OSFA rather than random or counter-tailored), along with the promising of multidimensional personalization. Therefore, **the goal of this paper was to**

[1] Or *static adaptation* according to Hallifax et al. [25]
[2] That is, personalizing to multiple dimensions (criteria).

understand the effect of gamification personalized to multiple dimensions, compared to the OSFA approach, on users' motivations in assessment learning tasks.

To achieve that goal, we conducted a mixed-methods sequential explanatory study [15]. Software Engineering students (N = 26) were randomly assigned into experimental (i.e., multidimensional personalized gamification) or control (i.e., OSFA gamification) groups to complete two classroom assessments, and feedback was collected through standard human-computer interaction (HCI) methods: the Situational Motivation Scale (SIMS) [22] and semi-structured interviews [6]. The findings indicate a positive effect of the multidimensional personalization of gamification on intrinsic motivation and identified regulation compared to OSFA gamification and suggest the personalized gamification designs featured game elements suitable to users' preferences, a possible reason for students perceiving it as motivating and need-supporting, and for the positive quantitative results. Therefore, contributing initial promising evidence on the effectiveness of a personalization approach that, compared to the standard, OSFA gamification, is perceived as more motivating by the students. Thus, our contribution informs: i) designers of educational systems on how to gamify them towards enhancing students' experiences, which will likely improve their learning due to its relationship with motivation and ii) researchers of gamification with indication that multidimensional personalization can improve the OSFA approach, expanding findings from [67] that are similar but compared to random instead of OSFA gamification, hence suggesting previous personalization strategies did not work compared to the OSFA approach because they considered a single personalization criterion.

2 RELATED WORKS

This section provides background information on gamification in education, and customization and personalization of gamification.

2.1 Gamification in Education

Education is the domain with the most gamification research [35]. Meta-analyses of gamification's effect within the educational domain show its effectiveness, compared to no gamification, while demonstrating that gamification might affect different types of learning outcomes (e.g., psychological and behavioral) and that its effect depends on several moderators [4, 62]. For instance, gamification has been used in education to improve students' grades, satisfaction, motivation, and lecture attendance, among other goals [33]. Accordingly, understanding how gamification works is important to enable a proper evaluation of whether it is working as expected.

While the Theory of Gamified Learning [37] has been considered suitable to understand gamification applied to educational scenarios [62], it does not account for psychological states, which play an essential role in gamification [35]. Differently, the Gamification Science framework [38] does, which is built around four main constructs: *predictors* (game elements); *criteria* (the distal outcome to be affected; e.g., students' knowledge retention); *mediators* (users' psychological states, e.g., intrinsic motivation; and behaviors, e.g., completing quizzes); and *moderators* (independent factors that might increase/decrease predictors' or mediators' effects) [38]. Based on those, the framework suggests the following:

- Predictors affect Psychological states;
- Psychological States affect both Behaviors and Distal Outcomes;
- Behaviors affect Distal Outcomes;
- Moderators change the effect of all previous connections.

Accordingly, regardless of the behavior/distal outcome targeted, gamification must first affect users' psychological state. This is aligned to claims that low performance (distal outcome) and drop-out rates (behavior) are related to students' lack of motivation (psychological state) [55, 78]. For instance, consider using gamification to improve class attendance and, consequently, students' grades. Failing to improve grades (distal) might be because gamification did not motivate (psychological state) them to attend class (behavior) or because they felt motivated but instruction was poor. Importantly, not considering the psychological state could lead to a misinterpretation that gamification failed, whilst it worked as expected and the problem was not related to it [38]. Such context shows the importance of motivation, and other psychological states, on gamification in education research as it is on the starting point of gamification functioning as well as among the most important factors for learning [27, 55, 57, 78]; a possible explanation for motivation being the most studied construct in gamification research [81]. Thus, considering motivation in gamification in education studies is pertinent because of both its value to learning and its role in preventing misinterpretations of whether it worked.

2.2 Customized Gamification

Customized gamification occurs when users are in charge of choosing the game elements the system presents [72]. Research on that approach has implemented it by allowing users to select the game elements they want from a predefined list (also referred to as *bottom-up* gamification) [74] or to create their gamification (i.e., participants writing their idea of a gamification design) [66]. Experimental studies comparing customized and OSFA gamification suggest positive effects of customization on behavioral outcomes (e.g., number of tasks solved) [42, 74]. While customization benefits have been discussed in terms of the freedom of choice it provides, it did not affect participants' feelings of autonomy, enjoyment, competence, and pressure [66]. Consequently, this raises the question of whether the performance improvements found are related to the customized gamification design or due to the effort participants put into creating them.

Customization's advantage likely emerges from providing individualized gamification designs. That is, all of someone's characteristics and preferences, as well as the context in which gamification will be used, are considered when selecting the game elements. On the other hand, the need for selecting game elements might become a burden. The effort for a one-time selection might be acceptable. However, discussions that gamification must be aligned to the task [25, 43, 56] suggest that, ideally, user selection would have to be done for each task. Hence, implementing customization might end up leading to substantial efforts from users, and designers/developers when users create their gamification.

2.3 Personalized Gamification

Personalization of gamification is when designers or the system itself chose game elements based on user information [72]. Predominantly, studies have personalized gamification by capturing user information and using it as an input to the gamified system, which provides a tailored design accordingly. For such approach, researchers often rely on preference-based recommendations. For instance, [73] and [24] show insights on the most suitable game elements depending on one's HEXAD and BRAINHEX types (i.e. motivational profiles), personality traits, gender, and age. Common to most recommendations is that the game elements are to be defined based on a single criterion [32], whereas there is empirical evidence demonstrating multiple factors affect users' preferences and, therefore, gamification success [19, 26, 31].

Furthermore, the effectiveness of personalization recommendations, when applied with users, are unclear [60]. Research comparing personalized gamification to random, counter-tailored, or no game elements are mostly positive (e.g., [1, 40, 67]). Similarly, some studies that compared personalized and OSFA gamification are also positive, while they failed to properly isolate the analyzed conditions (i.e., featuring both personalized learning and gamification for the experimental group [16, 28]). Nevertheless, most findings from comparing personalization and OSFA gamification are inconclusive. Mora et al. compared gamification personalized to students' HEXAD user types to the OSFA approach in the context of undergraduate learning. Their findings suggested a positive but statistically non-significant effect of personalization on psychological and behavioral outcomes [46]. Oliveira et al. also compared personalized and OSFA experimentally, tailoring the game elements according to participants' player types [48]. While findings were inconclusive as well, the negligible effect size suggested a null effect. In contrast, Hajarian et al. described positive (e.g., time-on-system and number of page views) outcomes when comparing personalized gamification to the OSFA design [23]. Besides the different context (i.e., social networks rather than educational), their personalization strategy was built from interaction preferences and implemented in the same system, whereas other studies [46, 48] implemented recommendations from surveys in real systems.

From that context, it is evident that there is a lack of empirical studies comparing personalized and OSFA gamification. Nevertheless, those studies have different characteristics that can be used as insights to improve personalization strategies. First, they differ in context (i.e., learning activities [46, 48] and social networks [23]), which might be a reason for the contradictory findings. Second, personalizing based on interaction instead of survey preferences might be another one [2, 54]. Consequently, Hajarian et al. [23] personalized based on data from the same context and activity, while others [46, 48] implemented recommendations from general preferences. Another point is that interaction data capture users' behaviors based on several of their dimensions, if not all, similar to what customization enables. The other empirical studies, however, personalized to a single dimension [46, 48]. Therefore, raising the question of whether personalizing gamification to multiple dimensions of a specific context would improve gamification, contrary to approaches based on a single dimension.

To personalize gamification to multiple dimensions, one needs recommendations that consider two or more information simultaneously, which are scarce [32]. Oyibo and colleagues analyzed how users' culture, age, and gender affect their persuasiveness to social influence's constructs (e.g., rewards and competition) in a series of studies [50–52]. In common, their recommendations do not consider any situational dimension. Bovermann and Bastiens suggest the most suitable game elements considering users' HEXAD user type and the learning activity [7]. However, their recommendations are limited to common Moodle activities. Differently, Baldeon et al. provide recommendations to personalize gamification based on users' characteristics (i.e., HEXAD user type, personality traits, and learning style) and the teaching activity (e.g., brainstorm and group discussion) [5]. Rodrigues et al. provide a recommender system that suggests game elements based on users' information (e.g., gender, preferred game genre and playing style, weekly playing time, and geographic location and experience with gamification), as well as the cognitive process to be worked in the learning activity [59]. A limitation common to all of those is that neither has been applied with users to evaluate its effects compared to the OSFA approach. Hence, there are some limited options to aid in a multidimensional personalization of gamification considering user and contextual/situational information. Mainly, they vary in dimensionality (i.e., the number of criteria considered) and generalization of the situational factor (i.e., whether it guides on predefined [5, 7] or any [59] task).

2.4 Summary and Hypothesis

Despite empirical evidence suggest customized gamification's benefits, it might require unfeasible efforts from users and designers/developers to be implemented. Differently, personalized gamification takes that burden away from users but demand predefined strategies on how to tailor gamification to them. Despite much research on personalized gamification, most compare it to random and/or counter-tailored approaches and few compare it to the OSFA [60]. That is problematic because

Table 1. Comparison between this study and main related work (either compared to OSFA gamification or applied multidimensional personalization). For a complete comparison of empirical research on personalized gamification, see [25, 59].

Work	Compared to OSFA gamification?	Multidimensional personalization?
[46]	Yes	No
[48]	Yes	No
[67]	No	Yes
This one	Yes	Yes

to contribute to understanding personalization's effects we need evidence from comparing it to standard gamification designs. However, there is inconclusive evidence on whether personalization leads to improvements compared to the OSFA approach from the few studies available [25, 60]. Perhaps, the fact that prior research in the educational domain [46, 48] personalized gamification to a single dimension, when comparing to the OSFA approach, might be the reason for the inconclusive findings. In that line, gamification personalized to two dimensions resulted in positive results when applied with learners, but compared to random, not OSFA gamification [67]. Compared to those, our study mainly differs by evaluating the effects of gamification personalized to multiple dimensions compared to the OSFA approach (see Table 1). Therefore, we focused on personalization aiming to achieve results similar to those of customization while preventing the burden on users and developers/designers. In doing so, we implement the recommendations from Rodrigues et al. [59] because it has the highest dimensionality (i.e., eight factors/criteria), the more generic contextual factor (i.e., task's cognitive process, which is independent of the system, e.g., Moodle, and teaching activity, e.g., discussion), and does not require users completing long questionnaires (e.g., to assess personality traits or learning style). Thus, testing the hypothesis that **a multidimensional personalization of gamification, considering user and contextual information, improves the one-size-fits-all approach in terms of learners' motivations.**

3 METHOD

The goal of this study was to understand the effect of gamification personalized to multiple dimensions, compared to the OSFA approach, on users' motivations in assessment learning tasks. Therefore, we conducted a mixed-methods sequential explanatory study [15]. In the first phase, we compared OSFA and personalized gamification through a 2x2 mixed factorial experiment. We manipulated *gamification design* (between-subject) to create two versions of the system where students would complete the assessments. Those versions featured either an OSFA or the personalized gamification design. Participants engaged in two sessions that differed by the assessment *discipline* (within-subject): Programming Techniques and Object-Oriented Analysis and Design. Thus, we were able to compare the gamification designs based on two applications. In the second phase, we conducted semi-structured interviews to understand participants' motivations to use and engage with the gamified system. Figure 1 presents an overview of the study, which was reviewed by the institution's ethical board. It shows a flow chart demonstrating the steps followed during the preparation and execution of the experiment, indicating the time interval in which they happened as well as descriptions of particular aspects of those steps.

3.1 Context and Participants

Sampling was made by convenience according to the willingness of the instructor to apply gameful interventions in their lessons. Accordingly, this study was conducted with second-period students

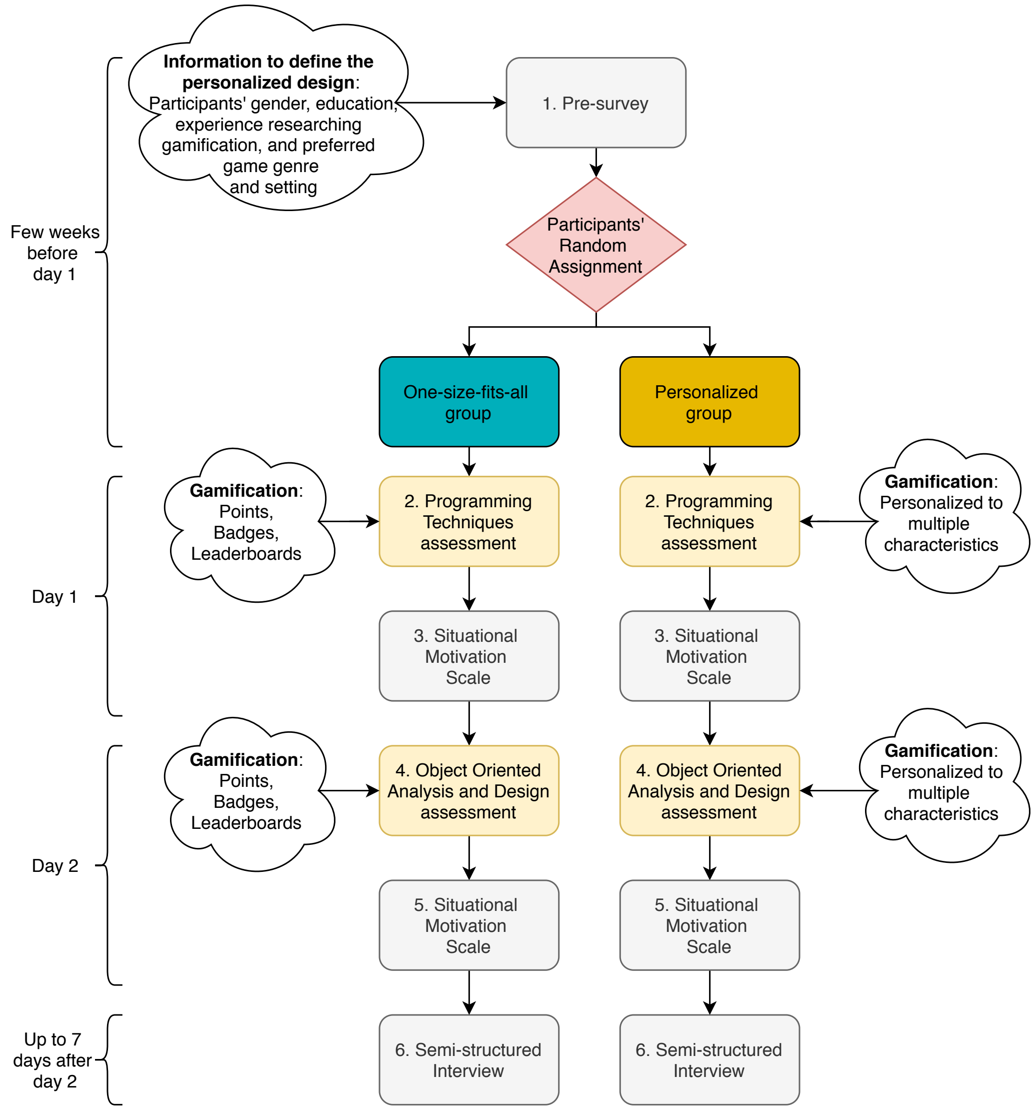

Fig. 1. Study methodological overview. The figure shows each step of the experiment, along with particular characteristics of some of those steps as well as the time interval in which they happened.

of the undergraduate Software Engineering course of SENAI Londrina, a small, private institution from Brazil. Despite being conducted in a natural context (i.e., the assessment activities were part of the disciplines' programs), participation in this experiment was voluntary. Twenty-six out of 27 learners agreed to participate in this experiment (all men, with an average - M - of 21.92 years and a standard deviation - SD - of 3.77), a sample size similar to those seen in overall HCI studies [11] and CHI papers [10]. Twenty-three subjects attended the Programming Techniques discipline during the first day, and 22 attended the Objected-Oriented Analysis and Design discipline during the second day. Of all participants, 4 agreed to participate in the interviews (Age - M: 23.75; SD: 4.65; all had been randomly assigned to the personalized gamification design). The same instructor (male, 31 years old, holds an MsC. in Computer Science, 6 years of teaching experience) taught both disciplines. All participants provided written consent through an online form since classes were held online due to the COVID-19 pandemic.

3.2 Gamified Assessments

We worked with two classroom assessments, composed of 30 multiple-choice items each. Specifically, each assessment can be seen as a test composed of 30 multiple-choice items. Notice that there are two forms of assessment, this experiment relies on assessment *for* learning, as opposed to assessment *of* learning [34]. Because students completed the assessments towards the end of the semester and before the final exam (which is an assessment of learning), this study's assessments aimed to make students' remember concepts introduced during the semester to complete its items. Hence, as aligned with the instructor, this would help the students to recall important information about the course. Accordingly, learners had to work on the remembering dimension of the cognitive process, according to Bloom's taxonomy [36], which was further used as one of the personalization criteria of the gamification designs (see Section 3.3).

To ensure the assessment was exclusively within that dimension, all items were developed by the class instructor and one researcher, while being validated by an independent instructor. For instance, one item reads as follows: *If 'p' is a pointer, the 'p' and '*p' commands access, respectively: a) a memory address and the value stored in the address; b) a value and the memory address where the value is stored; c) a memory address and the name of the variable to which 'p' is pointing; d) a value and the name of the variable that 'p' is pointing to.* The full version of both original (i.e., Brazilian Portuguese) and English versions of these assessments are available as supplementary material.

These assessments were implemented in the gamified system Eagle-Edu[3] since our participants had experience with it because other instructors used the system in their classes. Eagle-Edu was also suitable for this study because it allows creating multiple-choice activities, such as the ones from the proposed assessments. Additionally, this system enables instructors to choose which game elements (e.g., objectives, progress, storytelling, points, badges, leaderboards, and time pressure) they want to enable or disable, assisting in creating multiple gamification designs. In practice, each assessment translated into 10 quizzes of 3-multiple-choice items, providing rewards when the learner completed each quiz. Table 2 describes Eagle-Edu's implementation of the game elements used in this study according expert-validated definitions [70]. Next, we explain the rationale for selecting game elements.

3.3 Experimental Conditions

When designing the gamification versions (OSFA and personalized), we ensured they featured the same number of game elements. We made this choice to increase the study's internal validity based on discussions that the number of game elements might affect gamification's success [39]. In the **OSFA condition**, we used Points, Badges, and Leaderboards (PBL). Although PBL have been criticized, a recent meta-analysis indicates that this combination has positive effects in gamified educational settings [4], while being among the most used ones [21, 68, 77, 81]. Therefore, we considered it a suitable baseline to represent an OSFA gamification approach.

For the **personalized condition**, we relied on recommendations for a multidimensional personalization of gamification [59]. Specifically, the recommendations are given by three decision trees that receive user and contextual information as input and output ratings for different game elements, with each tree indicating one game element. Together, they recommend the top-three most suitable game elements to a user performing some type of learning activity according to the highest rating of each tree. While each tree outputs an independent recommendation to form the top-three, they all receive the same information about the user - preferred game genre (Action, Adventure, RPG, Strategy or Other), preferred playing setting (Multiplayer or Singleplayer), weekly

[3] eagle-edu.com.br/

Table 2. Description of Eagle-Edu's implementation of the game elements used in this study, along with the indication of which of them were available in each gamification design. For further description of gamification design definitions, see Section 3.3

Game element	Description	OSFA	P1	P2
Acknowledgment	Badges for achieving predefined goals (e.g., completing a quiz with no error).	X	X	
Chance	Surprise item, such as virtual goods, that appears due to randomness and the user might collect.			X
Competition	Leaderboard showing users sorted by the points they have earned; highlights the first and last two users.	X		X
Objective	Skills tree demonstrating course's topics/skills to guide users towards short-term goals (i.e., completing each topic at a time)		X	
Points	Numerical feedback for achieving predefined goals (e.g., completing 10 quizzes).	X		
Progression	Progress bar indicating one's progress within quizzes and on topics/skills.		X	
Time Pressure	Decreasing timer showing the time left to improve in the leaderboard.			X

OSFA = One-size-fits-all condition; PN = Version N of the personalized gamification condition

playing time (in hours), gender (Female, Male, Other), educational background (High School, Technical Education, Undergraduate, MsC, Ph.D., Other), and experience researching gamification (Yes or No) - and about the context in which the gamification will be used: country and learning activity type (remembering, understanding, applying, analyzing, evaluating, creating).

With such information, we analyzed the trees to define the personalized gamification design for each participant of the personalized condition. Note that some of the trees' inputs were fixed due to the characteristics of this experiment: the assessments exclusively focused on remembering activities and all participants were Brazilians with no experience in researching gamification. Then, when analyzing the trees with those fixed inputs, we found that the factor affecting the recommendations was a participant's preferred game genre, with the other factors only impacting the recommendations for people from other countries, with experience researching gamification or when completing other learning activity types. Therefore, we derived a simplified personalization algorithm (Algorithm 1) from such analyses, which is shown next and was executed to define the personalized gamification design of each participant of the personalized condition.

Lastly, we call attention to the following aspects. First, in cases where the same element was recommended by different decision trees, we selected the following most suitable one to ensure all gamification designs had three different game elements (Algorithm 1 states this situation). Second, because the decision trees have dozens of nodes and leafs, we do not present them here due to the lack of space to present them in detail and because a partial presentation could mislead their recommendations. Instead, we provide them as supplementary materials and refer to [59] for details. Finally, we highlight one can see images of both system versions in the supplementary material.

3.4 Instruments

For the personalization of gamification, participants completed a pre-survey, which captured information to design the personalized version (i.e., preferred game genre and playing setting,

ALGORITHM 1: Personalization algorithm derived from [59] given that for all participants: country is Brazil; experience researching gamification is No; and learning activity type is Remembering.

Data: preferred game genre
Result: List of game elements forming the personalized gamification design
```
/* Some elements might differ from the first recommendation from [59] to
   avoid repeated elements and ensure there are three game elements in all
   gamification designs                                                  */
if preferred game genre = Action then
    /* The second element would originally be Competition. It was replaced
       by the second option - Chance - to comply with the need for three
       game elements in all conditions                                   */
    return [Competition, Chance, Time Pressure];
else
    /* The third element would originally be Objective. It was replaced by
       the second option - Progression - to comply with the need for three
       game elements in all conditions                                   */
    return [Acknowledgment, Objective, Progression];
end
```

weekly playing time, gender, previous experience researching gamification, education background). For the quantitative data collection, we used the SIMS [22] since it captures participant's motivations to engage with an activity based on four constructs (i.e., intrinsic motivation, identified and external regulations, and amotivation) aligned to the self-determination theory [61]. Furthermore, the SIMS has been empirically validated in the educational context and in several languages, including that of this study's participants [20]. For our data, reliability (i.e. consistency among items of the same construct) was good for intrinsic motivation (0.94), identified regulation (0.82), and amotivation (0.90), but questionable for external regulation (0.51), according to Cronbach's alpha. Additionally, note that despite suitable, we did not use assessments' outcomes as a dependent variable because previous knowledge would play a major role as covariate and because we did not plan ahead of the experiment execution to examine such variable, consequently, we did not capture any measure of previous knowledge. For the qualitative data collection, we conducted semi-structured interviews [6], wherein the interviewer relied on the following questions as the initial source of discussions:

(1) Could you introduce yourself and tell me a little bit about you and your hobbies?
(2) What do you think about games, either digital or analogical?
(3) Overall, what do you think about going to college?
(4) What do you think about the disciplines' assessment activities?
(5) What did you think about the assessment activities you made last week?
    (a) How would you compare those activities to other assessment activities you did before?
(6) What did you think about doing the assessment activity in a system unlike those you regularly use?
    (a) How would you compare that system to others used in classes?
    (b) Would you compare the experience of doing the assessment activities in that system to what other experience?
    (c) After you began the activity, what were your reasons to keep doing it?
(7) What did you think about the system you made the assessment activities?

(a) Did you note elements uncommon to other systems that might be used to perform similar activities?
(b) Overall, would you compare that system to another one?
(c) What did you think about the game elements available in the system?
(d) How would you compare the experience of doing the assessment activity with and without game elements?
(e) What changes do you suggest?

Generally, the goal was to understand participants' subjective experiences while performing the assessment activity in terms of how the gamified system affected their motivations. We sought to elicit answers about those aspects especially with items 6-7. However, we began the interview with items 1-5 to capture an overview of who the participants were and their view and overall motivations to go to college and perform assessment activities. The item's generality (e.g., what do you think about) was to reduce bias, leaving for the participant the option of mentioning specific aspects (e.g., *I liked the competition*) early in the interview if relevant to them. This often happened, avoiding the need to explicitly ask subsequent questions (e.g., item 7.c). Interviews (M: 45 minutes) were all conducted by the same person (male, Ph.D. student, 26 years old) using Google Meet. The codebook and quotes supporting the qualitative results are available in the supplementary material. Interviews' transcriptions are not available due to sensitive information.

3.5 Procedure

The study was conducted on consecutive days of the last week of the semester, following six steps. First, learners completed the pre-survey a few weeks before the experiment. Second, they completed the assessment activity of the Programming Techniques discipline during class-time[4] (around one hour). Third, students completed the SIMS right after finishing the first assessment. Fourth, on the second day of execution, participants completed the second assessment activity (Object-Oriented Analysis and Design discipline), also lasting around one hour of class-time. Fifth, they completed the SIMS again, right after the fourth step. The sixth step was participating in a semi-structured interview to talk about their experiences with the gamified system during the assessment activities. Despite all learners were invited, only four of them completed this last step.

3.6 Data Analysis

For the quantitative analysis, given the experimental design (2x2 mixed factorial) and the study goal, Mixed ANOVAs were applied. As the first recommended step of the data analysis [79], we tested the assumption of residuals' (normal) distribution. Since there were violations, we analyzed our data with robust methods (i.e., 20% trimming) [11, 80]. We conducted the robust Mixed ANOVAs and main effect analyses (effect of one independent variable) with the *WRS2* R package [45]. Additionally, we calculated effect sizes through the explanatory measure of Effect Size (ES), which is a robust location measure that handles non-normal data as well as groups with unequal variances; values of 0.1, 0.3, and 0.5 correspond to small, moderate, and large effects, as suggested by the package authors [45]. We executed that process for each motivation regulation type, considering a 0.05 alpha level. As recommended [3, 79], we do not correct p-values because each test concerns a different planned comparison (i.e., a different motivation/regulation), similar to prior research [1, 74]. The dataset and data analysis procedure are available in the supplementary material.

[4]Due to the COVID-19 pandemic, classes were online - streamed through a synchronous meeting service - despite the course was originally face-to-face.

The qualitative analysis aimed to understand learners' experiences in terms of how the gamification influenced their motivations compared to their motivations to perform traditional, non-gamified assessment activities; and, consequently, further explain the quantitative results. Thereby, we adopted the *discourse analysis* analytical framework [14] to enable the understanding of implicit and hidden meanings in participants' answers. We performed a thematic analysis on the semi-structured interviews' transcriptions. Following the thematic analysis procedure [9], one author got familiarized with the data and generated initial codes. Together with other two authors, they searched, reviewed, defined, and named themes to produce a report. During these steps, we adopted an *interpretivist semi-structured approach*[5] that, besides being aligned with our data collection method (i.e., semi-structured interviews), is one of the qualitative approaches most common in HCI research [6].

For coding, we adopted a mixture of inductive and deductive schemes. We used inductive coding especially in the first iterations of the analysis to understand learners' subject experiences. Although rare in HCI research [6], we used deductive coding in the latter steps to relate low-level codes and themes to motivation- and learning-related theories (e.g., Self-determination Theory, SDT). Hence, the codebook was developed throughout the analysis to allow the identification of emergent themes from interviewees' subjective experiences. To support results' validity, we quote participants' answers, whereas we rely on triangulation to support reliability because the multiple independent coders approach is inappropriate to validate rich subjective analyses [6]. Therefore, we employ methodological and theoretical triangulation. That is, relating qualitative to quantitative data and interpreting qualitative results in terms of multiple theoretical lenses, respectively.

4 RESULTS

Table 3 shows descriptive statistics of the quantitative results. It demonstrates the average and standard deviation, as well as the number of participants (N), for each motivation/regulation in each discipline and overall. Figure 2 presents further descriptive information through boxplots, comparing conditions in both disciplines for each motivation/regulation. Table 4 introduces the results of the statistical tests. Those reveal the main effect of design was significant for intrinsic motivation and identified regulation but nonsignificant for external regulation and amotivation. Differently, the effect of discipline, as well as that of its interaction with design, were nonsignificant for all motivations/regulations. We only conducted further analyses for the significant differences, which reveal large significant effects of the design factor for both intrinsic motivation, $F(1, 21.04) = 8.491$; $p = 0.00829$; ES = 0.64; ES 95% Confidence Interval (CI) [Lower CI; Upper CI] = [0.22; 0.98], and identified regulation, $F(1, 21.92) = 8.4093$; $p = 0.00833$; ES = 0.62; ES 95% CI = [0.21; 0.94]. Figure 3 shows boxplots comparing both design conditions, in terms of intrinsic motivation and identified regulation, regardless of the disciplines.

The qualitative results revealed two main themes, which concern learners' subjective experiences with either traditional (non-gamified) or the experiment's assessment activities. (See Tables 5 and 6 for quotes supporting all tags.) Concerning a *common assessment activity*, interviewees' perceived them as (subthemes): *pressing*, *unsatisfactory*, and *satisfactory*. Interviewees consider those activities as pressing because they might lead to bad results, one needs to perform them to actually learn, and they require significant preparation time. Also, the interviewees showed dissatisfaction because such activities are not the best evaluation method and obligatory. Differently, they felt satisfaction in terms of improving knowledge, having the chance for self-assessment, and gaining grades.

[5]Interpretivist refers to assuming a subjective view of reality; semi-structured concerns the fact themes will be covered to different extents, depending on the best line of inquiry [6].

Table 3. Descriptive statistics about participants' motivations overall or in either the Programming Techniques (PT) or the Object-Oriented Analysis and Design (OOAD) disciplines. Data are shown as Mean (Standard Deviation), which were collected in a seven-point Likert scale.

			Motivation/Regulation			
	Design	N	Intrinsic	Identified	External	Amotivation
Overall	One-size-fits-all	25	4.20 (1.40)	4.74 (1.23)	**4.65 (1.15)**	**3.35 (1.65)**
Overall	Personalized	20	**5.39 (1.46)**	**5.79 (1.16)**	4.49 (1.26)	2.65 (1.66)
PT	One-size-fits-all	12	4.48 (1.21)	5.00 (1.15)	4.56 (1.23)	**3.08 (1.57)**
PT	Personalized	11	**5.66 (1.06)**	**6.14 (0.68)**	**4.59 (1.43)**	2.64 (1.48)
OOAD	One-size-fits-all	13	3.94 (1.55)	4.50 (1.31)	**4.73 (1.12)**	**3.60 (1.74)**
OOAD	Personalized	09	**5.06 (1.86)**	**5.36 (1.50)**	4.36 (1.08)	2.67 (1.95)

N = Number of participants/responses.
The larger values, when comparing gamification designs, are **bolded**.

Table 4. Results of the robust two-way ANOVAs for different Intrinsic Motivation (IM), Identified Regulation (IR), External Regulation (ER), and Amotivation (AM) as Dependent Variables (DV) and gamification design (one-size-fits-all or personalized) and discipline (Programming Techniques and Object-Oriented Analysis and Design) as factors.

	Factor					
	Design		Discipline		Design:Discipline	
DV	F(df1, df2)	P-value	F(df1, df2)	P-value	F(df1, df2)	P-value
IM	5.2533(1, 11.3193)	**0.0420**	0.2657(1, 9.8951)	0.6175	0.0160(1, 11.3193)	0.9015
IR	6.3288(1, 9.7345)	**0.0312**	0.9237(1, 8.3351)	0.3635	0.0887(1, 9.7345)	0.7721
ER	0.0302(1, 11.5548)	0.8650	0.2175(1, 12.8690)	0.6487	0.5773(1, 11.5548)	0.4626
AM	1.8524(1, 12.7829)	0.1970	0.0217(1, 12.8194)	0.8852	0.2223(1, 12.7829)	0.6453

Significant p-values (< 0.05) in **bold**.

Concerning the *experiment's assessment activities*, interviewees' perceive them as (subthemes) *need-supporting*, *enjoyable*, and *unsatisfactory* at some extent, while also noting its *gamification*[6]. The need-supporting perception comes from interviewees feeling autonomous, competent, and related. The enjoyable subtheme emerged because the activity was something free of risk, relaxing, and novel. Also, the activity allowed self-assessment. On the other hand, interviewees felt dissatisfaction regarding some system bugs. Furthermore, they noted the gamification, which they considered better than not having it and suitable to their preferences while noting some missing game elements. Additionally, interviewees perceived the gamification as motivating because it made them want to perform well and motivated social comparison, as well as supported basic psychological needs. Figure 4 shows the thematic map linking themes and subthemes, with each subtheme's tag shown within its box and different colors for distinct kinds of experiences.

[6]We named the subtheme as *personalized gamification* because all interviewees were from the personalized condition.

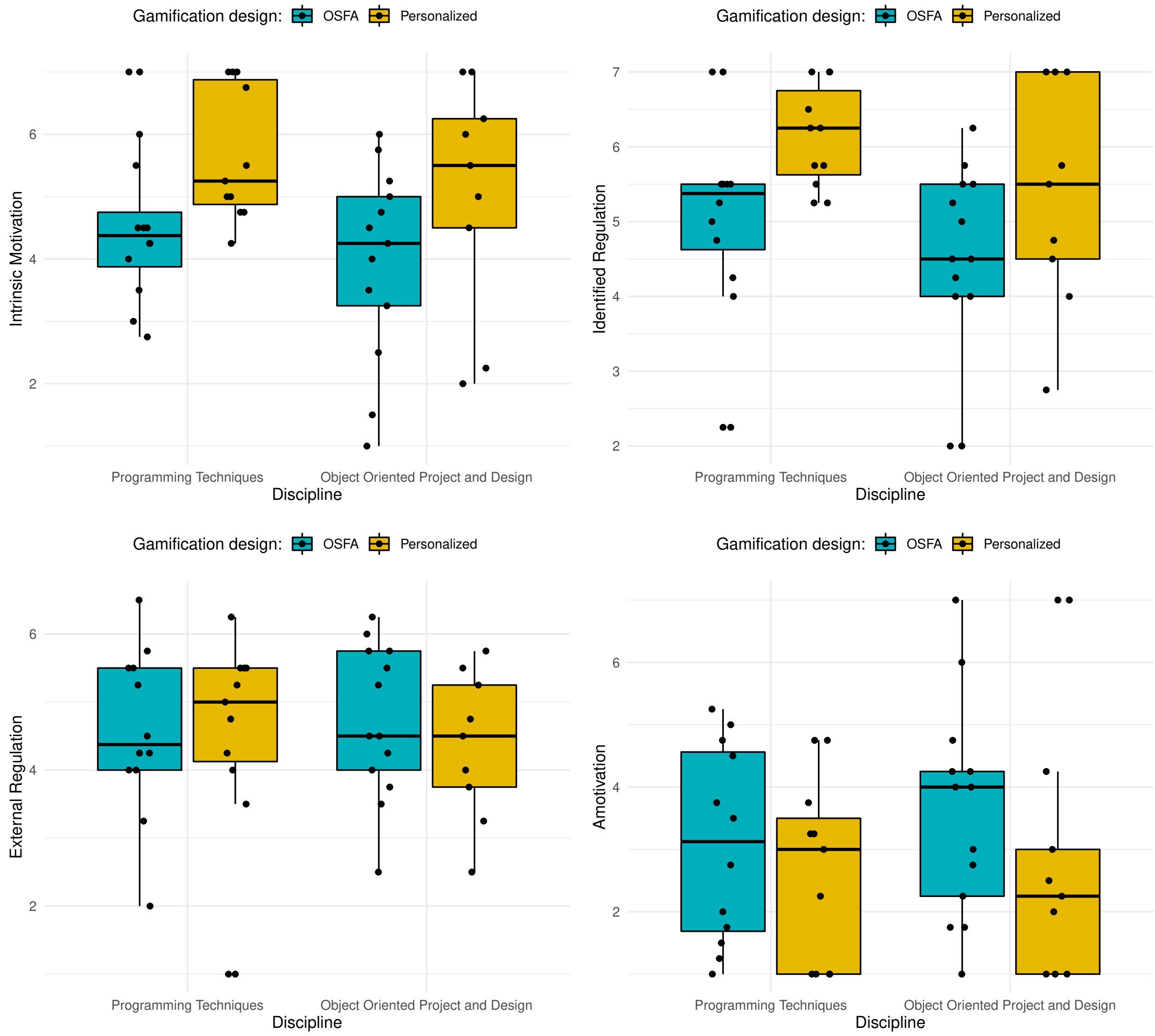

Fig. 2. Boxtplots comparing participants' motivations in both conditions (one-size-fits-all, OSFA, and personalized) for each discipline in which the experiment was executed. Intrinsic motivation, identified regulation, external regulation, and amotivation are shown on top-left, top-right, bottom-left, and bottom-right, respectively.

5 DISCUSSION

This section discusses our results, reasons for why they were positive, why they differ from previous research, findings' implications, and study limitations.

5.1 Overall Findings

Our quantitative findings show a large, significant difference in the perception of participants who used the personalized gamification design, compared to reports of those who used the OSFA gamification, in terms of intrinsic motivation and identified regulation. This significant difference supports the hypothesis that offering users a gamification design personalized to multiple characteristics (e.g., information of both users and task at hand - learning activity) is more effective than providing a single, general design for all. The magnitude of that difference might be interpreted from two perspectives. First, gamification applied to education is associated with overall small

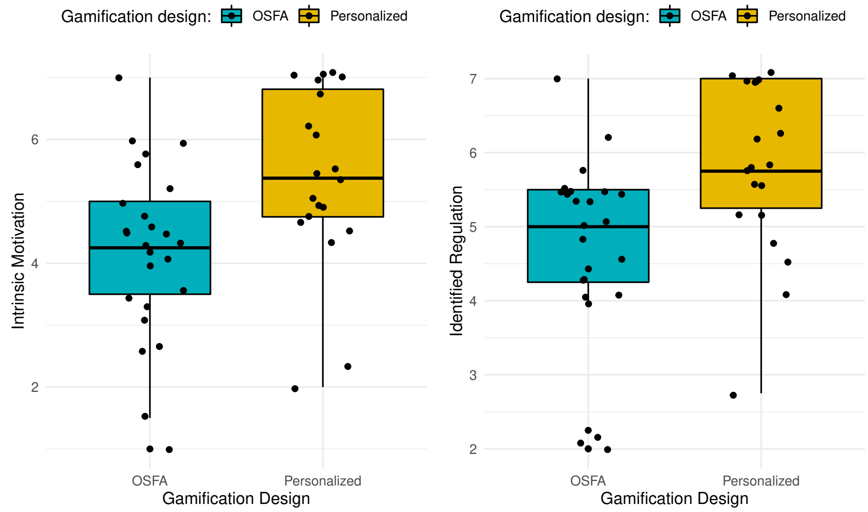

Fig. 3. Boxplots' comparing participants' overall intrinsic motivation (left) and identified regulation (right) among experimental conditions (i.e., one-size-fits-all, OSFA, and personalized).

Table 5. Quotes supporting the *Common assessment activity* theme found in the thematic analysis of semi-structured interviews. I*N* refers to Interviewee N (1-4).

Tag	Subtheme	Quote
Demands significant preparing	Pressing	*Because sometimes there is a test that has a lot of burden to study for that specific subject*, I1
Necessary to actually learn	Pressing	*I do [assessment activities] mainly because I have to learn*, I4
Might lead to bad results	Pressing	*So I want to do well [in the assessment] to prove that I really absorbed that [knowledge]*, I3
Is not the best method	Unsatisfactory	*... sometimes it doesn't test all of your content retention and everything*, I1
Obligatory	Unsatisfactory	*It is an obligation. Total obligation. In the real sense of the word, is to do it because otherwise you will not pass*, I1
Improve knowledge	Satisfactory	*But generally, when there is a test I feel good because I try to study hard and dedicate myself to doing well in the test, and also learning the content*, I2
Self-assessment	Satisfactory	*I feel good doing the activities..even because I have to do it to be graded, but I feel good doing because I will be able to follow my evolution*, I2
Gain grade	Satisfactory	*Assessment activity when you think, you will think about the return, right, the grade that you can take and take advantage of it*, I3

effects on motivational learning outcomes when compared to no gamification [62]. Second, the hinge-point for effect sizes of general educational interventions is a *moderate* effect (Cohen's d > 0.4 [13]) according to [29]. Our findings suggest a large effect size above the average in both

Table 6. Quotes supporting *Experiment's assessment activity* theme found in the thematic analysis of semi-structured interviews. I*N* refers to Interviewee N (1-4).

Tag	Subtheme	Quote
Allows self-assessment	Satisfactory	*[...] I could see that, for example, there was something I didn't learn well*, I2
Autonomy	Need-supporting	*I felt it was a little [...] out of class [...] because I think it was not an obligation*, I3
Competence	Need-supporting	*I think doing the activity was much more satisfying [...] because I have real-time feedback*, I3
Relatedness	Need-supporting	*[...] the little rank is just to make fun of each other when doing the activity*, I1
Bugs	Unsatisfactory	*Then an error goes to the screen and it broke.*, I4
Risk-free	Enjoyable	*[...] when you take it and even when you miss a question like that, [...] it feels lighter, I would tell you*, I1
Relaxing	Enjoyable	*[the activity was like] a form of relaxation and occupying part of a gap between something at work and another*, I1
Novel	Enjoyable	*So I think that compared to other activities of the day, I think it was something different [...] maybe because I don't do it often, I think it was a really cool thing to do ....*, I3
Suitable game elements	Personalized gamification	*If I were to talk about the game aspect, this [competition - the weekly leaderboard] would be the type of game that I like, interestingly*, I3
Better than no gamification	Personalized gamification	*with game element is much better*, I2; *[with gamification] It didn't get boring, right? you don't get tired*, I4
Missing game elements	Personalized gamification	*Eagle would be good if you have a progress bar*, I2; *It would be nice to show the medal maybe on the profile or display next to the name*, I3
To perform well	Motivating	*when I was there doing a question, I was very careful 'damn it, I have to get this one right [...]*, I1
To make social comparisons	Motivating	*I can compete with that ranking, with the people who also performed the activity*, I3

perspectives. Therefore, indicating the multidimensional personalization of gamification employed in this study represented an improvement over the OSFA approach that is above both the common gamification effects on motivation as well as the threshold for educational interventions.

Additionally, we found that completing the task with personalized gamification mitigated negative perceptions of common assessment activities while motivating and supporting the learners. From semi-structured interviews with participants that used the personalized gamification design, we found they acknowledge assessment activities are valuable to the learning process but perceived them as pressing and unsatisfactory. The effectiveness of the personalized gamification is suggested because participants considered the game elements available to them suitable to their preferences, as well as mentioned the personalized gamification supported basic psychological needs. Also, when using the personalized gamification, participants felt the activity was less pressing and more relaxing, compared to common assessments (see Tables 5 and 6), possibly because the gamification

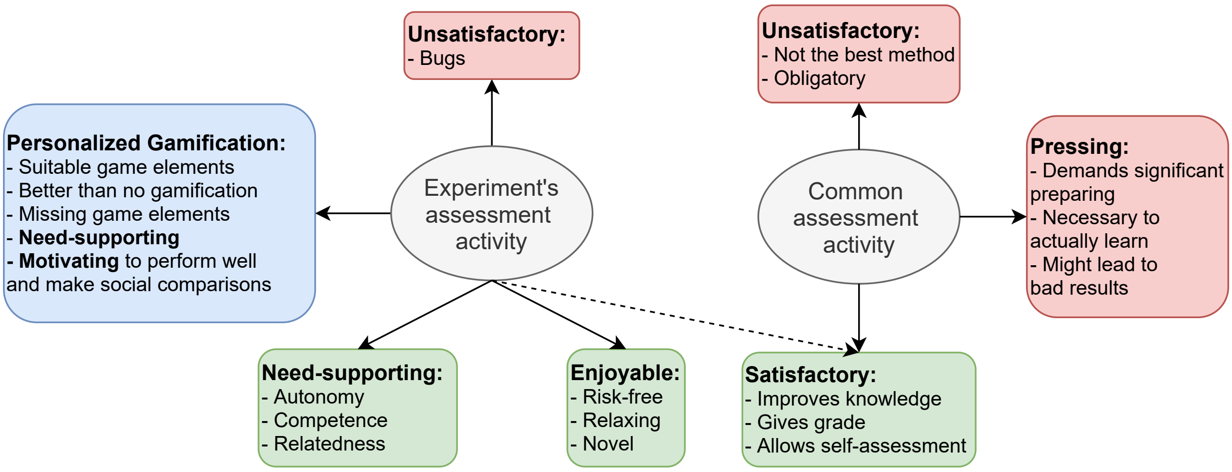

Fig. 4. Thematic map from the analysis of the semi-structured interviews showing learners' subjective experiences with assessment activities. Ovals represent the main themes, rectangles represent the subthemes, bolded text indicates subthemes' names, and regular text indicates tags. Red and green indicate negative and positive experiences, respectively. Solid connections from X to Y means X is/has Y and all of its tags; the dashed line means the connection is partial.

mitigated some of those negative feelings (e.g., of pressure), supporting basic psychological needs. Furthermore, the personalized gamification motivated them to perform well in the assessments as well as to make social comparisons, indicating its educational value.

Summarizing those findings, we have the following takeaways:

- Multidimensional personalization improved students' experiences with the gamified educational system, compared to the OSFA approach, in a degree above the hinge-point for educational interventions and OSFA gamification when compared to no gamification;
- The personalized design overcame the OSFA approach, in terms of autonomous motivation, by providing game elements suitable to learners' preferences that supported their needs and mitigated drawbacks from regular assessment activities.

5.2 Why Personalization Worked?

We discuss our findings in terms of why learners using the personalized gamification felt more motivated than those using the OSFA design from two perspectives. First, *preference satisfaction*. Previous research has advocated that one size does not fit all [47]. For instance, there is evidence showing different users have distinct preferences [24, 49, 73] and that gamification's effect varies from user to user [58, 75]. Hence, supporting the argument regards the suitability of the game elements offered to them. Second, *psychological needs support*. That is, offering users the right game elements supported their basic psychological needs. Consequently, improving their autonomous motivation compared to participants that received the OSFA design. Accordingly, the non-significant effect on participants' external regulation might be because the game elements were suitable to their preferences, instead of being perceived as external drivers. Note that while the identified regulation belongs to the extrinsic perspective, it is the closest regulation to intrinsic motivation, which together are seen as autonomous motivation [61, 76]. In contrast, the non-significant effect on amotivation might be attributed to the importance of the experimental assessment activities, which probably led to low amotivation levels for all.

5.3 Why Previous Personalization Strategies did not Work?

Regarding why previous empirical studies comparing personalized and OSFA designs failed to provide conclusive results, unlike this study, we discuss three perspectives.

First, the degree of personalization. While most prior research used a single personalization criterion [46, 48], the literature demonstrates that users' motivations/experiences differ in many factors (e.g., user types [24, 73], task familiarity/knowledge [59, 63], and gender [31], among others - see [32, 60] for reviews). Accordingly, a study that applied a dual personalization suggests it overcame the single dimension personalization - with random gamification as the baseline [67]. Similarly, this study employed a multidimensional approach but considered eight values as input for the personalization of gamification. Hence, the multidimensional strategy might be the explanation for this study finding a large, positive effect of personalization of gamification when compared to the OSFA approach, unlike prior research. Note that despite Mora et al. discussed their non-significant results might be a product of low statistical power [46], their sample size was larger than the one of this study, although more unbalanced. Then, if they had found an effect of similar magnitude, statistical significance would likely be found as well.

Second, considering the context. Whereas research has highlighted contextual information affects gamification's success [24], arguing factors such as the task to be done should be considered by personalization strategies [25, 56], most of the related research only focused on user information [46, 48]. The somewhat exception is Hajarian et al. that, differently, found that the personalized gamification overcame the OSFA design [23]. In that study, the personalization is driven by user information as well (their likes about the game elements), but such preferences were captured in one context, doing one activity (i.e., using a social network to arrange dates) and, then, the authors used those insights to personalize gamification for the same context and task. In our study, we used a personalization strategy focused on the educational context [59], unlike those used by research reporting inconclusive findings. Additionally, guidance from that strategy considers the type of the task/activity to be done, which was a remembering learning task in the case of this study. Therefore, it might be that considering contextual information (the domain and the task in our case) is the reason for this study and that of Hajarian et al. [23] finding positive effects from the personalization of gamification.

Third, approximating to what customization provides. The indications that customization of gamification is effective when compared to the OSFA approach (e.g., [42, 66]) support the rationales discussed previously, at least to some extent. That is because multiple information from the user and the context are considered, simultaneously, while one is customizing their gamification design. For instance, one's user type, gender, gaming preferences and habits, etc., as well as the domain and task, are taken into accounting when they are selecting the game elements they want to be available. Consequently, the more information the personalization strategy receives, the more its outcomes are expected to approximate to those of customization; similarly for considering contextual factors. Based on that relation, the lack of considering multiple factors simultaneously, including contextual ones, might be the reason for previous studies finding positive effects for customization of gamification but not for its personalization.

5.4 Summary of Implications

Based on our findings, we derive two main implications:

(1) *Multidimensional personalization of gamification improves students' motivations in assessment activities when compared to the OSFA approach.* Because motivation, especially autonomous, is among the most important factors for learning [27, 55, 57, 76, 78], this practical implication

provides valuable information for designers of gamified educational systems. Then, practitioners can follow the strategy used in our experiment to personalize gamified educational systems, which is likely to improve students' experiences and, according to the relationship between autonomous motivation and learning, improve students' overall learning. Consequently, future research needs to ground and expand the understanding of the effects of gamification personalized to multiple information. Therefore, we call for future research to run similar experiments with larger, varied samples and in other contexts (e.g., with other learning activity types and subjects). Those would increase the understanding of how the multidimensional personalization's benefits we found generalize.

(2) *If one wants to personalize, they cannot oversimplify it.* Because much research on personalization of gamification has been conducted [32], but with little evidence on its effectiveness compared to the OSFA approach [60], this implication is of special value to researchers. In this study, we hypothesized that the unexpected results of prior research comparing personalized and OSFA gamification were due to personalizing to a single user characteristic. While our positive findings might be due to the multidimensional personalization, considering the task to be done is a factor as well. In that context, we also discussed how the increased complexity of the multidimensional strategy approximated it to some of the benefits of customization of gamification. Accordingly, we need further research to better understand how such strategies compare to each other and what are the main reasons for the differences. Thus, we call for future research to experimentally compare tailoring strategies, such as single-dimension personalization, multidimensional personalization, and customization, as well as to investigate which are the most relevant criteria for their effectiveness when applied to users.

5.5 Limitations and Recommendations to Future Research

Some limitations need to be considered when interpreting our findings. First, about our sample. Subjects of our sample had already used the system before but with different game elements, and that might have affected their experiences during this study. To mitigate that risk, we created a new account for each user, aiming to ensure they would have a baseline for when the study began, and highlighted they should complete the motivation measure based on their experiences during the specific assessment activities. Furthermore, twenty-six learners from the same period participated in this study; a total of 45 data points considering both experiment days. This sample is similar to those seen in overall HCI studies [11] and CHI papers [10]. While that does not set off the small sample size limitations, it shows conducting large-scale studies in this field is challenging and expensive and reflects that small-sampled studies can also contribute to the literature. Note that our study was conducted in an ecological context (i.e., classroom assessment activities), which increases its validity and further supports the difficulty of recruiting large samples.

Also, all participants were males. Overall, around 20% of first-year students of computing-related courses are females [64, 65] - similarly, the rate is 14% in Brazil [17] - and such courses are known to have a high drop-out rate, which was 26% for first-years in Brazil in 2019 [8]. Therefore, while the convenience sampling limits our findings' generalization towards an overall learning domain, the impact of not having female subjects in our study is mitigated by the fact that the distribution is similar to that of similar courses. Thus, we advocate that the context above, along with the increased ecological validity by working within a real learning context, leads to a positive trade-off compared to an increased external validity due to a large sample recruited through, for instance, a crowdsourcing platform in which participants would complete a possibly meaningless task. To cope with generalization limitations, we call for replications with larger and varied samples to further ground our findings.

Second, about the study design. Because all participants were from the same class, most participants possibly discussed about having game elements that were not available to others. Those differences were unavoidable due to the personalization of gamification and participants were aware they would exist. While that might have affected their experiences, our findings suggest personalization worked overall, regardless of some participants wanting more game elements. Additionally, subjects did not know in which condition they were, preventing the hypothesis guessing threat. Moreover, our findings are based on two measurements, which were performed in subsequent days, and suggest that personalized gamification's effectiveness did not change from one day to another. However, we cannot ensure it would continue to work in the long run. Because the novelty effect is often considered to influence the effectiveness of OSFA gamification, it might also affect personalized gamification, especially considering the novelty of the approach was suggested by our qualitative findings. On one hand, students complete assessment activities like this experiment's ones on few occasions (e.g., once or twice per term), similar to our approach, which reduces the impact of that limitation. On the other hand, research assessing personalized gamification's effect based on more applications remains needed. For instance, applying it every semester during a class's full stay in college (e.g., 4 years) given that assessments such as the ones explored in this study are not frequent.

Third, about the instruments. To the multidimensional personalization, we followed the recommendations from [59]. Despite build from statistical analyses of survey data, similar to other recommendations [32], the authors acknowledge the recommendations demand empirical validation, such as what we did in this article, because it is based on user preferences. While this might be seen as a limitation, the only way to validate those recommendations is through research such as this one, which yielded encouraging results towards the recommendations' validity. Whereas we selected Eagle-Edu by convenience, its flexibility is valuable for this kind of research. The system is in its beta version, however, and a few bugs likely affected participants' experiences. Because all subjects used the same system, those bugs likely did not differ among experimental versions, consequently, having little impact on our findings. The external regulation construct from our quantitative results showed a *questionable* reliability. This might be related to some limitations the SIMS have in many of its versions [20, 22]. While that problem might not be so pertinent to HCI studies, our quantitative data analysis helped to handle it by controlling for between participants variation [11].

Lastly, about the data and its analysis. First, our dataset might suffer from hypothesis-guessing because some participants' answers were, for instance, all sevens and ones (maximum and minimum choices). While such patterns might be perceived as unreliable, the goal of Likert scales is to capture people's self-reports of their experiences while removing their ambiguities. Accordingly, one cannot ensure whether someone has experienced that motivation level or if they carelessly selected that option due to, e.g., hypothesis-guessing. Therefore, we chose not to systematically inspect or remove data based on those patterns following similar studies [2, 40, 46, 66, 67, 74] while we complemented our dataset with qualitative data from semi-structured interviews to enrich our findings [15]. Second, our sample size is limited to 45 data points, which is associated with low statistical power and often large confidence intervals. As discussed in [12], while analyzing small samples might lead to overestimation of effect sizes, such situations often happen along with p-hacking and publication bias. To cope with that thread, we followed recommendations from the same study by adhering to data and material transparency, which encourages replication and clear reporting - including effect sizes' confidence intervals. We also limited our analyses to the four constructs SIMS measures and limited further testing to significant effects. Additionally, we highlight gamification studies have not suffered from publication bias [62]. Thus, mitigating the limitations from our sample size on conclusion validity by following guidelines to address the replication crisis. Third, a single author coded the semi-structured interviews. Having multiple independent coders to ensure validity is inadequate for our *interpretivist semi-structured approach*, while relying on multiple coders to achieve complementary views would be suitable. To mitigate that limitation, other researchers worked during subsequent analysis steps (e.g., gathering and reviewing themes) and reviewed the codebook to ensure findings were supported by the interviews' data. Lastly, while having four interviewees is within HCI common practice, none of them was from the OSFA condition, which limits our qualitative findings and points to the need for more mixed-methods and qualitative studies to examine users' subjective experiences with personalized gamification that despite elementary for HCI research, are missing in the personalized gamification field [60].

6 CONCLUSION

Personalization of gamification has attracted researchers' and practitioners' interest as the literature highlights the limitations of the OSFA - one-size-fits-all - approach. However, there is a lack of empirical evidence on how those approaches compare and the few studies comparing them personalized gamification to a single user characteristic. This study faces that gap with a mixed-methods experimental study comparing learners' motivations when performing two assessment activities, on different days, with either the OSFA approach or a gamification design personalized to multiple user and contextual information. The quantitative results showed learners who used the personalized design felt higher levels of intrinsic motivation and identified regulation compared to those who used the OSFA alternative; differences in the external regulation and amotivation were nonsignificant. Additionally, the results suggest personalization's effect did not change from the first to the second assessment. Accordingly, the qualitative findings revealed the students who used the personalized gamification considered that the game elements were suitable to their preferences and that they motivated them to perform well and make social comparisons, besides supporting their basic psychological needs.

From those findings, we derive two main implications. First, to the design of gamified educational systems. Our empirical evidence suggests that gamifying educational systems with the strategy for multidimensional personalization we used can improve students' autonomous motivation (intrinsic and identified) compared to using the standard, OSFA approach. Thus, given that increased motivation is related to learning gains, applying multidimensional personalization is likely to enhance students' learning more than OSFA gamification. Second, to research on personalized gamification. Our findings provide promising evidence that multidimensional personalization can improve OSFA gamification, a result that has not been found by other studies personalizing gamification using a single dimension. Thereby, contributing indication that personalizing to a single criterion might explain why related research found inconclusive results. Thus, suggesting the need for more complex personalization strategies such as the one used in this paper, and providing an initial empirical validation of that strategy. Nevertheless, such findings must be interpreted with caution mainly because of the study's limited sample size and the experiment being limited to remembering learning activities. Therefore, the need for replications, which can rely on our open materials for planning, execution, and data analysis.

ACKNOWLEDGMENTS

This research was partially funded by the Brazilian National Council for Scientific and Technological Development (CNPq - 141859/2019-9; 308395/2020-4), Brazilian Coordination for the Improvement of Higher Education Personnel (CAPES - Finance code 001), São Paulo Research Foundation (FAPESP - 2018/15917-0; 2018/07688-1; 2020/02801-4), and Santa Catarina State Research and Innovation

Support Foundation (FAPESC - T.O. No.: 2017TR1755). Additionally, the authors are thankful to SENAI Londrina, Brazil, for enabling this research.

REFERENCES

[1] Maximilian Altmeyer, Pascal Lessel, Subhashini Jantwal, Linda Muller, Florian Daiber, and Antonio Krüger. 2021. Potential and effects of personalizing gameful fitness applications using behavior change intentions and Hexad user types. *User Model User-Adap Inter* (2021). https://doi.org/10.1007/s11257-021-09288-6

[2] Maximilian Altmeyer, Gustavo F Tondello, Antonio Krüger, and Lennart E Nacke. 2020. HexArcade: Predicting Hexad User Types By Using Gameful Applications. In *Proceedings of the Annual Symposium on Computer-Human Interaction in Play*. 219–230.

[3] Richard A Armstrong. 2014. When to use the B onferroni correction. *Ophthalmic and Physiological Optics* 34, 5 (2014), 502–508.

[4] Shurui Bai, Khe Foon Hew, and Biyun Huang. 2020. Is gamification "bullshit"? Evidence from a meta-analysis and synthesis of qualitative data in educational contexts. *Educational Research Review* (2020), 100322.

[5] Johan Baldeón, Inmaculada Rodríguez, and Anna Puig. 2016. LEGA: A LEarner-centered GAmification Design Framework. In *Proceedings of the XVII International Conference on Human Computer Interaction* (Salamanca, Spain) *(Interacción '16)*. ACM, New York, NY, USA, Article 45, 8 pages. https://doi.org/10.1145/2998626.2998673

[6] Ann Blandford, Dominic Furniss, and Stephann Makri. 2016. Qualitative HCI research: Going behind the scenes. *Synthesis lectures on human-centered informatics* 9, 1 (2016), 1–115.

[7] Klaudia Bovermann and Theo J Bastiaens. 2020. Towards a motivational design? Connecting gamification user types and online learning activities. *Research and Practice in Technology Enhanced Learning* 15, 1 (2020), 1–18. https://doi.org/10.1186/s41039-019-0121-4

[8] Brasscom. 2019. *Estudo Brasscom – Formação Educacional e Empregabilidade em TIC (Brasscom Study - Educational Training and Employability in ICT)*. https://brasscom.org.br/estudo-brasscom-formacao-educacional-e-empregabilidade-em-tic-achados-e-recomendacoes/

[9] Virginia Braun and Victoria Clarke. 2006. Using thematic analysis in psychology. *Qualitative research in psychology* 3, 2 (2006), 77–101.

[10] Kelly Caine. 2016. Local standards for sample size at CHI. In *Proceedings of the 2016 CHI conference on human factors in computing systems*. 981–992.

[11] Paul Cairns. 2019. *Doing better statistics in human-computer interaction.* Cambridge University Press.

[12] Andy Cockburn, Pierre Dragicevic, Lonni Besançon, and Carl Gutwin. 2020. Threats of a replication crisis in empirical computer science. *Commun. ACM* 63, 8 (2020), 70–79.

[13] Sarah Cohen, Werner Nutt, and Yehoshua Sagic. 2007. Deciding equivalances among conjunctive aggregate queries. *J. ACM* 54, 2, Article 5 (April 2007), 50 pages. https://doi.org/10.1145/1219092.1219093

[14] Adrian Coyle. 1995. Discourse analysis. *Research Methods in Psychology. London* (1995).

[15] John W Creswell and J David Creswell. 2017. *Research design: Qualitative, quantitative, and mixed methods approaches.* Sage publications.

[16] Lamya F Daghestani, Lamiaa F Ibrahim, Reem S Al-Towirgi, and Hesham A Salman. 2020. Adapting gamified learning systems using educational data mining techniques. *Computer Applications in Engineering Education* 28, 3 (2020), 568–589.

[17] Sociedade Brasileira de Computação (SBC Brazilian Computing Society). 2019. *Higher Computing Education - Statistics 2019*. https://www.sbc.org.br/documentos-da-sbc/summary/133-estatisticas/1324-educacao-superior-em-computacao-estatisticas-2019

[18] Sebastian Deterding, Dan Dixon, Rilla Khaled, and Lennart Nacke. 2011. From game design elements to gamefulness: defining gamification. In *Proceedings of the 15th international academic MindTrek conference: Envisioning future media environments*. ACM, 9–15. https://doi.org/10.1145/2181037.2181040

[19] Christo Dichev and Darina Dicheva. 2017. Gamifying education: what is known, what is believed and what remains uncertain: a critical review. *International journal of educational technology in higher education* 14, 1 (2017), 9.

[20] Vítor Gamboa, Sandra T Valadas, and Olímpio Paixão. 2013. Validação da versão portuguesa da Situational Motivation Scale (SIMS) em contextos académicos. In *Atas do XII Congresso Galego-Português de Psicopedagogia, 11-13 de setembro de 2013*. Universidade do Minho, 4868–4882.

[21] MN Gari, GS Walia, and AD Radermacher. 2018. Gamification in computer science education: A systematic literature review. In *American Society for Engineering Education*.

[22] Frédéric Guay, Robert J Vallerand, and Céline Blanchard. 2000. On the assessment of situational intrinsic and extrinsic motivation: The Situational Motivation Scale (SIMS). *Motivation and emotion* 24, 3 (2000), 175–213.

[23] Mohammad Hajarian, Azam Bastanfard, Javad Mohammadzadeh, and Madjid Khalilian. 2019. A personalized gamification method for increasing user engagement in social networks. *Social Network Analysis and Mining* 9, 1 (2019), 47.

[24] Stuart Hallifax, Serna Audrey, Marty Jean-Charles, Lavoué Guillaume, and Lavoué Elise. 2019. Factors to Consider for Tailored Gamification. In *CHI Play*. ACM, Barcelona, Spain, 559–572. https://hal.archives-ouvertes.fr/hal-02185647

[25] Stuart Hallifax, Audrey Serna, Jean-Charles Marty, and Élise Lavoué. 2019. Adaptive Gamification in Education: A Literature Review of Current Trends and Developments. In *Transforming Learning with Meaningful Technologies*, Maren Scheffel, Julien Broisin, Viktoria Pammer-Schindler, Andri Ioannou, and Jan Schneider (Eds.). Springer International Publishing, Cham, 294–307. https://doi.org/10.1007/978-3-030-29736-7_22

[26] Juho Hamari, Jonna Koivisto, and Harri Sarsa. 2014. Does Gamification Work?-A Literature Review of Empirical Studies on Gamification.. In *HICSS*, Vol. 14. 3025–3034.

[27] Michael D. Hanus and Jesse Fox. 2015. Assessing the effects of gamification in the classroom: A longitudinal study on intrinsic motivation, social comparison, satisfaction, effort, and academic performance. *Computers & Education* 80 (2015), 152 – 161. https://doi.org/10.1016/j.compedu.2014.08.019

[28] Muhammad Awais Hassan, Ume Habiba, Fiaz Majeed, and Muhammad Shoaib. 2019. Adaptive gamification in e-learning based on students' learning styles. *Interactive Learning Environments* (2019), 1–21.

[29] John Hattie. 2012. *Visible learning for teachers: Maximizing impact on learning*. Routledge.

[30] Karl M Kapp. 2012. *The gamification of learning and instruction*. Wiley San Francisco.

[31] Ana Carolina Tomé Klock, Isabela Gasparini, and Marcelo Soares Pimenta. 2019. Designing, Developing and Evaluating Gamification: An Overview and Conceptual Approach. In *Data Analytics Approaches in Educational Games and Gamification Systems*. Springer, 227–246.

[32] Ana Carolina Tomé Klock, Isabela Gasparini, Marcelo Soares Pimenta, and Juho Hamari. 2020. Tailored gamification: A review of literature. *International Journal of Human-Computer Studies* (2020), 102495.

[33] Ana Carolina Tomé Klock, Aline Nunes Ogawa, Isabela Gasparini, and Marcelo Soares Pimenta. 2018. Does gamification matter?: a systematic mapping about the evaluation of gamification in educational environments. In *Proceedings of the 33rd Annual ACM Symposium on Applied Computing*. ACM, 2006–2012.

[34] Selay Arkn Kocadere and Şeyma Çağlar. 2015. The design and implementation of a gamified assessment. *Journal of e-Learning and Knowledge Society* 11, 3 (2015).

[35] Jonna Koivisto and Juho Hamari. 2019. The rise of motivational information systems: A review of gamification research. *International Journal of Information Management* 45 (2019), 191–210. https://doi.org/10.1016/j.ijinfomgt.2018.10.013

[36] David R Krathwohl. 2002. A revision of Bloom's taxonomy: An overview. *Theory into practice* 41, 4 (2002), 212–218. https://doi.org/10.1207/s15430421tip4104_2

[37] Richard N Landers. 2014. Developing a theory of gamified learning: Linking serious games and gamification of learning. *Simulation & gaming* 45, 6 (2014), 752–768.

[38] Richard N Landers, Elena M Auer, Andrew B Collmus, and Michael B Armstrong. 2018. Gamification science, its history and future: Definitions and a research agenda. *Simulation & Gaming* 49, 3 (2018), 315–337.

[39] Richard N. Landers, Kristina N. Bauer, Rachel C. Callan, and Michael B. Armstrong. 2015. *Psychological Theory and the Gamification of Learning*. Springer International Publishing, Cham, 165–186. https://doi.org/10.1007/978-3-319-10208-5_9

[40] Elise Lavoué, Baptiste Monterrat, Michel Desmarais, and Sébastien George. 2018. Adaptive gamification for learning environments. *IEEE Transactions on Learning Technologies* 12, 1 (2018), 16–28.

[41] Joey J Lee and Jessica Hammer. 2011. Gamification in education: What, how, why bother? *Academic exchange quarterly* 15, 2 (2011), 146.

[42] Pascal Lessel, Maximilian Altmeyer, Marc Müller, Christian Wolff, and Antonio Krüger. 2017. Measuring the effect of" bottom-up" gamification in a microtask setting. In *Proceedings of the 21st International Academic Mindtrek Conference*. 63–72.

[43] De Liu, Radhika Santhanam, and Jane Webster. 2017. Toward Meaningful Engagement: A Framework for Design and Research of Gamified Information Systems. *MIS quarterly* 41, 4 (2017), 1011–1034.

[44] Kevin Loughrey and Daire Broin. 2018. Are We Having Fun Yet? Misapplying Motivation to Gamification. In *2018 IEEE Games, Entertainment, Media Conference (GEM)*. IEEE, 1–9. https://doi.org/10.1109/GEM.2018.8516535

[45] Patrick Mair and Rand Wilcox. 2018. Robust statistical methods using WRS2. *The WRS2 Package* (2018).

[46] A. Mora, G. F. Tondello, L. E. Nacke, and J. Arnedo-Moreno. 2018. Effect of personalized gameful design on student engagement. In *2018 IEEE Global Engineering Education Conference (EDUCON)*. 1925–1933. https://doi.org/10.1109/EDUCON.2018.8363471

[47] Lennart E Nacke and Christoph Sebastian Deterding. 2017. The maturing of gamification research. *Computers in Human Behaviour* (2017), 450–454.

[48] Wilk Oliveira, Armando Toda, Paula Toledo, Lei Shi, Julita Vassileva, Ig Ibert Bittencourt, and Seiji Isotani. 2020. Does Tailoring Gamified Educational Systems Matter? The Impact on Students' Flow Experience. In *Proceedings of the 53rd Hawaii International Conference on System Sciences*. ScholarSpace, 1226–1235.
[49] Rita Orji, Gustavo F Tondello, and Lennart E Nacke. 2018. Personalizing persuasive strategies in gameful systems to gamification user types. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*. ACM, 435. https://doi.org/10.1145/3173574.3174009
[50] Kiemute Oyibo, Rita Orji, and Julita Vassileva. 2017. The influence of culture in the effect of age and gender on social influence in persuasive technology. In *Adjunct publication of the 25th conference on user modeling, adaptation and personalization*. 47–52.
[51] Kiemute Oyibo, Rita Orji, and Julita Vassileva. 2017. Investigation of the Persuasiveness of Social Influence in Persuasive Technology and the Effect of Age and Gender.. In *PPT@ PERSUASIVE*. 32–44.
[52] Kiemute Oyibo, Rita Orji, and Julita Vassileva. 2017. Investigation of the social predictors of competitive behavior and the moderating effect of culture. In *Adjunct publication of the 25th conference on user modeling, adaptation and personalization*. 419–424.
[53] Paula Palomino, Armando Toda, Luiz Rodrigues, Wilk Oliveira, and Seiji Isotani. 2020. From the Lack of Engagement to Motivation: Gamification Strategies to Enhance Users Learning Experiences. In *19th Brazilian Symposium on Computer Games and Digital Entertainment (SBGames)-GranDGames BR Forum*. 1127–1130.
[54] Paula T Palomino, Armando M Toda, Wilk Oliveira, Luiz Rodrigues, and Seiji Isotani. 2019. Gamification journey: A Novel approach for classifying gamer types for gamified educational systems. *Proceedings of SBGames* (2019), 165–173.
[55] Paul R Pintrich. 2003. A motivational science perspective on the role of student motivation in learning and teaching contexts. *Journal of educational Psychology* 95, 4 (2003), 667.
[56] Luiz Rodrigues, Wilk Oliveira, Armando Toda, Paula Palomino, and Seiji Isotani. 2019. Thinking Inside the Box: How to Tailor Gamified Educational Systems Based on Learning Activities Types. In *Proceedings of the Brazilian Symposium of Computers on Education*. SBC, 823–832. https://doi.org/10.5753/cbie.sbie.2019.823
[57] Luiz Rodrigues, Armando M Toda, Wilk Oliveira, Paula T Palomino, Anderson Paulo Avila-Santos, and Seiji Isotani. 2021. Gamification Works, but How and to Whom? An Experimental Study in the Context of Programming Lessons. In *Proceedings of the 52nd ACM Technical Symposium on Computer Science Education*. 184–190.
[58] Luiz Rodrigues, Armando M Toda, Wilk Oliveira, Paula T Palomino, and Seiji Isotani. 2020. Just beat it: Exploring the influences of competition and task-related factors in gamified learning environments. In *Anais do XXXI Simpósio Brasileiro de Informática na Educação*. SBC, 461–470.
[59] Luiz Rodrigues, Armando M. Toda, Wilk Oliveira, Paula T. Palomino, Julita Vassileva, and Seiji Isotani. 2021. Automating Gamification Personalization: To the User and Beyond. arXiv:2101.05718 [cs.HC]
[60] Luiz Rodrigues, Armando M. Toda, Paula T. Palomino, Wilk Oliveira, and Seiji Isotani. 2020. Personalized gamification: A literature review of outcomes, experiments, and approaches. In *Proceedings of the 8th International Conference on Technological Ecosystems for Enhancing Multiculturality (TEEM 2020) (Salamanca, Spain, October 21-23, 2020)*, F. J. García-Peñalvo (Ed.).
[61] Richard M Ryan and Edward L Deci. 2017. *Self-determination theory: Basic psychological needs in motivation, development, and wellness*. Guilford Publications.
[62] Michael Sailer and Lisa Homner. 2020. The Gamification of Learning: a Meta-analysis. *Educ Psychol Rev* 32 (2020), 77–112. https://doi.org/10.1007/s10648-019-09498-w
[63] Diana R. Sanchez, Markus Langer, and Rupinder Kaur. 2020. Gamification in the classroom: Examining the impact of gamified quizzes on student learning. *Computers & Education* 144 (2020), 103666. https://doi.org/10.1016/j.compedu.2019.103666
[64] Sharon Sassler, Katherine Michelmore, and Kristin Smith. 2017. A tale of two majors: explaining the gender gap in stem employment among computer science and engineering degree holders. *Social Sciences* 6, 3 (2017), 69.
[65] Christian Schindler and Matthias Müller. 2019. Gender gap? a snapshot of a bachelor computer science course at Graz University of Technology. In *Proceedings of the 13th European Conference on Software Architecture-Volume 2*. 100–104.
[66] Marc Schubhan, Maximilian Altmeyer, Dominic Buchheit, and Pascal Lessel. 2020. Investigating User-Created Gamification in an Image Tagging Task. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. 1–12.
[67] Hallifax Stuart, Elise Lavoué, and Audrey Serna. 2020. To tailor or not to tailor gamification? An analysis of the impact of tailored game elements on learners' behaviours and motivation. In *21th International Conference on Artificial Intelligence in Education*.
[68] Sujit Subhash and Elizabeth A Cudney. 2018. Gamified learning in higher education: A systematic review of the literature. *Computers in Human Behavior* 87 (2018), 192–206.
[69] Armando Toda, Wilk Oliveira, Lei Shi, Ig Bittencourt, Seiji Isotani, and Alexandra Cristea. 2019. Planning Gamification Strategies based on User Characteristics and DM: A Gender-based Case Study. In *12th International Conference on*

Educational Data Mining. 438–443.
[70] Armando M Toda, Ana CT Klock, Wilk Oliveira, Paula T Palomino, Luiz Rodrigues, Lei Shi, Ig Bittencourt, Isabela Gasparini, Seiji Isotani, and Alexandra I Cristea. 2019. Analysing gamification elements in educational environments using an existing Gamification taxonomy. *Smart Learning Environments* 6, 1 (2019), 16. https://doi.org/10.1186/s40561-019-0106-1
[71] Armando M. Toda, Pedro H. D. Valle, and Seiji Isotani. 2018. The Dark Side of Gamification: An Overview of Negative Effects of Gamification in Education. In *Higher Education for All. From Challenges to Novel Technology-Enhanced Solutions*, Alexandra Ioana Cristea, Ig Ibert Bittencourt, and Fernanda Lima (Eds.). Springer International Publishing, Cham, 143–156. https://doi.org/10.1007/978-3-319-97934-2_9
[72] Gustavo Fortes Tondello. 2019. *Dynamic Personalization of Gameful Interactive Systems*. Ph.D. Dissertation. University of Waterloo.
[73] Gustavo F Tondello, Alberto Mora, and Lennart E Nacke. 2017. Elements of gameful design emerging from user preferences. In *Proceedings of the Annual Symposium on Computer-Human Interaction in Play*. ACM, 129–142. https://doi.org/10.1145/3116595.3116627
[74] Gustavo F Tondello and Lennart E Nacke. 2020. Validation of User Preferences and Effects of Personalized Gamification on Task Performance. *Frontiers in Computer Science* 2 (2020), 29.
[75] Rob Van Roy and Bieke Zaman. 2018. Need-supporting gamification in education: An assessment of motivational effects over time. *Computers & Education* 127 (2018), 283–297. https://doi.org/10.1016/j.compedu.2018.08.018
[76] Maarten Vansteenkiste, Eline Sierens, Bart Soenens, Koen Luyckx, and Willy Lens. 2009. Motivational profiles from a self-determination perspective: The quality of motivation matters. *Journal of educational psychology* 101, 3 (2009), 671.
[77] Marisa Venter. 2020. Gamification in STEM programming courses: State of the art. In *2020 IEEE Global Engineering Education Conference (EDUCON)*. IEEE, 859–866.
[78] Cathrin Vogel, Jana Hochberg, Sarah Hackstein, Alina Bockshecker, Theo J Bastiaens, and Ulrike Baumöl. 2018. Dropout in Distance Education and how to Prevent it. In *EdMedia+ Innovate Learning*. Association for the Advancement of Computing in Education (AACE), 1788–1799.
[79] Jan B Vornhagen, April Tyack, and Elisa D Mekler. 2020. Statistical Significance Testing at CHI PLAY: Challenges and Opportunities for More Transparency. In *Proceedings of the Annual Symposium on Computer-Human Interaction in Play*. 4–18.
[80] Rand R Wilcox. 2011. *Introduction to robust estimation and hypothesis testing*. Academic press.
[81] Zamzami Zainuddin, Samuel Kai Wah Chu, Muhammad Shujahat, and Corinne Jacqueline Perera. 2020. The impact of gamification on learning and instruction: A systematic review of empirical evidence. *Educational Research Review* (2020), 100326.

A SUPPLEMENTARY MATERIAL

To cope with transparent research, we attached at https://osf.io/grzhp/: the assessment activities (original Portuguese version and a version translated to English); the decision trees used to drive the personalization of gamification; the SIMS (Portuguese); the thematic analysis codebook with quotes for each tag; and the quantitative dataset along with the data analysis process. The interview transcriptions will not be made available due to sensitive data.

Received February 2021; revised June 2021; accepted July 2021

ANNEX

I

HOW PERSONALIZATION AFFECTS MOTIVATION IN GAMIFIED REVIEW ASSESSMENTS

How Personalization Affects Motivation in Gamified Review Assessments

Luiz Rodrigues[1*], Paula T. Palomino[1,2], Armando M. Toda[1,3], Ana C. T. Klock[4,5], Marcela Pessoa[6,7], Filipe D. Pereira[8,3], Elaine H. T. Oliveira[7], David F. Oliveira[7], Alexandra I. Cristea[3], Isabela Gasparini[9] and Seiji Isotani[1]

[1]ICMC, University of São Paulo, São Carlos, Brazil.
[2]HCI Games Group, University of Waterloo, Stratford, Canada.
[3], Durham University, Durham, United Kingdom.
[4]Institute of Informatics, Federal University of Rio Grande do Sul, Porto Alegre, Brazil.
[5]Gamification Group, Tampere University, Tampere, Finland.
[6]Amazonas State University, Manaus, Brazil.
[7]Institute of Computing, Federal University of Amazonas, Manaus, Brazil.
[8]Department of Computer Science, Federal University of Roraima, Roraima, Brazil.
[9]Department of Computer Science, Santa Catarina State University, Joinville, Brazil.

*Corresponding author(s). E-mail(s): lalrodrigues@usp.br; Contributing authors: paulatpalomino@usp.br; armando.toda@usp.com; actklock@inf.ufrgs.br; msppessoa@uea.edu.br; filipe.dwan@ufrr.br; elaine@icomp.ufam.edu.br; david@icomp.ufam.edu.br; alexandra.i.cristea@durham.ac.uk; isabela.gasparini@udesc.br; sisotani@icmc.usp.br;

Author Contributions: L. Rodrigues: Conceptualization, Methodology, Formal Analysis, Investigation, Data Curation, Writing - Original Draft, Visualisation; **P. T. Palomino:**

Conceptualization, Methodology, Validation, Data Curation, Writing - Review & Editing; **A. M. Toda:** Conceptualization, Methodology, Validation, Data Curation, Writing - Review & Editing; **A. C. T. Klock:** Conceptualization, Methodology, Validation, Writing - Review & Editing; **M. Pessoa:** Resources, Data Curation, Writing - Review & Editing; **F. D. Pereira:** Resources, Data Curation, Writing - Review & Editing; **E. H. T. Oliveira:** Resources, Writing - Review & Editing; **D. F. Oliveira:** Resources, Writing - Review & Editing; **A. I. Cristea:** Validation, Writing - Review & Editing; **I. Gasparini:** Validation, Writing - Review & Editing; **S. Isotani:** Conceptualization, Methodology, Writing - Review & Editing, Supervision.

Abstract

Personalized gamification aims to address shortcomings of the one-size-fits-all (OSFA) approach in improving students' motivations throughout the learning process. However, studies still focus on personalizing to a single user dimension, ignoring multiple individual and contextual factors that affect user motivation. A few exceptions explored multidimensional personalization, but failed to use OSFA as the baseline or analyze multiple institutions' students. Thus, we conducted a controlled experiment in three institutions, comparing gamification designs (OSFA and Personalized to the learning task and users' gaming habits/preferences and demographics) in terms of 58 students' motivations to complete assessments for learning. Our results suggest multidimensional personalization reduces motivation sensitivity to covariates (e.g., assessment performance), indicate user groups demanding further preference modeling, and show students perceive gamified assessments positively, besides considering them effective learning tools. Our contribution benefits designers, suggesting how personalization works; instructors, demonstrating how to design assessments for learning; and researchers, providing future research directions.

Keywords: Gamification, gameful, tailoring, education, self-determination theory

1 Introduction

Virtual Learning Environments (VLE) play a crucial role in education, especially during the COVID-19 pandemic, which forced remote learning in many countries (Dash, 2019; Mpungose, 2020). VLE enable managing educational materials and deploying assessments (Kocadere & Çağlar, 2015; Pereira et al., 2021, 2020), activities many instructors value during face-to-face and remote learning considering they improve learning outcomes (Batsell Jr, Perry, Hanley, & Hostetter, 2017; Mpungose, 2020; Rowland, 2014). However, educational activities are not motivating oftentimes, and that is problematic because motivation is tightly associated with learning (Hanus & Fox, 2015; Pereira et al., 2020; Pintrich, 2003; Rodrigues, Toda, Oliveira, Palomino, Avila-Santos, & Isotani, 2021). Gamification might improve motivation to engage with assessments, considering empirical evidence demonstrates gamifying learning improves motivational outcomes (Deterding, Dixon, Khaled, & Nacke, 2011; Sailer & Homner, 2020). Unfortunately, such effect varies from person to person and context to context (Hallifax, Audrey, Jean-Charles, Guillaume, & Elise, 2019; Hamari, Koivisto, & Sarsa, 2014; Rodrigues, Toda, Oliveira, Palomino, & Isotani, 2020). To mitigate such variations, researchers are exploring personalizing gamification, especially for educational purposes (Klock, Gasparini, Pimenta, & Hamari, 2020).

Specifically, personalization of gamification is when designers or the system itself define the game elements (Tondello, 2019). In practice, that is often realized by offering different motivational affordances - implemented by game elements - to different users of a gamified system (Hallifax, Serna, Marty, & Lavoué, 2019). Accordingly, personalized gamification acknowledges people have different preferences and, thereby, are motivated differently (Altmeyer, Lessel, Muller, & Krüger, 2019; Tondello, Mora, & Nacke, 2017; Van Houdt, Millecamp, Verbert, & Vanden Abeele, 2020). Despite it has been widely researched, the understanding of how personalized gamification compares to the one-size-fits-all (OSFA) approach is limited because prior research mostly compared personalization to counter-tailored or randomly defined designs (Rodrigues, Toda, Palomino, Oliveira, & Isotani, 2020). Initial empirical evidence suggests personalized gamification can overcome the OSFA approach within social networks and health domains (Hajarian, Bastanfard, Mohammadzadeh, & Khalilian, 2019; Lopez & Tucker, 2021), but results from experiments comparing those approaches within the educational domain are mostly inconclusive (Rodrigues, Toda, Palomino, et al., 2020). Having gamification personalized to a single dimension might explain related work's inconclusive findings (Mora, Tondello, Nacke, & Arnedo-Moreno, 2018; Oliveira et al., 2020). Recent research highlighted the need for considering multiple dimensions simultaneously (Klock et al., 2020), and preliminary evidence supports the value of multidimensional personalization. However, such empirical evidence is limited by not comparing multidimensional personalization to the OSFA approach (Stuart, Lavoué, & Serna, 2020) or by low external validity (Rodrigues, Palomino, et al., 2021).

Thus, **our goal is to test whether the effect of gamification personalized to the learning task and users' gaming habits/preferences and demographics generalizes to other samples/contexts, as well as investigate possible moderators[1] of that effect.** We accomplish that goal

[1] Moderators are factors that increase/decrease an intervention's effect (Landers, Auer, Collmus, & Armstrong, 2018).

with an experimental study conducted in three institutions. Thereby, differing from prior research in three directions. First, unlike most studies (e.g., Lavoué, Monterrat, Desmarais, and George (2018); Roosta, Taghiyareh, and Mosharraf (2016); Stuart et al. (2020)), our baseline is OSFA gamification, implemented with points, badges, and leaderboards (PBL). That is important to contribute to practitioners because PBL is the game elements set used the most by research on gamification applied to education and, overall, have effects comparable to other sets (Bai, Hew, & Huang, 2020). Second, we differ from Hajarian et al. (2019); Lopez and Tucker (2021) in terms of context (i.e., education instead of dating/exercise). That is important because context affects gamification's effect (Hallifax, Audrey, et al., 2019; Hamari et al., 2014; Liu, Santhanam, & Webster, 2017). Third, we mainly differ from Rodrigues, Palomino, et al. (2021) by i) involving three, instead of one institution, ii) sampling northwestern Brazilian students instead of southwesterns, and iii) capturing repeated-measures with four to six weeks of spacing instead of a one-day interval. Hence, we expand the literature by testing the external validity of state-of-the-art findings. Additionally, we present exploratory analyses to understand variations in multidimensional personalization's effect. This is important to advance the field from *whether* to *when/to whom* personalization works, especially because standard gamification's effect depends on several factors (Bai et al., 2020; Huang et al., 2020; Sailer & Homner, 2020). Differently, Rodrigues, Palomino, et al. (2021) is limited to a confirmatory analysis.

Therefore, as contributions, our findings suggest that using gamification personalized to multiple dimensions (i.e., learning task and users' gaming habits/preferences and demographics) is more effective than the OSFA approach because it reduces outcomes' sensibility to covariates (e.g., assessment performance). That is, personalization users' motivations depended on fewer covariates than those of OSFA users. Additionally, qualitative analyses showed students of both conditions considered the gamified review assessments effective learning tools that provide positive experiences. Hence, highlighting the value of assessments focused on theoretical/conceptual aspects. Furthermore, exploratory analyses suggested when/to who personalization was more or less effective and how to improve the recommendations we followed to personalize gamification. Thus, our findings inform *designers* what to expect from personalized gamification, *instructors* the value of gamified review assessments, and *researchers* directions and hypotheses to be explored in future research.

2 Background

This section provides background information on VLE and assessments for learning, gamification's effect and sources of its variation, and tailored gamification. Then, it reviews related work.

2.1 Virtual Learning Environments and Assessments for Learning

VLE are essential for nowadays education. Such tools provide better access to materials and supplementary resources, and facilitate feedback and learning outside the class (Dash, 2019; Pereira et al., 2020). They have been especially important during the COVID-19 pandemic, which forced the adoption of remote learning in many countries (Mpungose, 2020). Regardless, instructors still value students completing assignments and assessments (Mpungose, 2020; Pereira et al., 2021), which are only enabled through VLE during these times.

A theoretical perspective to understand those activities' value comes from Bloom's Taxonomy (Bloom, 1956). It classifies educational outcomes, helping instructors in defining what they expect/intend students to learn. Considering how learning objectives are commonly described, the taxonomy was revised and split into two dimensions: knowledge and cognitive (Krathwohl, 2002). The former relates to learning terminologies, categories, algorithms, and strategies knowledge. The latter refers to whether the learners are expected to remember, understand, apply, analyze, evaluate, or create. Accordingly, instructors might use assignments to encourage students in applying algorithms to new contexts or provide assessments to help students fix terminologies/strategies by remembering them. Thus, instructors likely value such activities because they allow exploring and encouraging varied learning outcomes.

From a practical perspective, those activities' value is supported by the testing effect: the idea that completing tests (e.g., assessments/quizzes) improves learning (Roediger-III & Karpicke, 2006). Empirical evidence supports that theory, showing completing tests positively affects learning outcomes in general (Adesope, Trevisan, & Sundararajan, 2017; Batsell Jr et al., 2017; Rowland, 2014) and in gamified (Rodrigues, Toda, Oliveira, Palomino, Avila-Santos, & Isotani, 2021; Sanchez, Langer, & Kaur, 2020) settings. On one hand, that is important because most educational materials are not motivating for students (Hanus & Fox, 2015; Palomino, Toda, Rodrigues, Oliveira, & Isotani, 2020; Pintrich, 2003). Hence, gamifying those assessments likely improves student motivation to complete a task known to enhance learning. On the other hand, research shows there are several factors that decrease gamification's effectiveness (i.e., moderators, such as age (Polo-Peña, Frías-Jamilena, & Fernández-Ruano, 2020) and being a gamer (Recabarren, Corvalán, & Villegas, 2021)), leading to cases wherein effects end up negatively affecting users (Hyrynsalmi, Smed, & Kimppa, 2017; Toda, Valle, & Isotani, 2018) and, consequently, harming their learning.

2.2 Gamified Learning: Effects and Moderators

Gamification is the use of game design elements outside games (Deterding et al., 2011), which has been researched the most in the educational domain (Koivisto & Hamari, 2019). Meta-analyses summarizing the effects of gamification applied to education found positive effects on cognitive, behavioral, and motivational learning outcomes (Bai et al., 2020; Huang et al., 2020; Sailer & Homner, 2020). However, these studies also reveal substantial variation within gamification's effects due to varied characteristics, such as geographic location, educational subject, and intervention duration. For instance, studies have found gamification's effect ranged from positive to negative within the same sample, depending on the user (Rodrigues, Toda, Oliveira, et al., 2020; Van Roy & Zaman, 2018). Similarly, empirical evidence shows cases wherein gamification's impact changed depending on specific characteristics, such as gender (Pedro, Lopes, Prates, Vassileva, & Isotani, 2015), age (Polo-Peña et al., 2020) and being a gamer (Recabarren et al., 2021).

These findings reveal that gamification only affects behavioral and cognitive outcomes mediated by motivational ones, as defined by the Gamification Science framework (Landers et al., 2018). That is, gamification affects motivation; motivation affects behavior; behavior affects cognitive outcomes. Consequently, analyzing behavior/cognitive outcomes without considering motivational ones is problematic. Gamification might improve motivation, but that improved motivation might not lead to the desired behavior. In that case, the problem was not gamification, but motivating some other behavior, which cannot be observed by a study limited to analyzing behavior. Therefore, not observing motivational outcomes, or applying gamification designs that motivate behaviors other than those expected, might explain studies reporting gamification failed. Thus, gamification studies must prioritize measuring motivational outcomes aligned to gamification's goals to prevent misleading conclusions (Landers et al., 2018; Tondello & Nacke, 2020). On the other hand, the problem might be the gamification design itself (Loughrey & Broin, 2018; Toda, do Carmo, da Silva, Bittencourt, & Isotani, 2019). Empirical evidence demonstrates that different users are motivated differently (Hallifax, Audrey, et al., 2019; Tondello, Mora, & Nacke, 2017) and that gamified designs must be aligned to the task wherein it will be used (Hallifax, Serna, et al., 2019; Liu et al., 2017; Rodrigues, Oliveira, Toda, Palomino, & Isotani, 2019). Thereby, gamified systems should offer gamification designs aligned to the task and the users. However, most gamified systems present the same game elements for all users, regardless of the task they will do (Dichev & Dicheva, 2017; Liu et al., 2017); the OSFA approach. Thus, the OSFA design might explain variations on gamification's outcomes and cases wherein it only works for some users.

In other words, while selecting adequate motivational outcomes to measure is rather straightforward, ensuring a system provides suitable game elements, considering who will use it to accomplish which task, concerns tailored gamification (Klock et al., 2020).

2.3 Tailored Gamification

Fundamentally, tailoring gamification leads to different gamification designs depending on who/to what it will be used (Klock et al., 2020; Rodrigues, Toda, Palomino, et al., 2020). When users define the design, it is known as *customization* (Tondello, 2019). Consequently, users define the design based on all of their preferences, as well as considering the task they will do. Empirical evidence supports customization's effectiveness compared to the OSFA approach in terms of behavior (Lessel, Altmeyer, Müller, Wolff, & Krüger, 2017; Tondello & Nacke, 2020). However, there is no evidence supporting improvements are due to increased motivation or, for instance, the effort put on to define their designs (Schubhan, Altmeyer, Buchheit, & Lessel, 2020). Additionally, customization is subject to the burden of making users select their designs for each task, if literature suggestions of matching gamification and task are followed (Hallifax, Serna, et al., 2019; Liu et al., 2017; Rodrigues et al., 2019).

Differently, when designers or the system itself define the tailored design, it is known as *personalization* (Tondello, 2019). In that case, one needs to model users/tasks to understand the most suitable gamification design for each case. Commonly, that is accomplished by gathering user preferences via surveys (Rodrigues, Toda, Palomino, et al., 2020) and, then, analyzing those to derive recommendations on which game elements to use when (e.g., Tondello, Orji, and Nacke (2017)). Most recommendations, however, guide on how to personalize gamification to a single or few dimensions (Klock et al., 2020), while several factors affect user preferences (see Section 2.2). Accordingly, empirical evidence from comparing gamification personalized through such recommendations to the OSFA is mostly inconclusive (Rodrigues, Toda, Palomino, et al., 2020). In contrast, the few recommendations for multidimensional personalization of gamification (e.g., Baldeón, Rodríguez, and Puig (2016); Bovermann and Bastiaens (2020)) have not been experimentally compared to OSFA gamification. The exception is Rodrigues, Toda, Oliveira, Palomino, Vassileva, and Isotani (2021), which has been validated in an initial study that yielded promising results (Rodrigues, Palomino, et al., 2021).

Summarizing, while customization naturally leads to gamification tailored to multiple user and contextual dimensions, it requires substantial effort from the users. Personalization mitigates that burden by using predefined rules to define gamification designs, but those rules are mostly driven by a single user dimension. That is problematic because several user and contextual dimensions affect gamification's effectiveness and, to our best knowledge, the only recommendation for multidimensional personalization to be empirically tested was analyzed in a small experimental study. Thus, highlighting the need for studies grounding the understanding of whether mitigating the burden of customization through multidimensional personalization improves OSFA gamification.

2.4 Related Work

Most publications applying personalized gamification with users compare it to random, counter-tailored, or no gamification (Rodrigues, Toda, Palomino, et al., 2020). Consequently, those studies do not add to the understanding of whether personalization improves the state-of-the-art: well-designed, OSFA gamification (Bai et al., 2020; Huang et al., 2020; Sailer & Homner, 2020). Therefore, we limit our related work review to experimental studies comparing personalized and OSFA gamification, considering those studies provide reliable evidence to understand personalization's contribution to practice. To our best knowledge, five studies meet such criteria, which were found by screening recent literature reviews (Hallifax, Serna, et al., 2019; Klock et al., 2020; Rodrigues, Toda, Palomino, et al., 2020) and through ad-hoc searches in recent studies not included in these reviews. Those are summarized in Table 1.

Table 1 shows prior research mostly personalized gamification to a single user dimension: HEXAD (Lopez & Tucker, 2021; Mora et al., 2018) and BRAINHEX (Oliveira et al., 2020) typologies. Exceptions are Rodrigues, Palomino, et al. (2021), which considered user (i.e. gaming habits/preferences and demographics) and contextual dimensions (i.e., the learning task), and Hajarian et al. (2019) that designed a interaction-based the personalization strategy. For the latter, we understand it considers contextual information because data were collected and used to personalize in the same context and task. Also, due to interaction data, we cannot stipulate how many dimensions were considered because all of a person's characteristics are involved when interacting with a system. Therefore, we see that related work using multidimensional personalization either applied it to non-educational ends (Hajarian et al., 2019) or has limited external validity (Rodrigues, Palomino, et al., 2021). This study addresses that gap with an experimental study conducted in three institutions that compares the OSFA approach to gamification personalized to multiple user and contextual characteristics. Thus, this study differs from Hajarian et al. (2019); Lopez and Tucker (2021), Mora et al. (2018); Oliveira et al. (2020), and Rodrigues, Palomino, et al. (2021) in terms of domain, personalization dimensionality, and external validity, respectively.

As Rodrigues, Palomino, et al. (2021) is the most similar research, we further discuss how this study differs from it. On one hand, Rodrigues, Palomino, et al. (2021) conducted an experiment (N = 26) in a single, southwestern Brazilian university. That experiment had two sessions, which happened in subsequent days, and was focused on a confirmatory analysis. That is, identifying *whether* students' motivations differed when comparing OSFA and personalized gamification. On the other hand, this study (N = 58) involved three institutions of the same country but from another geographical location (northwestern), increased the spacing between sessions from one day to four to six weeks, and extended to both confirmatory and exploratory analyses. Thereby, we expand the literature by testing the external validity of state-of-the-art results with a new, larger sample from different institutions of a distinct region. Considering Rodrigues, Palomino, et al. (2021) found personalization had a positive effect on students' autonomous motivation[2], testing whether those findings hold with new students and in other contexts is imperative to ground such results (Cairns, 2019; Seaborn & Fels, 2015). Furthermore, we extend prior research's contribution by analyzing if personalization's effect on student motivation depends on contextual (e.g., subject under study) and user

[2] Autonomous motivation encompasses intrinsic motivation and identified regulation (Vansteenkiste, Sierens, Soenens, Luyckx, & Lens, 2009).

Table 1 Related work compared to this study in terms of the personalization strategy and the study design.

		Personalized to ...		Study			
Ref	ND	user?	context?	Domain	NI	NP	Setting: task
(Mora et al., 2018)	1	HEXAD	None	Education	1	81	*Ecological*: lab practices and assessments
(Hajarian et al., 2019)	NA	Data log	Data log	Social Networks	1	2102	*Ecological*: free usage
(Oliveira et al., 2020)	1	BRAINHEX	None	Education	1	121	*Laboratory*: studying and question-answering
(Lopez & Tucker, 2021)	1	HEXAD	None	Exercise	1	35*	*Laboratory*: physical tasks
(Rodrigues, Palomino, et al., 2021)	8	Gaming habits/preferences and demographics	Learning task	Education	1	26	*Ecological*: classroom assessments
This study	8	Gaming habits/preferences and demographics	Learning task	Education	3	58	*Ecological*: classroom assessments

ND = Number of dimensions; NI = Number of institutions; NP = Number of participants; NA = Not applicable; *Considering participants that used either personalized or OSFA gamification.

characteristics, such as gender and age (i.e., works for some but not others). This understanding is important because the effectiveness of OSFA gamification is known to depend on such factors (Bai et al., 2020; Huang et al., 2020; Sailer & Homner, 2020). Thus, our exploratory analyses are imperative to shed light on *when and to whom* personalized gamification works.

In summary, expanding the understanding of personalization effect according to those aspects is important because i) context affects gamification (Hallifax, Audrey, et al., 2019; Hamari et al., 2014; Liu et al., 2017), ii) multidimensional personalization likely overcomes single dimension strategies (Rodrigues, Toda, Oliveira, Palomino, Vassileva, & Isotani, 2021; Stuart et al., 2020), and iii) the relevance of expanding and replicating gamification studies' findings (Cairns, 2019; Seaborn & Fels, 2015). Furthermore, we present exploratory analyses to further understand factors affecting the personalization impact. Only Hajarian et al. (2019) presents such analysis, which is done in a different context and limited to gender's role. Thus, our overall contribution is **empirically analyzing whether and how multidimensional personalization of gamification improves students' motivations, in the context of classroom assessments, and exploring which user and contextual information moderates that effect.**

3 Apparatus

To enable this study, we designed and deployed learning assessments in a VLE. All assessments featured 30 multiple-choice, four-alternative items. Items were designed so that students could correctly solve them if they were able to recall information from lessons. Therefore, according to the revision of Bloom's taxonomy, the experimental task was limited to the remembering cognitive dimension while it explored varied knowledge dimensions intentionally (Krathwohl, 2002). To ensure items' suitability, one researcher developed and revised all assessments instructors' guidance. A sample item, which concerns the Introduction to Computer Programming subject, reads: *About operations with strings, indicate the wrong alternative: a) 'size' returns the number of characters in a string; b) 'str' converts a number to a string; c) 'replace(a, b)' creates a string by replacing 'a' with 'b'; d) 'upper' turns all characters in the string to uppercase.* All assessments are available in the supplementary materials.

We deployed the assessments in the gamified system Eagle-Edu[3] because the system developers granted us access to use it for scientific purposes. Eagle-Edu allows creating courses of any subject, which have missions composed of activities, such as multiple-choice items. For this study, all courses feature 10 3-item missions considering students gave positive feedback about that design in Rodrigues, Palomino, et al. (2021). Missions' items, as well as items' alternatives, appeared in a random order for all courses. Because those were assessments *for* learning, students could redo items they missed until getting it right. In terms of gamification design, this study used the nine game elements described next, which are based on definitions for educational environments (Toda, Klock, et al., 2019):

- **Acknowledgment**: Badges awarded for achieving mission-related goals (e.g., completing a mission with no error); shown at the course's main page and screen's top-right;
- **Chance**: Randomly provides users with some benefit (e.g., extra points for completing a mission);
- **Competition**: A Leaderboard sorted by performance in the missions completed during the current week that highlights the first and last two students; shown at the course's main page;
- **Objectives**: Provide short-term goals by representing the course's missions as a skill tree;
- **Points**: Numeric feedback that functions similar to Acknowledgment; shown at the screen's top-right and within Leaderboards when available;
- **Progression**: Progress bars for missions; shown within missions and in the skill tree (when Objectives is on);
- **Social Pressure**: Notifications warning that some student of the same course completed a mission;

[3] http://eagle-edu.com.br/

- **Time Pressure**: Timer indicating the time left to climb in the Leaderboard before it resets (week's end);

Note that each Eagle-edu course features its gamification design. Accordingly, for the OSFA condition, we implemented a single course for each educational subject. For the personalized condition, however, we had to create one course for each gamification design. Then, if the gamification design of users A and B differed by a single game element, they would be in different Eagle-edu courses even though they study the same subject. Nevertheless, students of the same subject always completed the same assessment, and all courses had the same name to mitigate that restriction. Thus, ensuring gamification design was the only difference, although that affected the Leaderboards appearance (see Section 8). The supplementary material provides a video as well as screenshots of the system.

4 Method

According to our goal, this study investigated the following:

- **Hypothesis 1 - H1:** Multidimensional personalization of gamification improves autonomous motivation but not external regulation and amotivation, compared to the OSFA approach, in gamified review assessments.
- **Research Question 1 - RQ1:** What are students' perceptions of gamified review assessments?
- **RQ2:** Do user and contextual characteristics moderate the effect of multidimensional personalization of gamification, in gamified review assessments?
- **RQ3:** How does the variation of students' motivations change when comparing gamification personalized to multiple dimensions to the OSFA approach, in gamified review assessments?

H1 is derived from Rodrigues, Palomino, et al. (2021), which found such results in a small, single-institution study. Therefore, we aim to test whether those hold for different users, from other institutions, completing assessments of different subjects. **RQ1** is related to research demonstrating gamification is perceived positively, overall (Bai et al., 2020; Huang et al., 2020; Sailer & Homner, 2020) and within assessments (Rodrigues, Palomino, et al., 2021). While that supports expecting positive results, we frame it as an RQ because did not plan such analysis *apriori*. **RQ2** is based on research showing user and contextual characteristics moderate gamification's effect (Hallifax, Audrey, et al., 2019; Huang et al., 2020; Rodrigues, Toda, Oliveira, Palomino, Avila-Santos, & Isotani, 2021; Sailer & Homner, 2020). Thereby, we want to test whether the same happens for personalized gamification, and identify which factors are responsible for it. Similarly, **RQ3** is based on research showing gamification's effect vary from user-to-user and context-to-context (Hamari et al., 2014; Rodrigues, Toda, Oliveira, et al., 2020; Van Roy & Zaman, 2018). Thus, we want to understand if personalization can mitigate such variation. Based on those, we designed and conducted a multi-site, experimental study, following a mixed factorial design: gamification *design* (levels: OSFA and Personalized) and session (levels: 0 and 1) were the between-subject and within-subject factors, respectively. For *design*, participants were randomly assigned to a condition at each trial, while sessions 0 and 1 refer to mid-term and end-term assessments, respectively. Figure 1 summarizes this study, which received an ethical committee approval (CAAE: 42598620.0.0000.5464).

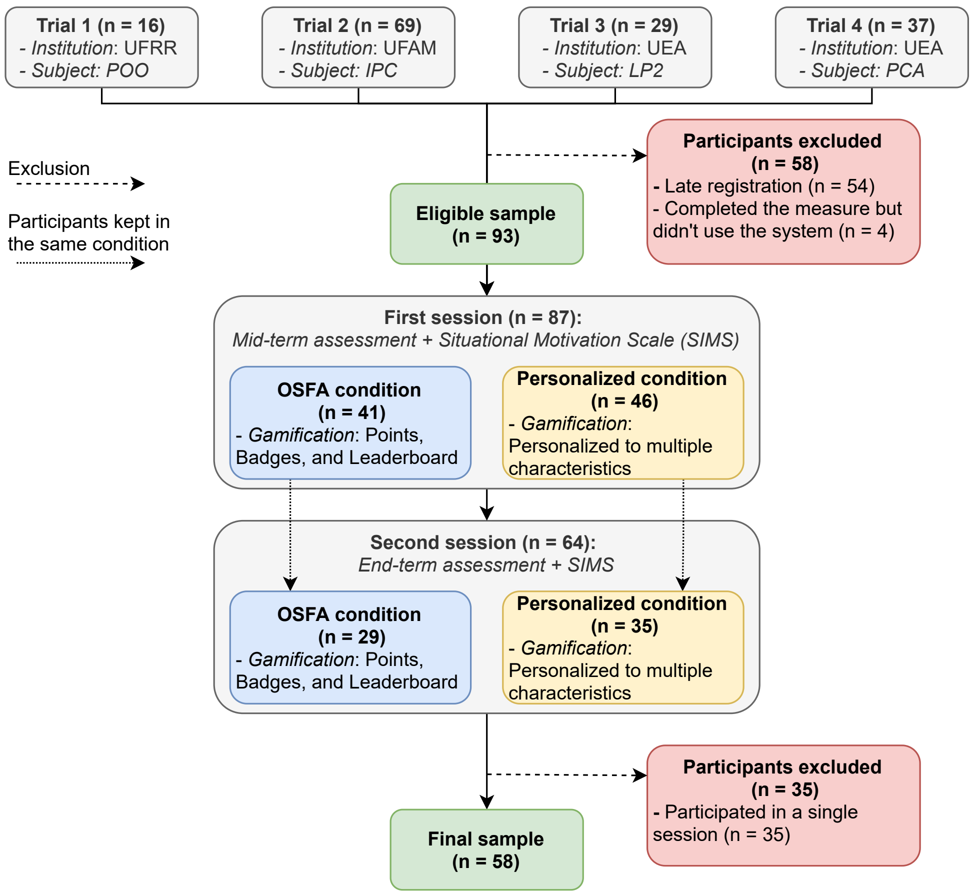

Fig. 1 Study Overview. Institutions are Federal University of Roraima (UFRR), Federal University of Amazonas (UFAM), and Amazonas State University (UEA). Subjects are Object Oriented Programming (POO), Introduction to Computer Programming (IPC), Programming Language 2 (LP2), Computer and Algorithms Programming (PCA).

4.1 Sampling

We relied on convenience sampling. Researchers contacted four fellow instructors, presented the research goals, and proposed applying review assessments for their students in two lessons. All contacted instructors agreed without receiving any compensation. Those worked in three institutions (Federal University of Roraima, Federal University of Amazonas, and Amazonas State University) and were responsible for four different subjects (Object Oriented Programming, Introduction to Computer Programming, Programming Language 2, and Computer and Algorithms Programming). Thus, all eligible participants were enrolled in one of the four subjects from one of the three

Table 2 Participants' demographic information.

Information	Overall	OSFA	Personalized
	Mean (Standard Deviation)		
Age	20.10 (1.99)	20.04 (1.99)	20.16 (2.02)
Weekly Playing Time	8.28 (9.66)	8.23 (9.05)	8.31 (10.28)
	Count (Percentage)		
Gender			
Female	21 (36%)	12 (46%)	9 (28%)
Male	37 (64%)	14 (54%)	23 (72%)
Preferred game genre			
Action	18 (31%)	10 (38%)	8 (25%)
Adventure	8 (14%)	3 (12%)	5 (16%)
RPG	13 (22%)	5 (19%)	8 (25%)
Strategy	16 (28%)	5 (19%)	11 (34%)
Others	3 (5%)	3 (12%)	0 (00%)
Preferred playing setting			
Singleplayer	23 (40%)	10 (38%)	13 (41%)
Multiplayer	35 (60%)	16 (62%)	19 (59%)
Highest degree			
High School	37 (64%)	18 (69%)	19 (59%)
Technical	10 (17%)	5 (19%)	5 (16%)
Undergraduate	11 (19%)	3 (12%)	8 (25%)
Researched gamification?			
Yes	50 (86%)	3 (12%)	5 (16%)
No	8 (14%)	23 (88%)	27 (84%)

institutions (see Figure 1). Then, for each trial, we sent the characterization survey, about a month before session 0, and asked students to complete it by the weekend before the first session to enable registering students into the system.

4.2 Participants

After the four trials, 151 students completed the measure. Four of those were excluded because they did not use the system. Another 54 were excluded due to late registration (i.e., completing the characterization form after the deadline), which made random assignment unfeasible. Nevertheless, they participated in the activity with no restriction as it was part of the lesson. Finally, 35 students were excluded because they participated in a single session, leading to our sample of 58 participants: 26 and 32 in the OSFA and Personalized conditions, respectively (see Table 2 for demographics). Students from Federal University of Amazonas and Amazonas State University received points (0.5 or 1%) towards their grades as compensation for participating in each session. We let that choice up to instructors.

4.3 Experimental Conditions

We designed two experimental conditions, *OSFA* and *Personalized*, which differ in terms of the motivational affordances they present to users. We implement

Table 3 Game elements used in the Personalized condition according to user characteristics based on recommendations from Rodrigues, Palomino, et al. (2021).

PGG	ERG	Gender	Bdg	Obj	Prog	SP	Comp	TP	Cnc
Action	No	Female	X				X	X	
Action	No	Male					X	X	X
Not action	Yes	Both	X	X		X			
Not action	No	Both	X	X	X				

PPG = Preferred game genre; ERG = Experience researching gamification; Bdg = Badges; Obj = Objectives; Prog = Progression; SP = Social Pressure; Comp = Competition; TP = Time Pressure; Cnc = Chance

such affordances through game elements, changing the game elements available considering it is the most common approach (Hallifax, Serna, et al., 2019). The *OSFA* condition featured Points, Badges, and Leaderboards (PBL), similar to Rodrigues, Palomino, et al. (2021). That combination is among the most used in gamification research and provides effects comparable to other combinations (Bai et al., 2020). Thus, we believe PBL offer external validity as an implementation of the standard, OSFA gamification.

The *Personalized* condition provided game elements according to recommendations from Rodrigues, Toda, Oliveira, Palomino, Vassileva, and Isotani (2021), following the same procedure that Rodrigues, Palomino, et al. (2021). According to Rodrigues, Toda, Oliveira, Palomino, Vassileva, and Isotani (2021), such recommendations are based on a survey that captured people's top-three preferred game elements. With these data, the authors generated decision trees, one for each top-three element. These trees receive an eight-value input set concerning user and contextual information. For the former, inputs are the user's gender, highest educational degree, weekly playing time, preferred game genre and playing setting, and whether they already researched gamification. For the latter, inputs are the country where gamification will be used and the cognitive process worked while doing the learning task, according to the processes described in the revision of Bloom's taxonomy (Krathwohl, 2002). Because contextual information is fixed for this study (all participants were from the same country and the learning task concerned the remembering cognitive process), we analyzed the decision trees and identified that, for our sample and experimental task, defining the game elements to be available for each user depends on one's preferred game genre, gender, and whether they already researched gamification. That analysis led to Table 3, which summarizes the game elements available for each user of the personalized condition.

Note that, in some cases, two decision trees (e.g., top-one and top-two) recommended the same game elements. In those cases, we selected the next recommended game element to ensure the system presented three game elements for all participants. We made that choice to avoid confounding factors of participants interacting with different numbers of game elements, which might affect gamification's effect (Landers, Bauer, Callan, & Armstrong, 2015).

4.4 Measures and Moderators

To measure our dependent variable - motivation - we used the Situational Motivation Scale (SIMS) (Guay, Vallerand, & Blanchard, 2000). It is aligned to Self-Determination Theory (Deci & Ryan, 2000), has been used in similar research (e.g., Lavoué et al. (2018); Rodrigues, Palomino, et al. (2021)), and has a version in participants' language (Gamboa, Valadas, & Paixão, 2013). Thus, its suitability to this study. Using the recommended seven-point Likert-scale, the SIMS captured motivation to engage with the VLE through four constructs: intrinsic motivation, identified regulation, external regulation, and amotivation[4] (Deci & Ryan, 2000). Additionally, we provided an open-text field so that participants could make comments about their experiences.

Note that assessment performance is not considered a dependent variable. Because the experimental task was completing assessments *for* learning, its effect on participants' knowledge would only be properly measured after the task, not during it. Instead, our exploratory analyses inspect performance in the assessments - operationalized as the number of errors per item - as a possible covariate of the personalization effect based on research showing performance-related measures might moderate gamification's effect (e.g., Rodrigues, Toda, Oliveira, Palomino, Avila-Santos, and Isotani (2021); Sanchez et al. (2020)). We also analyze user characteristics (i.e., age, gender, education, preferred game genre, preferred game setting, and weekly playing time) as possible moderators because we followed a strategy that personalizes the gamification design to them and is based on research discussing those might moderate gamification's effect (Rodrigues, Toda, Oliveira, Palomino, Vassileva, & Isotani, 2021). Additionally, we study the role of contextual information (i.e., assessment subject and usage interval) as this study differs from similar work (Rodrigues, Palomino, et al., 2021) in those. Thereby, investigating within personalized gamification moderators that have demanded attention in the standard approach.

4.5 Procedure

First, participants were invited to participate in the study. Second, they had to complete the characterization survey by the deadline, which captured identifying information plus those described in Section 4.3. Such information and respecting the deadline were essential to enable personalization. Third, around mid-term, participants had to complete the first session's assessment and the SIMS. Fourth, towards the term's end, participants completed the second session's assessment and the SIMS again. One researcher participated in both sessions and provided clarifications as required (e.g., explaining how to use the system), but at the start of session 0, the researcher presented the study goal and procedure and a tutorial on how to use the system.

[4]Intrinsic and external motivations refer to being driven by internal (e.g., enjoyment) and external (e.g., money) rewards, respectively, while identified regulation lies between them and amotivation refers to no motivation at all (Deci & Ryan, 2000).

4.6 Data Analysis

For the confirmatory analyses (**H1**), we used the same method as Rodrigues, Palomino, et al. (2021) because we are testing the generalization of their findings. Therefore, we applied robust (i.e., 20% trimmed means) mixed ANOVAs (Wilcox, 2011), according to our study design, which handle unbalanced designs and non-normal data (Cairns, 2019). We do not apply p-value corrections because each ANOVA tests a planned analysis (Armstrong, 2014). For the exploratory analyses (**RQs**), we follow recommendations from Vornhagen, Tyack, and Mekler (2020): instead of relying on p-values, which might be misleading in that context, we analyze confidence intervals (CIs) to compare participants' motivations among subgroups. That is, whether one subgroup's CI overlaps that of another subgroup. To cope with non-normal data, we calculated CIs using the bias-corrected and accelerated bootstrap, as recommended in Cairns (2019); Carpenter and Bithell (2000). Throughout those analyses, we consider user data captured by the characterization survey and *subject* and *interval* between sessions, besides the number of *errors per item* (see Section 4.4), as contextual information. Confidence levels are 95% and 90% for confirmatory and exploratory analyses, respectively (Hox, Moerbeek, & Van de Schoot, 2010; Rodrigues, Toda, Oliveira, Palomino, Avila-Santos, & Isotani, 2021). We ran all analyses using the *WRS2* (Mair & Wilcox, 2018) and *boot* (Canty & Ripley, 2021) R packages.

Our qualitative analysis concerns open-text comments. Because commenting was optional, we expect such feedback to reveal the most important perceptions/suggestions/critique from students' perspectives. The analysis process involved four steps and five researchers. First, one author conducted a thematic analysis (Braun & Clarke, 2006), familiarizing with the data, generating and reviewing codes, and grouping them into themes. Acknowledging the subjective nature of participants' comments, he followed the interpretivist semi-structured strategy (Blandford, Furniss, & Makri, 2016). Accordingly, he applied inductive coding due to participants' freedom to mention varied aspects. Next, a second author reviewed the codebook. Third, three other authors independently tagged each comment through deductive coding, using the codebook developed and reviewed in previous steps. According to the interpretivist approach, the goal with multiple coders was to increase reliability through complementary interpretations, although the importance of others inspec ting such interpretations (Blandford et al., 2016). Therefore, in the last step, the author who conducted the first step reviewed step three's results. Here, he aimed for a wider, complementary interpretation of participants' comments, rather than seeking for a single definitive tag for each comment. That step led to the consolidated, final report we present in the next section.

5 Results

This section analyzes the comparability of the experimental conditions, in terms of participants' information, and presents data analyses' results.

5.1 Preliminary Analyses - Do groups differ?

These analyses compare each conditions' participants to identify possible covariates of our results. We compare continuous variables using robust ANOVAs (see Section 4.6) and categorical ones using independence chi-squared tests. When counts are lower than 5, we simulate p-values using bootstrap through R's base *chisq* function. In those cases, degrees of freedom (df) are *NA*.

The results showed nonsignificant differences in terms of demographics, gaming preferences/habits, and experience researching gamification. We omit those results for simplicity but provide them in the supplementary material. For performance, design's main effect was nonsignificant (F(1 24.8985) = 0.9042; p = 0.35) but session's main effect (F(1, 29.8645); p = 0.0112), as well as their interaction (F(1, 29.8645); p = 0.0041) were significant. Because both main factors have two levels, the significant interaction suggests a difference among conditions on session 1 but not on session 0. Descriptive statistics show participants of the OSFA condition made less mistakes per assessment item (M = 0.964; SD = 0.509) than those of the Personalized condition (M = 1.19; SD = 0.459). That suggests personalized users had lower performance than OSFA ones in the end-term assessment. Thus, preliminary analyses indicate a single statistically significant difference among conditions - session 1's *performance* - when comparing possible covariates of our main results, despite descriptive statistics show uneven distributions for some user information (see Section 2).

5.2 Quantitative Analysis of H1

Table 4 presents descriptive statistics for all motivation constructs. Constructs' reliability, measured by Cronbach's alpha, was acceptable (≥ 0.7) for all but external regulation (0.59), which was questionable (Gliem & Gliem, 2003). Table 5 shows the results from testing **H1**, which assumes a positive effect of personalization on autonomous motivation but not on external regulation and amotivation. They reveal no statistically significant difference for all motivation constructs (all p-values > alpha), suggesting personalization did not affect any construct. Thus, partially supporting **H1**.

Table 4 Descriptive statistics, overall (Ovr) and per session (S0 and S1), for motivation constructs. Data shown as Mean (Standard Deviation).

	Design	N	Intrinsic Motivation	Identified Regulation	External Regulation	Amotivation
Ovr	One-size-fits-all	52	5.60 (1.36)	5.86 (1.31)	4.25 (1.12)	2.16 (1.10)
	Personalized	64	5.56 (1.30)	5.87 (1.03)	4.52 (1.28)	2.38 (1.60)
S0	One-size-fits-all	26	5.68 (1.23)	6.02 (1.07)	4.24 (0.93)	2.17 (1.07)
	Personalized	32	5.75 (1.21)	6.02 (0.94)	4.72 (1.34)	2.28 (1.56)
S1	One-size-fits-all	26	5.51 (1.49)	5.70 (1.51)	4.26 (1.30)	2.14 (1.15)
	One-size-fits-all	32	5.37 (1.38)	5.73 (1.10)	4.31 (1.20)	2.47 (1.65)

Table 5 Confirmatory analyses, conducted through robust mixed ANOVAs with alpha set to 0.05, of **H1**: personalization affects autonomous motivation (intrinsic and identified) but not external regulation and amotivation.

	Design		Session		Design:Session	
M	F(df1, df2)	P-val	F(df1, df2)	P-val	F(df1, df2)	P-val
IM	0.073(1, 29.914)	0.789	3.261(1, 28.141)	0.082	1.478(1, 28.141)	0.234
IR	0.157(1, 27.974)	0.695	1.692(1, 23.562)	0.206	0.128(1, 23.562)	0.723
ER	0.854(1, 29.793)	0.363	0.760(1, 29.730)	0.390	1.200(1, 29.730)	0.282
AM	0.007(1, 29.977)	0.934	0.038(1, 29.835)	0.847	0.523(1, 29.835)	0.475

5.3 Qualitative Analysis of RQ1

Thirty-two of the 58 participants provided 52 comments (participants could comment on each session). The thematic analysis found seven codes that were grouped into two themes. In step three, researchers attributed 114 codes to the 52 comments. Lastly, the consolidation step updated the codes of 13 comments, leading to the final average of 2.19 codes per comment. Table 6 describes codes and themes, exemplifying them with quotes.

One theme concerns comment specifically referring to the **assessments** and/or their items. Within that context, some comments noted the *complexity* of assessments (e.g., *"[...] I missed a lot [of questions], [an] experience I've never had in my life, very cool"*). That might be related to students' perceptions that some items had a *bad presentation* (e.g., *"Some [items] were written strangely."*). In contrast, most comments noted the assessments were *well designed*, highlighting that *"the idea of the activity is wonderful, as we don't always focus on theory"* and that it was *"[...] a good way to practice concepts [...]"*.

The second theme concerns comments referring to perceptions from participating in the **activity**. Within that context, most students had a *positive experience* (e.g., *"Very interesting and fun this approach to the activities [...] I found it a great way to motivate students, especially in the current [pandemic] scenario."*) that was *good for learning* (e.g., *"The activity was very interesting, it helped to review the exercises."; "[...] this means of activity is very interesting and cool, as it further intensifies the personal demand of becoming a better student [...]"*). However, they also experienced some *usage bugs* (e.g., *"There was a question about identifiers where there were two equal options, I clicked on the second one and got it wrong, but the first one that was identical got it right."*) and considered the *gamification demands improvement*.

Concerning the gamification designs, students suggested improvements in terms of adding other game elements (e.g., *"In addition to points and objectives, bringing competitiveness is something important to motivate the use of these gamified technologies."; "I would find it interesting for the platform to have some indication of my progress [...]"*), changing elements' mechanics (e.g., *"I also think it's cool if it had rewards for my progress and punishments for my mistakes [...]"*), and making it more meaningful (e.g., *"Gamification should be*

Table 6 Themes and codes attributed to participants' comments after conducting and validating a thematic analysis. Codes shown as: (Number of commenters/Percentage of commenters).

Code	Refers to:	Quote
Theme: Assessment		
Bad presentation (6/19%)	The way assessments/items were designed/appeared in the system should be improved	*"Make it more visible if the question asks for CORRECT or WRONG alternatives."; "I would like the questions to be formatted, I found some with errors and had difficulty understanding"*
Complexity (3/9%)	The assessments' length and/or items' complexity/difficulty level	*"Slightly improve the drafting of the questions. Some were written strangely."*
Well designed (20/63%)	Positive perceptions about the structure, the topic, and/or the presentation of the assessments/items	*"the idea of the activity is wonderful because we don't always focus on theory."; "A great way to get out of the everyday of learning and see what you know and what you don't know about the content."*
Theme: Activity		
Good for learning (20/63%)	Providing learning-related experiences, such as need-supporting, self-efficacy, self-assessment, etc.	*"very fun activity and very good to practice knowledge"; "I found the platform fun and very useful to help with my studies. I really liked it"*
Positive experience (21/66%)	Providing positive experiences, such as fun and enjoyment), not directly linked to learning	*"Excellent and super fun activity! The interaction with the discipline is very dynamic and fulfills its purpose."; "Very inviting to answer the form, besides being intuitive and simple to use."*
Gamification demands improvement (9/28%)	Suggestions about changing the gamification design (e.g., changing mechanics; adding game elements)	*"Very good, if it had a scoring system, more questions, more types of achievements and a sound it would be much better"; "refactoring the achievement system to give simpler achievements for students who have a degree of growth so they feel more excited to be able to complete them, and will not find them impossible right away"."*
Usage Bug (2/6%)	Perceiving bugs while using the system	*"When changing the platform language and clicking on an activity, the language reverts to what it was previously."*

as dynamic as possible, it's important that you bring some important aspect of your personal life to the program [...]"). These results show most students perceived the activity positively, with few notes on how it could be improved (**RQ1**).

5.4 Exploratory Analyses (RQ2 and RQ3)

To answer RQ2 and RQ3, we analyze conditions' differences based on CIs. Accordingly, differences are suggested by CIs not including zero for correlations and not overlapping for group comparisons. Tables 7 and 8 summarizes the results, which are described next.

Table 7 Exploratory analyses based on 90% Confidence Intervals (CIs) calculated through bootstrap. In comparing columns, CIs that do not overlap indicate an effect of personalization (Pers.) compared to one-size-fits-all (OSFA) gamification. Green and red backgrounds illustrate positive and negative effects, respectively. In comparing rows, CIs that do not overlap (highlighted by *) indicate student motivation varied according to that characteristic (e.g., gender on amotivation). Data shown as [Lower CI;Upper CI].

	IM		IR		ER		AM	
	OSFA	Pers.	OSFA	Pers.	OSFA	Pers.	OSFA	Pers.
Gen								*
Fem.	[4.74;5.75]	[5.36;6.14]	[5.07;6.03]	[5.60;6.30]	[4.02;4.66]	[3.89;4.89]	[1.95;2.70]	[1.31;1.78]
Male	[5.40;6.12]	[5.09;5.77]	[5.48;6.29]	[5.54;6.06]	[3.77;4.52]	[4.23;4.83]	[1.74;2.39]	[2.34;3.16]
Edu		*	*	*				
HS	[5.12;5.86]	[4.97;5.63]	[5.19;5.98]	[5.41;5.93]	[3.94;4.47]	[3.97;4.62]	[1.69;2.21]	[1.96;2.66]
Tech	[5.17;6.25]	[6.25;6.80]	[6.10;6.65]	[6.25;6.80]	[3.44;5.09]	[3.92;5.38]	[1.89;3.00]	[1.27;3.48]
Grad	[4.25;6.42]	[4.88;5.97]	[5.08;6.79]	[5.23;6.19]	[3.50;5.00]	[4.53;5.48]	[1.74;3.58]	[2.18;3.79]
PGG	*		*				*	
Act	[4.96;5.99]	[4.88;6.13]	[5.08;6.10]	[5.47;6.28]	[4.06;4.88]	[4.11;4.95]	[2.06;2.92]	[1.83;3.35]
Adv	[6.01;6.83]	[4.58;6.03]	[6.50;6.88]	[5.08;6.25]	[3.42;4.58]	[4.25;5.03]	[1.00;1.00]	[1.60;3.35]
RPG	[4.60;5.70]	[4.91;5.97]	[4.47;6.28]	[4.92;5.90]	[4.03;4.90]	[4.42;5.42]	[1.38;2.08]	[1.91;2.97]
Stg	[4.33;6.04]	[5.17;6.02]	[4.80;6.08]	[5.82;6.40]	[3.23;4.20]	[3.65;4.75]	[1.94;3.08]	[1.77;2.98]
PPS	*		*					
Mult	[5.83;6.53]	[4.97;5.79]	[6.17;6.60]	[5.33;6.06]	[4.06;4.89]	[4.18;4.94]	[1.96;2.76]	[1.88;2.80]
Sing	[4.73;5.54]	[5.29;5.97]	[4.98;5.91]	[5.68;6.18]	[3.75;4.39]	[4.10;4.80]	[1.77;2.37]	[2.03;2.93]
Sub			*	*	*			*
POO	[3.79;6.12]	[4.88;6.00]	[3.28;5.83]	[5.75;6.22]	[4.54;5.17]	[4.62;5.45]	[1.29;3.00]	[2.70;4.47]
IPC	[4.95;5.87]	[4.90;5.86]	[5.34;6.21]	[5.47;6.26]	[3.54;4.30]	[4.03;4.91]	[1.71;2.47]	[1.59;2.65]
LP2	[5.83;6.42]	[5.40;6.53]	[6.22;6.80]	[5.97;6.62]	[3.92;4.97]	[3.75;5.25]	[1.55;2.73]	[1.45;3.98]
PCA	[4.68;6.17]	[4.95;5.86]	[5.12;6.21]	[5.12;5.93]	[3.85;4.98]	[3.81;4.74]	[1.98;2.73]	[1.79;2.51]
Int					*			
0w	[5.28;6.06]	[5.39;6.06]	[5.59;6.30]	[5.70;6.26]	[3.91;4.52]	[4.33;5.09]	[1.83;2.51]	[1.91;2.84]
4w	[4.75;6.16]	[5.18;6.04]	[4.59;6.20]	[5.32;6.09]	[4.39;5.21]	[3.81;4.71]	[1.66;2.59]	[1.92;3.09]
6w	[4.58;6.04]	[4.05;5.57]	[4.98;6.29]	[4.98;6.14]	[3.04;4.22]	[3.73;4.82]	[1.67;2.85]	[1.82;3.55]

Gen = Gender; Fem. = Female; Edu = Education; HS = High School; Tech = Technical; Grad = Graduated; PGG = Preferred game genre; Act = Action; Adv = Adventure; RPG = Role-playing game; Stg = Strategy; PPS = Preferred playing setting; Mult = Multiplayer; Sing = Singleplayer; Sub = SUbject; IPC = Introduction to Computer Programming; LP2 = Programming Language 2; PCA = Computers and Algorithms Programming; Int = Interval; Nw = Number of weeks.

5.4.1 RQ2: Moderators of Personalization's Effect

For continuous variables, moderations are indicated when CIs from the OSFA condition do not overlap with those of the Personalized condition. Accordingly, Table 8 indicates *performance* was the single continuous moderator, which had an effect on the relationship between student performance and external motivation. For categorical variables, moderations are indicated when CIs suggest a difference for a variable's subgroup but not for others (compare columns in

Table 8 Exploratory Analyses for continuous variables.

IM		IR		ER		AM	
OSFA	Pers.	OSFA	Pers.	OSFA	Pers.	OSFA	Pers.
Performance							
[-0.10;0.29]	[-0.14;0.23]	[-0.03;0.36]	[-0.10;0.31]	[0.40;0.70]	[-0.24;0.16]	[0.09;0.45]	[-0.19;0.22]
Age							
[-0.43;0.11]	[-0.30;0.10]	[-0.31;0.21]	[-0.48;-0.06]	[-0.36;0.16]	[-0.15;0.25]	[-0.11;0.40]	[-0.02;0.44]
Weekly Playing time							
[-0.31;0.09]	[-0.34;0.18]	[-0.11;0.18]	[-0.14;0.26]	[-0.28;0.14]	[-0.35;0.16]	[-0.13;0.27]	[-0.14;0.34]

Table 7 for an overview). This is the case of *gender*. Females' CIs do not overlap when comparing the amotivation of the personalized [1.31;1.78] and the OSFA [1.95;2.70] conditions. Differently, males' CIs overlap when comparing the personalized [2.34;3.16] and the OSFA [1.74;2.39] conditions in terms of amotivation. Thus, suggesting gender moderated personalization's effect, which was only positive for females. *Education* appears to be another moderator. Students with a technical degree who used the personalized design experienced higher intrinsic motivation than those who used the OSFA design. CIs for the remainder subgroups of *education* overlapped. *Preferred game genre* also appears to be a moderator. When considering participants that prefer *adventure* games, those of the OSFA condition reported better identified regulation and amotivation[5] than those of the personalized condition. Results suggested no other differences among *preferred game genre* subgroups. *Preferred playing setting* seems to be another moderator. Those who prefer singleplayer reported higher intrinsic motivation and identified regulation than those of the personalized condition. The analyses suggested no motivation differences among conditions for those who prefer multiplayer. The results indicate the *assessment's subject* and *usage interval* did not moderate personalization's effect on any construct, in contrast to *gender*, *education*, *preferred game genre and playing setting*, *performance*, and *age* (**RQ2**).

5.4.2 RQ3: Differences in Motivation Variation Among Conditions

Based on Tables 7 and 8 (comparing rows), student motivation varied according to six characteristics for users of the **OSFA** design. First, *performance* was positively correlated to external regulation and amotivation. Second, people whose *preferred game genre* is adventure reported higher intrinsic motivation than those who prefer action and RPG games and higher identified regulation, as well as lower amotivation, than those who prefer any other genre analyzed. Third, participants whose *preferred playing setting* is singleplayer reported higher intrinsic motivation and identified regulation than those who prefer multiplayer. Fourth, *education*. Those with a technical degree reported higher identified regulation than those with high school. Fifth, *assessment's*

[5]The amotivation of all participants of this subgroup was 1.

subject. Identified regulation was higher for LP2 students than that of all other subjects' students and External regulation was higher for POO students than that of IPC students. Sixth, external regulation was lower when *usage interval* was six or more weeks than up to four weeks. Differently, student motivation varied according to four characteristics for users of the **Personalized** design. First, *age* was negatively correlated to identified regulation. Second, amotivation differed depending on *gender*. Third, *education*. Students with a technical degree reported higher intrinsic motivation and identified regulation compared to those with other degrees. Fourth, *assessment's subject*. Identified regulation was higher for LP2 students than that of PCA students and amotivation of POO students was higher than that of IPC and PCA ones. These results suggest motivation from personalized gamification varied according to fewer factors than that from the OSFA design (**R3**).

5.5 Summary of Results

We summarize our results as follows:

- Preliminary analyses showed participants of the personalized condition experienced more difficulty in the second session's assessment.
- **H1** is partially supported. Surprisingly, results do not confirm the personalization positive effect on autonomous motivation - instead, indicating a non-significant difference - while they corroborate the non-significant effect on external regulation and amotivation.
- **RQ1**: Qualitative results indicated the gamified assessments provided positive experiences that students perceived as well designed and good for their learning, although a few of them mentioned gamification demands improvement and considered the assessments complex and badly presented.
- **RQ2**: Exploratory analyses suggested *gender* and *education* positively moderated the personalization effect, in contrast to *preferred game genre* and *preferred playing setting*. Specifically, personalization was positive for *females* and those holding a *technical* degree, but negative for people whose preferred game genre is *adventure* and for those whose preferred playing setting is *singleplayer*.
- **RQ3**: Exploratory analyses revealed motivation varied according to six characteristics for students who used the OSFA design: performance, preferred game genre, preferred playing setting, education, assessment's subject, and usage interval. The analyses indicate the motivation of students who used personalized gamification varied according to only four factors, two common to OSFA (education and assessment's subject) and two uncommon: age and gender.

6 Discussion

For each hypothesis/RQ we studied, this section interprets its results, relates them to the literature, and discusses possible explanations that we present as testable hypotheses (TH): hypotheses that emerged from our findings and can be tested in future research.

6.1 How does personalization affect student motivation?

Confirmatory analyses revealed no significant differences among conditions for all motivation constructs (see Table 3). However, in testing groups' comparability, we found participants of the personalized one had lower performance than those of the OSFA in the second assessment, and research shows performance might affect gamification's effect (Rodrigues, Toda, Oliveira, Palomino, Avila-Santos, & Isotani, 2021; Sanchez et al., 2020). If so, participants of the personalized condition would report lower motivation than those of the OSFA condition in that session. In contrast, our results show comparable motivation levels. Taken together, those results show personalized gamification provided motivation levels comparable to those of the OSFA approach even though participants experienced higher difficulty during the second session's task. On one hand, it might be interpreted that personalized gamification contributed to student motivation by making it less sensitive to their performances, and not by increasing. Past research (Sanchez et al., 2020) argues gamification might distract students with low knowledge, based on research about seductive details (Rey, 2012). For instance, that could translate to complicating students' performances during the assessment and, consequently, affecting their motivations negatively. Here, personalization might have addressed those distractions by offering game elements suitable to the student and the task. Another rationale is that OSFA gamification's benefits for students with low initial knowledge decrease over time (Rodrigues, Toda, Oliveira, Palomino, Avila-Santos, & Isotani, 2021). Then, personalization might have addressed that time effect on low-knowledge students. Thus, by preventing distractions and avoiding time effects for low-knowledge students, one might assume that:

- **TH1:** Personalization makes user motivation less sensitive to user performance.

On the other hand, it might be that personalization increased student motivation after a first-time experience (i.e., at session 1), but participants' lower performance decreased it, which prevented any differences from appearing. That assumption builds upon the lack of longitudinal studies evaluating the personalization effect. Scholars advocate OSFA gamification suffers from the novelty effect (Bai et al., 2020; Koivisto & Hamari, 2014). For personalized gamification, however, prior research has only started to address that question (see Section 2.4). Most related work compared OSFA and personalized gamification through cross-sectional studies (Hajarian et al., 2019; Lopez & Tucker, 2021; Mora et al., 2018; Oliveira et al., 2020), while only Rodrigues, Palomino, et al. (2021) used a repeated-measures design. However, that study is limited to two measurements with a one-day spacing. While our study also captured two measurements, spacing varied between four to six weeks. Performance differences, however, limited our findings' contribution to understanding how personalization's effect change over time. Thus, while personalization might mitigate the novelty effect's impact on gamification, regardless of students knowledge level, empirical research is needed to test TH2:

- **TH2:** Personalization increases user motivation after a first-time experience.

6.2 How do students perceive gamified review assessments?

Qualitatively analyzing students open-text comments revealed two themes (Table 6). Those showed the large majority of students had positive experiences while doing the assessments, besides considering them well designed and good for learning. Themes also concerned a few comments contenting about the assessments' complexity and presentation and that gamification demands improvement. That is, students perceived the gamified assessments positively, but noted room for improving gamification as well as the educational content.

On one hand, these results extend past research on the testing effect; that is, the value of assessments, such as quizzes and tests, to learning (Batsell Jr et al., 2017; Roediger-III & Karpicke, 2006). Specifically, our analysis showed that students considered the activity good for their learning process and that they considered it well designed in the sense of approaching a perspective rarely explored in their studies; that is, taking time to review theoretical/conceptual aspects of computing education. Hence, demonstrating that students understand and value gamified review assessments focused on theories and concepts. Empirical evidence shows the testing effect improves learning, overall (Adesope et al., 2017; Rowland, 2014) and in gamified settings (Rodrigues, Toda, Oliveira, Palomino, Avila-Santos, & Isotani, 2021; Sanchez et al., 2020). However, those gamification studies have not inspected users' perceptions about the learning activities, which is important because those and other educational tasks are not motivating oftentimes (Hanus & Fox, 2015; Palomino et al., 2020; Pintrich, 2003). Thereby, we expand the literature with results that encourage the use of gamified review assessments to explore the testing effect in practice while providing overall positive experiences to students.

On the other hand, the results corroborate gamification literature by showing it is mostly perceived positively, but not always (Bai et al., 2020; Huang et al., 2020; Hyrynsalmi et al., 2017; Sailer & Homner, 2020; Toda et al., 2018). Prior research demonstrating different people are motivated by different game elements (e.g., Bovermann and Bastiaens (2020); Jia, Xu, Karanam, and Voida (2016); Orji, Tondello, and Nacke (2018); Tondello, Mora, and Nacke (2017)) corroborate those results. Consequently, pointing to the need for improving the personalization strategy applied. A possible reason is that we limited gamification to feature three game elements, and the literature discusses that number predetermines gamification's effectiveness (Landers et al., 2015). Another possible explanation is that our personalization mechanism was *changing the game elements available*, whereas some comments suggested *changing how* *game elements work*. While the latter approach has been discussed, the former is researched more often (Hallifax, Serna, et al., 2019; Rodrigues, Toda, Palomino, et al., 2020). A third perspective is that the recommendations on how to personalize (i.e., which game elements to use when) demands refinement to model users/tasks better. While we further inspect that latter perspective through RQ2 and RQ3, the other rationales we discussed suggest:

- **TH3:** Designs with more than three game elements improve users' perceptions about gamification.
- **TH4:** Successful personalization of gamification requires tailoring the mechanics of game elements as well as which of them should be available.

6.3 Which factors moderate personalization's effect?

Exploratory analyses indicated four moderators of personalized gamification's effect: gender, education, preferred game genre, and preferred playing setting (Tables 7 and 8). Because we considered those factors in defining personalized designs (Rodrigues, Toda, Oliveira, Palomino, Vassileva, & Isotani, 2021), a straightforward expectation, which our findings confront, is finding no moderator effect from them. That contrast is somewhat expected, however. Research on OSFA gamification shows several factors (e.g., gender) moderate its effectiveness (Bai et al., 2020; Huang et al., 2020; Sailer & Homner, 2020), even though scholars have striven to develop methods for designing it over the last decade (Mora, Riera, Gonzalez, & Arnedo-Moreno, 2015). Accordingly, one should expect the need for updating such models as they are empirically tested, as with every theory (Landers et al., 2018). Therefore, we discuss two research lines to explain moderators of personalization's effect.

First, how game elements are recommended depending on those moderators' levels. For instance, results indicate personalization mitigated females' amotivation, but did not work for males. The strategy we used (Rodrigues, Toda, Oliveira, Palomino, Vassileva, & Isotani, 2021) does not consider gender for selecting game elements within contexts such as the one of this study: Brazilian students doing remembering learning activities. The same happens for education and preferred playing setting. Differently, preferred game genre is one of the most influential factors. However, the strategy simplifies that factor to either one prefers action genre or not, which might explain the preferred game genre's moderator role. Thereby, further modeling those factors is likely to mitigate moderations and, consequently, improve personalization's effects.

- **TH5:** Further modeling user demographics and gaming-related preferences will improve personalization's effect.

Another possible reason is that we used a preference-based personalization strategy. Despite that approach is the most used in personalized gamification research (Klock et al., 2020; Rodrigues, Toda, Palomino, et al., 2020), the literature advocates user preference often fails to reflect user behavior (Norman, 2004). That is, what users *tell* they prefer might not be what motivates them the most in practice. To face that issue, researchers started to investigate data-driven personalization strategies (e.g., Hajarian et al. (2019)). That is, inspecting user behavior to determine the most suitable game elements (Tondello, Orji, & Nacke, 2017). Therefore, assuming that by relying on interaction data, they will reliably identify users' preferences and, consequently, improve gamification's effectiveness. Thus, given the limitations from preference-based strategies, one might expect that:

- **TH6:** Data-driven personalization strategies are more effective than preference-based ones.

Nevertheless, one must note preference- and data-driven strategies are complementary approaches. The latter is based on true usage, but is challenging in terms of data collection and inferring preference for specific game elements while several of them are available in the system, simultaneously affecting user experience. In contrast, preference captures user perceptions about each game element, is cheaper to collect, and allows personalizing since the first usage. However, those are limited by preference reliability. Thus, each approach has its advantages and drawbacks, and should be used/researched accordingly.

6.4 How do user motivation variation change when comparing OSFA and personalized gamification?

Exploratory analyses revealed that motivation from using OSFA gamification varied according to six characteristics (Tables 7 and 8). Those are expected considering that the OSFA design does not consider any user or contextual characteristics. Prior research has shown substantial homogeneity of gamification's outcomes in terms of user-to-user variation (Rodrigues, Toda, Oliveira, et al., 2020; Van Roy & Zaman, 2018) and other characteristics (Bai et al., 2020; Huang et al., 2020; Sailer & Homner, 2020). Thereby, our findings corroborate the overall gamification literature, showing gamified assessments' outcomes vary depending on the following user and contextual characteristics: performance, preferred game genre, preferred playing setting, education, assessment's subject, and usage interval. Differently, motivation from users of the personalized condition did not vary due to performance, preferred game genre, preferred playing setting, and usage interval, but similarly varied according to assessments' subject and education. We personalized gamification to eight dimensions (involving user and contextual information; see Section 4.3), including preferred game genre and playing setting. Previous studies show that user preference and motivation change depending on their characteristics (Bovermann & Bastiaens, 2020; Jia et al., 2016; Orji et al., 2018; Tondello, Mora, & Nacke, 2017). Thereby, one might expect reduced variation in outcomes if gamification designs provide game elements suitable to everyone's preferences/motivations (Hallifax, Serna, et al., 2019; Klock et al., 2020; Rodrigues, Toda, Palomino, et al., 2020). Then, the rationale is that providing personalized game elements made motivation less sensible to performance and reuse issues. Thus:

- **TH7:** Personalization mitigates OSFA outcomes' sensibility to user and contextual factors.

Nevertheless, personalization did not tackle variations from education and assessments subjects. Moreover, the motivation from personalization users varied due to gender and age. Assessments' subject and age are not considered by the personalization strategy we used. Hence, those are likely explained by not considering them during the personalization. Additionally, while the strategy considers gender and education, it assumes those factors do not play a role in game elements selection for Brazilians doing remembering learning activities (Rodrigues, Toda, Oliveira, Palomino, Vassileva, & Isotani, 2021). Hence, the rationale for such findings might be that gender and education were not completely modeled by the personalization strategy in Rodrigues, Toda, Oliveira, Palomino, Vassileva, and Isotani (2021). Thus:

- **TH8:** Gender and Education require further modeling to properly suggest the most suitable game elements.

7 Implications

This section discusses our findings' implications to design and theory.

7.1 Design Contributions

First, our findings inform the design of gamified education system on **how personalization contributes to gamification**. We investigated personalization's role in making student motivation less sensitive to user (e.g., gaming-preferences) and situational (e.g., performance) characteristics, and the results suggested personalization tackled motivation variations compared to the OSFA approach. Hence, designers might not see direct effects, such as increased motivation, but personalization might be acting by minimizing the extent to which one user group benefits more/less than others. Such findings expand the literature, which shows gamification's effect varies from person-to-person and context-to-context (Hallifax, Audrey, et al., 2019; Rodrigues, Toda, Oliveira, et al., 2020; Van Roy & Zaman, 2018), but is limited to testing whether, and not how personalization works (see Section 2.4). Notice that employing the personalization strategy we did (Rodrigues, Toda, Oliveira, Palomino, Vassileva, & Isotani, 2021) only requires turning on/off some game elements. Thus, **designers might use multidimensional personalization to offer more even experiences to their systems' users**.

Second, our findings provide considerations on **how to design personalized gamification**. We found some students would like more game elements or that the game elements behaved differently (i.e., having other mechanics). On one hand, we personalized by changing game elements available considering that strategy has received more literature support than changing mechanics (Hallifax, Serna, et al., 2019; Rodrigues, Toda, Palomino, et al., 2020). On the other hand, we defined all designs would have three game elements to ensure comparability with the number of game elements of the OSFA condition. However, student comments question whether those are the best choices for deploying and personalizing gamified designs. Therefore, contributing considerations that **gamified designs with more than three game elements might improve users' perceptions about gamification** (TH3) and that **successful personalization of gamification requires tailoring the game elements' mechanics as well as their availability** (TH4).

Third, our results inform instructors on the **design of learning assessments**. We found most students perceived the gamified review assessments positively, often mentioning that they had positive experiences during the task and that completing the assessments contributed to their learning. That is important because educational activities are not motivating for students oftentimes, and low motivation harms learning performance (Hanus & Fox, 2015; Palomino et al., 2020; Pintrich, 2003; Rodrigues, Toda, Oliveira, Palomino, Avila-Santos, & Isotani, 2021; Vogel et al., 2018). Those findings extent gamification studies' results to the context of review assessments, contributing to research on how to improve engagement with assessments through positive experiences. Thus, informing designers how they might successfully use such learning activities in practice, considering that **gamified review assessments are valued and positively perceived by students**.

7.2 Theoretical Contributions

Our first theoretical contribution, similar to the first design one, relates to **how personalization contributes to gamification**. Empirical studies comparing personalized and OSFA gamification (Hajarian et al., 2019; Lopez & Tucker, 2021; Rodrigues, Palomino, et al., 2021) are limited to testing whether the former improves the latter in terms of increased outcomes (e.g., motivation/performance). Differently, our findings suggested personalization improved gamification by offering more even experiences for the different user groups, instead of increasing the outcome's average. Thus, contributing to researchers the question of **what is the exact mechanism through which personalization contributes to gamification?**. In exploring answers to that question, our findings led to considerations suggesting that **either personalization mitigates the sensibility of OSFA gamification's outcomes to user and contextual factors** (TH1 and TH7) or **personalization increases user motivation after a first-time experience** (TH2) when samples' characteristics are comparable.

Second, our results inform researchers on **predeterminants of personalized gamification's success**. Because personalized gamification is a recent field study (Klock et al., 2020; Tondello, Orji, & Nacke, 2017), understanding predeterminants of its success is important to inform future research efforts (Koivisto & Hamari, 2019). Nevertheless, related work has not inspected, for instance, when or to whom personalization strategies worked more or less. This study addressed that gap through exploratory analyses that indicated when/to whom personalization was more or less effective (Tables 7 and 8). Hence, advancing our understanding of moderators of personalization's effect and subgroups in which personalized gamification offered distinct motivation levels. Consequently, those findings provide theoretical considerations that the **personalization strategy from** Rodrigues, Toda, Oliveira, Palomino, Vassileva, and Isotani (2021) **might benefit from considering other information and further modeling some characteristics it already considers** (i.e., TH5 and TH8).

Third, our analyses revealed a theoretical consideration on **how to develop personalization strategies**. Preference-based strategies allow personalizing systems from user's first use, which is hard for data-driven approaches. Differently, data-driven strategies rely on true usage while preference-based ones are based on intentions. While both approaches are complementary, the limitations from user preference and evidence from Hajarian et al. (2019) supporting data-driven personalization effectiveness led to the theoretical consideration that **data-driven personalization strategies are more effective than preference-based ones** (TH6).

Lastly, **we share our data and materials**. That complies with open science guidelines and literature recommendations toward mitigating the replication crisis in computer science (Cockburn, Dragicevic, Besançon, & Gutwin, 2020; Vornhagen et al., 2020). Thus, extending our contribution.

8 Limitations and Future Work

This section discusses study limitations and presents future research directions accordingly.

First, our 58-participants sample size is not far but below related works' median. That might be partly attributed to attrition, which is common for longitudinal designs and caused a loss of 38% of the eligible participants. Because this study was conducted during the COVID-19 pandemic, instructors mentioned they are witnessing unseen drop-out rates, which might explain the attrition rate. While that size affects findings' generalization, we believe that having conducted a multi-site study in ecological settings leads to a positive trade-off. Nevertheless, sample size directly affects our confirmatory analyses' validity as it implies low statistical power. We sought to mitigate that issue by only conducting planned comparisons, but that limitation plays a larger role in exploratory analyses because they rely on sub-samples. Then, we adopted recommendations of using 90% CIs measured through bootstrap to increase results' reliability while avoiding misleading conclusions that could emerge from p-values (Cairns, 2019; Carpenter & Bithell, 2000; Vornhagen et al., 2020). Thus, the limited statistical power suggests the need for similar studies, especially considering our findings confront those of similar research (Rodrigues, Palomino, et al., 2021), and the hypotheses that emerged from our exploratory analyses provide directions for future studies.

Second, students of the same class used a new system wherein the gamification design varied from one to another. We informed participants they would use different gamification designs, but they did not know PBL was the control group. Therefore, we believe contamination did not substantially affect our results. Moreover, all participants had never used Eagle-Edu, and they only used it twice during the study. Consequently, there is no evidence on how participants' motivations would change when completing gamified review assessments over multiple terms. Additionally, how students' motivations would behave in the context of more regular uses, such as reviewing each class' contents, remains an open question. Based on those, we recommend future studies to analyze how personalized and OSFA gamification compare, in the context of review assessments, when used across multiple terms as well as more frequently.

Third, there are three limitations related to our study's instruments/apparatus. Concerning our measure, the external regulation construct showed questionable reliability, similar to results from other studies (Gamboa et al., 2013; Guay et al., 2000; Rodrigues, Palomino, et al., 2021). Aiming to mitigate that limitation, we controlled for between participants variations in our quantitative analyses, despite some argue such issue might not be pertinent to HCI research (Cairns, 2019). Concerning Eagle-Edu, we needed to create one course shell for each gamification design. Consequently, some of those had few students. That technical issue affected designs featuring leaderboards, leading to cases wherein students had few peers to compete against. Concerning the personalization strategy, it was developed based on user preference, which is often criticized (Norman, 2004) compared to data-driven approaches (Khoshkangini, Valetto, & Marconi, 2017; Van Houdt et al., 2020), and was only initially validated compared to OSFA gamification (Rodrigues, Palomino, et al., 2021). In summary, those limitations suggest the need for i) further inspecting SIMS' validity in the context of learning assessments, ii) explicitly studying how *small* competition affects students' motivations, and iii) extending the external validity of the personalization strategy introduced in Rodrigues, Toda, Oliveira, Palomino, Vassileva, and Isotani (2021).

9 Conclusion

Educational systems play a crucial role in enabling assessment for learning. That approach has strong support for its positive effect on learning gains, but is not motivating for students oftentimes. While standard, one-size-fits-all (OSFA) gamification can improve motivation, effects' variation inspired research on personalizing gamified designs. However, there is little knowledge on how personalization contributes to OSFA gamification. Therefore, we conducted a multi-site experimental study wherein students completed gamified assessments with either personalized or OSFA gamification. Our results reveal a new way of seeing personalization's role in gamification and inform designers, instructors, and researchers:

- We show whereas personalization might not increase outcome's average, it likely improves gamification by reducing its outcome's variation;

- We show gamified review assessments provide positive experiences considered good learning means from students' perspectives;
- Our discussions provide design and research directions towards advancing the field study;

Our results inform i) designers interested in personalized gamification, showing what benefits to expect from it; ii) instructors using interactive systems to deploy assessments for learning on the value of gamifying them; and iii) personalized gamification researchers with guidance on how to advance the field study. Additionally, we extend our contribution by sharing our data and materials.

Competing Interests

The authors have no relevant financial or non-financial interests to disclose.

References

Adesope, O.O., Trevisan, D.A., Sundararajan, N. (2017). Rethinking the use of tests: A meta-analysis of practice testing. *Review of Educational Research*, *87*(3), 659–701.

Altmeyer, M., Lessel, P., Muller, L., Krüger, A. (2019). Combining behavior change intentions and user types to select suitable gamification elements for persuasive fitness systems. *International conference on persuasive technology* (pp. 337–349).

Armstrong, R.A. (2014). When to use the b onferroni correction. *Ophthalmic and Physiological Optics*, *34*(5), 502–508.

Bai, S., Hew, K.F., Huang, B. (2020). Is gamification "bullshit"? evidence from a meta-analysis and synthesis of qualitative data in educational contexts. *Educational Research Review*, 100322.

Baldeón, J., Rodríguez, I., Puig, A. (2016). Lega: A learner-centered gamification design framework. *Proceedings of the xvii international conference on human computer interaction* (pp. 45:1–45:8). New York, NY, USA: ACM. Retrieved from http://doi.acm.org/10.1145/2998626.2998673 10.1145/2998626.2998673

Batsell Jr, W.R., Perry, J.L., Hanley, E., Hostetter, A.B. (2017). Ecological validity of the testing effect: The use of daily quizzes in introductory psychology. *Teaching of Psychology*, *44*(1), 18–23.

Blandford, A., Furniss, D., Makri, S. (2016). Qualitative hci research: Going behind the scenes. *Synthesis lectures on human-centered informatics*, *9*(1), 1–115.

Bloom, B.S. (1956). *Taxonomy of educational objectives: The classification of educational goals.* New York: Longman.

Bovermann, K., & Bastiaens, T.J. (2020). Towards a motivational design? connecting gamification user types and online learning activities. *Research and Practice in Technology Enhanced Learning*, *15*(1), 1-18. Retrieved from 10.1186/s41039-019-0121-4

Braun, V., & Clarke, V. (2006). Using thematic analysis in psychology. *Qualitative research in psychology*, *3*(2), 77–101.

Cairns, P. (2019). *Doing better statistics in human-computer interaction.* Cambridge University Press.

Canty, A., & Ripley, B.D. (2021). boot: Bootstrap r (s-plus) functions [Computer software manual]. (R package version 1.3-28)

Carpenter, J., & Bithell, J. (2000). Bootstrap confidence intervals: when, which, what? a practical guide for medical statisticians. *Statistics in medicine*, *19*(9), 1141–1164.

Cockburn, A., Dragicevic, P., Besançon, L., Gutwin, C. (2020). Threats of a replication crisis in empirical computer science. *Communications of the ACM*, *63*(8), 70–79.

Dash, S. (2019). Google classroom as a learning management system to teach biochemistry in a medical school. *Biochemistry and molecular biology education*, *47*(4), 404–407.

Deci, E.L., & Ryan, R.M. (2000). The "what" and "why" of goal pursuits: Human needs and the self-determination of behavior. *Psychological Inquiry*, *11*(4), 227-268. Retrieved from 10.1207/S15327965PLI1104_01

Deterding, S., Dixon, D., Khaled, R., Nacke, L. (2011). From game design elements to gamefulness: defining gamification. *Proceedings of the 15th international academic mindtrek conference: Envisioning future media* *environments* (pp. 9–15). 10.1145/2181037.2181040

Dichev, C., & Dicheva, D. (2017). Gamifying education: what is known, what is believed and what remains uncertain: a critical review. *International journal of educational technology in higher education*, *14*(1), 9.

Gamboa, V., Valadas, S.T., Paixão, O. (2013). Validação da versão portuguesa da situational motivation scale (sims) em contextos académicos. *Atas do xii congresso galego-português de psicopedagogia, 11-13 de setembro de 2013* (pp. 4868–4882).

Gliem, J.A., & Gliem, R.R. (2003). Calculating, interpreting, and reporting cronbach's alpha reliability coefficient for likert-type scales..

Guay, F., Vallerand, R.J., Blanchard, C. (2000). On the assessment of situational intrinsic and extrinsic motivation: The situational motivation scale (sims). *Motivation and emotion*, *24*(3), 175–213.

Hajarian, M., Bastanfard, A., Mohammadzadeh, J., Khalilian, M. (2019). A personalized gamification method for increasing user engagement in social networks. *Social Network Analysis and Mining*, *9*(1), 47.

Hallifax, S., Audrey, S., Jean-Charles, M., Guillaume, L., Elise, L. (2019, October). Factors to Consider for Tailored Gamification. *CHI Play* (p. 559–572). Barcelona, Spain: ACM. Retrieved from https://hal.archives-ouvertes.fr/hal-02185647

Hallifax, S., Serna, A., Marty, J.-C., Lavoué, É. (2019). Adaptive gamification in education: A literature review of current trends and developments. M. Scheffel, J. Broisin, V. Pammer-Schindler, A. Ioannou, & J. Schneider (Eds.), *Transforming learning with meaningful technologies* (pp. 294–307). Cham: Springer International Publishing. 10.1007/978-3-030-29736-7_22

Hamari, J., Koivisto, J., Sarsa, H. (2014). Does gamification work?-a literature review of empirical studies on gamification. *Hicss* (Vol. 14, pp. 3025–3034).

Hanus, M.D., & Fox, J. (2015). Assessing the effects of gamification in the classroom: A longitudinal study on intrinsic motivation, social comparison, satisfaction, effort, and academic performance. *Computers & Education*, *80*, 152 - 161. Retrieved from http://www.sciencedirect.com/science/article/pii/S0360131514002000 https://doi.org/10.1016/j.compedu.2014.08.019

Hox, J.J., Moerbeek, M., Van de Schoot, R. (2010). *Multilevel analysis: Techniques and applications.* Routledge.

Huang, R., Ritzhaupt, A.D., Sommer, M., Zhu, J., Stephen, A., Valle, N., ... Li, J. (2020). The impact of gamification in educational settings on student learning outcomes: a meta-analysis. *Educational Technology Research and Development*, 1–27.

Hyrynsalmi, S., Smed, J., Kimppa, K. (2017). The dark side of gamification: How we should stop worrying and study also the negative impacts of bringing game design elements to everywhere. *Gamifin* (pp. 96–104).

Jia, Y., Xu, B., Karanam, Y., Voida, S. (2016). Personality-targeted gamification: a survey study on personality traits and motivational affordances. *Proceedings of the 2016 chi conference on human factors in computing systems* (pp. 2001–2013).

Khoshkangini, R., Valetto, G., Marconi, A. (2017). Generating personalized challenges to enhance the persuasive power of gamification. *Personalization in persuasive technology workshop.*

Klock, A.C.T., Gasparini, I., Pimenta, M.S., Hamari, J. (2020). Tailored gamification: A review of literature. *International Journal of Human-Computer Studies*, 102495.

Kocadere, S.A., & Çağlar, Ş. (2015). The design and implementation of a gamified assessment. *Journal of e-Learning and Knowledge Society*, *11*(3).

Koivisto, J., & Hamari, J. (2014). Demographic differences in perceived benefits from gamification. *Computers in Human Behavior*, *35*, 179 - 188. Retrieved from http://www.sciencedirect.com/science/article/pii/S0747563214001289

Koivisto, J., & Hamari, J. (2019). The rise of motivational information systems: A review of gamification research. *International Journal of Information Management*, *45*, 191–210. Retrieved from 10.1016/j.ijinfomgt.2018.10.013

Krathwohl, D.R. (2002). A revision of bloom's taxonomy: An overview. *Theory into practice*, *41*(4), 212–218. Retrieved from 10.1207/s15430421tip4104$_2$

Landers, R.N., Auer, E.M., Collmus, A.B., Armstrong, M.B. (2018). Gamification science, its history and future: Definitions and a research agenda. *Simulation & Gaming*, *49*(3), 315–337.

Landers, R.N., Bauer, K.N., Callan, R.C., Armstrong, M.B. (2015). Psychological theory and the gamification of learning. In T. Reiners & L.C. Wood (Eds.), *Gamification in education and business* (pp. 165–186). Cham: Springer International Publishing. Retrieved from 10.1007/978-3-319-10208-5_9

Lavoué, E., Monterrat, B., Desmarais, M., George, S. (2018). Adaptive gamification for learning environments. *IEEE Transactions on Learning Technologies*, *12*(1), 16–28.

Lessel, P., Altmeyer, M., Müller, M., Wolff, C., Krüger, A. (2017). Measuring the effect of" bottom-up" gamification in a microtask setting. *Proceedings of the 21st international academic mindtrek conference* (pp. 63–72).

Liu, D., Santhanam, R., Webster, J. (2017). Toward meaningful engagement: A framework for design and research of gamified information systems. *MIS quarterly*, *41*(4), 1011-1034.

Lopez, C.E., & Tucker, C.S. (2021). Adaptive gamification and its impact on performance. *International conference on human-computer interaction* (pp. 327–341).

Loughrey, K., & Broin, D. (2018). Are we having fun yet? misapplying motivation to gamification. *2018 ieee games, entertainment, media conference (gem)* (pp. 1–9). 10.1109/GEM.2018.8516535

Mair, P., & Wilcox, R. (2018). Robust statistical methods using wrs2. *The WRS2 Package*.

Mora, A., Riera, D., Gonzalez, C., Arnedo-Moreno, J. (2015). A literature review of gamification design frameworks. *2015 7th international conference on games and virtual worlds for serious applications (vs-games)* (pp. 1–8).

Mora, A., Tondello, G.F., Nacke, L.E., Arnedo-Moreno, J. (2018, April). Effect of personalized gameful design on student engagement. *2018 ieee global engineering education conference (educon)* (p. 1925-1933). 10.1109/EDUCON.2018.8363471

Mpungose, C.B. (2020). Emergent transition from face-to-face to online learning in a south african university in the context of the coronavirus pandemic. *Humanities and Social Sciences Communications*, *7*(1), 1–9.

Norman, D.A. (2004). *Emotional design: Why we love (or hate) everyday things.* Basic Civitas Books.

Oliveira, W., Toda, A., Toledo, P., Shi, L., Vassileva, J., Bittencourt, I.I., Isotani, S. (2020). Does tailoring gamified educational systems matter? the impact on students' flow experience. *Proceedings of the 53rd hawaii international conference on system sciences* (p. 1226-1235). ScholarSpace.

Orji, R., Tondello, G.F., Nacke, L.E. (2018). Personalizing persuasive strategies in gameful systems to gamification user types. *Proceedings of the 2018 chi conference on human factors in computing systems* (p. 435). 10.1145/3173574.3174009

Palomino, P., Toda, A., Rodrigues, L., Oliveira, W., Isotani, S. (2020). From the lack of engagement to motivation: Gamification strategies to enhance users learning experiences. *19th brazilian symposium on computer games and digital entertainment (sbgames)-grandgames br forum* (p. 1127-1130).

Pedro, L.Z., Lopes, A.M., Prates, B.G., Vassileva, J., Isotani, S. (2015). Does gamification work for boys and girls? an exploratory study with a virtual learning environment. *Proceedings of the 30th annual acm symposium on applied computing* (pp. 214–219).

Pereira, F.D., Fonseca, S.C., Oliveira, E.H., Cristea, A.I., Bellhäuser, H., Rodrigues, L., ... Carvalho, L.S. (2021). Explaining individual and collective programming students' behaviour by interpreting a black-box predictive model. *IEEE Access*.

Pereira, F.D., Oliveira, E.H., Oliveira, D.B., Cristea, A.I., Carvalho, L.S., Fonseca, S.C., ... Isotani, S. (2020). Using learning analytics in the amazonas: understanding students' behaviour in introductory programming. *British journal of educational technology*, *51*(4), 955–972.

Pintrich, P.R. (2003). A motivational science perspective on the role of student motivation in learning and teaching contexts. *Journal of educational Psychology*, *95*(4), 667.

Polo-Peña, A.I., Frías-Jamilena, D.M., Fernández-Ruano, M.L. (2020). Influence of gamification on perceived self-efficacy: gender and age moderator effect. *International Journal of Sports Marketing and Sponsorship*.

Recabarren, M., Corvalán, B., Villegas, M. (2021). Exploring the differences between gamer and non-gamer students in the effects of gamification on their motivation and learning. *Interactive Learning Environments*, 1–14.

Rey, G.D. (2012). A review of research and a meta-analysis of the seductive detail effect. *Educational Research Review*, *7*(3), 216–237.

Rodrigues, L., Oliveira, W., Toda, A., Palomino, P., Isotani, S. (2019). Thinking inside the box: How to tailor gamified educational systems based on learning activities types. *Proceedings of the brazilian symposium of computers on education* (p. 823-832). SBC. 10.5753/cbie.sbie.2019.823

Rodrigues, L., Palomino, P.T., Toda, A.M., Klock, A.C.T., Oliveira, W., Avila-Santos, A.P., ... Isotani, S. (2021, oct). Personalization improves gamification: Evidence from a mixed-methods study. *Proc. ACM Hum.-Comput. Interact.*, *5*(CHI PLAY). Retrieved from 10.1145/3474714

Rodrigues, L., Toda, A.M., Oliveira, W., Palomino, P.T., Avila-Santos, A.P., Isotani, S. (2021). Gamification works, but how and to whom? an experimental study in the context of programming lessons. *Proceedings of the 52nd acm technical symposium on computer science education* (pp. 184–190).

Rodrigues, L., Toda, A.M., Oliveira, W., Palomino, P.T., Isotani, S. (2020). Just beat it: Exploring the influences of competition and task-related factors in gamified learning environments. *Anais do xxxi simpósio brasileiro de informática na educação* (pp. 461–470).

Rodrigues, L., Toda, A.M., Oliveira, W., Palomino, P.T., Vassileva, J., Isotani, S. (2021). *Automating gamification personalization: To the user and beyond.*

Rodrigues, L., Toda, A.M., Palomino, P.T., Oliveira, W., Isotani, S. (2020, oct). Personalized gamification: A literature review of outcomes, experiments, and approaches. F.J. García-Peñalvo (Ed.), *Proceedings of the 8th international conference on technological ecosystems for enhancing multiculturality (teem 2020) (salamanca, spain, october 21-23, 2020).*

Roediger-III, H.L., & Karpicke, J.D. (2006). Test-enhanced learning: Taking memory tests improves long-term retention. *Psychological science*, *17*(3), 249–255.

Roosta, F., Taghiyareh, F., Mosharraf, M. (2016). Personalization of gamification-elements in an e-learning environment based on learners' motivation. *2016 8th international symposium on telecommunications (ist)* (pp. 637–642). 10.1109/ISTEL.2016.7881899

Rowland, C.A. (2014). The effect of testing versus restudy on retention: a meta-analytic review of the testing effect. *Psychological Bulletin*, *140*(6), 1432.

Sailer, M., & Homner, L. (2020). The gamification of learning: a meta-analysis. *Educ Psychol Rev*, *32*, 77-112. Retrieved from 10.1007/s10648-019-09498-w

Sanchez, D.R., Langer, M., Kaur, R. (2020). Gamification in the classroom: Examining the impact of gamified quizzes on student learning. *Computers & Education*, *144*, 103666. Retrieved from 10.1016/j.compedu.2019.103666

Schubhan, M., Altmeyer, M., Buchheit, D., Lessel, P. (2020). Investigating user-created gamification in an image tagging task. *Proceedings of the 2020 chi conference on human factors in computing systems* (pp. 1–12).

Seaborn, K., & Fels, D.I. (2015). Gamification in theory and action: A survey. *International Journal of human-computer studies*, *74*, 14–31. Retrieved from 10.1016/j.ijhcs.2014.09.006

Stuart, H., Lavoué, E., Serna, A. (2020). To tailor or not to tailor gamification? an analysis of the impact of tailored game elements on learners' behaviours and motivation. *21th international conference on artificial intelligence in education.*

Toda, A.M., do Carmo, R.M., da Silva, A.P., Bittencourt, I.I., Isotani, S. (2019). An approach for planning and deploying gamification concepts with social networks within educational contexts. *International Journal of Information Management*, *46*, 294 - 303. Retrieved from 10.1016/j.ijinfomgt.2018.10.001

Toda, A.M., Klock, A.C., Oliveira, W., Palomino, P.T., Rodrigues, L., Shi, L., ... Cristea, A.I. (2019). Analysing gamification elements in educational environments using an existing gamification taxonomy. *Smart Learning Environments*, *6*(1), 16. Retrieved from 10.1186/s40561-019-0106-1

Toda, A.M., Valle, P.H.D., Isotani, S. (2018). The dark side of gamification: An overview of negative effects of gamification in education. A.I. Cristea, I.I. Bittencourt, & F. Lima (Eds.), *Higher education for all. from challenges to novel technology-enhanced solutions* (pp. 143–156). Cham: Springer International Publishing. 10.1007/978-3-319-97934-2_9

Tondello, G.F. (2019). *Dynamic personalization of gameful interactive systems* (Unpublished doctoral dissertation). University of Waterloo.

Tondello, G.F., Mora, A., Nacke, L.E. (2017). Elements of gameful design emerging from user preferences. *Proceedings of the annual symposium on computer-human interaction in play* (pp. 129–142). 10.1145/3116595.3116627

Tondello, G.F., & Nacke, L.E. (2020). Validation of user preferences and effects of personalized gamification on task performance. *Frontiers in Computer Science*, *2*, 29.

Tondello, G.F., Orji, R., Nacke, L.E. (2017). Recommender systems for personalized gamification. *Adjunct publication of the 25th conference on user modeling, adaptation and personalization* (pp. 425–430). 10.1145/3099023.3099114

Van Houdt, L., Millecamp, M., Verbert, K., Vanden Abeele, V. (2020). Disambiguating preferences for gamification strategies to motivate pro-environmental behaviour. *Proceedings of the annual symposium on computer-human interaction in play* (pp. 241–253).

Van Roy, R., & Zaman, B. (2018). Need-supporting gamification in education: An assessment of motivational effects over time. *Computers & Education*, *127*, 283–297. Retrieved from 10.1016/j.compedu.2018.08.018

Vansteenkiste, M., Sierens, E., Soenens, B., Luyckx, K., Lens, W. (2009). Motivational profiles from a self-determination perspective: The quality of motivation matters. *Journal of educational psychology*, *101*(3), 671.

Vogel, C., Hochberg, J., Hackstein, S., Bockshecker, A., Bastiaens, T.J., Baumöl, U. (2018). Dropout in distance education and how to prevent it. *Edmedia+ innovate learning* (pp. 1788–1799).

Vornhagen, J.B., Tyack, A., Mekler, E.D. (2020). Statistical significance testing at chi play: Challenges and opportunities for more transparency. *Proceedings of the annual symposium on computer-human interaction in play* (pp. 4–18).

Wilcox, R.R. (2011). *Introduction to robust estimation and hypothesis testing.* Academic press.

Statements Declarations

Funding: This research received financial support from the following agencies: Brazilian National Council for Scientific and Technological Development (CNPq) - processes 141859/2019-9, 163932/2020-4, 308458/2020-6, 308395/2020-4, and 308513/2020-7; Coordination for the Improvement of Higher Education Personnel (CAPES) - Finance Code 001; and São Paulo State Research Support Foundation (FAPESP) - processes 2018/15917- 0 and 2013/07375-0. Additionally, this research was carried out within the scope of the Samsung-UFAM Project for Education and Research (SUPER), according to Article 48 of Decree nº 6.008/2006 (SUFRAMA), and partially funded by Samsung Electronics of Amazonia Ltda., under the terms of Federal Law nº 8.387/1991, through agreements 001/2020 and 003/2019, signed with Federal University of Amazonas and FAEPI, Brazil.

ANNEX

J

GARFIELD: A RECOMMENDER SYSTEM TO PERSONALIZE GAMIFIED LEARNING

GARFIELD: A Recommender System to Personalize Gamified Learning

No Author Given

No Institute Given

Abstract. Motivation is strongly related to learning. Nevertheless, students often lack it while working in educational activities. Gamification has the potential to increase (intrinsic and extrinsic) motivation in several contexts, including education. Unfortunately, it does not always work. According to recent findings, the main reason for gamification failing to achieve its full potential is due to poor design. Consequently, researchers have been developing strategies to improve gamification designs through personalization. To date, most personalization strategies for gamification are based on theoretical understanding of game elements and their impact on students without considering real interaction data. In this work, we proposed a novel approach to personalize gamification designs built upon data from real students' experiences with a learning environment and create the basis for a recommender system: GARFIELD - Gamification Automatic Recommender for Interactive Education and Learning Domains. We followed the CRISP-DM methodology to develop personalization strategies by collecting self-reports from Brazilian students (N = 221) who have used one out of our five gamification designs. Then, we regressed from such data to obtain recommendations of which design is the most suitable to achieve a desired motivation level. Using Cohen's Kappa, we found an agreement between GARFIELD's recommendations and the ground truth (K = 0.43). This value shows a moderate performance and demonstrates the potential of our approach. To the best of our knowledge, GARFIELD is the first model to guide practitioners and instructors on how to personalize gamification based on empirical data.

Keywords: Tailored gamification · Data-driven · Education · e-Learning

1 Introduction

Intrinsic motivation (IM) to study is central for learning: it has a strong, positive relationship with academic performance [15, 8] and is considered ideal for educational purposes [27]. In this regard, gamification (i.e., the use of game design elements in non-gaming contexts [4]) is one method with strong potential to improve motivational learning outcomes [19]. However, studies indicate that gamification's effects often varies from person to person [16, 26], leading to potential adverse effects (e.g., performance loss and addiction) for some people on some situations [22]. Research shows that if gamified designs are not tailored to users and contexts, they are likely to not achieve their full potential (e.g., [14]), which encourages new studies on how to tailor gamification [10].

Most often, gamification is tailored through personalization: designers or the system itself change the gamified design according to predefined information [25]. For instance, changing the game elements shown according to the learning task. However, personalization demands a user/task model, such as those developed in [23] and [17]. In common, those and other similar models are based on *potential* experiences: they were built from data captured through surveys or after seeing mock-ups [18]. Thereby, they are limited because *potential* experiences might not reflect *real* experiences [13]. For instance, [20] developed a model based on both learners' profiles and motivation before using the gamification, but with no information of learners' real experiences (e.g., after actually using gamification). Hence, to the best of our knowledge, there is no data-driven model, based on users' real (instead of potential) experiences, for personalizing gamification designs.

To address that gap, this paper presents GARFIELD - Gamification Automatic Recommender for Interactive Education and Learning Domains, a recommender system for personalizing gamification built upon data from real experiences. Our goal was to promote the most suitable gamification design according to students' intrinsic motivation because of its positive relationship with learning [27]. For this, we followed the CRISP-DM methodology [30], since it is suggested for goal-oriented projects [12], using a two-step reverse engineering approach. First, we collected self-reports of users' intrinsic motivations from *actually* using one out of five gamification designs. Then, we regressed from such data (N = 221) to obtain recommendations of which design is the most suitable to achieve a desired motivation level given the user's information. To the best of our knowledge, GARFIELD is the first model that guides practitioners and instructors on how to personalize gamification based on empirical data from *real* usage. Therefore, this paper contributes by creating and providing a motivation-based model for personalizing gamification. Thus, our results inform educators on how to personalize their gamified practices and researchers by performing a first step towards developing experience-driven models for designing gamification.

2 Method: CRISP-DM

Because we had an apriori goal, we followed the CRISP-DM reference model, which is suggested for goal-oriented projects [30]. Next, we introduce each CRISP-DM phase and describe how we implemented it. For transparency, our appendix shows all steps and their results.

CRISP-DM's first phase is **business understanding**. In this phase, we first defined the project's goal: creating a model based on students' intrinsic motivation captured after real system usage to allow the personalization of gamified educational systems. Additionally, we defined two requirements: i) the model must consider user characteristics and ii) the model must be interactive. The former is based on research showing users characteristics affect their experiences with gamified systems [15, 23]. The latter aims to facilitate practical usage.

The second phase is **data understanding**. It involves collection, quality analysis, and exploration of the dataset. Openly sharing data extends a paper's contribution because it enables cheaper, optimized exploratory analyses [29]. Those benefits are especially valuable for educational contexts wherein data collection is expensive. Accordingly, we opted to work with a dataset collected and made available by [removed for blind review]. This dataset has data from Brazilian students enrolled in STEM (Science, Technology, Engineering, and Mathematics) undergraduate courses of three northwestern universities. Those students self-reported their motivations after using a gamified system to complete lecture assessments of computing-related classes (e.g., Introductory and Object-oriented programming). These assessments had 30 multiple-choice questions organized into 10 three-question missions. Additionally, those were assessments *for* learning [6]. Hence, students could retry questions until getting them right. An ethical committee reviewed and approved this data collection [ID REMOVED FO REVIEW].

Importantly, each student interacted with one of five possible designs: i) points, acknowledgments, and competition (PBL)[1], ii) acknowledgments, objectives, and progression (AOP), iii) acknowledgments, objectives and social pressure (AOS), iv) acknowledgments, competition, and time pressure (ACT), and v) competition, chance, and time pressure (CCT). We selected those designs by convenience, as we used data shared by a previous study, which employed such designs aiming to tailor gamification to user characteristics and learning activity type to be done [removed for blind review].

Specifically, each game element functions as follows, but only when available. Students received *points* after completing any mission as a positive feedback because, according to the testing effect, the simple act of trying contributes to learning [6]. After each finished mission, they would be *acknowledged* with a badge depending on their performances (e.g., getting all items right). Also, students could *compete* with each other based on a leaderboard that ranked them based on the points they made during the week. With the leaderboard, a clock provoked *time pressure* by highlighting the time available to climb the leaderboard before the week's end. Additionally, a progress bar indicated student's *progression* within missions, a notification aimed to provoke *social pressure* by warning that peers just completed a mission, and a skill tree represented short-term *objectives* (i.e., completing each of the 10 missions).

In terms of quality, the data's ecological validity is high because it was collected during lectures. Additionally, intrinsic motivation reports have a high reliability (Cronbach's alpha = 0.9). Lastly, we conducted an initial, exploratory analysis of the dataset. Mainly, we identified some unbalanced columns (e.g., the number of observations per educational background). Additionally, we identified

[1] We consider Badges and Leaderboards implementations of Acknowledgments and Competition, respectively [21], but use PBL here to maintain the standard nomenclature

some extreme values likely to be outliers (e.g., people who reported to play 80 hours per week) and some observations from students that reported their motivation without having used the system. Then, aware of this aspects, we moved to the next phase.

The third phase is **data preparation**. In this stage, we first made the *attribute selection* task: we chose columns related to students' characteristics, intrinsic motivation, status, and the game elements they interacted with. Next, we proceeded to *data cleasing*. At first, we removed answers from students with less than 18 years (N = 1) due to ethical aspects. Then, we removed participants that provided their motivations but did not use the system (N = 4). Notice that some observations had extreme values likely to be outliers in terms of age and weekly playing time. However, we opted not to remove them because we had no reason to believe those values emerged from technical errors or typos. Accordingly, those are likely to be real students valuable to the analysis [1].

Subsequently, we started the *data transformations*. In this stage, we first transformed the intrinsic motivation variable. That was captured through a seven-point Likert-scale using the respective subscale of the Situational Motivation Scale (SIMS) [7]. Thereby, we changed its range from one to seven to zero to six, aiming to facilitate the interpretation of regression coefficients. Next, we examined categorical attributes' levels/values to identify those highly unbalanced. Then, for levels representing less than 5% of the dataset, we removed their observations unless grouping them with another level was feasible. We made this choice to avoid overfitting due to high unbalance [1]. Accordingly, observations that indicated FPS, MOBA, Platform, or Run & Gun as their preferred game genre were grouped into the Action, RPG, Adventure, and Action genres, respectively. Those who indicated Sports and "Action and Adventure"[2] were grouped into the Other category. These groupings were defined based on [5]. We also removed categories that represented less than 1% of the dataset (i.e., game genre Other (N = 3) and gender 'rather not say' (N = 5)) For education background, eight observations answered 'undergraduate student'. Because we are interested in students' highest degrees, we grouped them with the 'high school' level as such degree is needed to be an undergraduate student.

Finishing *data preparation*, we *constructed new attributes* because age and weekly playing time are continuous and highly skewed. Therefore, we decided to categorize them. For weekly playing time, we categorized it into whether the student plays an average of at least one hour per day (i.e., weekly playing time ≤ 7) or more than that (i.e., weekly playing time > 7). For age, we considered that Brazilian undergraduate STEM students have around 21 years [3]. Thereby, we categorized age into those below that average (i.e., < 21) and those at or above it (i.e., ≥ 21). Lastly, we analyzed the 'game elements' column. As this is our dependent variable, we analyzed it at last. In this regard, we found a single observation of the ACT design. Consequently, we removed it, leading to the prepared dataset featuring 221 observations (see Table 1 for a summary).

[2] We chose not to group this value with Action and/or Adventure to respect students' choice of not doing so.

Table 1. Descriptive statistics of the prepared dataset. Data are shown as *Mean (Standard Deviation)* for intrinsic motivation and *Count (%)* for the remainder.

	PBL	**AOP**	**AOS**	**CCT**
Intrinsic Motivation	4.70 (1.39)	4.57 (1.13)	4.48 (1.16)	4.64 (1.37)
Age				
Below 21	48 (66)	47 (65)	47 (85)	12 (57)
21 or more	25 (34)	25 (35)	08 (15)	09 (43)
Gender				
Male	30 (41)	22 (31)	14 (25)	00 (00)
Female	43 (59)	50 (69)	41 (75)	21 (100)
Educational Background				
High School	51 (70)	47 (65)	40 (73)	14 (67)
Technical	11 (15)	11 (15)	06 (11)	04 (19)
Undergraduation	11 (15)	14 (19)	09 (16)	03 (14)
Preferred Game Genre				
Action	31 (42)	04 (06)	20 (36)	20 (95)
Adventure	13 (18)	20 (28)	11 (20)	01 (05)
RPG	16 (22)	21 (29)	11 (20)	00 (00)
Strategy	13 (18)	27 (38)	13 (24)	00 (00)
Preferred Playing Setting				
Multiplayer	48 (66)	44 (61)	31 (56)	16 (76)
Singleplayer	25 (34)	28 (39)	24 (44)	05 (24)
Daily Playing Time				
Up to 1 hour	48 (66)	45 (62)	35 (64)	09 (43)
More than 1 hour	25 (34)	27 (38)	20 (36)	12 (57)

The fourth phase is **modeling**. Here, we worked with a regression modeling technique, which is a form o machine learning: the Multinomial Logistic Regression [11]. This technique enables working with a nominal dependent variable such as gamification designs. Additionally, it is based on the null hypothesis significance testing framework. Hence, it allows us to evaluate coefficients' contributions to the model based on their significance. Specifically, this technique works similarly to standard Logistic Regression. However, it compares the dependent variable's reference value to all other ones. That is, a *one versus all* approach. In our analysis, we defined the PBL design as the reference value. We made this choice because it is the most used gamification design in educational contexts [2]. As independent variables, we opted to start with all of those of the prepared dataset. Additionally, the model should consider how students' intrinsic motivation from using a gamification design change depending on their characteristics. Therefore, our model assumes intrinsic motivation interacts with all other variables. To perform this modeling, we used the *nnet* R package. We kept all parameters as standard but the maximum number of iterations. Instead, we set it to 1000 to ensure the algorithm's convergence. Note that R automatically dummy codes nominal variables with more than two levels. Hence, R transforms educational background and preferred game genre into two and three coefficients, respectively. Next, we present how we conducted the

CRISP-DM's fifth phase, *evaluation*. Then, we present the results of phases four and five in the next section.

In the fifth phase, one should **evaluate** modeling alternatives to determine the best option. To do so, we decided to use recursive feature elimination. Notice that we followed the standard of working within the null hypothesis significance testing framework. Hence, we used p-values as the criteria for feature elimination. However, this project has an exploratory nature. Thereby, we assumed a 90% confidence level, following similar research [15, 9]. Therefore, we would eliminate coefficients with p-values greater than 0.1. After selecting the final model, we evaluate it based on its predictions. In this case, we rely on Cohen's Kappa and F-measure, calculated using R packages *vcd* and *caret*, respectively. We use those metrics because they are reliable for multi-class problems wherein data is unbalanced. Lastly, note we discuss the CRISP-DM's sixth phase, **deployment**, in the next session, after introducing results of modeling and evaluation.

3 Results

After running the Multinomial Logistic Regression, we found significant interactions between all user's characteristics and intrinsic motivation. Consequently, we removed no features and defined the initial model as the final one. Then, in evaluating this model, we found the Cohen's Kappa for the agreement between its predictions and the ground truth is 0.43. This value is significantly different from zero ($p < 0.001$), with its 95% confidence interval ranging from 0.34 to 0.52. Thus, revealing a moderate agreement [28]. Furthermore, to help understand the model's predictions, Table 2 shows the confusion matrix along with the F-measure of each category. The table demonstrates the model performed the best for designs AOP and CCT. Differently, its performance for designs PBL and AOS were slightly worse. Additionally, the confusion matrix reveals the model's misclassifications (e.g., wrongly predicting AOS design should be PBL and AOP 13 and 18 times, respectively). Therefore, the evaluation phase shows the model recommends gamified designs with moderate performance, despite variations from one design to another. Thus, demonstrating the model's potential as well as room for improvement.

Table 2. Confusion Matrix of the models predictions against the ground truth.

	PBL	**AOP**	**AOS**	**CCT**	**Balanced Accuracy**	**F-measure**
PBL	40	11	13	04	0.68	0.57
AOP	17	54	18	01	0.75	0.67
AOS	14	06	24	02	0.65	0.48
CCT	02	01	00	14	0.83	0.74

In terms of deployment, Figure 1 provides design recommendations that can be used to put our model into practice. Aiming to have highly motivated students, we used our model to make predictions for students with varied characteristics but intrinsic motivation fixed at its maximum (six). In the figure, the X-axis, Y-axis, and circle colors represent preferred game genre, preferred playing setting, and the recommended design, respectively (other inputs were Gender male, Weekly Playing Time ≤ 7, and Age < 21). Additionally, the larger the circle, the higher the probability of recommending the design. Consequently, larger circles indicate our model is more certain that students using the recommended design will report high intrinsic motivation. However, Figure 1 is limited to variations in three inputs. Therefore, Figure 2 introduces GARFIELD's interface, our interactive recommender system (available in the appendix). This interface receives user input and passes it to our model. Then, our model predicts the probability of recommending each possible design and presents it as a barplot. Accordingly, practitioners can use it to get recommendations for personalizing their gamified design in a simple, interactive way. Thus, attending to our project's second requirement.

4 Discussion

Overall, our goal was to facilitate the personalization of gamification with a model that recommends a gamified design given an expected intrinsic motivation level. Additionally, we aimed that such recommendations considered user

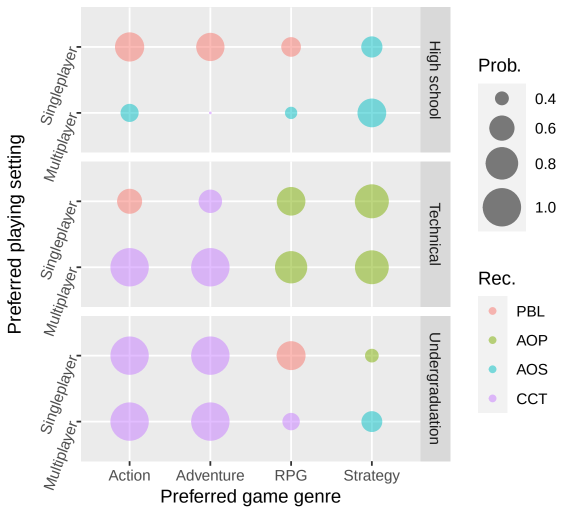

Fig. 1. Predictions from our model aiming to achieve highly motivated students while considering their characteristics. Recommendations are points, badges, and leaderboards (PBL), acknowledgments, objectives, and progression (AOP), acknowledgments, objectives and social pressure (AOS), and competition, chance, and time pressure (CCT).

characteristics and could be used interactively. Ultimately, our recommender system - GARFIELD - achieves these goals, allowing educators to use it in an interactive, web-based way to get design recommendations based on the aforementioned input (see Figure 2). For instance, consider a male student that plays less than one hour a day and is less than 21 years old. Additionally, he has a high school degree and prefers playing strategy games with other people. Then, if an instructor wants him to be highly motivated, they could select the AOS (acknowledgments, objective, and social pressure) design (see Figure 1). Similarly, instructors could select designs for students with other preferences, either using recommendations from Figure 1 or GARFIELD. Thus, this paper contributes a data-driven instrument that enables performing a motivation-based personalization of gamification.

Consequently, this research expands the literature by creating personalization guidelines from feedback collected after *real* experiences. In contrast, prior research developed personalization guidelines based on potential experiences [18]. For instance, recommendations from [23] and [17] are based on surveys. That is, based on what participants think will be the best for them. Differently, [20] provided recommendations derived after students used a gamified system. However, those consider users' motivations collected before such usage. Thereby, such recommendations received no information of students' experiences with gamification. Unlike related research, this paper employed a reverse engineering approach. That is, we used participants' motivations that were collected after real usage as input to our modeling phase. Thus, to our best knowledge,

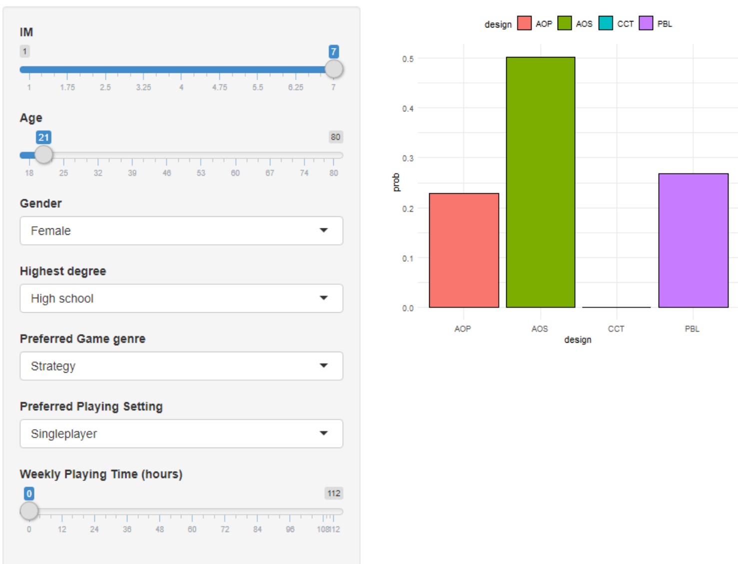

Fig. 2. Printscreen of the interactive recommender system.

this paper contributes the first recommender system for the personalization of gamified educational systems grounded in real rather than potential students' experiences.

As a practical implication, our contribution facilitates the deployment of personalized gamification. Previous research has advocated towards the development of recommender systems to help personalization of gamification [24]. However, most personalization strategies are conceptual [10], with few attempts to develop recommender systems (e.g., [17]). Unfortunately, neither of those attempts provided a concrete, data-driven recommender system that could be further used. Differently, we provide GARFIELD: a concrete, interactive recommender system. In doing so, practitioners can use it to get suggestions of which gamification design to deploy depending on users' characteristics and the expected motivation level.

Importantly, GARFIELD's suggestions are based on probabilities. Furthermore, practitioners become aware when recommendations are more or less uncertain. For instance, cases wherein two or more designs have similar probabilities lead to a somewhat ambiguous recommendation. Similarly, recommendations are somehow uncertain when the highest probability is low. In contrast, cases with highly discrepant probabilities indicate more stable recommendations. Therefore, the practical implication is to facilitate deploying personalized gamification in educational contexts with an interactive recommender system that provides transparent recommendations. Thus, we address the need for technological support to gamify educational systems [24].

As implications for future research, our contribution is twofold. First, the lack of data-driven strategies likely poses a challenge for researchers interested in developing similar approaches. In developing our approach, we demonstrate how one can create such personalization strategies step-by-step through the CRISP-DM reference model. Therefore, we also contribute a concrete example that future research can follow to implement data-driven personalization guidelines. Thus, we inform future research on how to use a well-established data mining framework to create a recommender system from empirical data.

Second, we understand that modeling users efficiently is challenging, especially for tasks that depend on people's subjective experiences (e.g., intrinsic motivation). In this paper, we created a model using 221 observations captured from students of three Brazilian institutions. Importantly, the model's inputs are self-reported intrinsic motivation plus easily obtainable demographic characteristics (e.g., age, gender, and gaming preferences). Yet, our model yielded a moderate predictive power (Cohen's Kappa = 0.43). Thus, our results inform future research that while such information contributes to understanding which gamification design to use, we likely need additional information to personalize gamification more accurately.

With that need in mind, we highlight that our findings must be interpreted with caution. The main reason is that our model's external validity is limited by the dataset we used. The dataset features observations from Brazilian students and ended up being reduced to 221 observations after preparation. Yet, some columns remained somewhat skewed (e.g., educational level). Those issues posed limitations in which modeling and evaluation techniques could be used. For instance, if a larger sample was available, we could have used other machine learning algorithms, along with their standard validation methods, aiming for higher predictive power. However, considering predictions involve intrinsic motivation - a highly subjective construct, the model would likely overfit data with our sample size. To cope with that issue, we opted to work within the null hypothesis significance testing framework, adopting a lenient alpha level due to our analysis' exploratory nature. Given that context, we call for research to update our model based on additional data and, thus, increase its recommendation's accuracy.

5 Conclusion

Personalization aims to mitigate cases in which standard gamification fails to improve students' motivations. Mitigating such cases is important because motivation, especially intrinsic motivation, is strongly related to learning performance. However, previous strategies to personalize gamification are based on potential rather than real experiences. Additionally, those strategies are mostly conceptual, which difficulties their usage.

To address those gaps, this paper introduced GARFIELD, an interactive recommender system that guides on how to personalize gamified educational systems. GARFIELD was built upon a conceptual model developed through a reverse engineering approach that followed the CRISP-DM reference model. Specifically, GARFIELD uses as input students' characteristics and self-reported intrinsic motivation collected after they actually used a gamified system. Then, it uses such input to predict recommended designs accordingly. Thus, allowing one to identify the most suitable gamification design given the student's characteristics and the intrinsic motivation level we expect them to achieve.

With our results, practitioners now have technological support to help them to personalize their gamified practices. This can be achieved using GARFIELD, an interactive, ready-to-use recommender system to get design suggestions. Additionally, with this paper, researchers have a concrete guide on how to use CRISP-DM for creating data-driven personalization strategies based on real (instead of potential) experiences. Note, however, that our recommender's predictions are limited to moderate predictive power. We understand that limits its practical usage as it is. Nevertheless, to our best knowledge, GARFIELD is the first tool to provide gamification design recommendations based on real experiences. Thus, we believe it provides practitioners with a reliable starting point and paves the way for researchers to expand and improve it in future research.

Appendix

Supplementary material available at[3] `t.ly/RdQG`.

[3] Full link: `https://osf.io/nt97s/?view_only=9b625f1347744bd1b5f2d098c0eba55e`

References

1. Aguinis, H., Gottfredson, R.K., Joo, H.: Best-practice recommendations for defining, identifying, and handling outliers. Organizational Research Methods **16**(2), 270–301 (2013)
2. Bai, S., Hew, K.F., Huang, B.: Is gamification "bullshit"? evidence from a meta-analysis and synthesis of qualitative data in educational contexts. Educational Research Review p. 100322 (2020)
3. do Carmo Reis, C.A., Russell, R.B., Polydoro, S.A., Fior, C.A.: As crenças de autoeficácia para a formação superior de estudantes de engenharia. In: Brazilian Congress on Engineering Education (2019)
4. Deterding, S., Dixon, D., Khaled, R., Nacke, L.: From game design elements to gamefulness: defining gamification. In: Proceedings of the 15th international academic MindTrek conference: Envisioning future media environments. pp. 9–15. ACM (2011). https://doi.org/10.1145/2181037.2181040
5. ESA.: 2021 essential facts about the video game industry (2021)
6. Goldstein, E.B.: Cognitive psychology: Connecting mind, research and everyday experience. Nelson Education (2014)
7. Guay, F., Vallerand, R.J., Blanchard, C.: On the assessment of situational intrinsic and extrinsic motivation: The situational motivation scale (sims). Motivation and emotion **24**(3), 175–213 (2000)
8. Hanus, M.D., Fox, J.: Assessing the effects of gamification in the classroom: A longitudinal study on intrinsic motivation, social comparison, satisfaction, effort, and academic performance. Computers & Education **80**, 152 – 161 (2015). https://doi.org/https://doi.org/10.1016/j.compedu.2014.08.019, http://www.sciencedirect.com/science/article/pii/S0360131514002000
9. Hox, J.J., Moerbeek, M., Van de Schoot, R.: Multilevel analysis: Techniques and applications. Routledge (2010)
10. Klock, A.C.T., Gasparini, I., Pimenta, M.S., Hamari, J.: Tailored gamification: A review of literature. International Journal of Human-Computer Studies p. 102495 (2020)
11. Kwak, C., Clayton-Matthews, A.: Multinomial logistic regression. Nursing research **51**(6), 404–410 (2002)
12. Martínez-Plumed, F., Contreras-Ochando, L., Ferri, C., Orallo, J.H., Kull, M., Lachiche, N., Quintana, M.J.R., Flach, P.A.: Crisp-dm twenty years later: From data mining processes to data science trajectories. IEEE Transactions on Knowledge and Data Engineering (2019)
13. Palomino, P., Toda, A., Rodrigues, L., Oliveira, W., Isotani, S.: From the lack of engagement to motivation: Gamification strategies to enhance users learning experiences. In: 19th Brazilian Symposium on Computer Games and Digital Entertainment (SBGames)-GranDGames BR Forum. pp. 1127–1130 (2020)
14. Rodrigues, L., Palomino, P.T., Toda, A.M., Klock, A.C., Oliveira, W., Avila-Santos, A.P., Gasparini, I., Isotani, S.: Personalization improves gamification: evidence from a mixed-methods study. Proceedings of the ACM on Human-Computer Interaction **5**(CHI PLAY), 1–25 (2021)
15. Rodrigues, L., Toda, A.M., Oliveira, W., Palomino, P.T., Avila-Santos, A.P., Isotani, S.: Gamification works, but how and to whom? an experimental study in the context of programming lessons. In: Proceedings of the 52nd ACM Technical Symposium on Computer Science Education. pp. 184–190 (2021)

16. Rodrigues, L., Toda, A.M., Oliveira, W., Palomino, P.T., Isotani, S.: Just beat it: Exploring the influences of competition and task-related factors in gamified learning environments. In: Anais do XXXI Simpósio Brasileiro de Informática na Educação. pp. 461–470. SBC (2020)
17. Rodrigues, L., Toda, A.M., Oliveira, W., Palomino, P.T., Vassileva, J., Isotani, S.: Automating gamification personalization: To the user and beyond (2021)
18. Rodrigues, L., Toda, A.M., Palomino, P.T., Oliveira, W., Isotani, S.: Personalized gamification: A literature review of outcomes, experiments, and approaches. In: García-Peñalvo, F.J. (ed.) Proceedings of the 8th International Conference on Technological Ecosystems for Enhancing Multiculturality (TEEM 2020) (Salamanca, Spain, October 21-23, 2020) (oct 2020)
19. Sailer, M., Homner, L.: The gamification of learning: a meta-analysis. Educ Psychol Rev **32**, 77–112 (2020). https://doi.org/10.1007/s10648-019-09498-w, `https://doi.org/10.1007/s10648-019-09498-w`
20. Stuart, H., Lavoué, E., Serna, A.: To tailor or not to tailor gamification? an analysis of the impact of tailored game elements on learners' behaviours and motivation. In: 21th International Conference on Artificial Intelligence in Education (2020)
21. Toda, A.M., Klock, A.C., Oliveira, W., Palomino, P.T., Rodrigues, L., Shi, L., Bittencourt, I., Gasparini, I., Isotani, S., Cristea, A.I.: Analysing gamification elements in educational environments using an existing gamification taxonomy. Smart Learning Environments **6**(1), 16 (2019). https://doi.org/10.1186/s40561-019-0106-1
22. Toda, A.M., Valle, P.H.D., Isotani, S.: The dark side of gamification: An overview of negative effects of gamification in education. In: Cristea, A.I., Bittencourt, I.I., Lima, F. (eds.) Higher Education for All. From Challenges to Novel Technology-Enhanced Solutions. pp. 143–156. Springer International Publishing, Cham (2018). https://doi.org/10.1007/978-3-319-97934-2_9
23. Tondello, G.F., Mora, A., Nacke, L.E.: Elements of gameful design emerging from user preferences. In: Proceedings of the Annual Symposium on Computer-Human Interaction in Play. pp. 129–142. ACM (2017). https://doi.org/10.1145/3116595.3116627
24. Tondello, G.F., Orji, R., Nacke, L.E.: Recommender systems for personalized gamification. In: Adjunct Publication of the 25th Conference on User Modeling, Adaptation and Personalization. pp. 425–430. ACM (2017). https://doi.org/10.1145/3099023.3099114
25. Tondello, G.F.: Dynamic Personalization of Gameful Interactive Systems. Ph.D. thesis, University of Waterloo (2019)
26. Van Roy, R., Zaman, B.: Need-supporting gamification in education: An assessment of motivational effects over time. Computers & Education **127**, 283–297 (2018). https://doi.org/10.1016/j.compedu.2018.08.018
27. Vansteenkiste, M., Sierens, E., Soenens, B., Luyckx, K., Lens, W.: Motivational profiles from a self-determination perspective: The quality of motivation matters. Journal of educational psychology **101**(3), 671 (2009)
28. Viera, A.J., Garrett, J.M., et al.: Understanding interobserver agreement: the kappa statistic. Fam med **37**(5), 360–363 (2005)
29. Vornhagen, J.B., Tyack, A., Mekler, E.D.: Statistical significance testing at chi play: Challenges and opportunities for more transparency. In: Proceedings of the Annual Symposium on Computer-Human Interaction in Play. pp. 4–18 (2020)
30. Wirth, R., Hipp, J.: Crisp-dm: Towards a standard process model for data mining. In: Proceedings of the 4th international conference on the practical applications of knowledge discovery and data mining. vol. 1. Springer-Verlag London, UK (2000)

ICMC USP
SÃO CARLOS