User Modeling and User-Adapted Interaction (2021) 31:1–34
https://doi.org/10.1007/s11257-019-09255-2

Automatic generation and recommendation of personalized challenges for gamification

Reza Khoshkangini[1] · Giuseppe Valetto[1] · Annapaola Marconi[1] · Marco Pistore[1]

Received: 27 October 2018 / Accepted in revised form: 18 December 2019 / Published online: 24 May 2020

Abstract
Gamification, that is, the usage of game content in non-game contexts, has been successfully employed in several application domains to foster end users' engagement and to induce a change in their behavior. Despite its impact potential, well-known limitations concern retaining players and sustaining over time the newly adopted behavior. This problem can be sourced from two common errors: basic game elements that are considered at design time and a one-size-fits-all strategy in generating game content. The former issue refers to the fact that most gamified applications focus only on the superficial layer of game design elements, such as points, badges and leaderboards, and do not exploit the full potential of games in terms of engagement and motivation; the latter relates to a lack of personalization, since the game content proposed to players does not take into consideration their specific abilities, skills and preferences. Taken together, these issues often lead to players' boredom or frustration. The game element of *challenges*, which propose a demanding but achievable goal and rewarding completion, has empirically proved effective to keep players' interest alive and to sustain their engagement over time. However, they require a significant effort from game designers, who must periodically conceive new challenges, align goals with the objectives of the gamification campaign, balance those goals with rewards and define assignment criteria to the player population. Our hypothesis is that we can overcome these limitations by automatically generating challenges, which are personalized to each individual player throughout the game. To this end, we have designed and implemented a fully automated system for the dynamic generation and recommendation of challenges, which are personalized and contextualized based on the preferences, history, game status and performances of each player. The proposed approach is generic and can be applied in different gamification application contexts. In this paper, we present its implementation within a large-scale and long-running open-field experiment promoting sustainable urban mobility that lasted 12 weeks and involved more than 400 active players. A comparative evaluation is performed, considering challenges that are generated and assigned fully automatically through our system versus

Extended author information available on the last page of the article

analogous challenges developed and assigned by human game designers. The evaluation covers the acceptance of challenges by players, the impact induced on players' behavior, as well as the efficiency in terms of rewarding cost. The evaluation results are very encouraging and suggest that procedural content generation applied to the customization of challenges has a great potential to enhance the performance of gamification applications and augment their engagement and persuasive power.

Keywords Gamification · Recommender systems · Procedural content generation · Personalization

1 Introduction

A significant effort has been put in the last few years among researchers as well as practitioners to understand how interactive technologies can be leveraged to encourage and promote more sustainable lifestyles (Di Salvo et al. 2010; Huang 2011; Hojer and Wangel 2015). Among others, gamification is emerging as a persuasive technology (Fogg 2002) with significant potential and applications in numerous domains (Hamari et al. 2014b).

We have been developing and experimenting with a gamification framework for Smart Cities (Kazhamiakin et al. 2015, 2016). Smart Cities are an eminent socio-technical system, with a constant flow of interactions between citizens and a myriad of technological affordances—including information systems, services and service providers, mobile computing environments and applications, sensors, smart objects, etc. For that reason, and because of their scale, the number of different aspects of city life that can be positively impacted, and the sheer magnitude of the potential upside, they are an ideal environment for experimenting with, and operationalizing, advanced persuasive technologies for behavioral change, including gamification (Hojer and Wangel 2015).

While gamification applications that promote sustainable behaviors have shown a remarkable impact potential in a number of critical Smart City domains, including recycling (Lessel et al. 2015), energy conservation (Simon et al. 2016; Brewer et al. 2015) and personal mobility (Froehlich et al. 2009; Kazhamiakin et al. 2015; Ferron et al. 2019), a shortcoming that is often observed is that their effects tend to diminish and taper off over time (Farzan et al. 2008; Hamari et al. 2014a). That is especially unfortunate, because a behavioral change cannot be immediately internalized; rather, it must be reinforced long enough to successfully form new sustainable habits (Weiser et al. 2015). This problem can be sourced from two common design errors characterizing most gamified applications: the dominance of simple and static game elements (e.g., points, badges and leaderboards) and a one-size-fits-all strategy in proposing game content to players during the game (Chou 2015; Charles et al. 2005; Das et al. 2015; Hamari et al. 2014b; Weiser et al. 2015).

In our previous research, we have investigated approaches to counter these issues, and we have experimented with the dynamic injection of a variety of fine-grained playable units, in the form of individually personalized *challenges*. Challenges are units of playable content that set a demanding goal that a player should achieve—

under temporal or other constraints—in exchange for an in-game prize or reward. Challenges have empirically proved effective to "nudge" players toward desirable behaviors according to the goals of the gamification campaign, to keep players' interest alive and to sustain their engagement over time (Kazhamiakin et al. 2016; Khoshkangini et al. 2017).

While injecting automatically in our gamification framework the software code to define, present and evaluate those challenges is rather straightforward, they require a significant effort on the part of game designers, who must periodically conceive challenges, set and align goals with the objectives of the gamification campaign, balance those goals with rewards, define assignment criteria to the player population and so on. This design work not only is laborious, but also requires specialized knowledge of the game mechanics as well as the domain being gamified; therefore, such a design effort can hardly be repeated long term, for example in support of permanent or semipermanent campaigns.

Motivated by all the above considerations, we developed a fully automated procedural content generation (PCG) and recommendation approach (Hendrikx et al. 2013) for the dynamic and personalized injection of challenges, which is a new component of our gamification framework.

PCG techniques are used increasingly often in contemporary electronic games to computationally generate a wide variety of game elements and greatly increase the diversity of game play without incurring huge editorial, design and implementation costs. They can keep players engaged and interested with a diversified and enhanced game experience, which can be dynamically adapted to the personal preferences, abilities and style of each individual player (Lopes and Bidarra 2011).

Our work is among the first to introduce forms of PCG also in gamification applications, coupled with recommendations upon the generated game elements. Our system produces challenges that are personalized (based on the preferences and the past game history and performance of each player) and contextualized (based on the current state of the player in the game and her game objectives). The proposed solution addresses both the identification of a tailored challenge goal, in terms of the effort required to the player and of the deriving difficulty, and the computation of a commensurate reward. Moreover, when recommending challenges, the system takes into account any policies or objectives, which the entity administering the game (in our case, the administration of the Smart City) means to promote. In that way, we can incentivize specific citizens' behaviors that are in line with those policies/objectives, which can dynamically change during the game. Our PCG and recommendation system fully automates the generation of compelling, personalized and varied playable units; it thus slashes the effort necessary to administer long gamification campaigns and supports the engagement and long-term retainment of participants.

The following research questions (RQ), covering three complementary aspects, were defined to evaluate the effectiveness and efficiency of the proposed fully automated solution:

– *RQ1—Player acceptance* How does the player acceptance rate for the automatically generated challenges compare to the acceptance rate of the same types of challenges assigned manually (i.e., via expert judgment)?

– *RQ2—Challenge impact* How does the improvement recorded on the target goal for the automatically generated challenges compare to the improvement recorded for the challenges assigned via expert judgment?
– *RQ3—Reward efficiency* How do the rewards computed for the automatically generated challenges compare to those of challenges assigned via expert judgment?

Taken together, these research objectives outline a comparative evaluation between playable units generated and assigned fully automatically, and analogous units developed and assigned by human game designers. The latter had already been shown empirically their efficacy in inducing significant as well as sustained behavioral change (Kazhamiakin et al. 2016; Khoshkangini et al. 2017). Our evaluation tries to investigate whether the former can lead to comparable or even superior persuasive outcomes.

In this paper, we describe in detail our approach and system for the fully automated generation of individually personalized challenges and their recommendation, which we have deployed in the Smart City of Trento, Italy, in the course of a large-scale (more than 400 citizen actively playing) and long-running (12 weeks) gamification campaign on sustainable urban mobility. The system evaluation was performed during the last 3 weeks of that game, by means of an A/B test that, guided by the aforementioned RQs, compares it with the semiautomatic approach that relied greatly on the game designers' experience and on their development effort for challenge assignment to segments of the player population (Kazhamiakin et al. 2016). We report hereby on the results of that A/B test, which shows how PCG applied to the generation of playable units of content has a great potential to enhance gamification applications and augment their persuasive power.

The rest of the paper is organized as follows: We present relevant related work in Sect. 2 and provide some background on our gamification framework for Smart Cities in Sect. 3. We describe our PCG and recommendation approach and system in Sect. 4 and discuss its evaluation in Sect. 5. Finally, we conclude by discussion and future works in Sect. 6.

2 Related work

The main application domain we have chosen for our gamification research is environmental sustainability in Smart Cities. This is an area in which gamification approaches have been widely applied in the last few years, ranging from energy saving (Shiraishi et al. 2009; Cowley et al. 2011; Orland et al. 2014; Du et al. 2014), waste recycling (Lessel et al. 2015; Zlatow and Kelliher 2007; Greengard 2010), sustainable mobility (Broll et al. 2012; Gabrielli et al. 2013; Kazhamiakin et al. 2015, 2016) and other environmental issues, such as community-wide environmental missions (Lee et al. 2013), participatory governance of urban neighborhoods (Coenen et al. 2013) or educational city discovery (Gordillo et al. 2013).

Although all aforementioned studies vary widely in terms of the motivational affordances they implement application context, nature of the gamified system and ways in which their effectiveness has been measured, in most cases gamification has proved to be successful as a persuasive technology and for raising awareness and

changing the behavior toward more sustainable habits. However, its impact is often transient and tends to diminish with time, unless it is reinforced with opportune motivational affordances (Weiser et al. 2015), and corresponding game design elements and mechanics (Khoshkangini et al. 2017). These diminishing effects are confirmed also from studies on gamification applications in other domains (Hamari et al. 2014b).

Our research aims at countering precisely the phenomenon mentioned above. Since PCG can generate game content that varies during the game and that is dynamically tailored to each player, we postulate that procedural content generation can play a key role in that, by producing a variety of diverse game elements that enhance the game experience and hence support players' retainment in the long term. We also maintain that the full potential of a PCG approach can be reached only if PCG is coupled with a recommendation system, in order to select for, and propose to, each individual player targeted game elements that are contextualized and personalized with respect to her preferences, skills and playing style.

The integration of these two elements together in the domain of gamification drives our research agenda and constitutes the main contribution of this paper. Therefore, before we dive into the details of our approach, we review some relevant results in the areas of procedural content generation for digital games and recommender systems, and position our own work with respect to those areas.

2.1 Procedural content generation

Online services, digital games and web interfaces are the most common applications in which PCG techniques are used. Among those, many attentions in PCG have been paid in digital games in order to optimize and improve players' experience within the game (Yannakakis and Togelius 2011).

PCG technologies automate the construction of game design elements (GDEs) to be delivered in electronic games. Elements may vary widely (Hendrikx et al. 2013), from sounds and textures, to buildings, maps and whole game level layouts, to items, equipment and other virtual goods, all the way to *playable units of content* such as riddles to solve, obstacles to overcome, encounters with non-playing characters, challenges, missions and quests to complete, and even entire story lines. Oftentimes, PCG algorithms for playable units take into account players' skills and game state, including taking into account the choices players make, tries and errors, and player's physiological signals which are logged throughout the game (Yannakakis and Hallam 2007; Yannakakis et al. 2010).

Advancements in PCG have largely been driven by the need to contain the time and cost for developing diversified game content at scale, while improving the experience of players, who dislike excessive repetition. PCG has also been used to automatically adapt the game experience to the characteristics of each player (Lopes and Bidarra 2011; Karpinskyj et al. 2014). For example, Zook et al. (2012a, b) have used PCG to tailor the game play to a model that captures the player's current know-how and performance, and have applied that to both missions in a role-playing fantasy game and training scenarios in a military serious game. This personalization of game content shows great potential to avoid frustrating video game players, and keep them

engaged (Charles et al. 2005; Das et al. 2015). For instance, Khajah et al. (2016) exploit Bayesian optimization approaches (in particular they used Gaussian process surrogate-based optimization) in order to construct games that maximize players engagement.

PCG is also a means to automatically adjust the difficulty of playable units of content, and ensure a balance between player's satisfaction and challenge over time according to the concept of "flow," which is recognized as a major factor for fun and retention (Chen 2007; Cowley et al. 2008). For instance, Van Lankveld et al. (2009) worked on game difficulty balancing based on the theory of incongruity, in which they assessed to what extend game adaptation can be exploited to enhance the entertainment value of a game (Glove). In the same context of game balancing, Cooper et al. (2016) exploited player rating systems [e.g., Elo which is a method to calculate the players' relative skills in competitor games like *Chess* (Elo 1978)] to select and sequence tasks, with the goal to balance difficulty in human computation games (HCGs). In Togelius et al. (2007), race tracks are generated according to acquired driving models of the players, to improve the entertainment value of an auto racing game, and Lora et al. (2016) proposed a way to dynamically change the difficulty of the *Tetris* game. Bakkes et al. (2014) developed an adaptation version of Infinite Mario by providing personalized challenges, while the player is playing the game. This version of the game has been improved based on facial expression analysis by Blom et al. (2014).

The power of computational techniques adopted in PCG has quickly grown with the demand for more and more sophisticated generated content in contemporary electronic games. All effective PCG contains two facets: *exploration*, which generates a potentially very large and diverse number of options, and *selection*, which picks among the generated options those that are fittest for purpose. For the former, some of the prevalent approaches are search-based techniques (e.g., genetic algorithms) (Togelius et al. 2011) and planning techniques, in particular when creating new gameplay scenarios or storylines (Pizzi et al. 2010; Soares de Lima et al. 2014). For the latter, some PCG approaches leverage insights from recommender systems (RS).

The above works demonstrate the recent attention in exploiting PCG techniques to foster players' experiences in digital entertainment games, which resonates with the goals of our research; at the same time, they underline how the rich potential of these techniques can be exploited and expanded in germane but different contexts such as (gamification). Furthermore, enhancing PCG with a recommender system that helps with the selection of the generated game elements may have a positive influence on personalization and, in turn, can improve players' engagement during gamification campaigns.

2.2 Recommender systems in games

In the domain of digital games, RS have often been used to recommend new games that fit well the player's preferences and game play. For instance, Sifa et al. (2014) proposed two different techniques of archetypal analysis for an RS, which, based on the player's activity as she plays, recommends new games, and Skocir et al. (2012) designed a multi-agent recommendation system that uses a number of different parameters elicited

during game play, to propose new mobile games that best suit the player's style and skills.

More similarly to our line of research, a recommender system is used in combination with PCG (Raffe et al. 2013), to take into account individual player characteristics and past performance, when proposing procedurally generated maps and content in a game platform. A similar collaborative system is designed in Cheng and Vassileva (2006) in which a novel "incentives adaptive reward mechanism" is exploited to target individual and group of students with the aim of enhancing the quality and quantity of students' contributions in their learning process. In Harrison et al. (2015), collaborative filtering based on past player achievements is used to propose achievements that a given player may enjoy most taking on next, for the *World of Warcraft* MMORPG. Earlier, another collaborative filtering approach—in the online community context—had been applied in Slashdot, to exploit the feedback (in terms of ratings) on the quality of comments users post to the site (Lampe 2006).

In the context of automating the generation of game content, Andersen et al. (2013) introduced a theoretical trace-based framework which was demonstrated on two domains—elementary and middle school mathematics, and well-known puzzle game *Refraction*—to estimate the difficulty of the procedural problem by mining the features that were traced in the game. Later on, a large-scale experiment has been evaluated on the same puzzle game *Refraction* by proposing a framework that automates the generation of "level progression" aimed to enhance the players' engagement in the game (Butler et al. 2015). In gamification proper, the work closest to ours is perhaps *PHESS* (Silva et al. 2013), which produces recommendations on sustainability actions based on monitored and accumulated user history, in the domain of home energy conservation. In that work, though, those actions are not automatically generated, but taken from a limited repertoire of static alternatives.

All in all, two main limitations can be observed in the previously presented works: (1) most of them are limited to a particular application domain and cannot be generalized, and (2) they still suffer from the lack of a systematic and fully automated solution for generating game content. To our knowledge, the combination of computational generation of playable content and recommender systems in gamified applications, as a strategy to increase engagement and ensure retainment of players in gamified applications, is novel. Our work addresses this, with a procedural content generation approach, which is generic, i.e., applicable to a variety of gamification applications, and a recommender system for personalized and contextualized selection and assignment of the generated units of playable contents. In the remainder of this paper, we describe our approach and evaluate its potential.

3 Background

3.1 Gamification platform in Smart City

This gamification platform has been designed and developed having in mind the following principles: (i) open-ended integration of existing IT systems and services in a Smart City, whose interactions with citizens must become part of the game as player

actions; (ii) support for dynamic city policies and objectives, as well as a set of extensible game design elements; (iii) sustain behavioral change through long-running games that can keep players engaged in the long term.

The architecture of our gamification platform supports the entire game life cycle (i.e., design, deployment, execution and analysis of games) and is organized in three layers (see Fig. 2 for an overview).

- The *gamification enablers* layer supports for the basic functionality related to the design, deployment (*game definition*) and execution (*gamification engine*) of games, and its integration with Smart Cities IT systems (*wrapping*). This layer is also responsible for the automatic generation of personalized and contextualized challenges (the *challenge generator* component). Being the focus of this work, a detailed description of this component can be found in the next section.
- The *gamification services* layer exposes the functionalities realized by the enablers as services, which can be exploited to build new gamification components and applications (e.g., services supporting the definition and deployment of games, services for accessing information about game and player state, services for supporting the configuration of players notifications).
- The *gamification front-end* layer contains end user applications for the different stakeholders: It provides applications supporting the definition and deployment of games (for game experts), the presentation of game state (for game players) and the analysis of game results and impact vis-a-vis the city objectives (for officials and decision makers).

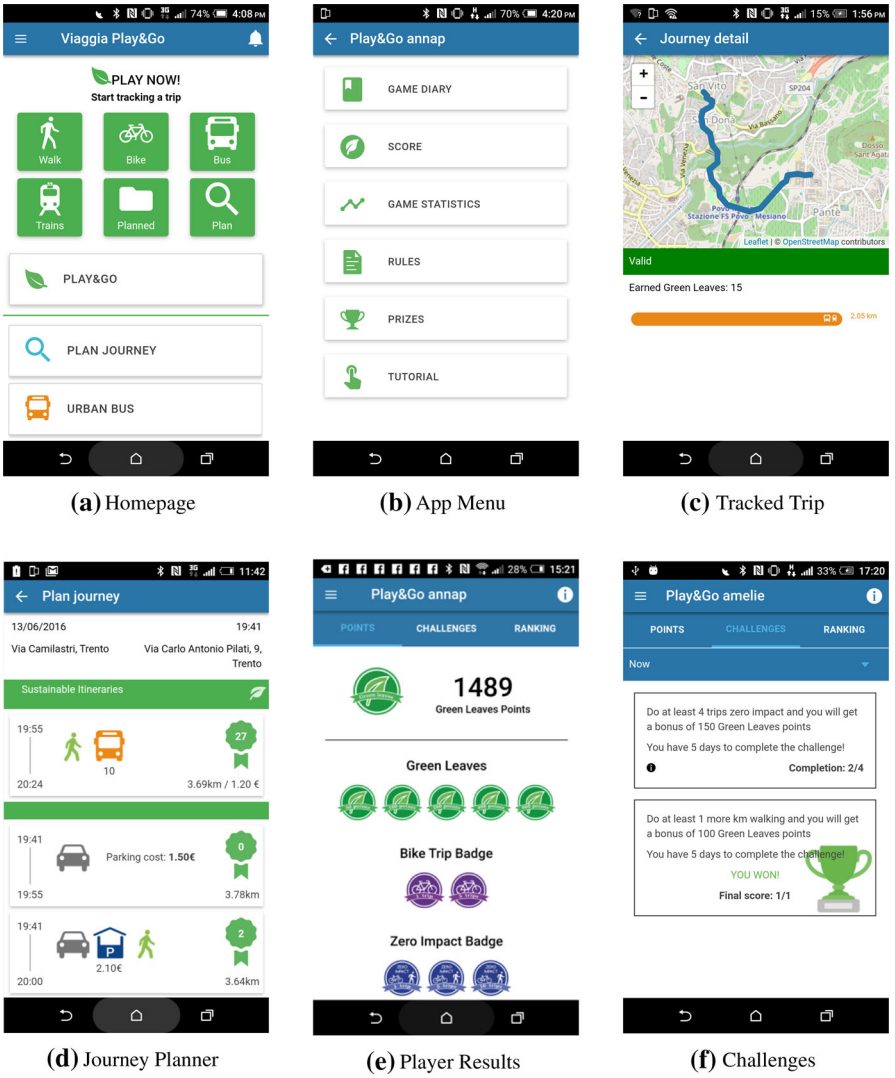
The developed platform has been released in GitHub under the Apache License version 2.0[1] and is available as a stand-alone application as well as a software-as-a-service. The gamification platform has been deployed and experimented in a variety of Smart City games around Europe. Although its main application domain so far has been sustainable urban mobility, the gamification platform is generic, and we are exploring its application to various other Smart City domains, including energy efficiency, participatory e-government, educational games, and health care and well-being. Within the sustainable mobility domain, we have developed a gamified App called `Viaggia Play&Go`[2] that players can install, register and interact with using their personal handheld Android and iOS devices.

3.2 Viaggia Play&Go gamified application

Trento Play&Go was a large-scale and long-running open-field gamification campaign that lasted 12 weeks (from September 10 to December 2, 2016). The game was targeting residents of the city of Trento in Italy, as well as commuters from the surrounding area of the Trentino province. Citizens could take part in the game using the Viaggia Play&Go mobile app (available on the official Android and Apple stores), which supported them in planning and tracking their journeys, checking their status in the game, sharing their results on social networks (i.e., Facebook or Twitter), inspecting the

[1] See https://github.com/smartcommunitylab/smartcampus.gamification.

[2] https://www.smartcommunitylab.it/apps/viaggia-trento-e-rovereto-playgo/.



(a) Homepage **(b)** App Menu **(c)** Tracked Trip

(d) Journey Planner **(e)** Player Results **(f)** Challenges

Fig. 1 Screenshots of the Viaggia Play&Go mobile app

rules of the game and being notified of weekly prizes. Figure 1a shows the homepage of the app and Fig. 1b the menu for accessing the various game-related app features.

The Trento Play&Go game contains different motivational gamification elements. The app assigns Green Leaves points for each tracked journey (see Fig. 1c). The apps supported the following sustainable transportation means: walking, bike, bus and train trips. The points obtained depend on the kilometers traveled and on the level of sustainability of the transportation means. Bonus points are associated with zero-impact trips (i.e., trips taken with transportation means that cause no CO_2 emissions). Each tracked journey is subjected to a semiautomatic validation procedure that assesses

whether the trip is legal in terms of minimum length and compatibility with the declared transportation. The validation system is based on the smartphone GPS data, which allow to compute position and speed, as well as on the knowledge about routes and timetables of the public transportation means. If the trip is valid, the computed amount of Green Leaves points is assigned to the player. If players have recurrent journeys, i.e., daily commutes, they can plan them through the journey planner included in the mobile app (see Fig. 1d), save them as recurrent journeys and use them anytime it suits them. The planned journeys allow to save multimodal trips (i.e., trips in which more than one transportation means is involved).

The game also supported weekly and global leaderboards, as well as a variety of badges and badge collections assigned when reaching certain achievements; for example, taking a certain number of trips in a specific mode or exploring new mobility resources (e.g., when trying the bike sharing service for the first time or exploring different bike sharing stations). The player can inspect the earned Green Leaves and badges through the app (see Fig. 1e).

On top of the basic game rules, we introduced weekly themes (e.g., a bike week, a zero-impact week, a public transport week, etc.) in conjunction with the city administration of Trento and its mobility objectives. Those objectives were associated with weekly challenges targeting the specific theme. Since a one-size-fits-all approach may be ineffective in the process of promoting a positive behavioral change in individuals, the main strength of the game is the introduction of highly personalized weekly challenges. These challenges aim at improving, or maintaining, the player's performance in the game by asking them to reach a target (e.g., number of trips, number of kilometers) with a certain transport means (e.g., by bike, foot, bus or train) within a certain time constraint (e.g., in 1 week). Upon completion, challenges award additional Green Leaves. This reward is calibrated to the difficulty of the challenge and based on the player performances. For instance, during the bike week, we had different types of challenges promoting the usage of the bicycle, which were targeted to the player's profile, i.e., players that were already employing bikes were asked to increase either kilometers or number of trips (e.g., "Do at least 30 km by bike during the current week to earn 200 Green Leaves"), while players that had not previously used this transportation means were asked to try it (e.g., "Do at least 1 trip by bike during the current week to earn 150 Green Leaves"). This was a design choice to make the game attractive to, and playable by, newcomers (who are encouraged to compete in the short-term challenges and ranks), as well as to sustain participation of already committed players in the long run. The app contains a section dedicated to the weekly challenges (Fig. 1f), in which players can keep track of the state of their current challenges and inspect the outcome of the previous ones.

Final and weekly tangible prizes are another important motivational aspect of the game. There were 8 appealing final prizes: For example, the final prize for the player who collected the most points over the entire 12-week period was a paid vacation of 3 days in a hotel. But the main intended emphasis was on weekly prizes, which were smaller but quantitatively much more numerous, since two to three prizes were awarded each week. A prize was assigned to the top player in the weekly leaderboard and the remaining prizes were assigned by draw on the weekly Top 50 (the 50 players with most point collected during the week). Weekly prizes included yearly subscriptions

to bike sharing and car sharing, tickets for music shows, sport events and museums. All the prizes were offered by local sponsors. Assigning weekly prizes was a design choice to give every player the opportunity to feel involved and have a chance to win something for her good behavior and her effort.

In the next section, we describe our proposed approach to show how the personalized challenges are generated and personalized, and how they are recommended to players in the Trento Play&Go game.

4 Technical approach

Borrowing from the taxonomy in Togelius et al. (2011), our procedural generation and recommendation approach can be characterized as an *online*—as it takes into account the current player state—*constructive*—as the generated instances are built all at once for every round of challenge administration, and are all valid by design—and *parameterized*—as the generative process works by choosing appropriate values in a parameter space—case of PCG applied to gamification, which results in the *automatic generation and injection of personalized and contextualized units of playable content*.

4.1 Challenge model

We define the playable unit as a tuple: $<P; G; C; D; R; W>$,where

- P refers to the individual *player* to whom the challenge is assigned. P has a profile, which contains her preferences, either *explicitly* or *implicitly* derived. For instance, in the context of sustainable mobility, *explicit* preferences refer to specific transport means that the player expressed as preferred during the registration survey, while *implicit* preferences are derived from the actual player's behavior which is captured during the game.
- G defines the *goal* that is a task or a performance target, which should be fulfilled to complete the challenge. For example, in our sustainable mobility game, a goal can be expressed as "take at least 6 public transport trips."
- C is the *constraint* for reaching the goal; a typical example is a *temporal deadline* that is used in this evaluation, e.g., player P must achieve goal G within 1 week.
- D represents the *difficulty* of the challenge for player P, considering goal G and constraint C. For D, we have been using a four-level scale: *{Easy, Medium, Hard or Very Hard}*. Notice that the difficulty level for the same challenge, that is, with same goal and constraint, may be different for—and tailored to—different players.
- R is the *reward* (aka prize) awarded for completing the challenge. We define rewards in terms of GDEs that are part of the game: for example, a prize can be a bonus or booster for points accumulation, a badge in a collection of achievements, etc. Rewards should be commensurate to the difficulty of the challenge.
- W is a numeric *weight* that captures how important challenge goal G—and the behavior it means to foster—are, according to the entity promoting the gamification campaign. For instance, in a sustainable mobility game, a Smart City administra-

tion may assign during its "public transportation promotion week" highest weights to challenges that increase usage of buses, trains, etc.

In the following, we list some examples of challenges that our proposed framework, which is depicted in Fig. 2, recommends to players:

- *Example 1*: "Increase <Walking> <Km> by at least <10%> during <next week> and receive <200> <Green Leaves>."
- *Example 2*: "Increase ⟨Train⟩ ⟨trips⟩ by at least ⟨30%⟩ during ⟨next month⟩ and receive additional ⟨20⟩ Green Leaves per ⟨trip⟩."
- *Example 3*: "Do at least ⟨1⟩ ⟨bike sharing⟩ ⟨trip⟩ ⟨next week⟩ and receive ⟨80⟩ ⟨Green Leaves⟩."

The challenge model is filled up at different stages by passing through the modules, as depicted in red font in Fig. 2, to generate challenge instances that can be injected in the game and proposed to players.

In particular, the *challenge generator* module is responsible of producing a set of candidate challenge instances for each player. The *challenge valuator* module computes, for each challenge instance, the player-specific difficulty and a commensurate reward. The *filtering and sorting* module receives the list of challenges from the *challenge valuator* module. Then, taking into consideration multiple parameters from player's profile (e.g., challenge history, game status), city's objectives, as well as the global game status, challenge recommendation takes place to propose the personalized challenges to the player (red line in Fig. 2).

In the next sections, we describe this PCG process, and the various modules responsible for it.

4.2 Challenge generator

This module is responsible for generating a set of candidate challenge instances that are then stored in the *challenge repository*, by instantiating the player (P), goal (G), constraint (C) and weight (W) parameters of the challenge model.

In our sustainable mobility game, G consists of two sub-elements: the mobility indicator MI and the target T that specifies the target value to be reached to win the challenge. As mobility indicators we considered both Km and trips to be done by different transport modes (i.e., walk, bike, bus, train, bike sharing, park and ride) or combination of modes (i.e., public transport, impact zero, no car). In the game we also set the constraint C to "1 week" for all generated challenge instances.

For example, a challenge instance with values, P = Bob, MI = Bus_Trips, $T = 8$ and $C = 1$ week, asks player Bob to take at least 8 trips by bus in the next week.

To instantiate challenges, the module takes into account players' history (particularly, previous week) and follows the below approach for parameterizing G:

- for each player P and each mobility indicator MI, if P has done any activity for MI during previous week, the algorithm fills T considering a set of target percentage improvements (i.e., 10%, 20%, 30%, …, 100% as defined in Table 1). For instance, if Bob has done 10 Walk_Km in the second week of the game, the module takes this number of kilometers traveled as the baseline to instantiate the target values for

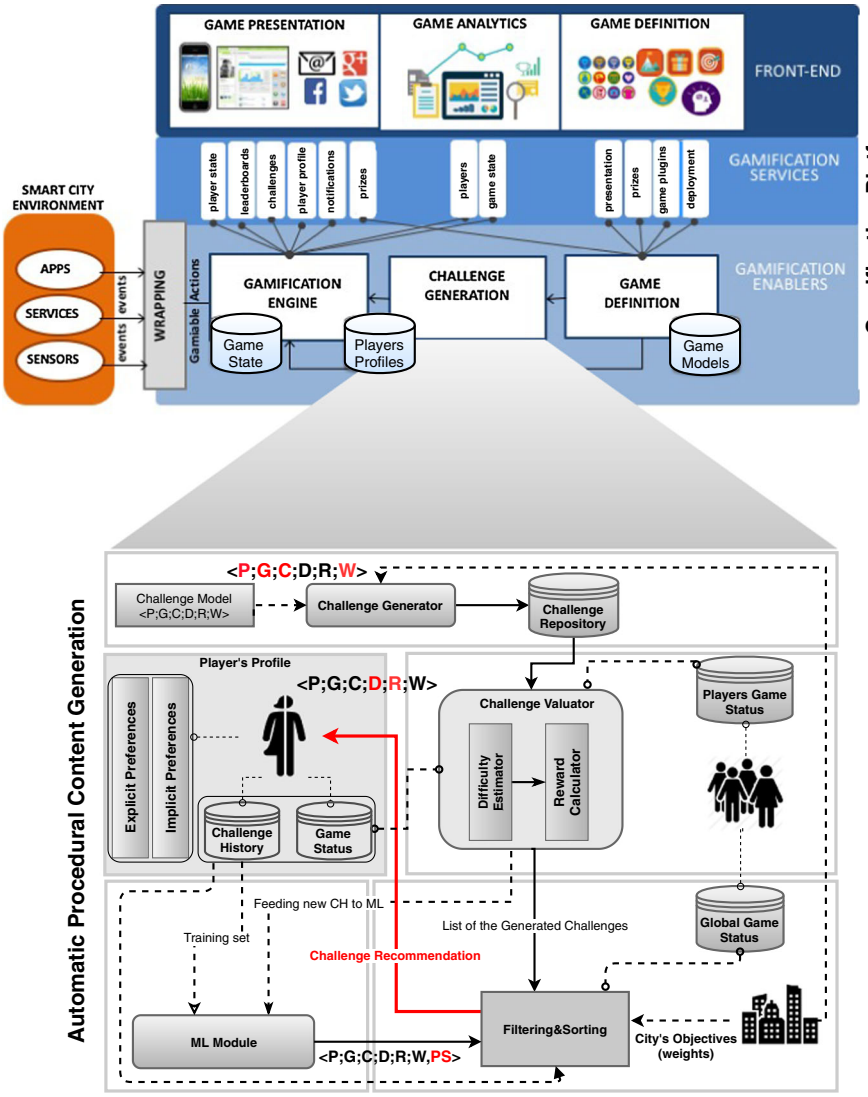


Fig. 2 Conceptual view of the gamification platform and automatic procedural challenge generator framework

candidate Walk_Km challenges in the third week. In this example, it will generate ten challenge instances: "Do at least $T \in \{11, 12, 13, \ldots, 20\}$ MI = Walk_Km during C = next week."

– If the player was not active for a specific mobility indicator MI during the previous game week, the algorithm sets T to 1, requiring the player to do at least one trip. For example, "Do at least $T = 1$ MI = Bike_Trips during C = next week."

More candidate challenges are similarly derived for all players and mobility indicators, and they are all stored in our challenge repository. This challenge generation

process is time-bound and is repeated weekly. Although for our mobility game we set constraint C to 1 week, the approach is generic and can support for different constraints on challenge duration.

4.3 Challenge valuator

The purpose of this module is to estimate the difficulty of each candidate challenge instance produced by the challenge generator for a given player and to compute a commensurate reward. Therefore, it is responsible to fill the D and R parameters of our challenge model. Accordingly, it consists of two sub-modules: *difficulty estimator* and *reward calculator*.

4.3.1 Estimation of difficulty

Accurately estimating the difficulty of a challenge for a given player is important for two reasons: firstly, it is a prerequisite for the system to assign a fair reward for the effort required to complete that challenge; secondly, it contributes to keep the player interested in the game by continually striking a good balance between the satisfaction for accomplishing goals and the stimulation of being challenged (Cowley et al. 2008; Tulloch 2014).

Since players' skills, performances and styles change during game execution, affected by internal or external factors (e.g., behavioral changes induced by the game or external reasons like weather or personal/physical issues of the player), the estimation of difficulty needs to be contextualized on the player profile and dynamically changed during game execution (in our case study, week by week). This idea of dynamic difficulty adjustment (DDA) within a game-based, for example, on the playing ability manifested by each individual player—has been introduced in Hunicke (2005) and Liu et al. (2009) (among others), and is becoming increasingly widespread as an engagement technique (Beau and Bakkes 2016).

The way we conceptualize difficulty for our challenges follows loosely from Aponte et al. (2011), in which difficulty is defined as the conditional probability for a player of losing the current challenge, given the results on previous challenges faced by the same player in the game. Practically speaking, we define some discrete difficulty levels, and from the past performance of a player, we compute a player's current "comfort zone" within the distribution of past results by the player population; the difficulty level of a proposed challenge depends on how distant the goal of the challenge is from that comfort zone in the same data distribution. The farthest that distance, the strongest the "nudging effect" we wish to apply to the player with the challenge, and the higher its difficulty level.

For Trento Play&Go, we have defined *four* difficulty levels, labeled as *Easy, Medium, Hard* and *Very Hard*. This decision has been taken exploiting an *unsupervised clustering* algorithm to differentiate the amount of players' activities for each mobility indicator. We considered the output of the clustering as the number of difficulty levels. Thus, we implemented the *expectation maximization (EM)* algorithm considering only one feature (mobility indicator) to find out the maximum likelihood

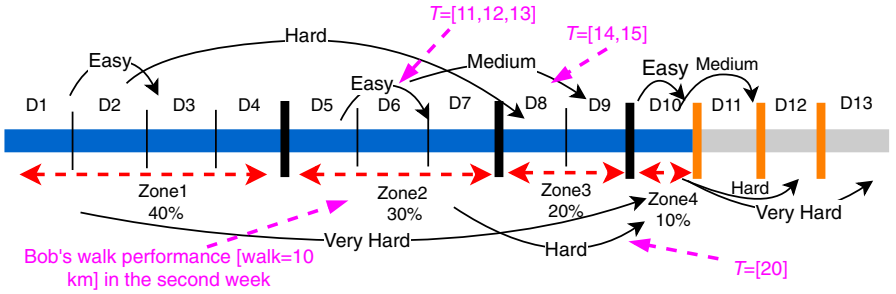


Fig. 3 Difficulty assignment method

among the players' activities in each specific transportation mode. By conducting the *EM* algorithm on the recorded data for each transportation mode, we have identified *Four* different clusters that express the different levels of players' activities.

Hence, given a challenge goal and a candidate player, our difficulty evaluator module assigns a difficulty label among [Easy, Medium, Hard and Very Hard] by taking into account the distribution of all players' past performances related to that goal. For example, if the goal of a challenge requires player *P* to walk X Km. during the next week of the game, the relevant distribution is that of weekly walked Km among all players.[3] We divide that distribution in 10 equal intervals, that is, *deciles*, and 4 "zones," as follows (see Fig. 3). Zone 1: the first zone contains the first four deciles, or 40% of the data. This zone has more players inside w.r.t the other deciles, and the players in this zone have done less activities compared to other deciles. Zone 2: the second zone contains the following three deciles, or 30% of the data. Zone 3: the third zone covers two deciles, or 20% of the data. Zone 4: the forth zone, which is the last zone, includes the last 10% of the data.

The difficulty evaluator module then evaluates the position in the distribution of the past performance of the candidate player versus the performance required by the challenge goal, according to the following rules:

- Easy: difficulty is set to *Easy*, if completing the challenge does not move the player's performance from the current zone to the next one;
- Medium: difficulty is set to *Medium*, if completing the challenge moves the player's performance from the current zone to the next one;
- Hard: difficulty is set to *Hard*, if completing the challenge moves the player's performance two zones higher;
- Very Hard: difficulty is set to *Very Hard*, if completing the challenge moves the player's performance three zones higher;
- If the challenge goal would move the player's performance beyond the tenth decile, we consider the range *(V)* of distribution values in the tenth decile to build a dynamic threshold to set the challenge difficulty. The next zone is set at the maximum of the distribution plus *1*V* (representing, in a sense the eleventh decile)

[3] This may configure, at the very beginning of the game, a sort of *cold start problem*; to bypass that, one can have an initial phase of the game without injection of challenges (2 weeks, in our game), in which sufficient game data are collected; alternatively, one can leverage data from previous instantiations of the same game, or any other suitable statistics, to establish an initial baseline distribution.

Table 1 Prize table for a challenge with a prize range of 100–250 points

Dif	Imp									
	10%	20%	30%	40%	50%	60%	70%	80%	90%	100%
Easy	**100**	106	111	119	125	130	135	140	145	**150**
Medium	133	139	144	150	156	161	165	170	175	183
Hard	166	172	177	180	186	192	197	203	210	217
Very Hard	197	205	211	219	225	230	235	240	245	**250**

IMP improvement, *Dif* difficulty
Bold values indicate the ones set by the game designer to calibrate the rewards

and so on. The previous rules apply then to this extended range. For example, if the current player performance is already in the highest zone, and the goal would move at about *2*V* beyond the maximum value in the observed distribution, the challenge is considered *Hard*.

For instance, looking at Bob's history in the second week of the game (Walk_Km = 10) and considering the distribution of all players activities in that mode, Bob's performance is located in Zone 2 (see Fig. 3). Hence, the module assigns difficulty $D = Easy$ for the challenge instances with $T \in \{11, 12, 13\}$, since these improvements do not move Bob's current position to the next zones. While the other challenges, $T \in \{14, 15\}$ and $T \in \{20\}$, move Bob current position of one and two next zones, respectively, $D = Medium$ and $D = Hard$ will be assigned for these challenge instances.

4.3.2 Computation of rewards

Challenge rewards are construed as "in-game" rewards: winning a challenge gets rewarded with some GDE, such as a badge for an achievement and a point bonus, which has the potential to improve the player's status in the game. In its current implementation, the reward calculator module works with point bonuses as the chosen form of challenge prizes.

The value of the assigned reward/prize R depends on the estimated difficulty level, and the amount of behavioral improvement required by a given challenge, according to a two-dimensional function f: $R = f(\text{dif}, \text{imp})$. dif and imp refer to the difficulty of the challenge and the improvement, respectively. We use for f a type of linear function called "plane flat," which has constant slopes in both the $x = \text{imp}$ direction and the $y = \text{dif}$ direction, defined as follows:

$$\begin{aligned} R &= m(\text{imp}) + n(\text{dif}) \\ m &= (\text{imp}_{\max} - \text{imp}_{\min})/(N-1) \\ n &= (\text{dif}_{\max} - \text{dif}_{\min})/(M-1) \end{aligned} \tag{1}$$

where m is the slope along the (imp) dimension, when dif is constant, and n indicates the slope along the (dif) dimension, when imp is constant; N and M are the number of

different improvement brackets (e.g., 10%, 20%, etc.) and the number of difficulty levels, respectively. Hence, $M-1$ and $N-1$ specify the maximum number of movements in improvement (9) and in level of difficulty (3) in the plane formula, respectively. The above formula can produce a table of prizes for all improvement / difficulty combinations for a given challenge (see Table 1 for an example); it simply requires the game designer to provide a range with minimum and maximum prize, plus one additional point value (e.g., the value for the top/right corner in the table), in order to set the m and n slopes as desired. By associating those values with each mobility indicator, the designer can modulate the relative importance assigned to behavioral improvement (a general characteristic, e.g., in our application scenario, promoting an improvement in the usage of one or more transportation means) versus difficulty (a personalized characteristic, e.g., promoting challenges that are more difficult to achieve).

As an example, we recall the challenge instances generated for Bob related to MI = Walk_Km. The module calculates the rewards R by taking into account the difficulty and the percentage improvement and assigns the following rewards: $R = \{100, 106, 111\}$ for 10%, 20%, 30% improvements (labeled as *Easy*), $R = \{150, 156\}$ for 40% and 50% improvements (*Medium* challenges) and $R = 217$ for 100% improvement (*Hard* challenge).

4.4 Filtering and sorting

Filtering and sorting is the final stage, which takes generated challenges that have been valuated for each individual player in terms of difficulty and reward, and recommends a subset of them. This step is critically important since we are in a multi-dimensional recommendation problem, in which there is a high probability of conflict between the objectives of the entity promoting the game (e.g., the city administration in our case) and the preferences of the player. The filtering and sorting module tries to find a *sweet spot* between the player's own interest and the interest of the entity promoting the gamification campaign.

We construe the player's interest by considering objectives for the player, which are typically tied to some GDE included in the game. For example, in a point-based game a game objective can be to climb at least one position in the leaderboard for a certain type of points; another objective can be earning enough points to reach a new level; in a game based on achievements, an objective could be instead earning a new badge, to come closer to complete a certain badge collection, or unlock some other new game element. In general, at any point during a game, a player may be interested to reach one or more such objectives. If we consider the distance between those objectives and the current state of the player, a good challenge for the player is one which offers a reward that enables that player to reach (or at least come as close as possible) one of her current game objectives. Among any such challenges, the ones that are less difficult to win are preferable from the player's point of view; however, some players do not always prefer less difficulty.

From the player's perspective, our recommender system tends thus to adopt an opportunistic stance and aims at providing immediate value and a measure of satisfaction, in return for the player's effort to complete the proposed challenges. Moreover,

the recommendations produced in this way are situational: a challenge that is preferable at a given juncture in the game, may become less valuable for the same player at some other time. The current implementation of the recommender system applies to all players the same filtering and sorting strategy. However, player's motivation and objectives are strongly related to her preferences, traits, abilities and personality (Hamari et al. 2014a; Bartle 2003, 1996) and an interesting extension of the presented work could exploit player modeling and play style analysis techniques (Valls-Vargas et al. 2015; Hsieh and Sun 2008; Weber et al. 2011; Khoshkangini et al. 2018) to tailor the filtering and sorting algorithm to the play style of each player.

From the Smart City's point of view, the preferable challenges are those most in line with the objectives currently being promoted, and among those, the ones that incentivize the most significant behavior improvement for the player to whom they are assigned. To identify the best candidate challenges, the RS uses the concept of *weighted improvement* (WI) , that is, the product between the weight property W of the challenge and the amount of improvement mandated by its goal. *Weight* W is a static value between [1–5] and is assigned to each mobility indicator on the basis of the theme of the week by the game designer. For instance, a challenge which requires a 20% improvement and has a weight $W = 5$ has a WI $= 100$, and is preferable to another challenge which requires a 30% improvement, but has $W = 3$ WI $= 90$).

To implement the above principles, the filtering and sorting module operates for each individual player according to the following algorithm:

1. separate the challenges in the repository in two subsets: those whose prize is sufficient for the player to reach some of the player's current game objective and the others;
2. sort the first subset of challenges by least difficulty;
3. within each difficulty level, calculate the weighted improvement and sort challenges by highest WI;
4. sort the second subset of challenges by least difficulty, then highest prize and then weighted improvement;
5. append the sorted list from the second subset at the end of the first sorted subset to obtain a single ranking.

At the end of the process described above, the RS proposes to each player the K top challenges in her personalized ranking above (where K is a parameter that can be adjusted).

A further component of our system is the *ML module*. This module is currently under development; it leverages machine learning over the personal player history and performances on past challenges to extract the probability of success *PV* for each candidate challenge instance. This information can be exploited to further personalize and tune the filtering and sorting of challenges to be administered to players.

5 Evaluation

We implemented the proposed framework in a gamified urban mobility experiment (called Trento Play&Go) that aimed at evaluating the effectiveness of gamification in

general—and of personalized challenges in particular—in changing the behavior of players toward more sustainable transport means and maintaining their participation active in the long term.

5.1 Study setup

As it is described in Sect. 3.2, Trento Play&Go was a gamification campaign and an open-field study that lasted 12 weeks (from September 10 to December 2, 2016). The game was open to all residents of Trento, as well as commuters from the surrounding areas of Trentino. During the twelve game weeks, 1061 citizens downloaded the App, 785 registered to the game, and 410 actively participated in it (*active players* from now on). Within the 12 weeks, we collected through the players' App more than 20,000 trip traces, which were validated, in terms of mode and path, through a semiautomated itinerary validation system. Those traces enabled us to collect detailed statistics about the itineraries of App users when playing the game, such as trips, trip legs and kilometers traveled in each transport mode.

During Trento Play&Go, we carried out an A/B test in the last 3 weeks of the game, with the aim of comparing our system that automatically generates and proposes personalized challenges to players (*RS challenges*, from now on) with a semiautomatic approach, used throughout the 12 weeks of the game, according to which challenges were decided and administrated by the game designers using their expert judgment,[4] and then injected in the game rule set (*non-RS challenges*). In that semiautomatic approach, game experts had to decide each week which challenge models to instantiate (e.g., *bike Km percentage increment* or *bus trips absolute increment*) specified values for the parameter sets (i.e., percentage of improvement, amount of rewards), and defined which assignment criteria to apply (using logical clauses that predicate on the game state and on profiling variables collected for each player). The assignment criteria play a key role in that approach, in terms of challenges variation and personalization, and produce a coarse segmentation of the players' population based on players' performances and habits. With the new approach, instead, they only had to specify for the PCG systems a set of challenge models and a set of transport modes of interest for that week. (In our A/B test, of course, they were the same sets selected by the game experts.)

During the 3 weeks of the A/B test, our new PCG- and RS-based challenge generation system assigned 220 personalized challenges to 82 unique players ("RS players" from now on). Among those RS players, 60 were active in those 3 weeks, with 164 assigned challenges. The challenge generator used the *percentageIncrement* and *absoluteIncrement* challenge types, and applied them to an array of transport modes, covered by the following mobility indicators: *{Train_Trips, Bus_Trips, ZeroImpact_Trips, Walk_Trips, Walk_Km, Bike_Km, BikeSharing_Km}*.

RS players were randomly selected from a subset of the players' population, from which we excluded players who were *not active* in the previous week, as well as players

[4] Non-RS challenges were managed by Trento Play&Go game administration team, who were involved in the design and day-to-day management also of two previous editions of the game, and had become very knowledgeable of the game mechanics and dynamics, as well as of the urban mobility gamification domain.

who had very high performance in the previous week, (typically top 10 to top 12) whom we called "*weekly champions*". Weekly champions were among a group of players who proved to be extremely motivated over all game weeks, and constantly played to win the weekly leaderboard, or even the final global leaderboard. We observed that this group was quite likely to complete almost any challenge they were given. Therefore, we excluded them from the A/B test to avoid that—in case they happened to be randomly selected disproportionately among the RS players for a given week—they could skew the A/B test comparison in favor of the recommendation system. This random draw of RS players was repeated every week, in order to keep as close to constant as possible the proportion of RS to non-RS challenges administered each week, even if the population of active players changed from week to week. In fact, in Trento Play&Go new players could come into the game at any time, as well as existing players could leave, or simply interrupt their activity for some time and then resume the game (for instance, for personal reasons, unfavorable weather or any other factors not under our control). Those intrinsic characteristics of the Play&Go game made unfeasible to keep the same set of RS players across all A/B test weeks.

5.2 Evaluation objectives

We briefly recall and discuss here the research questions, introduced in Sect. 1, on which we based the evaluation of the proposed approach.

RQ1—Player acceptance speaks to the issue of user experience, since recommending playable units of contents that may be diverse, but are all in all well accepted, by each individual player, is the prerequisite to be able to leverage them as a mechanism for enhanced engagement and retainment. As a proxy measure of acceptance rate, we consider challenge completion rate, i.e., the proportion of challenges that players completed successfully, comparative within the A/B test.

RQ2—Challenge impact speaks to the efficacy of procedurally generated challenges as persuasive mechanisms that can appreciably impact players' behavior in a gamification campaign. To measure challenge-induced behavioral change, we consider those challenges whose goal requires to improve by a certain percentage the player's performance for a given mobility mode with respect to the previous game week.

RQ3—Reward efficiency speaks to whether the rewards that our system computes for the challenges it automatically proposes to players are commensurate and well balanced. This is an important consideration for game designers, as well as any organization promoting a gamification campaign: On the one hand, the reward should be valuable enough to induce the player to make an effort to complete the challenge; on the other hand, it should not be excessive, because that leads to inadequate, suboptimal exchanges between behavioral improvement and incentives. To measure this adequacy, we compute the ratio between the amount of improvement generated by challenges and the amount of rewards "attributed" by the game to players for challenge completion.

In the following sections, we present and discuss the evaluation results for each research question, as obtained from the performed A/B test.

5.3 Evaluation results

5.3.1 RQ1: Player acceptance

To evaluate RQ1, we have compared the proportion of players' success in automatically generated and recommended challenges versus challenges administered through expert judgment. In this analysis, we consider the 82 unique RS players who received challenges from our system during weeks 10, 11 and 12 of the Trento Play&Go game (a total of 220 RS challenges), and compare their challenge completion rate to that of non-RS players, considering the following groups.

- Group 1: includes both RS and non-RS players. Non-RS players include all players that were active at any point in the game and who were given, in game weeks 10–12, mobility challenges that are analogous to the RS challenges, that is, challenges of type *percentageIncrement* or *absoluteIncrement*, targeting the same set of transport modes covered in RS challenges. In total, RS and non-RS players in Group 1 were given 220 and 1296 challenges, respectively.
- Group 2: a subset of Group 1, including only non-RS and RS players who were active in game weeks 10–12, which were covered during the A/B test. Accordingly, non-RS players received 333 mobility challenges w.r.t the RS players who were given 164 challenges.
- Group 3: a subset of Group 2, from which we excluded top performers (typically top 10–12 players) of the week before that of challenge assignment (that is, weeks 9–11). This is the group that serves effectively as the **control group in our A/B test**, since—by eliminating these "weekly champions"—Group 3 reflects the same population from which we drew the RS players each week, as explained in Sect. 5.1. Non-RS players in Group 3 were assigned 280 mobility challenges.
- Group 4: as a final check, we examined the same 82 RS players "against themselves," that is, with respect to those challenges, which they received in weeks 10–12, but which were assigned to them by expert judgment, as opposed through our system. This challenge set included 151 mobility challenges analogous to RS challenges.

In Table 2, we show the results of a statistical test of equality for the challenge success rate of RS players on RS challenges versus the four groups defined above. We have exploited the equivalency testing to indicate that the differences that do exist between the above groups are small enough for practical purposes (Blackwelder 1982).

To this end, we used the two one-sided test procedure (TOST)[5] to check whether the challenge completion proportion of the various groups described above are close enough to the same proportion for the RS group to be considered equivalent (Schuirmann 1987; Robinson and Froese 2004). We have set the confidence level of the interval (CI) value to 0.95 and $\alpha = 0.05$, which are widely used in the state of the art.

Notice that, vis-a-vis Group 1, we tested whether RS players had similar success rate to the other players in general. That is clearly not the case, as evident from the

[5] We used TOST function which is available in "TOSTER" package in r: https://cran.r-project.org/web/packages/equivalence/equivalence.pdf.

Table 2 Proportion tests for challenge success rates

	Group 1				Group 2				Group 3				Group 4			
T-Ch	RS		Non-RS		RS		Non-RS		RS		Non-RS		RS		Non-RS	
Status	C	T	C	T	C	T	C	T	C	T	C	T	C	T	C	T
# Challenges	58	220	152	1296	58	164	153	333	58	164	113	280	58	220	49	151
AC ratio	26%		12%		35%		46%		35%		40%		26%		32%	
EqTest-tost()-	$p = 0.4435465$				$p = 0.1706046$				$p = 0.01809221$				$p = 0.04531761$			

T-Ch type of challenges such as *RS* and *non-RS* challenges, *AC ratio* challenge completion rate, *C* total numbers of the *completed* challenges, *T* total numbers of the *recommended* challenges, *EqTest* equivalency test

very large difference in percentage. Since a considerable number of players in Group 1 may not have been active through weeks 10–12, it is not surprising that RS challenges enjoy a better completion rate. In fact, in this case we use the test of equality mainly as a sanity check, and in fact, it fails (p value $= 0, 44$).

Instead, vis-a-vis the other groups, our hypothesis is that the RS challenge assignments should not differ from the challenges assigned via expert judgment; therefore, we wish to ascertain whether RS players had a statistically similar success rate than other players.

Notice also that in the tests versus Group 2 and Group 3, we had to eliminate 56 RS challenges, which were proposed to RS players who then chose not to be active (i.e., did not participate at all) to the game in the corresponding week. We did that to keep the testing conditions and the data sets congruent, since Group 2 and Group 3 contain only non-RS players who were *active* in the game weeks in which we tested our recommendation system.

In this setting, the test results show non-RS challenges perform better only in Group 2, since the test fails to reject the null hypothesis (p value $= 0.170606$). However, it is worth remarking that the player's population in Group 2 is not fully equivalent to the population from which we drew our RS players, since Group 2—as opposed to the RS group—*does include weekly champions*. Therefore, we had expected that Group 2 might perform overall somewhat better; in fact, weekly champions were responsible for 40 out of 153 challenges successes recorded for Group 2.

The equivalence tests of RS challenges versus Group 3—as well as Group 4—show that when samples are drawn from largely equivalent player populations, we can reject the null hypothesis with p value $= 0.01809221$ and p value $= 0.04531761$, respectively, at 95% confidence, which highlight that RS and non-RS challenges in these two groups are statistically equivalent.

Therefore, we can answer RQ1 by stating that: *There is no significant difference in acceptance between RS challenges and challenges assigned through expert judgment*.

5.3.2 RQ2: Challenge impact

To evaluate RQ2, we define improvement induced during a challenge relatively to the player's performance of the previous game period (in our case, a week). Given a performance indicator congruent with the challenge goal (in our mobility game, either number of trips or amount of Km traveled in a certain transport mode in a week), our normalized definition of improvement is as follows:

$$\text{Imp} = (\text{counter} - \text{base})/\text{base} \tag{1}$$

where counter is the numeric value of the performance indicator that starts from 0 to $+\infty$, and base is the value of the same indicator sampled at the moment in which the challenge has been administered to the player (i.e., at the end of the previous game week). According to the above formula, improvement (Imp) is a real number in the range $[-1, +\infty]$ (Fig. 4a, b y axes), where the range between -1 and 0 indicates that the player has performed *worse* during the challenge period, compared to the previous game week. More specifically, -1 means *nothing* was done by the player related to

the challenge goal, while 0 means that the player had *no improvement*, that is, she repeated the exact same performance of the previous game week.

We have considered weekly challenges of type *percentageIncrement*, since their goal is exactly to improve by a given percentage one's performance on a mobility indicator with respect to the previous week. Out of the 164 challenges automatically administered by our system during the A/B test to those players who were active in game weeks 10–12, 129 were of type *percentageIncrement*; the remaining 35 were of type *absoluteIncrement* and, in particular, those were all what we call "try mode" challenges; that is, they asked to players to do at least a single trip in a mode they had not used at all the previous week. Although 17 out of those 35 challenges were completed, their definition is clearly not conducive to measure relative improvement in a way that is congruent with our definition given above, and also with the *percentageImprovement* challenges. Therefore, we have excluded them from our analysis of RQ2.

The *percentageIncrement* RS challenges were subdivided in the following way:

- improvement on number of trips per week: 71 challenges, for modes *Train_Trips* (12), *Bus_Trips* (36) and *ZeroImpact_Trips* (23).
- improvement on amount of Km per week: 58 challenges, for modes *Walk_Km* (44) and *Bike_Km* (14).
- the challenge generator also produced candidate *percentageIncrement* challenges for modes *Walk_Trips* and *BikeSharing_Km*, but they were never picked by the RS.

We have compared the improvement induced by those RS challenges against equivalent non-RS challenges, i.e., *percentageImprovement* challenges covering the same set of transport modes, which were administered during weeks 10–12. There were 145 such challenges predicating on number of trips and 104 on amount of Km, for a total of 249 non-RS challenges.

One way to assess the difference in improvement between the RS and non-RS regimes is to sort the improvement metric for each challenge in ascending order, and plot this sequence of values at equally spaced intervals. We show those plots in Fig. 4a, b, separately for trip-related and Km-related challenges, respectively. In those figures, the y axis represents the improvement metric and the x axis the percentage of challenges having that improvement value or less; the black curve plots the results of the non-RS challenges, while the blue curve those of RS challenges. The amount of improvement collectively induced by those challenges can be visualized as the area in the chart comprised between those monotonic non-decreasing curves and the $y = 0$ axis (i.e., the *no improvement axis*). In fact, actual (positive) improvement occurs only in the area to the right of the curve intercept with the $y = 0$ axis (highlighted in green), while the area to the left of the intercept represents those cases in which we observed worse performance than the challenge baseline that can be seen as a *negative improvement* area (highlighted in red). In both figures, it is easy to appreciate that, while the negative improvement areas of the RS and non-RS curves are almost completely overlapping, the positive improvement areas under the blue curve are larger and contain almost everywhere the corresponding areas under the black curve.

To go beyond this intuitive assessment, we can also quantify the amount of improvement, by using a method for numerical integration of those curves, and estimate the

Table 3 RS Challenge improvement

	RS challenges					
	n (won)	Rewards attributed	AUiC	AUiC+	Players improved	Reward/AUiC + per capita
Trips						
Tot	**71 (27)**	**5650**	**0.385**	**0.643**	**30**	**293**
Train	12 (7)	1470	0.672	0.897	7	234
Bus	36 (13)	2860	0.204	0.536	13	410
Zero impact	23 (7)	1320	0.365	0.527	10	250
Walk	0	0	n/a	n/a	0	n/a
Km						
Tot	**58 (23)**	**5410**	**0.305**	**0.612**	**26**	**340**
Walk	44 (15)	3520	0.222	0.545	17	380
Bike	14 (8)	1890	0.516	0.729	9	288
Bike sharing	0	0	n/a	n/a	0	n/a

Bold indicate the total values

size of those areas.[6] We call this metric the area under the improvement curve, or AUiC.[7] We denote instead AUiC+ the estimate of the area of positive improvement only. Values for AUiC and AUiC+ for RS and non-RS challenges are reported in Tables 3 and 4, respectively (separately for trip-related and Km-related challenges, as well as for each single transport mode).

In almost all cases, the values of both AUiC and AUiC+ metrics are quite larger for RS challenges, with the single exception of *zero-impact* challenges, for which the AUiC value is somewhat larger in the non-RS case and the AUiC+ values are very similar.

It is noticeable how in several cases the total AUiC in the non-RS case is negative, which is due to many non-RS players not doing enough to reach their baseline of the previous week, or choosing not to take up that specific challenge at all, thus offsetting the actual improvement (AUiC+) from other players in the same group.

We can also look at the significance of the improvement differences signified by the AUiC and AUiC+ metric. Since the distribution of the data in both cases does not follow a normal distribution, we used the nonparametric Wilcoxon test for comparison of two data distributions (Gehan 1965); we used it to investigate whether the improvements from RS challenges are statistically larger than those from non-RS challenges, and we applied it to both trip-related and Km-related challenges (the same data sets visualized in Fig. 4a, b, respectively).

The results of the Wilcoxon test for Km-related improvement are:

```
W = 3280.5 p value = 0.1317
```

[6] For those estimates, we took advantage of a function made available to the R statistical suite by Jurasinski and Günther (2014).

[7] In order to avoid confusion with AUC, which usually indicates the area under the curve in a receiver operating characteristic (ROC) plot.

Table 4 Non-RS challenge improvements

	Non-RS challenges					
	n (won)	Rewards attributed	AUiC	AUiC+	Players improved	Reward/AUiC + per capita
Trips						
Tot	**145 (69)**	**15,600**	**0.101**	**0.366**	**76**	**561**
Train	31 (13)	3400	− 0.237	0.147	13	1779
Bus	52 (19)	5250	− 0.135	0.264	21	947
Zero impact	57 (34)	6200	0.466	0.532	39	299
Walk	5 (3)	750			3	
Km						
Tot	**104 (42)**	**11,300**	**− 0.017**	**0.301**	**48**	**782**
Walk	80 (30)	8150	− 0.006	0.314	36	721
Bike	17 (8)	2100	− 0.089	0.204	8	1287
Bike sharing	7 (4)	1050			4	

Bold indicate the total values

The result of the Wilcoxon test for trip-related improvement are:

```
W = 5344 p value = 0.2941
```

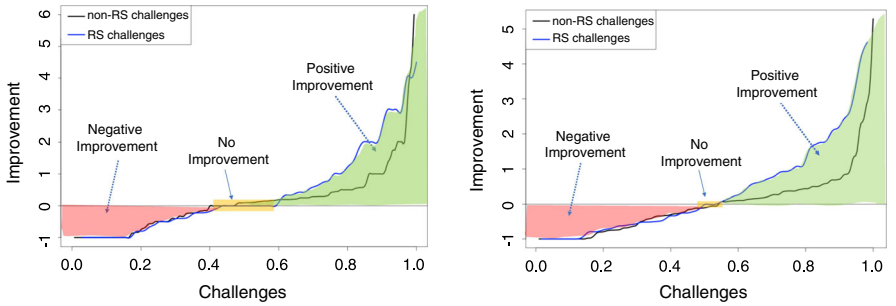
In both cases, the p value is too large to indicate that the RS improvement (and therefore the corresponding AUiC score) is significantly larger than the non-RS improvement. However, the observant reader looking at Fig. 4a, b may notice that the plots are very similar in the negative improvement areas (for example, 15–20% of the data consist of -1 values), and start to clearly divaricate after they reach the intercept with the $y = 0$ axis. In fact, if we repeat the Wilcoxon test only for the data that represent positive improvement, we obtain, for Km-related challenges:

```
W = 571 p value = 0.0003194
```

and for trip-related challenges:

```
W = 979.5 p value = 6.549e-06
```

That means that the difference in the AUiC+ scores is highly statistically significant. In turn, that suggests that automatically generated challenges may have induced superior outcomes for those players who embraced their assigned challenge goal, elevated their game, and put in place a positive effort to improve to some extent on their baseline mobility behavior. One possible interpretation is that the selection of transport modes and goals proposed by our recommendation system could have been better suited to the personal inclination of the individual players than the selection proposed through expert judgment to the non-RS players, making more congenial for many RS players to approach, reach or even go beyond the goals set in their personalized challenges.



(a) Improvement on trips number (all modes) **(b)** Improvement on Km travelled (all modes)

Fig. 4 Improvement induced by RS and non-RS challenges

On the basis of the above data, we can thus answer RQ2 by stating that *challenges assigned by our system may be conducive to higher level of improvement than analogous challenges assigned through expert judgment*.

5.3.3 RQ3: Reward efficiency

To evaluate RQ3, we put in relation the amount of improvement for the various challenge types to the rewards "attributed" by the game for challenge completion. Operationally, we characterize improvement via the data about AUiC+ in Table 3, since otherwise the total AUiC of non-RS challenges would often be negative. That means that we considered only the performance of those players who were able to achieve *some* amount of actual improvement in their challenges. Moreover, since the cardinality of the player sets yielding some improvement is different in RS versus non-RS cases, we further normalize the data considering only the number of players who contributed to that improvement (i.e., the *players improved* column in Table 3); we thus compute the *per capita* reward attributed by the system per unit of AUiC+ ($\text{Reward}_{\text{pc}}$). The corresponding formula is as follows:

$$\text{Reward}_{\text{pc}} = (\text{Reward}_{\text{tot}}/\text{players}_{\text{imp}})/\text{AUiC+} \tag{2}$$

By looking at those data, we can clearly see that the challenge proposals by our system are across the board *more economical* in terms of rewards attributed as incentives per unit of improvement. The difference is always in favor of RS challenges, and it is quite evident; in fact, the rewards attributed to non-RS players per unit of improvement are—in almost in all cases—about double of those attributed to RS players, or more. For example, the per capita reward attributed per unit of improvement in non-RS trip-based challenges is 1.91 times higher (561/293) than in the RS case, while in non-RS Km-based challenges is 2.3 times higher (782/340) than in the RS case.

It is worth mentioning that in our experiment and evaluation, we have many cases in which players could improve their performance even without completing the given challenges. For example, player "Bob" did 3 km in week "x," so a new challenge asked him to improve his walking activity up to 5 km in week "$x + 1$." Bob increases his

walking performance to around 4 km in week "$x+1$." Although he could not complete the given challenge, he still worked on improving his walking performance. Since the goal of every gamified system is to change players' behavior and the challenges are the main mechanism we use to encourage this improvement, further data normalization (using Eq. 2) is essential for calculating the unit of improvement, independently from the fact that they complete the challenge or not.

Therefore, we can answer RQ3 by stating that *challenges assigned by our system may yield better improvement for the sane amount of per capita reward, which effectively translates to the same level of improvement for less reward.*

6 Discussion and future work

We have presented an approach and system for the procedural generation of playable content units in gamification, which take the form of individual challenges. Our system generates challenges that are personalized with respect to the history and habits of each individual player, and contextualized with respect to her game state, as well as the objectives, principles and policies that underlie the gamification application.

Empirical results from games promoting sustainable urban mobility (Kazhamiakin et al. 2016; Khoshkangini et al. 2017; Ferron et al. 2019) have shown that such personalized units of playable content have a significant positive effect on players' engagement as well as retention over time. The fully automatic PCG approach presented in this work allows to create to mobility challenges that are tailored —with minimal additional design work—to players' habits and preferences. In fact, in this personalized PCG approach, such habits and preferences are inferred from their ability to play the game (e.g., skill), choices, attempts and errors that they made throughout the game over time.

We have evaluated our challenge generation and recommendation system by means of an A/B test conducted within a long-running sustainable mobility game involving more than 400 active players. The main objective of the test was to compare the fully automatic approach for challenge generation and assignment presented in this paper (RS) with an earlier semiautomatic approach based on expert judgment (non-RS). Regarding the research questions introduced in Sect. 1, we evaluated three key aspects, which express the effectiveness and efficiency of our framework: (i) comparison of the acceptance rate of challenges generated through the two approaches; (ii) comparison of the improvement in mobility habits of the players obtained through RS versus non-RS challenges; and (iii) comparison of the economicity, in terms of attribution of in-game rewards, of the RS versus the non-RS approach. For all three aspects, the evaluation results show that the RS-based automated approach, thanks to challenges that are tailored to each player's profile, is not only comparable to the one based on expert judgment, but may even be more effective and efficient.

Considering challenge acceptance (Ob1), measured in terms of completion rate, our experiments show that there is no significant difference between RS and non-RS challenges when samples are drawn form equivalent player populations. However, the relatively low number of challenges (and players) in the control and intervention

groups (164 RS challenges, 280 non-RS challenges) might be a potential threat to validity, and these encouraging results require confirmation in larger trials.

The results in terms of behavioral improvement—Ob2—are very promising: Conducted experiments show a statistically significant positive difference in the improvement induced by RS challenges with respect to non-RS challenges. However, both RS and non-RS challenge recommendation approaches fail to affect the behavior of a portion of the players population, i.e., players that—notwithstanding being proposed some challenge—did not reach even the baseline performance of the previous week or, in some extreme cases, did not take up that challenge at all. This aspect, although possibly influenced by other confounds, questions the effectiveness of the supported challenge mechanism itself (independently from the RS or non-RS system used to generate and recommend challenges) and needs to be further investigated.

Concerning the efficiency of the approach with respect to manual challenge generation, in our experiments we observed that RS-based challenges are far more economic (i.e., they induce more improvement in mobility behavior for less reward). However, more detailed research is needed to fully capture the relation between the reward associated with a challenge and the induced impact on player's behavioral outcomes.

Although being evaluated in the Viaggia Play&Go game, promoting sustainable mobility, the PCG approach proposed in this manuscript is general purpose and can be applied in any gamified system aiming at injecting personalized challenges in the game. For instance, in the mobility domain, it has already been used to implement other challenge-based mechanics and to cover a wider range of reward types: In Kids Go Green Marconi et al. (2018), an educational game promoting active and sustainable home–school mobility in primary schools, the proposed solution has been exploited to generate group challenges (i.e., class level or school level) rewarding home–school sustainable mobility behaviors with in-game virtual resources.

The same solution could be exploited in any other game promoting sustainable behaviors (e.g., waste reduction and recycling, energy saving) or healthy habits (e.g., healthy eating and nutrition games, exercising and fitness games). In the educational domain, such system could be used to recommend various assignments in the form of personalized challenges to students with different levels of ability or knowledge.

6.1 Limitations and directions for future research

The findings of this study have brought to light some limitations, which suggest some new directions for future research.

The first limitation pertains to the issue of the evaluation period. In this study, the limited duration of the A/B test brings about a threat to validity that should be addressed. Although we could observe a sustained engagement of the players receiving RS challenges over the 3-week experiment, and could appreciated the effectiveness of the proposed system to derive players performance w.r.t the non-RS players, 3 weeks is a period too short to positively draw conclusions on long-term engagement. Due to the fact that the proposed approach was under the construction during the game, it was not feasible for us to evaluate it from the beginning of the project for 12 weeks. Thus, further A/B experiments are needed to test the proposed solution in a longer period.

The second limitation is associated with personalization. Our approach considers some essential factors (i.e., game status, players' skills, leaderboard, objectives, etc.) to generate the personalized unit of playable content; however, it does not include players' characteristics and physiological signals, which are widely used in digital games for advanced personalization and to increase players' engagement (Yannakakis and Togelius 2015; Khoshkangini et al. 2018; Loria and Marconi 2018). Physiological signals could be expressed in different fashions, e.g., how fast or slow are players at handling the given challenges; or how precise are they at selecting the challenges to complete and improve their behaviors against other players in the game. This limitation strongly motivates us to extend our framework by integrating an analysis module to classify players' types from their behavior in the gamified system, and learn to adapt the generated challenges based upon that classification (Machado et al. 2011; Monterrat et al. 2015).

The third limitation of the proposed challenge generation approach, which might affect the impact in terms of engagement and behavioral change on broader demographics of players, concerns the fact that it considers the elevation of player's status as the key player objective. This objective guides the sorting and filtering algorithm, which promotes challenges whose reward maximizes the chances to elevate the player's status in the game (e.g., reaching a certain level in the game, rise in rankings, win a badge). Moreover, rewards themselves are currently limited to point bonuses. An interesting extension of the proposed work could also exploit player modeling and play style analysis techniques, to guide the generation and recommendation of challenges, better tailoring administered challenges and corresponding rewards to the player's motivation and objectives.

In addition, follow-up work could consider various challenge-based game mechanics: single-player challenges (as the one used in this game), player-to-player challenges, as well as team-level challenges. This would allow to compare different motivational affordances, from purely competitive ones to more collaborative ones, and to further modulate their effectiveness on different player types. In addition, we plan to extend our approach with a machine learning algorithm, which will augment our recommendation component, in order to optimize and tune the selection of challenges to the individual player based on her track record with challenges proposed in the past.

Finally, we plan to apply our solution to other large-scale gamification experiments in the mobility domain and beyond.

References

Andersen, E., Gulwani, S., Popovic, Z.: A trace-based framework for analyzing and synthesizing educational progressions. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, CHI '13, pp. 773–782 (2013). ISBN 978-1-4503-1899-0

Aponte, M.-V., Levieux, G., Natkin, S.: Measuring the level of difficulty in single player video games. Entertain. Comput. **2**(4), 205–213 (2011)

Bakkes, S., Whiteson, S., Li, G., Vişniuc, G.V., Charitos, E., Heijne, N., Swellengrebel, A.: Challenge balancing for personalised game spaces. In: 2014 IEEE Games Media Entertainment, pp. 1–8. IEEE (2014)

Bartle, R.: Hearts, clubs, diamonds, spades: players who suit muds. J. MUD Res. **1**(1), 19 (1996)
Bartle, R.: Designing Virtual Worlds. New Riders Games, Indianapolis (2003)
Beau, P., Bakkes, S.: Automated game balancing of asymmetric video games. In: 2016 IEEE Conference on Computational Intelligence and Games (CIG), pp 1–8. IEEE (2016)
Blackwelder, W.C.: Proving the null hypothesis in clinical trials. Control. Clin. Trials **3**(4), 345–353 (1982)
Blom, P.M., Bakkes, S., Tan, C.T., Whiteson, S., Roijers, D., Valenti, R., Gevers, T.: Towards personalised gaming via facial expression recognition. In: Tenth Artificial Intelligence and Interactive Digital Entertainment Conference (2014)
Brewer, R.S., Verdezoto, N., Holst, T., Rasmussen, M.K.: Tough shift: exploring the complexities of shifting residential electricity use through a casual mobile game. In: Proceedings of the 2015 Annual Symposium on Computer–Human Interaction in Play. ACM (2015)
Broll, G., Cao, H., Ebben, P., Holleis, P., Jacobs, K., Koolwaaij, J., Luther, M., Souville, B.: Tripzoom: An app to improve your mobility behavior. In: Proceedings of the 11th International Conference on Mobile and Ubiquitous Multimedia, pp. 57:1–57:4. ACM (2012). ISBN 978-1-4503-1815-0
Butler, E., Andersen, E., Smith, A.M., Gulwani, S., Popović, Z.: Automatic game progression design through analysis of solution features. In: Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems, CHI '15, pp. 2407–2416. ACM (2015)
Charles, D., Kerr, A., McNeill, M., McAlister, M., Black, M., Kcklich, J., Moore, A., Stringer, K.: Player-centred game design: player modelling and adaptive digital games. In: Proceedings of the Digital Games Research Conference, vol. 285 (2005)
Chen, J.: Flow in games (and everything else). Commun. ACM **50**(4), 31–34 (2007). https://doi.org/10.1145/1232743.1232769. ISSN 0001-0782
Cheng, R., Vassileva, J.: Design and evaluation of an adaptive incentive mechanism for sustained educational online communities. User Model. User Adapt. Interact. **16**(3–4), 321–348 (2006)
Chou, Y.-K.: Actionable Gamification: Beyond Points, Badges, and Leaderboards. Octalysis Media, Milpitas (2015)
Coenen, T., Merchant, P., Laureyssens, T., Claeys, L., Criel, J.: Zwerm: stimulating urban neighborhood self-organization through gamification. In: Using ICT, Social Media and Mobile Technologies to Foster Self-Organisation in Urban and Neighbourhood Governance (2013)
Cooper, S., Deterding, C.S., Tsapakos, T.: Player rating systems for balancing human computation games. In: Proceedings of 1st International Joint Conference of DiGRA and FDG (2016)
Cowley, B., Charles, D., Black, M., Hickey, R.: Toward an understanding of flow in video games. Comput. Entertain. **6**, 1–27 (2008)
Cowley, B., Moutinho, J.L., Bateman, C., Oliveira, A.: Learning principles and interaction design for 'green my place': a massively multiplayer serious game. Entertain. Comput. **2**, 103–113 (2011)
Das, S., Zook, A., Riedl, M.O.: Examining game world topology personalization. In: Proceedings of the 33rd ACM Conference on Human Factors in Computing Systems, CHI '15, pp. 3731–3734 (2015)
Di Salvo, C., Sengers, P., Brynjarsdóttir, H.: Mapping the landscape of sustainable HCI. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, pp. 1975–1984. ACM (2010)
Du, J., Feng, Y., Zhou, C.: Gamification for behavior change of occupants in campus buildings to affect improved energy efficiency (2014)
Elo, A.E.: The Rating of Chessplayers, Past and Present. Arco Pub., Nagoya (1978)
Farzan, R., DiMicco, J.M., Millen, D.R., Dugan, C., Geyer, W., Brownholtz, E.A.: Results from deploying a participation incentive mechanism within the enterprise. pp. 563–572. ACM (2008)
Ferron, M., Loria, E., Marconi, A., Massa, P.: Play&go, an urban game promoting behaviour change for sustainable mobility. Interact. Des. Architect. J. **40**, 24–45 (2019)
Fogg, B.J.: Persuasive Technology: Using Computers to Change What We Think and Do. Morgan Kaufmann Publishers Inc., Burlington (2002)
Froehlich, J., Dillahunt, T., Klasnja, P.V., Mankoff, J., Consolvo, S., Harrison, B.L., Landay, J.A.: Ubigreen: investigating a mobile tool for tracking and supporting green transportation habits. In: Proceedings of the 27th International Conference on Human Factors in Computing Systems, CHI 2009, April 4–9, 2009, pp. 1043–1052 (2009)
Gabrielli, S., Maimone, R., Forbes, P., Masthoff, J., Wells, S., Primerano, L., Haverinen, L., Bo, G., Pompa, M.: Designing motivational features for sustainable urban mobility. In: CHI'13 Extended Abstracts on Human Factors in Computing Systems, pp. 1461–1466. ACM (2013)
Gehan, E.A.: A generalized wilcoxon test for comparing arbitrarily singly-censored samples. Biometrika **52**(1–2), 203–224 (1965)

Gordillo, A., Gallego, D., Barra, E., Quemada, J.: The city as a learning gamified platform. In: Frontiers in Education Conference, pp. 372–378. IEEE (2013)
Greengard, S.: Tracking garbage. Commun. ACM **53**(3), 19–20 (2010). ISSN 0001-0782
Hamari, J., Koivisto, J., Pakkanen, T.: Do persuasive technologies persuade?—a review of empirical studies. In: International Conference on Persuasive Technology, pp. 118–136. Springer (2014a)
Hamari, J., Koivisto, J., Sarsa, H.: Does gamification work?—a literature review of empirical studies on gamification. In: 47th Hawaii International Conference on System Sciences. IEEE (2014b)
Harrison, B., Ware, S.G., Fendt, M.W., Roberts, D.L.: A survey and analysis of techniques for player behavior prediction in massively multiplayer online role-playing games. IEEE Trans. Emerg. Top. Comput. **3**(2), 260–274 (2015)
Hendrikx, M., Meijer, S., Van Der Velden, J., Iosup, A.: Procedural content generation for games: a survey. ACM Trans. Multimedia Comput. Commun. Appl. **9**(1), 1–22 (2013). ISSN 1551-6857
Hojer, M., Wangel, J.: Smart sustainable cities: definition and challenges. Volume 310 of Advances in Intelligent Systems and Computing, pp. 333–349. Springer International Publishing (2015)
Hsieh, J.-L., Sun, C.-T.: Building a Player Strategy Model by Analyzing Replays of Real-time Strategy Games. pp. 3106–3111. IEEE (2008)
Huang, E.M.: Building outwards from sustainable HCI. ACM Interact. **18**(3), 14–17 (2011)
Hunicke, R.: The case for dynamic difficulty adjustment in games. In: Proceedings of the 2005 ACM SIGCHI International Conference on Advances in Computer Entertainment Technology, pp. 429–433. ACM (2005)
Jurasinski, G., Günther, A.: Flux: Flux Rate Calculation from Dynamic Closed Chamber Measurements. R package version 0.1-4, Rostock (2014)
Karpinskyj, S., Zambetta, F., Cavedon, L.: Video game personalisation techniques: a comprehensive survey. Entertain. Comput. **5**(4), 211–218 (2014)
Kazhamiakin, R., Marconi, A., Perillo, M., Pistore, M., Valetto, G., Piras, L., Avesani, F., Perri, N.: Using gamification to incentivize sustainable urban mobility. In: 2015 IEEE First International Smart Cities Conference (ISC2), pp. 1–6 (2015)
Kazhamiakin, R., Marconi, A., Martinelli, A., Pistore, M., Valetto, G.: A gamification framework for the long-term engagement of smart citizens. In: 2016 IEEE International Smart Cities Conference (ISC2), pp. 1–7. IEEE (2016)
Khajah, M.M., Roads, Brett D., Lindsey, R.V., Liu, Y.-E., Mozer, M.C.: Designing engaging games using bayesian optimization. In: Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems, pp. 5571–5582 (2016). ISBN 978-1-4503-3362-7
Khoshkangini, R., Valetto, G., Marconi, A.: Generating personalized challenges to enhance the persuasive power of gamification. In: International Workshop on Personalized Persuasive Technologies (2017)
Khoshkangini, R, Ontañón, S., Marconi, A., Zhu, J.: Dynamically extracting play style in educational games. In: EUROSIS Proceedings, GameOn (2018)
Lampe, C.A.: Ratings use in an online discussion system: The slashdot case (2006)
Lee, Joey J., Matamoros, E., Kern, R., Marks, J., de Luna, C., Jordan-Cooley, W.: Greenify: fostering sustainable communities via gamification. In: CHI'13 Extended Abstracts on Human Factors in Computing Systems, pp. 1497–1502. ACM (2013)
Lessel, P., Altmeyer, M., Krüger, A.: Analysis of recycling capabilities of individuals and crowds to encourage and educate people to separate their garbage playfully. In: Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems, CHI 2015 (2015)
Liu, C., Agrawal, P., Sarkar, N., Chen, S.: Dynamic difficulty adjustment in computer games through real-time anxiety-based affective feedback. Int. J. Hum. Comput. Interact. **25**(6), 506–529 (2009)
Lopes, R., Bidarra, R.: Adaptivity challenges in games and simulations: a survey. IEEE Trans. Comput.Iintell. AI Games **3**(2), 85–99 (2011)
Lora, D., Sánchez-Ruiz, A.A., González-Calero, P.A., Gómez-Martín, M.A.: Dynamic difficulty adjustment in tetris. In: FLAIRS Conference, pp. 335–339 (2016)
Loria, E., Marconi, A.: Player Types and player behaviors: analyzing correlations in an on-the-field gamified system. In: Proceedings of the 2018 Annual Symposium on Computer–Human Interaction in Play Companion Extended Abstracts—CHI PLAY '18 Extended Abstracts, pp. 531–538. ACM Press, Melbourne, VIC, Australia (2018)
Machado, M.C., Fantini, E.P.C., Chaimowicz, L.: Player modeling: towards a common taxonomy. In: 2011 16th International Conference on Computer Games (CGAMES), pp. 50–57. IEEE (2011)

Marconi, A., Schiavo, G., Zancanaro, M., Valetto, G., Pistore, M.: Exploring the world through small green steps: improving sustainable school transportation with a game-based learning interface. In: Proceedings of the 2018 International Conference on Advanced Visual Interfaces, AVI '18, pp. 24:1–24:9. ACM, New York, NY, USA (2018)

Monterrat, B., Desmarais, M., Lavoué, E., George, S.: A player model for adaptive gamification in learning environments. In: International Conference on Artificial Intelligence in Education, pp. 297–306. Springer (2015)

Orland, B., Ram, N., Lang, D., Houser, K., Kling, N., Coccia, M.: Saving energy in an office environment: a serious game intervention. Energy Build. **74**, 43–52 (2014). ISSN 0378-7788

Pizzi, D., Lugrin, J.-L., Whittaker, A., Cavazza, M.: Automatic generation of game level solutions as storyboards. IEEE Trans. Comput. Intell. AI Games **2**, 149–161 (2010)

Raffe, W.L., Zambetta, F., Li, X.: Neuroevolution of content layout in the PCG: angry bots video game. In: 2013 IEEE Congress on Evolutionary Computation, pp. 673–680 (2013)

Robinson, A.P., Froese, R.E.: Model validation using equivalence tests. Ecol. Model. **176**(3–4), 349–358 (2004)

Schuirmann, D.J.: A comparison of the two one-sided tests procedure and the power approach for assessing the equivalence of average bioavailability. J. Pharmacokinet. Biopharm. **15**(6), 657–680 (1987)

Shiraishi, M., Washio, Y., Takayama, C., Lehdonvirta, V., Kimura, H., Nakajima, T.: Using individual, social and economic persuasion techniques to reduce Co_2 emissions in a family setting. In: Proceedings of the 4th International Conference on Persuasive Technology, pp. 1–13. ACM (2009)

Sifa, R., Bauckhage, C., Drachen, A.: Archetypal game recommender systems. In: LWA, pp. 45–56. Citeseer (2014)

Silva, F., Analide, C., Rosa, L., Felgueiras, G., Pimenta, C.: Gamification, social networks and sustainable environments. Int. J. Interact. Multimedia Artif. Intell. **2**, 52–59 (2013)

Simon, J., Jahn, M., Al-Akkad, A.: Saving energy at work: the design of a pervasive game for office spaces. In: Proceedings of the 11th International Conference on Mobile and Ubiquitous Multimedia. ACM (2016)

Skocir, P., Marusic, L., Marusic, M., Petric, A.: The MARS–A Multi-Agent Recommendation System for Games on Mobile Phones, pp. 104–113. Springer, Berlin Heidelberg (2012)

Soares de Lima, E., Feijó, B., Furtado, A.L.: Hierarchical generation of dynamic and nondeterministic quests in games. In: Proceedings of the 11th Conference on Advances in Computer Entertainment Technology, pp. 24–1. ACM (2014)

Togelius, J., De Nardi, R., Lucas, S.M.: Towards automatic personalised content creation for racing games. In: 2007 IEEE Symposium on Computational Intelligence and Games, pp. 252–259 (2007)

Togelius, J., Yannakakis, G.N., Stanley, K.O., Browne, C.: Search-based procedural content generation: a taxonomy and survey. IEEE Trans. Comput. Intell. AI Games **3**(3), 172–186 (2011)

Tulloch, R.: Reconceptualising gamification: play and pedagody. Digit. Cult. Educ. **6**(4), 317–333 (2014)

Valls-Vargas, J., Ontanón, S., Zhu, J.: Exploring player trace segmentation for dynamic play style prediction. In: Proceedings of the Eleventh AAAI Conference on Artificial Intelligence and Interactive Digital Entertainment, pp. 93–99 (2015)

Van Lankveld, G., Spronck, P., Van Den Herik, H.J., Rauterberg, M.: Incongruity-based adaptive game balancing. In: Advances in Computer Games, pp. 208–220. Springer (2009)

Weber, B.G., Mateas, M, Jhala, A.: Using data mining to model player experience. In: FDG Workshop on Evaluating Player Experience in Games (2011)

Weiser, P., Bucher, D., Cellina, F., De Luca, V.: A taxonomy of motivational affordances for meaningful gamified and persuasive technologies. In: Proceedings of the 3rd International Conference on ICT for Sustainability (ICT4S), Volume 22 of Advances in Computer Science Research, pp. 271–280. Atlantis Press (2015)

Yannakakis, G.N., Hallam, J.: Towards optimizing entertainment in computer games. Appl. Artif. Intell. **21**(10), 933–971 (2007)

Yannakakis, G.N., Togelius, J.: Experience-driven procedural content generation. IEEE Trans. Affect. Comput. **2**(3), 147–161 (2011)

Yannakakis, G.N., Togelius, J.: A panorama of artificial and computational intelligence in games. IEEE Trans. Comput. Intell. AI Games **7**(4), 317–335 (2015)

Yannakakis, G.N., Martínez, H.P., Jhala, A.: Towards affective camera control in games. User Model. User Adapt. Interact. **20**(4), 313–340 (2010)

Zlatow, M., Kelliher, A.: Increasing recycling behaviors through user-centered design. In: Proceedings of the 2007 Conference on Designing for User eXperiences, pp. 1–27. ACM (2007). ISBN 978-1-60558-308-2

Zook, A., Lee-Urban, S., Drinkwater, M.R., Riedl, M.O..: Skill-based mission generation: A data-driven temporal player modeling approach. In: Proceedings of the the Third Workshop on Procedural Content Generation in Games, pp. 1–8 (2012a)

Zook, A., Lee-Urban, S., Riedl, M.O., Holden, H.K., Sottilare, R.A., Brawner, K.W.: Automated scenario generation: toward tailored and optimized military training in virtual environments. In: Proceedings of the International Conference on the Foundations of Digital Games, pp. 164–171. ACM (2012b)

Reza Khoshkangini received his Ph.D. in computer science from the University of Padova and Fondazione Bruno Kessler (FBK) in Italy in 2018. He is currently a postdoc researcher at Halmstad University (Sweden) with a main research focus on data mining in the context of predictive maintenance. His research interests include gamification, data mining, machine learning, artificial intelligence and recommender systems.

Giuseppe Valetto received a Laurea degree in electronic engineering from Politecnico di Torino, Turin, Italy, in 1992, an MS in computer science from Columbia University, New York, NY, USA, in 1994, and a Ph.D. in computer science, again from Columbia University, in 2004. In his career as a researcher, he has mostly worked on collaborative software engineering and distributed and self-adaptive systems.

Annapaola Marconi received her Ph.D. in computer science at the University of Trento in 2008. She is currently Senior Researcher at Fondazione Bruno Kessler (FBK), where she directs the "Distributed Adaptive Systems" Research Unit. Her research interests include distributed systems, self-adaptive software systems, collective adaptive systems, service-oriented and cloud computing, gamification techniques, AI planning.

Marco Pistore is a research director at Fondazione Bruno Kessler (FBK), where he heads the Digital Society research line. He has an h-index of 36 and more than 100 publications in international journals, conferences and symposiums. He has been local coordinator of various national and European research projects including ALLOW and SLA@SOI.

Affiliations

Reza Khoshkangini[1] · Giuseppe Valetto[1] · Annapaola Marconi[1] · Marco Pistore[1]

✉ Reza Khoshkangini
reza.khoshkangini@hh.se

Giuseppe Valetto
valetto@fbk.eu

Annapaola Marconi
marconi@fbk.eu

Marco Pistore
pistore@fbk.eu

[1] Fondazione Bruno Kessler (FBK), Trento, Italy