How Personalisation 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

Personalised 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 personalising to a single user dimension, ignoring multiple individual and contextual factors that affect user motivation. Unlike prior research, we address this issue by exploring multidimensional personalisation compared to OSFA, based on a multi-institution sample. Thus, we conducted a controlled experiment in three institutions, comparing gamification designs (*OSFA* and *Personalised* 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 no significant differences between OSFA and Personalised designs, despite indicating that user motivation depends on fewer user characteristics when using personalisation. Additionally, exploratory analyses suggest personalisation is positive for females and those holding a technical degree, but negative for those who prefer adventure games and those who prefer single-playing. Our contribution benefits designers, suggesting how personalisation works; practitioners, demonstrating for whom the personalisation strategy is suitable or not; 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. For instance, they enable managing educational materials and deploying assessments (Kocadere & Çağlar, 2015; Pereira et al., 2021), which is critical for successful learning experiences (Batsell Jr, Perry, Hanley, & Hostetter, 2017;

Mpungose, 2020; Rowland, 2014). However, educational activities are oftentimes not motivating (Palomino, Toda, Rodrigues, Oliveira, & Isotani, 2020; Pintrich, 2003). This is problematic, because motivation is positively correlated with learning (Hanus & Fox, 2015; Rodrigues, Toda, Oliveira, Palomino, Avila-Santos, & Isotani, 2021). Consequently, that lack of motivation jeopardises learning experiences.

Empirical evidence demonstrates that gamification might improve motivational outcomes (Sailer & Homner, 2020). However, such effect varies from person to person and context to context (Hallifax, Audrey, Jean-Charles, Guillaume, & Elise, 2019; Hamari, Koivisto, & Sarsa, 2014). To mitigate such variations, researchers are exploring personalised gamification, especially for educational purposes (Klock, Gasparini, Pimenta, & Hamari, 2020). By definition, personalisation of gamification is having designers (or the system, automatically) tailor[1] the gamification design to different users/context, instead of presenting the same design for all (i.e., the one-size-fits-all (OSFA) approach) (Tondello, 2019). In practice, that is often conducted by offering different game elements to distinct users (Hallifax, Serna, Marty, & Lavoué, 2019); thus acknowledging people have different preferences and are motivated differently (Altmeyer, Lessel, Muller, & Krüger, 2019; Tondello, Mora, & Nacke, 2017; Van Houdt, Millecamp, Verbert, & Vanden Abeele, 2020).

Despite being widely researched, the understanding of how personalised gamification compares to the OSFA approach is limited. Initial empirical evidence suggests personalised gamification can overcome the OSFA approach within social networks and health domains (Hajarian, Bastanfard, Mohammadzadeh, & Khalilian, 2019; Lopez & Tucker, 2021). However, results within the educational domain are mostly inconclusive (Rodrigues, Toda, Palomino, Oliveira, & Isotani, 2020). Having gamification personalised to a single dimension might explain such inconclusive findings (Mora, Tondello, Nacke, & Arnedo-Moreno, 2018; Oliveira et al., 2020), given recent research highlighted the need for personalising to multiple dimensions simultaneously (Klock et al., 2020). While preliminary evidence supports the potential of multidimensional personalisation, empirical evidence is limited by either not comparing it to the OSFA approach (Stuart, Lavoué, & Serna, 2020), or low external validity (Rodrigues, Palomino, et al., 2021).

In light of these limitations, **our objective is to test the generalisation of the effect of multidimensional personalisation of gamification[2] as well as investigate possible moderators[3] of that effect**. We accomplish that goal 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); Stuart et al. (2020)), our

[1] We understand *tailoring* as an umbrella term that encompasses tailoring gamification through both personalisation (designer-based) and customisation (user-based) (Klock et al., 2020; Orji, Oyibo, & Tondello, 2017).

[2] That is, personalised to the learning task, users' gaming habits/preferences, and demographics, simultaneously; see Section 4.3 for details.

[3] Moderators are factors that increase/decrease an intervention's effect (Landers, Auer, Collmus, & Armstrong, 2018).

baseline is OSFA gamification, which we implemented with points, badges, and leaderboards (PBL). That is important because PBL is the game elements set used the most by research on gamification applied to education and, overall, has 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). Third, Rodrigues, Palomino, et al. (2021) studied multidimensional personalisation in a single institution based on a confirmatory analysis. Differently, this study involves three institutions and presents exploratory analyses to understand variations in multidimensional personalisation's effect, besides confirmatory ones. These are important to test prior research's external validity and advance the field from *whether* to *when/to whom* personalisation works (Bai et al., 2020; Huang et al., 2020; Sailer & Homner, 2020). Therefore, we contribute new empirical evidence on how multidimensional personalisation of gamification, implemented according to a decision tree-based recommender system, affects motivational learning outcomes in the context of real classrooms. Thus, we contribute to the design of gamified learning environments and to the understanding of when and for whom such personalisation is more (or less) suitable.

2 Background

This section provides background information on VLE and assessments for learning, motivational dimensions, 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. They 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). Additionally, instructors still valued students completing assignments and assessments (Mpungose, 2020; Pereira et al., 2021), which were only enabled through VLE during these times.

A theoretical perspective to understand those activities' relevance 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, analyse, 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.

From a practical perspective, those activities' relevance is supported, for instance, 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 (Batsell Jr et al., 2017; Rowland, 2014), as well as in gamified (Rodrigues, Toda, Oliveira, Palomino, Avila-Santos, & Isotani, 2021; Sanchez, Langer, & Kaur, 2020) settings. On the one hand, that is important because most educational materials are not motivating for students (Hanus & Fox, 2015; Palomino et al., 2020; Pintrich, 2003). Hence, gamifying 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 learning experiences (Hyrynsalmi, Smed, & Kimppa, 2017; Toda, Valle, & Isotani, 2018).

2.2 Motivational Dimensions

In a literature review, Zainuddin, Chu, Shujahat, and Perera (2020) found research on gamification applied to education has relied on several theoretical models, such as Flow Theory, Goal-setting Theory, and Cognitive Evaluation Theory. Despite that, the authors found Self-Determination Theory (SDT) is the most used to explain how gamification affects users, which is aligned to its goal of improving motivation (Sailer & Homner, 2020). Therefore, this section provides a brief introduction to SDT, as we also understand motivation according to it.

Following SDT (Deci & Ryan, 2000), one's motivation to do something (e.g., engage with an educational activity) varies in a continuum, from amotivation (i.e., no determination/intention at all) to intrinsic motivation (i.e., an internal drive due to feelings such as enjoyment and pure interest). In-between them, one experiences extrinsic motivation, which concerns four regulators: external (e.g., due to rewards), introjected (e.g., due to guilt), identified (e.g., due to personal values), and integrated (e.g., due to values incorporated in oneself). Respectively, those refer to exhibiting a behaviour due to external, somewhat external, somewhat internal, and internal drivers.

Within educational contexts, research advocates towards internal drivers. Based on SDT (Ryan & Deci, 2017), behaviours driven by avoiding punishments or aiming to receive rewards are likely to disappear once external drivers are no longer available. In contrast, SDT posits that internal regulators, such as doing something due to personal values or curiosity, are long-lasting. Accordingly, the literature considers that autonomous motivation, which encompasses

regulations connected to internal drivers, is ideal for learning (Vansteenkiste, Sierens, Soenens, Luyckx, & Lens, 2009).

2.3 Gamified Learning: Effects and Moderators

Gamification is the use of game design elements outside games (Deterding, Dixon, Khaled, & Nacke, 2011). Meta-analyses summarising the effects of gamification applied to education found positive effects on cognitive, behavioural, and motivational learning outcomes (Bai et al., 2020; Huang et al., 2020; Sailer & Homner, 2020). However, these studies also show limitations of gamification's effects, which vary due to geographic location, educational subject, and intervention duration. Moreover, experimental studies have found gamification's effect ranged from positive to negative within the same sample, depending on the user (Rodrigues, Toda, Oliveira, Palomino, & Isotani, 2020; Van Roy & Zaman, 2018). Additionally, empirical evidence shows cases wherein gamification's impact changed, depending on specific moderators, such as gender (Pedro, Lopes, Prates, Vassileva, & Isotani, 2015), age (Polo-Peña et al., 2020) and being a gamer (Recabarren et al., 2021).

These moderators are predicted by the Gamification Science framework (Landers et al., 2018). It claims gamification affects motivation; motivation affects behaviour; behaviour affects cognitive outcomes; and moderators affect each of those connections. Consequently, on the one hand, analysing behaviour/cognitive outcomes without considering motivational ones is problematic: gamification might improve motivation, but that improved motivation might not lead to the desired behaviour. In that case, the problem was not gamification, but motivating some other behaviour, which cannot be observed by a study limited to analysing behaviour. Therefore, 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 (Tondello, Mora, & Nacke, 2017) and that gamified designs must be aligned to the task wherein it will be used (Hallifax, Serna, et al., 2019; Rodrigues, Oliveira, Toda, Palomino, & Isotani, 2019). Thereby, due to the moderator effect of personal and task-related characteristics, gamification designs should be aligned to the specific task and the users. However, most gamified systems present the same design for all users, regardless of the task they will do (Dichev & Dicheva, 2017; Liu, Santhanam, & Webster, 2017); the OSFA approach. Therefore, the OSFA approach might explain variations on gamification's outcomes, and cases wherein it only works for some users, based on the role of moderator characteritics.

In summary, gamification studies should focus on measuring motivational outcomes known to affect the desired behavioural outcomes (Landers et al., 2018; Tondello & Nacke, 2020). Within the educational domain, for instance, strong empirical evidence supports autonomous motivation's positive role in

learning outcomes (Hanus & Fox, 2015; Rodrigues, Toda, Oliveira, Palomino, Avila-Santos, & Isotani, 2021; Vansteenkiste et al., 2009). Furthermore, ensuring systems provide gamification designs that consider *who* will use it, and to accomplish *which task*, is important ,to address the limitations of the OSFA approach (Klock et al., 2020). Based on that context, this article studies personalisation as a way to improve OSFA gamification. In doing so, we focus on measuring a single construct (i.e., motivation) to increase the study validity (Wohlin et al., 2012), given evidence supporting the positive relation between motivation and learning.

2.4 Tailored Gamification

Fundamentally, tailoring gamification leads to different gamification designs depending on who is to use it or for what (Klock et al., 2020; Rodrigues, Toda, Palomino, et al., 2020). Tailoring gamification can be achieved in two ways. First, when users define the design, it is known as *customisation* (Tondello, 2019). Empirical evidence supports customisation's effectiveness compared to the OSFA approach in terms of behaviour (Lessel, Altmeyer, Müller, Wolff, & Krüger, 2017; Tondello & Nacke, 2020). However, there is no evidence supporting improvements are due to increased motivation or the effort involved in defining their designs (Schubhan, Altmeyer, Buchheit, & Lessel, 2020). Additionally, customisation is subject to the burden of making users select their designs for each task, as per literature suggestions of matching gamification to task (Hallifax, Serna, et al., 2019; Liu et al., 2017; Rodrigues et al., 2019).

Second, when designers or the system itself defines the tailored design, it is known as *personalisation* (Tondello, 2019). In that case, one needs to model users/tasks to understand the gamification design most suitable for each case. Commonly, that is accomplished by gathering user preferences via surveys, then analysing 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 personalise gamification to a single or few dimensions (Klock et al., 2020). However, several factors affect user preferences (see Section 2.3). Accordingly, empirical evidence from comparing gamification personalised through such recommendations to the OSFA is mostly inconclusive (Rodrigues, Toda, Palomino, et al., 2020). Differently, the few recommendations for multidimensional personalisation 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).

Summarising, while customisation naturally leads to gamification tailored to multiple user and contextual dimensions, it requires substantial effort from the users. Personalisation 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. To the best of our knowledge, the only recommendation for multidimensional personalisation to be empirically tested was analysed in a small experimental study. This highlighted the need for studies grounding the understanding of whether multidimensional personalisation improves OSFA gamification.

2.5 Related Work

Empirical research often compares personalised gamification to random, counter-tailored, or no gamification (Rodrigues, Toda, Palomino, et al., 2020). Consequently, those studies do not add to the understanding of whether personalisation 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 personalised and OSFA gamification, considering those studies provide reliable evidence to understand personalisation's contribution to practice. To the best of our 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 summarised in Table 1, which shows that related work mostly personalised gamification to a single user dimension (e.g., HEXAD (Lopez & Tucker, 2021; Mora et al., 2018) typology) and that research using multidimensional personalisation 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, comparing the OSFA approach to gamification personalised 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, personalisation 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 the 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 personalised gamification. On the other hand, the current study (N = 58) involved three northwestern institutions of the same country. Besides encompassing more institutions, this contextual difference is relevant because Brazil is a continent-sized country. Consequently, the reality of the southwestern and northwestern regions is widely different. Amongst others, the northwestern region has nine out of the 12 Brazilian states with literacy rates below the average. On the contrary, Brazil's southwestern region has the highest literacy average in the country (Grin, Burgos, Fernandes, & Bresciani,

Table 1 Related work compared to this study in terms of the personalisation strategy and the study design.

		Personalised 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 personalised or OSFA gamification.

2021). Additionally, we increased the spacing between sessions from one day to four to six weeks and conducted confirmatory and exploratory analyses. Considering Rodrigues, Palomino, et al. (2021) found personalisation had a positive effect on students' autonomous motivation, 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 the prior research's contribution by analysing if personalisation's effect on student motivation depends on contextual (e.g., subject under study) and user 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, we expand the literature by testing the external validity of state-of-the-art results with a new, larger sample from institutions of a distinct region, as well as shed light on *when* and *for whom* personalised gamification works.

3 Apparatus

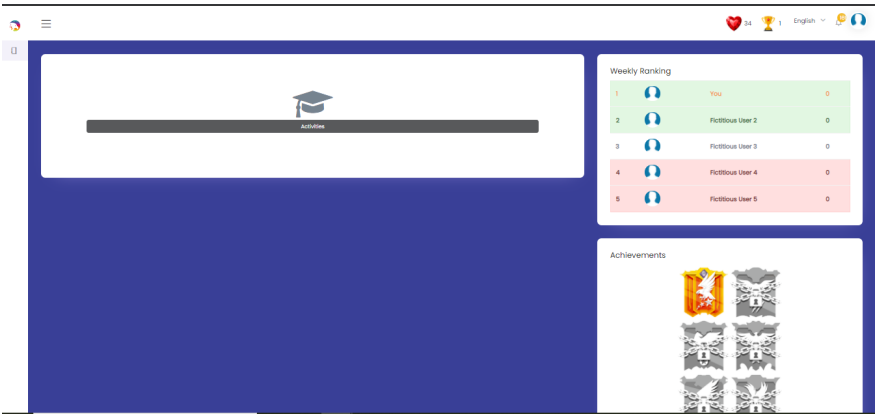
To enable this study, we designed and deployed learning assessments in a VLE. All assessments featured 30 multiple-choice, four-alternative items. Items were designed so that students could correctly solve them if they were able to recall information from lessons. Therefore, the experimental task was 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 under the 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[4], 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). Mission items, as well as item alternatives, appeared in a random order for all courses. Because those were assessments *for* learning, students could redo items they missed until getting them right. Figure 1 demonstrates the system, which can be seen as an assessment-based VLE. After logging in, the student selects the course to work on from the list that can be accessed from the top-left hamburger menu. Following, they can interact with game elements, such as checking leaderboard positions, or start a mission from the course home page (Figures 1a and 1b). In the latter case, students are emerged into each 3-item mission at a time, completing multiple-choice items individually (Figure 1c). Once a mission is finished, the system goes back to the course homepage and the usage flow restarts.

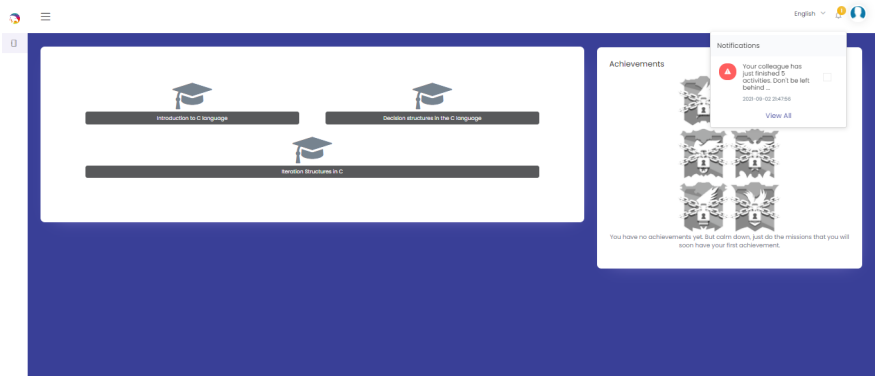
In terms of gamification design, this study used the nine game elements described next, which are based on definitions for educational environments from Toda's educational game taxonomy (Toda, Klock, et al., 2019):

- **Acknowledgment**: Badges awarded for achieving mission-related goals (e.g., completing a mission with no error); shown on 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 on the course's main page;
- **Objectives**: Provide short-term goals by representing the course's missions as a skill tree;

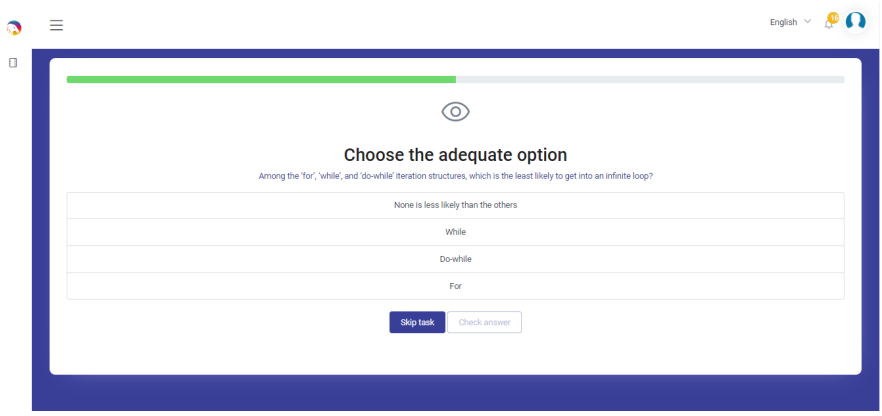
[4] http://eagle-edu.com.br/



(a) Course home page for the one-size-fits-all condition. It features Points, Badges, and a Leaderboard.



(b) Course homepage for the gamification design, personalised to people who have experience researching gamification and for whom the preferred game genre is not action (see Table 3). It features Badges, Objectives, and Social Pressure.



(c) Completing a mission item with the progress game element on. Otherwise, the progress bar would not be available.

Fig. 1 Screenshots of Eagle-edu.

- **Points**: Numeric feedback that functions similar to Acknowledgment; shown on 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;
- **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 personalised condition, however, we had to create one course for each personalised 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. Regardless of that, students of the same subject always completed the same assessment and all courses had the same name. For instance, consider students of the Object Oriented Programming subject. Those assigned to the OSFA condition would all be in a single Eagle-Edu course. Those assigned to the Personalisation Condition, for instance, could be attributed to two different Eagle-Edu courses, which is necessary because people in this condition will often use distinct designs depending on their characteristics (see Section 4.3 for details). However, all three courses would have the same assessment and name. This ensures that gamification design was the only difference among conditions, as needed for the experimental manipulation. Nevertheless, that affected the Leaderboards appearance (see Section 8). The supplementary material provides a video and images of the system.

4 Method

This study involves both confirmatory (i.e., testing assumptions) and exploratory (i.e., generating hypothesis) data analysis (Abt, 1987). Accordingly and based on our goal, we investigated the following:

- **Hypothesis 1 - H1:** Multidimensional personalisation of gamification in gamified review assessments improves autonomous motivation, but not external regulation and amotivation, when compared to the OSFA approach.
- **Research Question 1 - RQ1:** Do user and contextual characteristics moderate the effect of multidimensional personalisation of gamification, in gamified review assessments?
- **RQ2:** How does the variation of students' motivations change when comparing gamification personalised to multiple dimensions to the OSFA approach, in gamified review assessments?
- **RQ3:** What are students' perceptions of 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. The rationale for this hypothesis is twofold. First, autonomous motivation is considered ideal for learning purposes (Vansteenkiste et al., 2009). Second, although multidimensional personalisation of gamification holds potential to improve OSFA gamification, empirical evidence is limited (Rodrigues, Palomino, et al., 2021; Stuart et al., 2020). Thus, testing **H1** informs the effectiveness of equipping gamified educational systems with multidimensional personalisation to improve autonomous motivation, which is known to mediate improvements in learning outcomes according to empirical evidence (Hanus & Fox, 2015; Rodrigues, Toda, Oliveira, Palomino, Avila-Santos, & Isotani, 2021; Sanchez et al., 2020) and the Gamification Science framework (Landers et al., 2018).

RQ1 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 personalised gamification, and identify which factors are responsible for it. Similarly, **RQ2** is based on research showing that gamification's effect varies 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 personalisation can adapt to such variation. **RQ3** 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 we did not predict this result before our data analysis.

Based on those RQs and **H1**, we designed and conducted a multi-site, experimental study, following a mixed factorial design: gamification *design* (levels: OSFA and Personalised) and session (levels: 0 and 1) were the factors. For *design*, participants were randomly assigned to one of the two conditions (between-subject) at each trial, while *sessions* 0 and 1 refer to mid-term and end-term assessments (within-subject), respectively. Figure 2 summarises this study, which received an ethical committee approval (CAAE: 42598620.0.0000.5464).

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

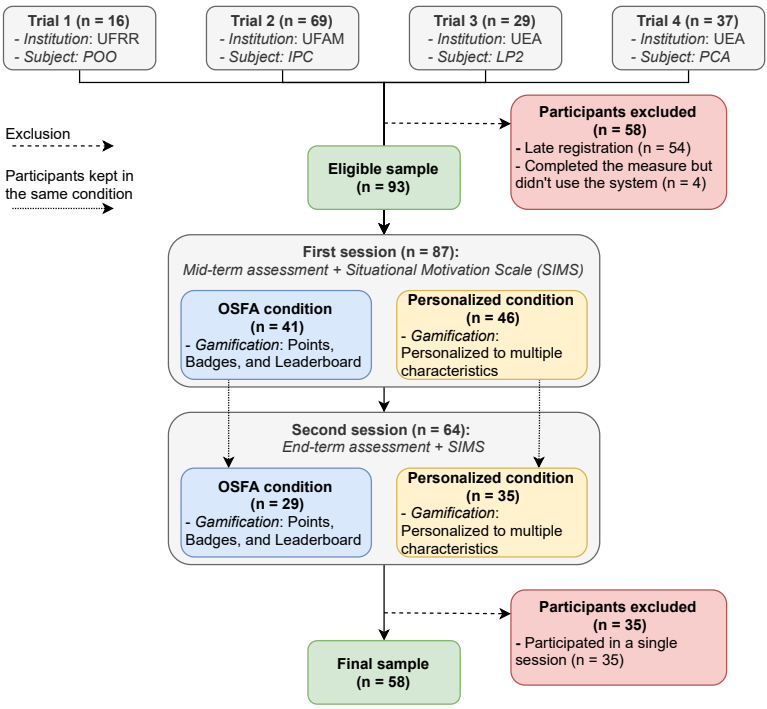


Fig. 2 Study Overview. Institutions are Federal University of Roraima (UFRR), Federal University of Amazonas (UFAM), and Amazonas State University (UEA). Subjects are Object Oriented Programming (POO), Introduction to Computer Programming (IPC), Programming Language 2 (LP2), Computer and Algorithms Programming (PCA).

institutions (see Figure 2). For each trial, we sent the characterisation 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 characterisation 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 Personalised 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 left that choice up to instructors. Additionally, note the within activity performance did not count towards any student's grade, to mitigate biases.

Table 2 Participants' demographic information.

Information	Overall	OSFA	Personalised
	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	8 (14%)	3 (12%)	5 (16%)
No	50 (86%)	23 (88%)	27 (84%)

4.3 Experimental Conditions

We designed two experimental conditions, *OSFA* and *Personalised*, which differ in terms of the game elements they present to users. We implement those by changing the game elements available, considering it is the common approach (Hallifax, Serna, et al., 2019). The *OSFA* condition featured Points, Badges, and Leaderboards (PBL), similar to Rodrigues, Palomino, et al. (2021). PBL are among the game elements used the most in gamification research and together provide effects comparable to other combinations (Bai et al., 2020; Gari, Walia, & Radermacher, 2018; Venter, 2020; Zainuddin et al., 2020). Therefore, we believe PBL offer external validity as an implementation of the standard OSFA gamification, because it is similar to most gamified systems used in practice (Wohlin et al., 2012). Accordingly, we defined the personalisation condition as having the same number of game elements as the OSFA. This ensures the only distinction was which game elements were available, mitigating confounding effects from comparing conditions with different numbers of game elements available (Landers, Bauer, Callan, & Armstrong, 2015).

The *Personalised* condition provided game elements according to recommendations from decision trees built by prior research (Rodrigues, Toda, Oliveira, Palomino, Vassileva, & Isotani, 2021), following the same procedure as Rodrigues, Palomino, et al. (2021). Such recommendations are based on conditional decision trees (Hothorn, Hornik, & Zeileis, Jan 2006) that were generated in two steps. First, the authors used a survey to capture people's

top-three preferred game elements for different learning activity types and their demographic information. According to the authors, this survey was disclosed through Amazon Mechanical Turk, a crowdsourcing platform that has been widely used to increase the external validity of such approach (Rodrigues, Toda, Oliveira, Palomino, Vassileva, & Isotani, 2021). Second, the authors generated conditional decision trees with these data. In particular, they created three trees, one for each of the top-three spots collected in the survey. Then, according to how conditional decision trees are created, the validation process relied on the null hypothesis significance testing framework (Hothorn et al., Jan 2006). Accordingly, they were validated based on whether an input significantly affected the output accuracy. Hence, this maximised generalisation based on the assumptions underlying inferential statistics (Sheskin, 2003).

After this validation process, the variables composing the trees' inputs, which are significant predictors of their outputs, are: user's i) gender, ii) highest educational degree, iii) weekly playing time, iv) preferred game genre, v) preferred playing setting, and vi) whether they already researched gamification (see Table 2), vii) the country where gamification will be used and viii) the learning activity type in terms of the cognitive process involved while doing it, according to the processes described in the revision of Bloom's taxonomy (Krathwohl, 2002). These were input according to the values presented in Table 2 (e.g., age was a numeric value, while preferred playing setting was either *singleplayer* or *multiplayer*), aiming to enable personalising to individual and contextual differences, as proposed in Rodrigues, Toda, Oliveira, Palomino, Vassileva, and Isotani (2021).

Additionally, Rodrigues, Toda, Oliveira, Palomino, Vassileva, and Isotani (2021) discuss that these variables were selected following research demonstrating that demographics, gaming preferences/experience, and contextual information should be considered when designing gamification designs for the educational domain (Hallifax, Serna, et al., 2019; Klock et al., 2020; Liu et al., 2017). For instance, take *researched gamification*. Rodrigues, Toda, Oliveira, Palomino, Vassileva, and Isotani (2021) first asked participants for how many years they had worked/scientifically researched gamification. Then, because most participants answered they had zero years of experience, the authors coded this variable as yes (more than zero) or no (zero). Despite that, their analyses revealed this binary variable plays a significant role on people's preferences. Hence, we similarly asked our participants if they had worked/scientifically researched gamification, considering, for example, it acknowledges the possibility that a student with experience in gamification design might have different preferences compared to people without such experience.

Importantly, *country* and *learning activity type* do not appear in Table 2. The reason is that - due to our experimental setting - those were fixed: all participants were Brazilians and the learning activity was limited to *remembering* . Considering those fixed information, we analysed the decision trees and identified that, for our sample and experimental task, selecting the game elements to be available for each user only depends on *whether one's preferred game*

Table 3 Game elements used in the Personalised 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

genre is action or not, their gender, and the variable researched gamification. That happened because, given our study's contextual information, other characteristics either led to the same game elements or were not part of the paths between the trees' root to a leaf. Hence, that analysis allowed us to only consider the three user characteristics Table 3 shows, which summarises the game elements available for each user of the personalised condition, according to their information. To exemplify the conditions' differences, consider a participant who has experience researching gamification and whose preferred game genre is adventure. If this person was assigned to the personalised condition, their gamification design would be the one shown in Figure 1b. In contrast, the same person would use the design shown in Figure 1a if they were assigned to the OSFA condition. For a complete view of how all designs differ, please refer to our supplementary material.

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. For instance, if the third tree's number one recommendation was Objectives, but one of the other trees had already recommended it, we would select the third tree's number two recommendation. We made that choice to avoid confounding factors of participants interacting with different numbers of game elements (Landers et al., 2015). Based on that, for each student, the personalisation process worked as follows. First, we analysed their information; particularly, those described in Table 3. Second, we identified which game element to offer for that student, according to their characteristics, following recommendations from the aforementioned decision trees. Finally, we assigned the student to use a gamification design that matches their characteristics.

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 SDT (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 the participants' language (Gamboa, Valadas, & Paixão, 2013). Using the recommended

seven-point Likert-scale (1: corresponds not all; 7: corresponds exactly), the SIMS captured motivation to engage with the VLE through four constructs: intrinsic motivation, identified regulation, external regulation, and amotivation (Deci & Ryan, 2000). Each construct was measured by four items, and these items' average led to the construct's final score. A sample prompt is *Why are you engaged with the system where the activity was made?* and a sample item was *Because I think that this activity is interesting.* Additionally, we provided an open-text field so that participants could make comments about their experiences.

The moderator analyses considered the following variables:

- Age (in years);
- Gender: male or female;
- Education: High School, Technical, or Graduated;
- Preferred game genre: Action, Adventure, RPG, or Strategy;
- Preferred game setting: Multiplayer or Singleplayer;
- Weekly playing time (in hours);
- Performance: the number of errors per assessment;
- Assessment subject: POO, IPC, LP2, or PCA;
- Usage interval (in weeks): 0 (first usage), 4, or 6.

In summary, variables one to six came from the trees' input, while the last three came from the experimental design. Note that 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. Accordingly, our exploratory analyses inspect performance as a possible moderator of personalisation's effect, based on research showing that 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 personalisation strategy grounded 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, we investigated within personalised gamification moderators that have demanded attention in the standard approach. Lastly, note that moderators' levels follow those in Rodrigues, Toda, Oliveira, Palomino, Vassileva, and Isotani (2021) and that gender is limited to male and female, because those were the ones our participants reported.

4.5 Procedure

First, participants were invited to participate in the study. Second, they had to complete the characterisation survey by the deadline, which captured identifying information plus those described in Section 4.3. Participants self-reported their age and weekly playing time through numeric, integer fields, and other information (e.g., preferred game genre) through multiple-choice items as collected in Rodrigues, Toda, Oliveira, Palomino, Vassileva, and Isotani (2021). Such information and respecting the deadline were essential to enable personalisation. Third, around mid-term, participants completed the first session's assessment and the SIMS. Fourth, towards the term's end, participants completed the second session's assessment and again the SIMS. One researcher participated in both sessions, providing clarifications as required (e.g., explaining how to use the system). Additionally, at the start of session 0, the researcher presented the study goal, the procedure, and an Eagle-Edu tutorial.

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 generalisation of their findings. Therefore, we applied robust (i.e., 20% trimmed means) mixed ANOVAs (Wilcox, 2011), 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).

Our exploratory analyses (**RQs**) follow recommendations for open science (Dragicevic, 2016; Vornhagen, Tyack, & Mekler, 2020). As suggested, we do not present (nor calculate) p-values, because they are often interpreted as conclusive evidence, which is contrary to exploratory analyses' goal (Abt, 1987). Instead, we limit our analyses to confidence intervals (CIs), which contribute to transparent reporting and mitigate threats to a replication crisis in empirical computer science (Cockburn, Dragicevic, Besançon, & Gutwin, 2020). To generate reliable CIs for non-normal data and avoid misleading inferences, our exploratory analyses rely on CIs calculated using the bias-corrected and accelerated bootstrap, as recommended in Cairns (2019); Carpenter and Bithell (2000). For categorical variables, we compare participants' motivations among subgroups, while for continuous variables we run correlations between them and the motivation constructs. For both categorical and continuous variables, we investigate whether one subgroup's CI overlaps with that of another subgroup, to understand their differences. Throughout those analyses, we consider all the moderators described in Section 4.4). 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 from students' perspectives. The analysis process involved four steps and five researchers. First, one author conducted a thematic analysis (Braun & Clarke, 2006), familiarising themselves with the data, generating and reviewing codes, and grouping them into themes. Acknowledging the subjective nature of participants' comments,s/he followed the interpretivist semi-structured strategy (Blandford, Furniss, & Makri, 2016). Accordingly, s/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 of having multiple coders was to increase reliability through complementary interpretations, despite it being important that others inspect such interpretations (Blandford et al., 2016). Therefore, in the last step, the author who conducted the first step reviewed step three's results. Here, s/he aimed for a wider, complementary interpretation of the 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 Section 5.

5 Results

This section analyses the comparability of the experimental conditions, in terms of participants' information, and presents the data analyses results.

5.1 Preliminary Analyses - Do groups differ?

These analyses compare the conditions' participants to identify possible covariates using robust ANOVAs (see Section 4.6) and independence chi-squared tests to compare continuous and categorical variables, respectively. When counts are lower than 5, we simulate p-values using bootstrap through R's *chisq* function, for increased reliability. In these cases, the degrees of freedom (df) are *NA* due to bootstrap.

We found nonsignificant differences for demographics, gaming preferences/habits, and experience researching gamification. Those results can be seen in our supplementary material for readability. For performance, the design's main effect was nonsignificant ($F(1\ 24.8985) = 0.9042$; $p = 0.35$) but the session's main effect ($F(1, 29.8645)$; $p = 0.0112$), as well as the factors' interaction ($F(1, 29.8645)$; $p = 0.0041$) were statistically significant. Accordingly, we ran post hoc comparisons for OSFA versus personalised, for both sessions 0 ($t = 0.2387$; $p = 0.78464$) and 1 ($t = -1.9778$; $p = 0.04341$) using Yuen's test, a robust alternative to compare two independent samples (Wilcox, 2011). The results provide evidence that participants of the OSFA condition made fewer mistakes per assessment item ($M = 0.964$; $SD = 0.509$) than those of the Personalised condition ($M = 1.19$; $SD = 0.459$). Thus, preliminary analyses indicate a single statistically significant difference among conditions -

session 1's *performance* - when analysing possible covariates, despite descriptive statistics showing uneven distributions for some demographics (Table 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). Additionally, Table 5 shows the results from testing **H1**. All p-values being larger than the 0.05 alpha level reveals no statistically significant difference for all motivation constructs. Thus, our findings partially support **H1**, because the expected significant differences in intrinsic motivation and identified regulation were not found, whereas our data support the nonsignificant differences in external regulation and amotivation.

Table 4 Descriptive statistics, overall (Ovr) and per session (S0 and S1). Data shown as Mean (Standard Deviation), with N referring to the number of data points. Accordingly, n = 64 refers to the two data points (sessions 0 and 1) of each participant of the personalised condition (32 in total).

	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)
	Personalised	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)
	Personalised	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)
	Personalised	32	5.37 (1.38)	5.73 (1.10)	4.31 (1.20)	2.47 (1.65)

Table 5 Confirmatory analyses of **H1**: personalisation 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 Exploratory Analyses (RQ1 and RQ2)

This section presents results for RQ1 and RQ2. As those are based on analyses of subgroups, each one's number of participants is available in Table 2. Note,

however, that all participants engaged in two sessions. Therefore, the number of data points in each subgroup is twice that in Table 2.

5.3.1 RQ1: Moderators of the Personalisation Effect

For continuous variables, moderations are indicated when CIs from the OSFA condition do not overlap with those of the Personalised condition. Accordingly, Table 6 indicates *performance* was the single continuous moderator. Results indicate that higher performance was associated with higher external motivation for OSFA users [0.4;0.7], but not for personalised users [-0.24;0.16], and that such correlations did not overlap. Hence, these results suggest performance moderated the effect of personalisation on 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 7 for an overview). This is the case of *gender*. Females' CIs do not overlap when comparing the amotivation of the personalised [1.31;1.78] and the OSFA [1.95;2.70] conditions. Differently, males' CIs overlap when comparing the personalised [2.34;3.16] and the OSFA [1.74;2.39] conditions in terms of amotivation. This suggests gender moderates the effect of personalisation, which was only positive for females. *Education* appears to be another moderator. Students with a technical degree who used the personalised design experienced higher intrinsic motivation than those who used the OSFA design. *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 personalised condition. *Preferred playing setting* seems to be another moderator. Those who prefer single-player reported higher intrinsic motivation and identified regulation than those of the personalised condition. Additionally, the results suggested no differences among subgroups not mentioned. Overall, our findings indicate the *assessment's subject* and *usage interval* did not moderate the effect of personalisation on any construct, in contrast to *gender*, *education*, *preferred game genre and playing setting*, *performance*, and *age* (**RQ1**).

Table 6 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]

[5] The amotivation of all participants of this subgroup was 1.

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 personalisation (Pers.) compared to one-size-fits-all (OSFA) gamification. Green and red text 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.

5.3.2 RQ2: Motivation Variation Among Conditions

Based on Tables 7 and 6 (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 analysed. Third, participants whose *preferred playing setting* is single-player reported higher intrinsic motivation and identified regulation than those who prefer multiplayer. Fourth relates to *education*: those with a technical degree

reported higher identified regulation than those with high school. Fifth, *assessment's 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, motivation varied according to four characteristics for users of the **Personalised** 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 personalised gamification varied according to fewer factors than that from the OSFA design (**RQ2**).

5.4 Qualitative Analysis of RQ3

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 8 describes codes and themes, exemplifying them with quotes.

5.5 Summary of Results

- Preliminary analyses showed participants of the personalised condition experienced more difficulty in the second session's assessment.
- **H1** is partially supported. Surprisingly, results do not confirm the personalisation positive effect on autonomous motivation - instead, indicating a non-significant difference - while they corroborate the non-significant effect on external regulation and amotivation.
- **RQ1**: Exploratory analyses suggested *gender* and *education* positively moderated personalisation's effect, in contrast to *preferred game genre* and *preferred playing setting*. Personalisation was positive for *females* and those holding a *technical* degree, but negative for people who prefer either the *adventure* game genre or the *single-player* playing setting.
- **RQ2**: 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 indicated the motivation of students who used personalised gamification varied according to only four factors: *education*, *assessment's subject* (common to OSFA), *age* and *gender* (uncommon).
- **RQ3**: Qualitative results indicated the gamified assessments provided positive experiences that students perceived as well designed and good for

Table 8 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."*

their learning, although a few of them mentioned gamification demands improvement and considered the assessments complex and badly presented.

Importantly, motivation is multidimensional and involves a number of constructs (see Section 2.2). Accordingly, changing one's feelings in terms of any of those constructs will inevitably affect that person's motivation. Based on that, note that we are not claiming, for instance, that gender moderates the personalisation effect on all motivation constructs. Instead, we are referring to our empirical finding where gender moderated the personalisation effect on

some motivation construct (amotivation in that case) which, thus, implies an effect on general motivation as well.

6 Discussion

For each hypothesis/RQ we studied, this section interprets its results, relates them to the literature, and discusses possible explanations, which we present as testable hypotheses (TH). We aim for those to be understood as hypotheses derived in discussing and interpreting our findings, not as conclusions drawn from or supported by our data, because TH emerged from exploratory analyses (Abt, 1987). Therefore, aiming to increase our contribution, we provide TH to inform future research that must empirically test them.

6.1 How does personalisation affect student motivation?

Confirmatory analyses (**H1**) revealed no significant differences among conditions for all motivation constructs. However, participants of the personalised condition had lower performance than those of the OSFA in the second assessment. Considering research shows performance might affect gamification's effect (Rodrigues, Toda, Oliveira, Palomino, Avila-Santos, & Isotani, 2021; Sanchez et al., 2020), participants of the personalised condition would report lower motivation than those of the OSFA condition in that session. In contrast, our results personalised gamification provided motivation levels comparable to those of the OSFA approach even though participants experienced higher difficulty during the second session's task. On the one hand, one might suspect that personalised gamification contributed to student motivation by making it less sensitive to their performances, and not by increasing. Based on research about seductive details (Rey, 2012), gamification might distract students with low knowledge and, consequently, affect their motivations negatively. Therefore, our suspicion is that personalisation might have addressed those distractions by offering game elements suitable to the student and the task. Another possibility is that OSFA gamification's benefits for students with low initial knowledge decrease over time (Rodrigues, Toda, Oliveira, Palomino, Avila-Santos, & Isotani, 2021). Then, personalisation might have addressed that time effect on low-performance students. Thus, research efforts should test whether personalisation works by preventing distractions and avoiding time effects for low-performance students:

- **TH1:** Personalisation makes user motivation less sensitive to user performance.

On the other hand, one might suspect that personalisation 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 suspicion builds upon the lack of longitudinal studies evaluating personalisation's effect. For instance, most related work used cross-sectional studies (e.g., Hajarian et al. (2019); Lopez and Tucker (2021); Mora et al.

(2018)). Only Rodrigues, Palomino, et al. (2021) used a repeated-measures design, which is limited to two measurements with a one-day spacing. Whereas our study also captured two measurements, spacing varied between four to six weeks. Performance differences, however, limited our findings' contribution to understanding how personalisation's effect change over time. Thus, while personalisation might mitigate the novelty effect's impact on gamification, regardless of students knowledge level, empirical research is needed to test TH2:

- **TH2:** Personalisation increases user motivation after a first-time experience.

6.2 How do students perceive gamified review assessments?

Qualitatively analysing open-text comments (**RQ3**) showed that students considered the activity good for their learning process and that they considered it well designed by approaching a perspective rarely explored in their studies: taking time to review theoretical/conceptual aspects of computing education. While empirical evidence shows the testing effect improves learning in gamified settings (Rodrigues, Toda, Oliveira, Palomino, Avila-Santos, & Isotani, 2021; Sanchez et al., 2020), studies have not inspected users' perceptions about such learning activities. This 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 while providing overall positive experiences to students.

Furthermore, the results corroborate gamification literature by showing it is mostly perceived positively, but not always (Sailer & Homner, 2020; Toda et al., 2018). Prior research demonstrating different people are motivated by different game elements (e.g., Bovermann and Bastiaens (2020); Tondello, Mora, and Nacke (2017)) corroborate those results. Consequently, this suggests the need to improve the personalisation 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 personalisation mechanism was *changing the game elements available*, whereas some comments suggested *changing how game elements work*. A third perspective is that the recommendations on how to personalise (i.e., which game elements to use when) demands refinement to model users/tasks better. While we further inspected the latter perspective through RQ1 and RQ2, we expect future research to test whether:

- **TH3:** Designs with more than three game elements improve users' perceptions about gamification.
- **TH4:** Successful personalisation of gamification requires tailoring the mechanics of game elements as well as which of them should be available.

6.3 Which factors moderate personalisation's effect?

Exploratory analyses (**RQ1**) indicated four moderators of personalised gamification's effect: gender, education, preferred game genre, and preferred playing setting. Because we considered those factors in defining personalised designs, we expected no moderator effect from them. The contrast is somewhat expected, however. Research on OSFA gamification shows several factors (e.g., gender) moderate its effectiveness (see Section 2.3), 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 personalisation's effect.

First, how game elements are recommended depending on those moderators' levels. For instance, results indicate personalisation mitigated females' amotivation, but did not work for males. The strategy we used (Rodrigues, Toda, Oliveira, Palomino, Vassileva, & Isotani, 2021) does not consider gender to select game elements within contexts such as the one of this study. That is, where the country is 'Brazil' and the learning activity type is 'remembering'. 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. Thereby, future research should test TH5:

- **TH5:** Further modeling user demographics and gaming-related preferences will improve the personalisation effect.

Another possible reason is that we used a preference-based personalisation strategy. Despite that approach being widely researched (Klock et al., 2020; Rodrigues, Toda, Palomino, et al., 2020), the literature advocates that user preference often fails to reflect user behaviour (Norman, 2004). To face that issue, researchers are investigating data-driven personalisation strategies (e.g., Hajarian et al. (2019)). That is, inspecting user behaviour to determine the most suitable game elements (Tondello, Orji, & Nacke, 2017). Therefore, assuming that by relying on interaction data, they will reliably identify user preferences and, consequently, improve the gamification effectiveness. Thus, given the limitations from preference-based strategies, future research should test TH6:

- **TH6:** Data-driven personalisation strategies are more effective than preference-based ones.

6.4 How does user motivation variation change when comparing OSFA and personalised gamification?

Exploratory analyses (**RQ2**) revealed that motivation from using OSFA gamification varied according to six characteristics: performance, preferred game genre, preferred playing setting, education, assessment's subject, and usage

interval. Those are expected considering 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). Differently, motivation from users of the personalised condition did not vary due to performance, preferred game genre, preferred playing setting, and usage interval, but similarly varied according to assessment subject and education. Due to personalisation, one might expect reduced variation in outcomes if it leads to gamification designs suitable to every user's preferences/motivations (Hallifax, Serna, et al., 2019; Klock et al., 2020; Rodrigues, Toda, Palomino, et al., 2020). Then, we suspect that providing personalised game elements made motivation less sensitive to performance and reuse issues (e.g., losing effectiveness after the novelty has vanished). This raises the need for testing whether:

- **TH7:** Personalisation mitigates OSFA outcomes' sensitivity to user and contextual factors.

Nevertheless, personalisation did not tackle variations from education and assessment subjects, besides varying due to gender and age. Assessment subject and age are not considered by the personalisation strategy we used, which might explain their role. 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 personalisation strategy in Rodrigues, Toda, Oliveira, Palomino, Vassileva, and Isotani (2021). Thus, research should test whether:

- **TH8:** Gender and Education require further modeling to properly suggest the most suitable game elements.

7 Implications

In light of our discussion, this section connects our findings to their implications to design and theory.

7.1 Design Contributions

First, our findings inform the design of gamified education system on **how personalisation contributes to gamification**. Results from **RQ2** suggested personalisation tackled motivation variations compared to the OSFA approach. Hence, designers might not see direct effects, such as increased motivation, but personalisation might be acting by minimising the extent to which one user group benefits more/less than others. However, findings from **RQ1** suggest there were moderators of the effect of personalisation, warning that caution is

needed when applying personalisation strategy as we did, because it might offset its contribution in some cases (e.g., improving amotivation but decreasing external motivation of females who prefer Adventure games). Thus, **designers might use multidimensional personalisation to offer more even experiences to their systems' users while paying attention to possible moderators of its effect**.

Second, our findings provide considerations on **how to design personalised gamification**. Results from **RQ3** question whether using only three game elements and personalising by changing the game elements available are the best choices for deploying and personalising gamified designs. Therefore, we contribute with considerations that **gamified designs with more than three game elements might improve users' perceptions about gamification** (TH3) and that **successful personalisation 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**. Results from **RQ3** also revealed that students had positive experiences while completing the gamified review assessments and that completing the assessments contributed to their learning. Such finding is important, because educational activities are not motivating for students oftentimes, and low motivation harms learning performance (Hanus & Fox, 2015; Palomino et al., 2020; Rodrigues, Toda, Oliveira, Palomino, Avila-Santos, & Isotani, 2021). Thus, this is informing designers of 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 relates to **how personalisation contributes to gamification**. By triangulating findings from **H1** and **RQ2**, it seems personalisation 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 personalisation contributes to gamification?** In exploring answers to that question, results from **RQ2** led to considerations suggesting that **personalisation mitigates the sensitiveness of OSFA gamification outcomes to user and contextual factors** (TH1 and TH7). Additionally, triangulating findings from **H1** and the preliminary analyses suggested **personalisation increases user motivation after a first-time experience** (TH2) when samples' characteristics are comparable.

Second, our results inform researchers on **predeterminants for the success of personalised gamification**. Results from **RQ1** suggested when/to whom personalisation was more or less effective. Consequently, providing theoretical considerations that the **personalisation 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 personalisation strategies**. In discussing **RQ1**, we compared preference-based and data-driven personalisation strategies. Note that they are complementary: the former allows personalising systems from user's first use, while the latter rely on true usage data (Tondello, 2019). Nevertheless, the limitations from user preference and evidence supporting data-driven personalisation's effectiveness (e.g., Hajarian et al. (2019)) led to the theoretical consideration that **data-driven personalisation 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 empirical computer science (Cockburn et al., 2020; Vornhagen et al., 2020). Additionally, we personalised gamification based on a freely available recommender system (Rodrigues, Toda, Oliveira, Palomino, Vassileva, & Isotani, 2021), besides using a gamified educational system (Eagle-Edu) that is education/research friendly. Thus, we are extending our contribution and facilitating replications.

8 Limitations and Future Work

This section discusses study limitations and presents future research directions accordingly. First, our sample size (n = 58) 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 our eligible participants. Because this study was conducted during the COVID-19 pandemic, instructors mentioned they witnessed unseen drop-out rates, which might explain the attrition rate. While that size affects findings' generalisation, we believe that having conducted a multi-site study in ecological settings leads to a positive trade-off. Despite sample size also affects our confirmatory analyses' validity as it implies low statistical power, we sought to mitigate that issue by only conducting planned comparisons.

On the one hand, we planned a one-factor experiment to compare OSFA and personalisation. Because personalisation is built upon the idea of having different people using different designs depending on, for instance, their characteristics, it is expected to have a distinct number of participants in each gamification design. Hence, we believe this aspect does not hinder our study validity. On the other hand, exploratory analyses' results are more prone to sample size limitations because they rely on subgroups. This is the reason our exploratory analyses were limited to one-factor comparisons (e.g., males versus females) - instead of multi-factor ones (e.g., females who prefer single-playing versus females who prefer multi-playing; then the same for males) - and explicitly presented them as findings *to be tested* (see Section 4). To further cope with this limitation, we used 90% CIs measured through bootstrap, aiming

to increase results' reliability while avoiding misleading conclusions that could emerge from reporting and interpreting p-values (Cairns, 2019; Carpenter & Bithell, 2000; Vornhagen et al., 2020).

Second, our study is limited to a single dimension of Bloom's taxonomy. That was a design choice aimed to increase internal validity, whereas a multi-factor experimental study (e.g., comparing multiple knowledge dimensions) would add numerous confounders and require a larger sample. While this limits the generalisability of our findings, we believe that choice leads to a positive trade-off in allowing us to approach a specific research problem with greater validity. Hence, given research discussing the learning task affects gamification' success (Hallifax, Serna, et al., 2019; Rodrigues et al., 2019), only further empirical evidence can answer whether our findings will be the same for other dimensions of Bloom's taxonomy. Thus, especially considering our findings confront those of similar research (Rodrigues, Palomino, et al., 2021), those sample size and research context limitations, along with our testable hypotheses, provide directions for future replication studies that must validate and test our results' generalisation.

Third, 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 not that 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 more often and over multiple terms. Based on those, we recommend future studies to analyse how personalised and OSFA gamification compare, in the context of review assessments, when used for longer periods of time.

Lastly, there are four limitations related to our study's instruments/apparatus. Concerning our measure, the external regulation construct showed questionable reliability, similar to prior 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). Additionally, students might have completed the instrument based on different experiences than they had with the gamified system. While we carefully instructed them on how to complete SIMS, we cannot ensure this due to human biases and subjectivity (Blandford et al., 2016; Wohlin et al., 2012). 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 personalisation strategy, it was developed based on user preference, which is often criticized compared to data-driven approaches (Norman, 2004), 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 a *small* competition affects students' motivations, and iii) extending the external validity of the personalisation strategy introduced in Rodrigues, Toda, Oliveira, Palomino, Vassileva, and Isotani (2021).

9 Conclusion

VLE 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 OSFA gamification can improve motivation, the effects' variation inspired research on personalised gamification. However, there is little knowledge on how personalisation contributes to OSFA gamification. Therefore, we conducted a multi-site experimental study wherein students completed gamified assessments with either personalised or OSFA gamification. Our results suggest a new way of seeing personalisation's role in gamification and inform designers, instructors, and researchers:

- We show that, whereas personalisation might not increase the 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 discussion provides design and research directions toward advancing the field study;

Our results inform i) designers interested in personalised 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) personalised gamification researchers with guidance on how to advance the field study. Also, 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

Abt, K. (1987). Descriptive data analysis: a concept between confirmatory and exploratory data analysis. *Methods of information in medicine*, *26*(02), 77–88.

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.

Dragicevic, P. (2016). Fair statistical communication in hci. *Modern statistical methods for hci* (pp. 291–330). Springer.

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).

Gari, M., Walia, G., Radermacher, A. (2018). Gamification in computer science education: A systematic literature review. *American society for engineering education.*

Gliem, J.A., & Gliem, R.R. (2003). Calculating, interpreting, and reporting cronbach's alpha reliability coefficient for likert-type scales..

Grin, E., Burgos, F., Fernandes, G., Bresciani, L. (2021, 09). O mapa regional das múltiplas desigualdades e do desenvolvimento humano no brasil. In (p. 99-122).

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

Hothorn, T., Hornik, K., Zeileis, A. (Jan 2006). Unbiased recursive partitioning: A conditional inference framework. *Journal of Computational and Graphical Statistics*, *15*(3), 651-674. Retrieved from https://doi.org/10.1198/106186006X133933

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).

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).

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., Oyibo, K., Tondello, G.F. (2017). A comparison of system-controlled and user-controlled personalization approaches. *Adjunct publication of the 25th conference on user modeling, adaptation and personalization* (pp. 413–418). New York, NY, USA: ACM. Retrieved from http://doi.acm.org/10.1145/3099023.3099116 10.1145/3099023.3099116

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.

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.

Ryan, R.M., & Deci, E.L. (2017). *Self-determination theory: Basic psychological needs in motivation, development, and wellness.* Guilford Publications.

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

Sheskin, D.J. (2003). *Handbook of parametric and nonparametric statistical procedures*. Chapman and Hall/CRC.

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.

Venter, M. (2020). Gamification in stem programming courses: State of the art. *2020 ieee global engineering education conference (educon)* (pp. 859–866).

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.

Wohlin, C., Runeson, P., Hst, M., Ohlsson, M.C., Regnell, B., Wessln, A. (2012). *Experimentation in software engineering.* Springer Publishing Company, Incorporated.

Zainuddin, Z., Chu, S.K.W., Shujahat, M., Perera, C.J. (2020). The impact of gamification on learning and instruction: A systematic review of empirical evidence. *Educational Research Review*, 100326.

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.