Dive Deeper

Empirical Analysis of Game Mechanics and Perceived Value in Serious Games

Kniestedt, Isabelle; Gómez Maureira, Marcello A.; Lefter, Iulia; Lukosch, Stephan; Brazier, Frances M.

DOI
10.1145/3474663

Publication date
2021

Document Version
Final published version

Published in
Proceedings of the ACM on Human-Computer Interaction

Dive Deeper: Empirical Analysis of Game Mechanics and Perceived Value in Serious Games

ISABELLE KNIESTEDT, Delft University of Technology, The Netherlands
MARCELLO A. GÓMEZ MAUREIRA, Leiden University, The Netherlands
IULIA LEFTER, Delft University of Technology, The Netherlands
STEPHAN LUKOSCH, University of Canterbury, New Zealand
FRANCES M. BRAZIER, Delft University of Technology, The Netherlands

Fig. 1. Treasure diving in *Pocket Odyssey*.

Validation of serious games tends to focus on evaluating their design as a whole. While this helps to assess whether a particular combination of game mechanics is successful, it provides little insight into how individual mechanics contribute or detract from a serious game's purpose or a player's game experience. This study analyses the effect of game mechanics commonly used in casual games for engagement, measured as a combination of player behaviour and reported game experience. Secondly, it examines the role of a serious game's purpose on those same measures. An experimental study was conducted with 204 participants playing several versions of a serious game to explore these points. The results show that adding additional game mechanics to a core gameplay loop did not lead to participants playing more or longer, nor did it improve their game experience. Players who were aware of the game's purpose, however, perceived the game as more beneficial, scored their game experience higher, and progressed further. The results show that game mechanics on their own do not necessarily improve engagement, while the effect of perceived value deserves further study.

CCS Concepts: • **Human-centered computing → Empirical studies in HCI**; • **Software and its engineering → Interactive games**.

Additional Key Words and Phrases: serious game, game user experience, engagement, game mechanics, game design, validation

Authors' addresses: Isabelle Kniestedt, i.kniestedt@tudelft.nl, Delft University of Technology, TPM, Delft, The Netherlands; Marcello A. Gómez Maureira, m.a.gomez.maureira@liacs.leidenuniv.nl, Leiden University, LIACS, Leiden, The Netherlands; Iulia Lefter, i.lefter@tudelft.nl, Delft University of Technology, TPM, Delft, The Netherlands; Stephan Lukosch, stephan.lukosch@canterbury.ac.nz, University of Canterbury, HIT Lab NZ, Christchurch, New Zealand; Frances M. Brazier, f.m.Brazier@tudelft.nl, Delft University of Technology, TPM, Delft, The Netherlands.

2573-0142/2021/9-ART236. https://doi.org/10.1145/3474663

ACM Reference Format:
Isabelle Kniestedt, Marcello A. Gómez Maureira, Iulia Lefter, Stephan Lukosch, and Frances M. Brazier. 2021. Dive Deeper: Empirical Analysis of Game Mechanics and Perceived Value in Serious Games. *Proc. ACM Hum.-Comput. Interact.* 5, CHI PLAY, Article 236 (September 2021), 25 pages. https://doi.org/10.1145/3474663

1 INTRODUCTION

Serious (digital) games (SG) are games with a purpose other than pure entertainment [34]. Designers of SGs apply game mechanics, not only to drive engagement with the game itself (resulting in an enjoyable experience), but also to achieve this purpose [24]. In academic literature, such games tend to be validated and evaluated *as a whole* both on their entertainment value and whether they fulfil their intended purpose [58]. So far, results have shown conflicting, but promising evidence for SG's effects across different applications [13, 14, 56]. While individual case studies help to assess whether a particular combination and implementation of mechanics was successful in achieving its goals, it provides little insight in how the individual mechanics contribute or detract from a game's purpose or a player's experience [33]. Games are systems with many interacting elements that give rise to complex emotional constructs [59] and that occupy a player's attention in varying ways from moment to moment [11]. While commercial game reviewers tend to make detailed distinctions between various aspects of a game's design and how they factor into the overall experience, such distinctions are generally lacking in SG evaluation. This study is based on the perspective that it is worthwhile to thoroughly test individual aspects of a serious game's design and ensure their effects are understood, especially if the goal is to reliably produce a measurable result rather than only provide an entertaining experience.

In addition to game mechanics and their influence on behaviour and experience, there is the question of how a game's purpose influences these factors as well. Models have been developed to guide SG design and evaluation in (theoretically) connecting general game aesthetics or feedback mechanisms to particular outcomes [2, 3, 5, 31, 60]. Researchers have also suggested that a game's purpose and its perceived value [29] play a role in a SG's success. It can be hypothesised that players regard a serious game more positively if they consider it to be beneficial. In turn, this may lead to increased enjoyment and engagement. To the authors' knowledge, however, there is limited work to validate such claims, and how this is meaningful for SG design.

This paper presents an empirical study using the web-game *Pocket Odyssey*. The game was developed as part of a multi-year EU-funded project aimed at developing technology that supports adult users in pursuing healthy habits and help them maintain their well-being. The game, in particular, was created to provide cognitive training. It aims to do so by engaging users in a cognitive game-based task (requiring memory and spatial navigation skills) on a regular basis. The game was designed with input from domain experts in Psychology and developed by a professional game designer. At present, the game has not been validated for its effects on cognition, nor is the presented study focused on examining those effects. Instead, this study examines the design of serious games using supplemental motivational game mechanics intended to make users more likely to perform a core 'serious' task on a regular and continuous basis. Operating under the assumption that the core task *could* be beneficial, its success is then measured in whether people play it, how much they play it, and whether they would continue to play it. *Pocket Odyssey* is used as a a representative example of a game with such a design philosophy. As such, this study tests the employed *game mechanics* and their effects on engagement, measured through player behaviour and game experience. Additionally, it examines *how players perceive the game depending on the purpose they are presented with*, and how this perception affects those same measures.

Participants played one of four possible game versions. The base version of the game focuses on the cognitive task alone, while each other version presents players with supplementary game

mechanics, adding up to a more diverse (and in turn potentially more engaging) game experience. The details of these versions are described in Section 3.5.

In having participants play different versions of the game, monitoring their behaviour, and assessing their game experience, this study explores the following hypothesis:

H1: When playing a game version with supplemental game mechanics, participants will play longer and rate their experience higher.

Participants who played *the most elaborate version of the game* were also presented with *different game purposes*. These included playing the game *for their own benefit, for the benefit of others*, or simply because they *participated in a research experiment*. With this additional data, the following hypotheses were examined:

H2a: When the game is presented as beneficial to the player or others, participants will value the game more and play longer, as opposed to those without an explicit purpose beyond contributing to research.

H2b: Players with higher awareness of the game's purpose will value the game more, play longer, and rate their experience higher.

Data was gathered using a mixed-methods approach, using surveys and game metrics to assess player behaviour and game experience. As such, comparative statistical tests were performed to assess differences between condition groups, while qualitative data provided additional context to interpret quantitative results.

Results indicate support for H2b, but not for H1 or H2a. The implications of these findings are discussed in Section 6. Through its results and discussion, this study exemplifies the need for in-depth analysis of individual aspects of a serious game's design, as important details in understanding their functionality may otherwise go undetected. It also provides an empirical perspective on the role of a serious game's purpose (and a player's awareness of said purpose) in the perceived value and experience of serious games. Additionally, reflections on the presented methodology can benefit others looking to evaluate their game designs more thoroughly, and the presented data can inform future empirical studies examining similar topics. Finally, *Pocket Odyssey* is presented as an artefact that can be used as a basis for further study. Support material and data of this study are available through OSF[1].

2 BACKGROUND

For the purposes of this study, *game mechanics* are defined as "methods invoked by agents, designed for interaction with the game state" [51]. In this definition, a method is understood as an action or behaviour available to an agent (e.g., the player) to interact with the game world. Methods are phrased as verbs, e.g., climb, take cover, shoot, or steer. They are invoked through input methods (e.g., pressing a button) and have visible effects on *game elements* (e.g., objects or characters in the game world), causing them to undergo designed changes and/or interact with one another. In turn, interactions are defined by the *rules* that apply to the game world (e.g., which surfaces are climbable). Individual game elements can be discerned from others via their unique properties, which "are often either rules or determined by rules" [51]. Together, game elements and rules define the game *system* and its *sub-systems* (e.g., a 'crafting' or 'cover' system).

Various models or frameworks exist connecting the inner workings of games to measurable effects aimed for with serious games. Literature reviews [2, 22] provide an overview of how learning outcomes and motivation have been connected to game elements in academic literature. Though game elements have been connected to various educational theories (e.g., Bloom's taxonomy), the definition of these game elements is extremely broad, ranging from feedback methods (e.g., points)

[1]Link to OSF repository: http://doi.org/10.17605/OSF.IO/27KUY

to concepts like 'uncertainty'. Specific models (e.g., [3, 31] suggest similar connections. However, such broad interpretations of mechanics make it difficult to relate any models back to concrete game design. This approach is not unique to serious games for learning either — more generic models for evaluating serious games revolve around testing the game's 'design' as a whole as well [15].

Studies into the (empirical) effects of specific game aspects do exist. However, what the concept of game mechanics means varies here as well. For example, Parnandi and Gutierrez-Osuna [45] assess the effect of manipulating properties in a racing game (speed, visibility, and steering jitter) on player arousal. Hew et al. [25] examine the effect of displaying feedback (through points, badges, and leader boards) on motivation and assignment quality in university students. Similarly, Cantador and Marczewski [20] examine rewards (primarily badges) in an e-learning environment. Matin et al. [37] examine the effect of a timer, top score, and leader board on performance and motivation in human computing games.

While efforts in examining broader game experiences are useful, they provide limited insight into how various aspects of the game influence a player's experience. This reliance on many individual case studies of 'complete' serious games means that frameworks are built upon incompatible and conflicting data [56]. Although the interactive nature of games appears to have a clear benefit over 'passive' modes of presentation (see e.g., [53]), merely framing an activity as a game has shown to be enough to increase interest and enjoyment [33] as well. As such, it is not clear how individual game mechanics affect player experience.

Digital games are complex systems with rules and mechanics that can interact in unexpected ways and induce a wide range of emotional states in players [59]. Players interacting with games focus their attention on them, an act that can lead to affective and behavioural states such as perceived 'flow', presence, and immersion [26]. However, attention is not simply focused on 'the game'. Instead, attention is a limited resource that is necessarily divided between different aspects of the game, such as its controls, aesthetics, narrative, social features, and more, depending on the game's design [11]. As such, engagement is influenced by many factors and may vary and be redirected during interaction with a game [42]. Entertainment game reviewers recognise this, and tend to evaluate various aspects of a game's design (e.g., individual mechanics, art, animation, sound, writing, atmosphere), as well as its mechanics and sub-systems on how they add to or detract from the overall game experience.

In academic evaluation of games, these distinctions are not always made. Measurement instruments, such as the Game User Experience Satisfaction Survey (GUESS) [46] and the Game Engagement Questionnaire (GEQ) [8] (as well as the disputed Game Experience Questionnaire (GEQ) [32]) tend to emphasise measuring the manifestation of emotional and behavioural states (e.g., 'losing track of time' and 'enjoyment'). Some also include items related to certain aspects of a game — the GUESS, for example, includes a module on narrative. However, instruments such as these are almost always used to evaluate a (serious) game as a whole. While the GUESS can give an indication on how rewarding a game's narrative was, details on sections of that narrative, as well as what other aspects of a game players found more or less enjoyable is not recorded. This aligns with the goal to provide a generalisable instrument that can be used to evaluate a wide range of games — individual questions relating to specific mechanics can hardly be included in a universal measuring tool. However, this means that details on the player experience inevitably get lost.

Which mechanics should then be investigated? Game design is not as simple as adding mechanics until the point of saturation. Ideally, games are designed around a 'core experience' [50], with every design decision working to enforce that core experience. In practice, however, this is not always the case. Commercial game developers often repeat or copy designs that have proven successful in the past and fulfil the expectations of audiences. As a result, recent years have resulted in many instances of third-person open-world exploration games with similar combat and crafting

mechanics, collectables, and a photo mode [6]. These trends are perhaps even more visible in the casual or mobile games sector, where game designs tend to be far less complex and copying what is successful is the norm rather than the exception. Searching for a 'Match 3' game in the iOS App Store results in a multitude of games that combine the basic design of *Bejeweled* [17] with a renovation- or decoration-simulation mechanic and narrative. Examples include *Gardenscapes* [47], *Match Town Makeover* [1], *RollerCoaster Tycoon®Story* [4], and many others. Considering the time and resources involved in developing such mechanics, these additions are likely used to add diversity to an otherwise simple base mechanic (adding to long-term engagement), expected by players, and necessary in distinguishing similar games in a crowded market. What sets these games apart from one another is how well the individual parts work on their own, how well they work together, and how well a particular 'flavor' (e.g., samurai, fantasy warrior, gardener, fashionista, and so on) resonates with audiences.

Serious games are, ideally, developed in a way that game designers suggest they should be — holistically, with every decision enhancing the 'core experience', and with the purpose thoroughly integrated into every aspect of that experience, all the while striking a balance between entertainment and 'meaning' [24]. Doing so, however, is no easy task. Given the complexity of game design, it makes sense that methods like gamification (i.e., applying game elements such as points and leaderboards to a non-gaming task) that are often not difficult to apply became popular some years ago [40]. While many serious game creators do not take such an approach, limitations and considerations unique to serious games still make that they are often smaller in scope than many commercial games, and that designs consisting of a core mechanic (possibly with a selection of supplemental mechanics and/or feedback mechanisms) are common (see [7, 9, 16, 38, 55] for some examples from recent years). At times, this simplicity is even preferable [39], as serious games need to be easily accessible to a broad (often non-gaming) audience, limiting their desired complexity. This is a requirement they have in common with casual games. Together with casual games' success in eliciting engagement, the mechanics of these types of games are a natural fit for SGs [35]. It is for this reason that this study attempts an empirical exploration of mechanics for SGs using commonly used mechanics within this sector of the market, using *Pocket Odyssey* (described in the next section) as a game that is considered comparable in complexity and design to others developed for similar purposes.

The focus on measurable outcomes and entertainment value of serious games as a whole [5] is valid for assessing individual games and their success. Yet, serious games, even relatively simple ones, not only have the same complex structure of interacting sub-systems and other elements as entertainment games — they also include a purpose that is ideally interwoven with a game's structure. Considering this is the case, understanding the inner workings of a serious game, how attention is drawn to different aspects of the design, and how such aspects add to or detract from player experience, becomes even more imperative. Therefore, this article poses, that it is desirable to study game mechanics in a more systematic manner, not only to understand how game mechanics affect player experience, but also to understand how well they support a SG's purpose as well. One approach to serious games is to use game mechanics as a distraction from a (possibly unpleasant) task (e.g., in health applications such as motivating physical exercise or distracting patients from treatment [39]). Depending on the purpose, however, it has also been suggested that reminding the player of the game's purpose, either through its framing or through the game design itself, may actually be beneficial as well. Through a review of literature, Hamari and Keronen [23] established a correlation between perceived usefulness and enjoyment among people playing games. In discussing long-term engagement in games for health, Kayali et al. [29] emphasise the importance of using game mechanics to increase the game's perceived value (e.g., by connecting game mechanics to everyday habits). Steinemann et al. [53] established a link between interactivity,

Fig. 2. Screenshots of submarine (left), ship (middle), and story (right) views.

appreciation (i.e., gratification not necessarily derived from media being 'fun', but rather thought-provoking or meaningful) and a player's inclination to donate after playing a serious game for change. In this vein, and to further explore the inner workings of game mechanics in emphasising a serious game's meaning, in-depth empirical studies dissecting game design implementations are warranted.

3 THE GAME: POCKET ODYSSEY

The game used in this study is *Pocket Odyssey*, a 2D game in which players buy an old boat and fix it up while hunting for treasure. It was developed to run in all modern browsers. For an offline version of the game (Windows and macOS), as well as a video showing gameplay, please refer to the supplementary files. The design of *Pocket Odyssey* is based on a popular type of mobile game (Section 2) and it was created to provide cognitive training for older adults, following guidelines regarding theme and complexity in line with this target audience [19, 30, 41].

3.1 Submarine

The submarine game (Figure 3) forms the core gameplay loop of *Pocket Odyssey*. The player goes diving for treasure, and needs to direct a submarine through an underwater cave. The player chooses the level they want to play from an overarching 'map' screen. They are then shown a map of the selected level, which shows its layout and the position of 'coin fragments'. A total of fifteen coin fragments are scattered throughout a level. Five fragments combine into one coin, for a maximum total of three complete coins per level. These coins act as a form of feedback (i.e., score) and are shown in the level select menu. The player can set the submarine's speed — a complete stop, a slow 'turtle' speed, or a faster 'rabbit' speed — which causes the submarine to move from left to right through the level. The player then navigates the level by directing the submarine up and down, either by click-and-dragging the mouse or using the arrow keys. Players also need to avoid obstacles, such as the boundaries of the maze and seaweed. If the submarine becomes too damaged, the player loses and needs to retry the level. If the player reaches the end of the level successfully, they unlock the next level and any complete coins they have collected will be added to their total. A maximum of three coins can be collected per level. Coins can only be collected once (e.g., if a player finishes the level with two coins first and replays it to get all three coins, they only get one additional coin). Levels become progressively harder and their aesthetic becomes darker as well.

The cognitive aspects of the submarine game include the steering of the submarine with accuracy, spatial navigation, and memorising and recalling the level layout. As stated before, these aspects have not been tested for cognitive benefits. For the purposes of this study, it was primarily important that players could believe playing the game might be beneficial.

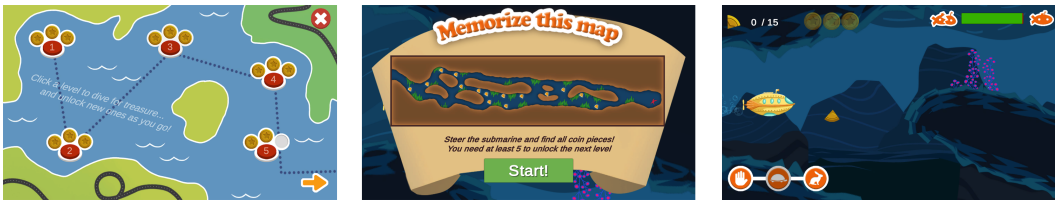


Fig. 3. Screenshots of level select (left), level map (middle), and submarine gameplay in a higher level (right).

3.2 Ship

The coins that the player earns by playing the submarine game are used to renovate their ship. Whenever the player completes a level, they return to the view of the ship. It is depicted from the side in a cross section, showing the different floors and rooms inside the ship (Figure 4). At the start, most of the ship is unfinished. The player can use the coins collected in the submarine game to unlock new rooms on the ship and renovate them step-by-step. There is a total of ten rooms, each of which has an adjustable wallpaper and three objects in it. In total, this means there are forty steps of progression until the ship has been fully restored. For every step, players have the choice of three decoration options. The steps unlock linearly, and always in the same order. The player can see what the next step will be (e.g., "Let's add a cupboard!") and how many coins they need. They can always change the decor of any previously unlocked step.

3.3 Narrative

As players progress and unlock new parts of the ship, they are presented with story scenes. Once the player has confirmed their decoration choice, the view of the ship becomes obscured by the story view (Figure 2). The story of *Pocket Odyssey* is presented over fourteen scenes, each of which are told through text on the screen. Players can choose their responses through text-based buttons. The story is largely linear, but has several 'closed branches' where the player's choices lead to different outcomes. The story is written from a second person point of view, i.e., "You see something on the horizon that catches your interest". The story posits the player as a person who decides to quit their job, buy a boat, and travel the world. Joining the player is the 'guide' character, who the player gets to know over the course of the story. The story incorporates themes of family and thinking of home. Over the course of the journey, the player and guide encounter different places, reminisce about the past, and experience the dangerous sides of the ocean. It ends with a newly formed friendship between them and both characters returning home.

3.4 Information and Loading Screens

Players login on the opening screen by putting in a unique code. An account is made upon the first login, and every consecutive login the player continues with their previously saved progress. The game shows a frame rate counter on this screen to ensure the game is running at an acceptable rate (at least 15 frames per second). Upon logging in, the player sees an information screen while the game loads (for approximately five seconds). The information explains the purpose of the study and has an additional tagline for the game. The exact information varies depending on the experiment condition (Section 3.5). The player can bring up the same information screen from either the ship screen or the submarine level select menu.

Whenever the player enters the submarine level select screen, a loading screen is visible until the player clicks on it to continue (Figure 4). The screen is visible for approximately two seconds before the player can click it. This screen shows instructions for the game and has a single encouraging

line of text that relates to the purpose of the game. Similar to the information screen, the tagline changes depending on the experiment condition (Section 3.5).

Fig. 4. Screenshots of loading screen in Cust (left), renovation option in the kitchen (middle), and nearly renovated ship (right).

3.5 Versions

For the purposes of this study, players were exposed to different versions of the game. To reiterate the goals of the study, they were (1) to examine the effects of supplemental game mechanics to enhance engagement, and (2) to examine the effects of the serious game's purpose and players' awareness of it. As such, the versions of the game exposed players to different combinations of mechanics, and adjusted the purpose of the game presented to players.

Participants played one of four possible versions of the game's mechanics, herein referred to as *Base*, *Cust*, *Narlin*, and *Narcho*:

Base: The version only includes the submarine mechanics, i.e., the core of *Pocket Odyssey*. In essence, it is what the result could be of any serious game project when a core task (e.g., cognitive training) is gamified. The activity itself is broken into manageable levels and the player is frequently rewarded with some form of points. It could be further expanded with other forms of feedback, but the minimum that can make the task be perceived as a game is incorporated (i.e., goal, win/lose conditions, feedback, aesthetics). In this version of the game, players go to the level select screen whenever they login or finish a level, instead of the ship. The loading screen shows instructions only related to the submarine game and the purpose-relevant tagline.

Cust (customize): This version includes the ship renovation mechanic in addition to the submarine game. Players return to the ship view whenever they complete a level. They can use the coins they have collected to renovate the ship. However, no story will trigger as they progress. The loading screen explains both the submarine and the ship mechanics.

Narlin (narrative linear): This version includes all sub-systems (submarine, ship, and story). In the story scenes, players are given a single option, making the story fully linear and reducing interactivity to letting players step through the narrative. The loading screen explains all sub-systems.

Narcho (narrative choice): Similar to Narlin, but provides players with choices at several points in the game's narrative scenes. Players can have up to three options at any point, some of which lead to differences in how details of the story unfold. The loading screen is the same as for Narlin; the presentation of choices to continue the narrative is presumed to be understood without explicit explanation.

Additionally, there are three possibilities for the game's purpose as presented to players. In this paper, they are referred to as *None*, *Self*, and *Other*.

None: Players are reminded they are participating in a study to examine digital games. The tagline and loading screen convey game related information only (e.g., 'Click to continue').

Self: Players are informed digital games can be used for cognitive training. The tagline and loading screen reminder reinforce this message.

Other: Players are informed digital games can collect behavioural data used to study cognitive decline. The tagline and loading screen reminder reinforce this message.

Each purpose changes the information that is shown on the menu and loading screens (Section 3.4). All participants, due to the nature of the recruitment platform (see Section 4.1), were aware that they were participating in a research study. Thus, the conditions contextualised their participation in various ways, as described below.

For example, the information in the Self condition reads: "This game trains your cognitive skills. Cognitive decline happens naturally with age, but may be delayed by some activities. Games can be such an activity. The more you play, the more you exercise your brain to stay healthy!". The tagline on the loading screen read: "Every second you play trains your brain!". For all included texts, please refer to the supplementary files.

With these differences in messaging, the aim of the study is to see whether people behave differently depending on the purpose of the game, assuming that they are aware of said purpose. The Self condition best represents the intended use case for which *Pocket Odyssey* was designed, i.e., to provide individual users with a way to train their cognition. In the context of the study, the game's purpose is most closely aligned with the None condition (although its description lacks specific details). The study design was approved by the Delft University of Technology Human Research Ethics Committee.

4 EXPERIMENT

The goal of the experiment is to examine the effects of game mechanics (H1) and game purpose (H2a and H2b) on player perception and behaviour. The following section describes the experiment design, the measurements, the pilot study, the procedure, and how data was processed.

4.1 Participants and Sampling

Participants were recruited using Prolific, an online platform focused on recruiting research participants. Participants were compensated for their time with ≈3.50 GBP. A total of 344 participants were recruited to participate in the study (of which 204 ultimately completed the study and provided complete data sets). Requirements for participation were a minimum age of 40 and fluent understanding of the English language, due to the amount of text present in the game. Although *Pocket Odyssey* was developed for an older target audience (55+), the minimum age was set lower to increase the potential sample size through the recruitment platform.

Formally, quantitative studies using frequentist statistics determine an appropriate sample size through a prospective power analysis. However, such an analysis requires existing quantitative data about the research topic [10]. In Bayesian statistics (used in this study, as described in Section 5) the concept of statistical power does not exist in the same manner, but previous work informs the priors used in analysis [43]. This study is an exploratory study on a topic that, to the authors' best knowledge, has limited quantitative work on which to base expectations, and uses a game not previously applied in experimental studies. As such, it is difficult to determine what sample size to aim for, what effect size to expect, or what priors to use. Instead, a minimum sample size was determined using *local standards*, i.e., sample sizes from comparable user studies published within the CHI community [10], as a guideline. Based on this information and taking into account the exploratory nature of this research, the choice was made to gather data from at least 30 participants per group. In the absence of well informed (and sourced) prior beliefs, the default priors of JASP are used for statistical tests. The limitations of this approach are discussed in Section 7.

4.2 Condition Groups

Participants are divided into condition groups, each of which plays a different combination of game and purpose (Section 3.5). The two variables (4 game versions and 3 purposes) make for a total of 12 possible combinations. However, the study focuses on a subset of these, leading to a total of 6 condition groups: Base_Self, Cust_Self, Narlin_Self, Narcho_Self, Narcho_Other, and Narcho_None.

By focusing on these groups in particular, it is possible to run statistical tests looking at each variable in isolation while also maximising available resources (i.e., time and budget for participant compensation). The study aims to examine differences in the game versions (with the same purpose) and differences in purpose (with the same game version). Statistical tests (see Section 5) focus on comparisons between these two separate data sets.

In deciding on which groups to focus resources, the versions of the game are chosen that have the highest ecological validity. As such, Narcho is chosen as a basis to compare purposes, as it is the most 'complete' version of the game and the version that its developers intend to be used as a serious game to train cognition. Similarly, the 'Self' purpose is used to compare game versions with, as this accurately reflects the intended purpose that the game was designed to fulfil.

As such, H1 is examined across the conditions Base_Self, Cust_Self, Narlin_Self, and Narcho_Self. H2a and H2b are examined across Narcho_Self, Narcho_Other, and Narcho_None.

The limitations of deciding to focus on a subset of conditions to maximise available resources are discussed in Section 7.

4.3 Pilot Study

A pilot study was conducted to test the experiment procedure (described below in Section 4.4). A total of 21 participants were recruited with Prolific, 52% of which were female (n=11). Results from the pilot showed that the game was well-received, with survey results well above the mid-point (see Section 4.5.3) and participants playing for longer than requested by the experiment instructions. This suggested that the game was of sufficient quality to be used in the larger study. Performing the pilot helped to increase the clarity of instructions given to participants, as well as uncover issues with the game that were fixed before the experiment took place (e.g., varying performance depending on browser).

4.4 Procedure

Participants are asked to play *Pocket Odyssey* for 3 days for a duration of at least 5 minutes per day. While it is necessary for them to play on three separate dates, completing the study is not hindered by playing less than the requested time. This duration was chosen based on the time it takes to complete the game content (i.e., getting perfect scores on each level), which is around 20 minutes for an experienced player playing the Base version of the game.

First, participants fill out a demographics survey and informed about the study through step-by-step instructions for each day of participation. They are then directed to the game's website and instructed to bookmark this page for subsequent days. The experiment instructions are repeated on the game page as well. Participants are invited randomly by Prolific according to the sampling restrictions listed above, and the pre-game survey takes around a minute to complete.

Participants automatically create a new server entry when they login for the first time with their Prolific ID. This ID is used to identify participants, and connect survey and gameplay data. At this point, the server randomly assigns participant to a condition group. The game then reads the condition from the server and changes which game mechanics are available and text is shown to the player (Section 3.5). Participants proceed to play the game according to the instructions and continue to do so independently for at least three days.

Players are invited for a second post-game survey when their game data shows they have played for three days. From day four, the game's information screen shows a code. This code is required to access the post-game survey and varies depending on experiment condition, therefore serving as an additional check to ensure data integrity. The post-game survey takes around 10 minutes to complete.

4.5 Measurements

Data is collected pre-, during, and post-game. Before playing the game, participants answer a general demographics pre-game survey. During gameplay, the player's in-game actions are logged and stored on the server (game metrics). Participants answer a second survey, the post-game survey, upon completing three days of gameplay. The post-game survey is made up of multiple parts. Each of the measures is described below. The surveys use a combination of open questions and those rated on a five-point Likert scale.

4.5.1 Pre-Game Survey. The pre-game survey covers basic demographic information. Participants report their age in years and gender ('female', 'male', 'not listed', 'prefer not to answer'). They also rate their previous experience playing games and time spent playing games on average (both on a 5-point Likert scale, with experience ranging from 'Novice' to 'Expert', and play time from 'Less than an hour per week' to '20+ hours per week'). Finally, they list the types of games they usually play as free text. For the full survey, please refer to the supplementary files.

4.5.2 Metrics. Pocket Odyssey logs interactions the player has with the game. Each event is logged with a timestamp, an event type, and a description line. This makes it possible to calculate the time spent playing, as well as group interactions of a similar type (e.g., interactions in the ship view, or during the submarine game). Examples of logged events include (but are not limited to): choices taken in the game narrative, unlocking ship progress steps, ship decoration choices, instances of redecorating, starting a submarine level, finishing a submarine level, changing speed during a submarine level, picking up a coin fragment during a submarine level, failing or succeeding at a submarine level, replaying a submarine level, time spent on loading screens, and opening/closing the information screen. General player statistics are also logged, such as maximum level reached, best score per level, and total number of coins collected. While these metrics partially allow to test the hypotheses (e.g., in showing how much time players spent playing the game), the metrics can further provide insight into a player's experience [21] when combined with other measures.

Game metrics were processed through custom Python scripts using the Pandas library [44]. Any data collected was only connected to an individual participant through their Prolific ID and, as such, was completely anonymous.

4.5.3 Post-Game Survey. The post-game survey aims to assess the player's experience with both the experiment and the game itself. It is split into three parts: game impressions and motivations, modules of the Game User Experience Satisfaction Scale (GUESS) [46], and (depending on condition) a questionnaire on agency in digital narratives.

Participants are reminded of the purpose of their game version and rate (1) how aware they were of this purpose while playing, as well as (2) how much this motivated them to play. Individual game aspects (coin collecting, ship decorating, and narrative) are similarly rated when applicable. Each of these questions is rated on a 5-point Likert scale with the descriptive options of 'Not at all', 'Slightly', 'Moderately', 'Very', and 'Extremely'. Participants also rate how beneficial and useful they considered the game using a similar 5-point scale. Finally, they rate how they experienced playing over 3 days, and having to stop playing using a 5-point Likert scale ranging from 'Extremely negative' to 'Extremely positive'. Open questions ask participants to elaborate on their primary

motivation for playing, why they did or did not play longer than the requested 5 minutes, and any other comments they have about their experience. The supplementary files provide all details of the survey.

The GUESS is a validated instrument to measure aspects of game user experience. The following modules are used in the study: Usability / Playability, Play Engrossment, Enjoyment, Personal Gratification, and Visual Aesthetics. Although the GUESS has more modules (e.g., social connectivity), these did not apply to the design of *Pocket Odyssey*. The chosen modules give an indication of the overall quality of the game (e.g., Usability providing an indication that negative experiences are not due to issues or difficulties with the game's controls) and the players' subjective experience.

Participants who are assigned either the Narlin or Narcho conditions also answered modules of an existing survey assessing agency in interactive narratives [48]. The included modules were: Effectance (i.e., sense of being able to influence the story), Presence (i.e., sense of 'being there'), Character Believability, Identification (i.e., feeling like the main character), Aesthetic Pleasantness, Curiosity, Suspense, and Enjoyment. Participants were informed these questions pertained to the story of the game in particular.

5 RESULTS

All statistical tests are performed using Bayesian methods in JASP [27, 36]. The value of the Bayes Factor (BF) indicates the likeliness that a given hypothesis (H1) is not equal to its null-hypothesis (H0), i.e., the assumption that different testing conditions can be considered equal. The Bayes Factor can be expressed as evidence for H1 relative to H0 (BF_{10}), or as evidence for H0 relative to H1 (BF_{01}). All BF values in this study are expressed in BF_{10} notation. A BF value of 1 indicates that there is an equal chance of the hypothesis being different from the null-hypothesis as there is of them being similar. A value lower than 1 indicates that the null-hypothesis is more likely. Unlike classical hypothesis testing, a Bayesian test can therefore be used to indicate likeliness of the null-hypothesis, rather than only reject it [43]. Only results with 'moderate' ($3<BF<10$) or 'strong' ($BF>10$) evidence for H1, or H0 ('moderate': $0.1<BF<0.3$; 'strong' $BF<0.1$) are reported (evidence labelling used in JASP based on [28]). In the absence of well informed (and sourced) prior beliefs, the default priors of JASP are used and are reported for each statistical test. All results are calculated using a repeatability seed of 1 in JASP.

The following section presents results relevant to exploring the hypotheses. To reiterate, these are:

H1: When playing game versions with supplemental game mechanics, participants will play longer and rate their experience higher.

H2a: When the game is beneficial to participants themselves or others, they will value the game more and play longer.

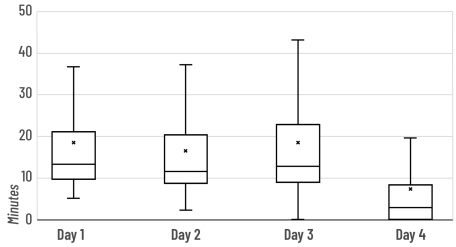
H2b: Participants more aware of the game's purpose will value it more, play longer, and rate their experience higher.

5.1 Descriptive Statistics

Overall, N=204 participants provided results for the study. A total of 344 participants started the game, with 211 of them playing for 3 days and completing the post-game survey. Data from 7 participants was found to be missing data in the submitted game log and was thus discarded. Out of the valid 204 participants, 51% identify as female (n=104), 48.5% as male (n=99), and 0.5% as non-binary (n=1). The median age is 48.5 (mean=49.7, SD=7.7, range 40-71). Reporting game playing frequency has a mean of 2.1, corresponding to "1 - 4 hours per week" (SD 1.2, range 1-5). Reported player experience has a mean of 2.4, corresponding to "casual" (SD=1.1, range 1-5).

	Mean	Median	SD	Min	Max
Day 1	18.49	13.29	18.16	5.14	202.84
Day 2	16.47	11.55	14.64	2.28	167.90
Day 3	18.51	12.83	18.06	0.0	189.16
Day 4	7.36	2.92	16.00	0.0	145.59
Total	63.33	48.66	55.77	15.28	514.36

(a) Playing times descriptive statistics (in minutes).



(b) Box plots for playing times per day.

Table 1. Aggregate playing times over multiple days.

Total playing times and playing times for each day are shown in Table 1. Days are counted relative for each participant, meaning that participants might have not played for a day between 'Day 1' and 'Day 2'. One participant logged in on 'Day 3' without playing the game, having played a total of 28.2 mins overall before. Data from this participant is kept due to providing play time of at least 15 minutes, despite not having actively played for three days.

Splitting up the play time per game segment, the total playtime for the submarine game has a mean of 36.4 mins (SD=36.6, range 5.42 - 335.71), playing time for the ship has a mean of 9.93 mins (SD=7.35, range 0.0 - 56.61), and playing time for the story has a mean of 4.88 mins (SD=4.78, range 0.0 - 38.61).

GUESS items are scored on a 1-7 Likert scale. Results per category are as follows: Usability — mean=5.7, SD=0.8; Play Engrossment — mean=4.7, SD=1.1; Enjoyment — mean=5.1, SD=1.3; Personal Gratification — mean=5.4, SD=0.9; and Visual Aesthetics — mean=5.3, SD=1.1.

On average, players are "moderately" aware of the game's purpose (mean=2.8, SD=1.1), and they are "moderately" motivated by the purpose (mean=3.0, SD=1.0). When asked about the game mechanics, players are "moderately" to "very" motivated by collecting all coins (mean=3.7, SD=1.0), "moderately" motivated to fix and decorate the ship (mean=3.1, SD=1.3), and "slightly" to "moderately" motivated to see the story (mean=2.7, SD=1.2). Participants consider the game "slightly" to "moderately" beneficial to themselves or others (mean=2.7, SD=1.0), and "slightly" to "moderately" useful (mean=2.7, SD=1.0). Each of these questions is scored on a 5-point Likert scale. Asked whether they would play the game again if it had more content, 78.9% (n=161) answered 'yes'.

5.2 Comparisons Between Condition Groups

Participants were sorted into condition groups with the following distribution: Base_Self (n=32), Cust_Self (n=34), Narcho_Self (n=35), Narcho_Other (n=36), Narcho_None (n=33), and Narlin_Self (n=34). As such, the amount of participants between condition groups was roughly equal and of large enough size to perform statistical analysis between groups.

To compare data between game versions, Bayesian ANOVA tests (with default priors of 0.5 for 'r scale fixed effects' and 1 for 'r scale random effects' [49]) were performed between Base_Self, Cust_Self, Narcho_Self, and Narlin_Self. Players in Base condition reached a higher maximum level (mean=14.3, SD=1.7, BF=8.332) than those in other conditions. Except for this, all other measures show either only anecdotal evidence, or evidence for condition groups having no meaningful impact. For example, strong evidence is found that total playing time is not impacted by the game condition (BF=0.141). Similarly, either moderate or strong evidence was found that GUESS categories are

not impacted either (Usability — BF=0.139, Play Engrossment — BF=0.107, Enjoyment — BF=0.080, Personal Gratification — BF=0.103, Visual Aesthetics — BF=0.193). A separate test was run between all condition groups except for Base_Self to examine progress on fixing up the ship, but results show no impact in this case either (BF=0.063).

The same tests were performed to compare data between purpose condition groups, i.e., Narcho_Self, Narcho_Other, and Narcho_None. No measures show moderate or strong evidence for H1. In this comparison, submarine playing time, level progress, and game perception and motivation are of particular interest. Strong evidence was found that submarine playing time (BF=0.091) and maximum level (BF=0.089) are not impacted by the game condition. The same lack of impact was found for purpose awareness (BF=0.222), GUESS Usability (BF=0.098), Play Engrossment (BF=0.130), Enjoyment (BF=0.113), and Visual Aesthetics (BF=0.145).

A Bayesian Independent Samples Student T-Test (prior cauchy scale 0.707) was performed between the condition groups Narcho_Self and Narlin_Self to examine differences between a linear narrative and one with player choice in regards to player agency. Most results again provide evidence that there is no impact. The only exception is Effectance, which shows strong evidence for H1 (BF=6.042).

5.3 Correlations

With evidence indicating that measures between testing conditions are largely similar, pair-wise Bayesian Pearson's ρ correlations were carried out across the entire study population. The investigation of correlations was motivated by the desire to explore potential impacts of player motivations and purpose awareness, as well as to identify metrics that should be studied in detail in future work.

Such an investigation of multiple comparisons typically involves adjustments to reduce the risk of false positives (e.g. Bonferroni correction). In Bayesian statistics, such adjustments are not necessary because evidence for (or against) a hypothesis is expressed directly as a probability of H1 versus H0, instead of the rejection of H0 with the probability of a type I error. As such, probabilities of multiple comparisons do not accumulate to increase the likelihood of a type I error. However, multiple Bayesian correlations require the adjustments of priors, as individual comparison pairs are likely not fully independent from one another [52].

Methods for establishing such priors, even in the absence of informed strong prior beliefs, have been described in literature [18], but have not yet found their way into JASP. Thus, the correlations that were carried out do not involve informed priors and treat each comparison as independent. The stretched beta prior width was kept at its default of 1 in JASP, indicating uninformed priors [57].

Correlations with 'strong' (BF>10), 'very strong' (BF>30), or 'decisive' (BF>100) support are reported below. For detailed correlation results, refer to the JASP report in the supplementary material.

- **GUESS** Usability correlates with maximum submarine level (r=0.225, BF>10), total collected coins (r=0.258, BF>30), and purpose awareness (r=0.290, BF>500). It correlates negatively to player age (r=-0.233, BF>10).
  Play engrossment correlates to total time (BF>50), time on day 3 (BF>50), time on day 2 (BF>10), maximum submarine level reached (BF>10), and number of submarine level attempts (BF>10).
  Enjoyment correlates with total submarine play time, maximum level reached, number of submarine level attempts (all BF>100), time on day 2 (BF>30), and total coins collected (BF>30).

Personal gratification correlates to total submarine play time, time on day 2, time on day 3, maximum submarine level reached, coins collected, number of submarine level attempts, and number of attempts failed (all BF>100).
All except Visual Aesthetics correlate with purpose motivation (BF>100). All GUESS measures correlate with considering the game beneficial and useful (BF>100, except beneficial—Usability with BF>10). They also correlate with coin and ship motivation (all BF>100, except Usability with BF>10). Story motivation only correlates with Play Engrossment, Enjoyment, and Personal Gratification (all BF>100). Play Engrossment (r=-0.242, BF>30) and Personal Gratification (r=-0.240, BF>30) both negatively correlate with playing experience.

- **Purpose awareness** correlates with purpose motivation (r=0.315, BF>1000), motivation for collecting all coins (r=0.217, BF>10), considering the game beneficial (r=0.302, BF>1000) and useful (r=0.363, BF>100k).
- **Purpose motivation** correlates with coin (r=0.342, BF>10k) and ship motivation (r=0.292, BF>100), considering the game beneficial (r=0.509, BF≫100k) and useful (r=0.492, BF≫100k), how players experienced playing for 3 days (r=0.313, BF>1000), and the maximum submarine level reached (r=0.227, BF>10). It also negatively correlates with having to stop playing at the end of the experiment (r=-0.339, BF>10k).
- **Considering the game beneficial and useful** both correlate with each other (r=0.838, BF≫100k), playing for 3 days (r=0.575, BF≫100k), and coin, ship, and story motivation (BF>1000). Both also correlated negatively with having to stop playing the game (BF>100). 'Beneficial' correlated positively with the maximum submarine level reached (r=0.242, BF>30).
- **Game motivations:** Coin motivation correlates with ship motivation (r=0.366, BF>10k), total time (r=0.337, BF>10k), and total coins collected (r=0.324, BF>1000). It correlates negatively with having to stop playing (r=-0.337, BF>10k). **Ship motivation** correlates with story motivation (r=0.348, BF>100). All three (coin, ship, and story motivation) correlate positively to playing over 3 days (BF>1000).
- **Age negatively** correlates with playing experience (r=-0.328, BF>1000), playing frequency (r=-0.245, BF>30), and purpose awareness (r=-0.217, BF>10). It also positively correlates with total submarine playing time (r=0.227, BF>10).
- **Playing experience** and **playing frequency** both negatively correlate with total times failing at a submarine level (experience: r=-0.286, BF>100; frequency: r=-0.230, BF>10), and correlate positively with each other (r=0.705, BF≫100k).

5.4 Qualitative Results

Participants could input text freely when asked (1) what their main **motivation** for playing was, (2) why they played **more or less** than the requested 5 minutes, and (3) whether they had any **other comments** about their experience. Although none of the fields were mandatory to complete the survey, all 204 participants answered the first question, 196 answered the second question, and 156 answered the third question. One of the researchers performed an analysis of the collected data, assessing comments in each of the three categories and coding them with recurring themes. This classification was reviewed by another researcher.

The coding protocol consisted of labelling responses specifically mentioning different game aspects (i.e., submarine navigation, coin collecting, ship decoration, story, or the game in general), those related to categories of the GUESS survey (i.e., personal gratification, enjoyment, play engrossment, usability/playability, aesthetics), and those mentioning the purpose of the game (e.g., training memory).

For questions two and three, responses for each of these categories could be either positive or negative (e.g., "I found the coin collecting quite fun" or "The coins were frustrating" were labelled

'coin (+)' and 'coin (-)' respectively). To determine whether a comment related to a GUESS category, the questions of the category were used as a guideline (e.g., comments relating to a sense of achievement from improving are closely aligned with the questions from the Personal Gratification module). Additional themes were identified throughout the coding process and comments were tagged accordingly. These themes include reflections on participation in an experiment or receiving monetary compensation (e.g., a motivation of "taking part in the experiment and earning some cash ;)"), and suggestions to improve the game (e.g., "I would like to use the WASD keys").

Comments could be labelled with multiple themes. For example, "I just liked the challenge of trying to navigate the submarine and collecting the pieces. I played because of enjoyment mainly" was coded as 'coin (+)/submarine (+)/personal gratification (+)/enjoyment (+)', while "It wasn't the most interesting game, so I probably wouldn't play it again, but I enjoyed it for the short time of the study" was coded as 'game (-)/enjoyment (+)/play engrossment (-)'.

The results are summarised in Table 2. For the complete list of coded responses, please refer to the supplementary files. The primary goal in collecting qualitative data for this study was to help contextualise and better understand the qualitative findings. Frequencies of theme occurrences were not used for statistical testing, but to serve as an indication of how often a certain sentiment occurred among the participants.

6 DISCUSSION

Overall, results suggest that the game was well received. GUESS measures were well above the mid-point and most participants said they would play the game again if it had more content. The weakest measure was that of Play Engrossment, relating to feeling 'absorbed' by an activity, which makes sense for this type of 'casual' game. Even so, Play Engrossment was still above the mid-point as well. Participants generally considered themselves casual game players and reported playing games for a few hours each week. Negative correlations between GUESS measures and playing experience, as well as comments made by participants, suggest the game was too simplistic for more advanced players. However, for the majority of less experienced players (i.e., the target audience for this type of design, should it be used for cognitive training) the game was enjoyable.

On average, participants played the game for ≈60 minutes, or ≈48 minutes when controlling for outliers. In both cases, this is well over the 15 minutes requested in the experiment instructions. Players were also generally aware of the game's purpose and motivated by it. Motivation to collect all coins, fixing up the ship, and seeing the story were all around the mid-point, with coin motivation being the highest. Not surprisingly, participants who became more engrossed by the game also enjoyed it more, experienced more gratification from playing, tended to play longer, progress further, and perform better. Older participants tended to be less aware of the game's purpose. They also tended to have less experience in playing games and scored the game lower in Usability. It is possible that, due to having to spend more time and energy understanding and mastering the game, older users had less attention for the game's purpose. However, despite difficulties, there is no evidence to suggest they played less far into the game or enjoyed the experience less.

Interestingly, participants played more on Day 3 of the experiment than on the other days. Given the comments of participants that they wanted to finish the game, get as far as they could, or collect all coins, it could be they played more on this day as the experiment was about to end, (despite having been informed that they could keep playing the game upon completing the final survey). Therefore, it's possible they put in more time to achieve their goals while they felt they still could.

6.1 H1: Effects of Game Mechanics

Statistical evidence suggests that the different game conditions did not meaningfully impact most measures. The only exception is that participants in the Base condition progressed further in the

Question	Theme	Count	Example
Motivation	Experiment	38	"To give good results for the test."
	Reward	23	"Cash reward."
	Coins	59	"It was interesting and I wanted to earn the maximum number of coins."
	Ship	36	"To deck the ship out as much as I could!"
	Personal Gratification	95	"The challenge to collect coins and get to the next stage."
	Enjoyment	31	"For some reason I actually enjoyed it."
	Purpose	5	"Knowing it was to train your memory."
More or Less	Coins (+)	34	"To retry to get the 3 coins."
	Submarine (+)	16	"I wanted to see the evolution of the levels."
	Ship (+)	11	"Because I wanted to finish the ship."
	Story (+)	13	"To find out how it ended."
	Personal Gratification (+)	88	"I wanted to complete all the levels by the end of the three days and replayed a number of stages to collect more coins."
	Enjoyment (+)	62	"It was nice to do this when taking a short break."
	Play Engrossment (+)	20	"Found it fun and lost track of time while while playing."
Other	Game (-)	18	"It was pretty dull and basic and reminded me of old games my kids played decades ago."
	Story (-)	16	"The story didn't really seem to go anywhere or have any relevance to the other elements."
	Enjoyment (+)	48	"I never got into computer games when I was younger and I'm usually bad at them. This game was accessible and quite enjoyable."
	Usability/Playability (-)	22	"All is good - apart from finding the submarine a little difficult to manoeuvre and not very responsive when negotiating tight places!"
	Purpose	9	"Found it really fun and did make me really think about strategy and help memory."

Table 2. A selected overview of identified themes per question.

game. This is not surprising, as the game had no other mechanics to potentially take up time. As such, participants spent the same amount of time playing, but with their attention focused on a single aspect of the gameplay. These findings go against expectations that a more elaborate game experience leads to desired player behaviour (e.g., playing longer) or enjoying the game experience more. **As such, H1 is rejected based on these results.**

Even though the submarine game is fairly simple in terms of design, it received similar scores across GUESS categories on its own as it did with additional mechanics. One explanation for this,

is that the submarine game formed the core of the game in every version – additional mechanics were used to enrich this experience. As evidenced by the widespread use of these mechanics, both in commercial and serious games, collecting items or reaching a perfect score are strong motivators with widespread appeal. The questionnaire results corroborate this interpretation. This mechanic therefore likely takes precedence over the others, and only those with an interest in the other mechanics specifically appreciated these.

Why then, do commercial games layer mechanics? It may be because they aid in long-term engagement, because they add variety and instil wonder on what will happen next. Although the length of this experiment was not enough to test for this, comments by participants do suggest some were motivated by this (e.g., wanting to see the story unfold). Another reason could be that it allows them to stand apart from the crowd, and appeal to different demographics. As stated in Section 2, at their core, many commercial games function the same. Extra mechanics are largely thematic, used to appeal to different groups of players. Players will search for the type of gameplay they want (e.g., a puzzle-type game), then choose one from the vast amount on offer based on graphics and theme (e.g., home make-over, garden renovation, haunted house, farming, animals). While different players may prefer certain mechanics (coin, ship, and story motivation correlated with different aspects of the GUESS in this experiment), they are not the main reason that players engage with the game. This appears to be supported by correlations between coin motivation, the amount of coins collected, and the total playing time — players collected coins primarily for the want to 'collect them all', rather than for what they could do with them.

Another reason that the submarine game did well on its own could be that, as the core mechanic, it was considered the most beneficial aspect of the game (i.e., training/testing the players' memory). Coin motivation was the only game-related motivation that correlated with purpose awareness. This could mean that participants perceived the submarine game to be the most important aspect, and the ship and story only additional.

Interestingly, additional interactivity to the story in the form of choices had very little impact, only showing a difference in one category of the agency questionnaire (Effectance). This difference is understandable, as Effectance measures the amount of impact users feel they have on the activity. However, this difference did not translate in different behaviour or appreciation of the game. Very few people commented on wanting more interaction in the narrative, and negative comments related more to the narrative itself than the lack of choice. This may have been different if choices had had larger effects or were more tied into other aspects of the game. However, it is also possible that interactive narratives are primarily something expected by those who play a lot of game, and not by the more 'casual' players that most of our participants considered themselves to be.

6.2 H2a: Effects of Different Purposes

Similar to the different game versions, no statistical differences were established between different purpose conditions. Based on previous research, it was anticipated that a more specific purpose would enhance participants' perception and appreciation of the serious game, potentially translating into measurable behaviour. However, the data does not suggest this to be the case. As it stands, **H2a is rejected**.

There are a number of possible reasons why no differences were found. First of all, all participants knew they were participating in a research study and were recruited from a platform with which they were acquainted. It can be presumed that people who voluntarily participate in such research studies do so because they see a certain value in them, either to assist research or because of monetary compensation. This, in itself, most likely already establishes a certain perception of value. Merely adding additional messages to the game to contextualise that base value was perhaps not impactful enough to change it. Second of all, it could be that the different purposes were simply too

similar, indicating a limitation of the study design. The Other purpose could be perceived as similar to the None purpose, as it only provides a clearer motivation for why data was being gathered. Data gathering is, however, quite standard in a remote research study, and it is possible this small amount of purpose integration did not affect participants' behaviour. The Self condition differed most from the other two, but also did not lead to measurable changes in behaviour. This could be related to the final reason why no effects were witnessed, namely that *Pocket Odyssey* was not tested for actual cognitive effects. Moreover, even if cognitive effects had been established, participants only played the game for 3 days. This is not long enough to truly experience cognitive improvement, even in the most successful serious game aimed at training it. Therefore, even though participants were generally aware of the game's purpose, their perception of it was likely not altered as they did not experience the benefits for themselves.

6.3 H2b: Effects of Purpose Awareness and Perceived Value

Although no differences were established between different purposes, participants who were aware of and motivated by the game's purpose, did consider the game more beneficial and useful. They also scored higher in GUESS categories. It can be argued that participants aware of the purpose were more forgiving of issues with usability, were more engrossed in the game, enjoyed it more, and experienced more gratification. Similarly, players who considered the game beneficial and useful had a more positive experience of the game. Participants who were motivated by the purpose or considered the game beneficial also progressed further in it, reaching a higher maximum level, and experienced the experiment ending more negatively. Generally, those players that were more aware of the purpose also felt more motivated by it. These findings are in line with work presented in Section 2 and, based on this data, **H2b is accepted**.

7 LIMITATIONS

The presented findings should be evaluated within the scope and limitations of the study. One aspect of the research design is that *Pocket Odyssey* was not validated for its intended use case of providing cognitive training. Although the game was considered useful and entertaining by the participants, having actual (confirmed) benefits could have impacted the findings. However, it is unlikely that the study time frame of 3 days was enough for players to notice any actual benefits, even if the game had been proven to provide them. Some users commented they had the feeling that playing the game was beneficial to them, while a few others questioned this. However, the participants that did, seemed to be questioning the supplemental mechanics more than the submarine. It is possible that the memory aspect was considered beneficial by users without further explanation. It is less likely that they would consider the submarine navigation to be beneficial without being told about it, as it is probably less directly associated with cognitive skills than memory. Although participants were generally aware of the purpose, if the game emphasised how its various aspects were designed to aid in training cognition (rather than only stating it could) it may affect the results. A follow-up study should examine the game's effects, but additional content to the game would be required (e.g., by making levels procedurally generated) so that it can be tested over a period of time long enough to measure cognitive effects.

In addition to the game's cognitive effects, the mechanics of the game were also not evaluated in terms of their relative quality to one another. Although participants were generally favourable towards the game, some comments suggest that primarily the story was lacking in quality. Some others mentioned that the integration between the story and the submarine gameplay was not meaningful. This is offset by comments from other users who found the story compelling and considered it a motivation. Results could have been impacted by the simplistic presentation of the story (text only, versus more 'animated' presentation modes common in casual games) and

the quality of the story itself. Performing the presented study helped to identify these issues, and thus shows the usefulness of assessing aspects of serious games over evaluating them as a whole. Regarding the presented results, however, and what they mean for future research, it is possible a 'better' story would have impacted the findings on different game purposes.

Purpose awareness, perceived value, and motivation were self-reported post-experiment using questions on a Likert scale. Considering the correlations found in this data, purpose awareness, and perceived value would ideally be explored further using additional methods (e.g., by examining where the game draws the player's attention and how they respond to it through biometric measures). In addition to this, game preferences were collected through a single question with a free-text response option. The data gathered in this way was difficult to meaningfully process, as participants had widely different ways of answering the question. Some answered with specific titles, others with game genres, or with even broader descriptions (e.g., 'basic games on the internet'). There is no single, unified understanding of game genres [12] and it was decided that attempting to code the gathered data with genres (in order to examine player's game preferences) was at risk of too much misinterpretation. Instead, the data gathered through Likert scale questions (i.e., amount of time spent playing and self-assessment of player experience) were deemed more reliable. The decision to condense previously played games to a single question was done to lessen the load on participants, who were already asked to invest time over the course of multiple days. However, future studies should consider including additional questions with the goal of establishing player preferences (e.g., following Tondello et al.'s work on player traits [54]).

This study investigated two variables: game version and game purpose. As such, it would have been possible to run a 4×3 between-subjects study with 12 condition groups. Given the exploratory nature of the study, and based on a cost/return on investment analysis, it was decided to first collect data from the 6 selected groups previously described. Results were assessed by treating the variables as separate, only running tests between groups that had a single variable change between them. Then, when results indicated groups were largely different, correlations were run to explore possible points of interest and worthwhile relationships to explore in further research. It was decided to halt data collection at this point, rather than spend resources in running the additional groups, and evaluate what could be learned from the gathered data. Although the results do not seem to indicate this, it is possible that the purpose of the game would have been perceived differently in, for example, the Base condition with less compounding variables. A future study could include the remaining groups to test for interaction effects and other possible findings not explored in the current approach, or to further contextualise the presented findings.

Ideally, sample sizes are derived from the statistical tests one plans to perform and based on previous work to calculate the required numbers. The sample size of this study was based on local standards in previously published CHI papers. Compared to similar study designs, the sample size in this study is adequate and the reported results indicate that they would not fundamentally change with additional testing. However, future studies should ideally use previous work (this article included) to make more precise calculations of appropriate sample sizes. Similarly, the Bayesian priors used in this study are uninformed and based on default settings. To compensate for this, a conservative approach was taken in interpreting the results, presenting them as interesting directions for further research rather than solid truths. Informed priors would have improved the study, and this study will be able to provide a basis for others to follow.

Finally, the experiment lasted for 3 days. As such, it is not possible to say how the tested mechanics (which, in a commercial setting, generally would be used to foster long-term engagement) would have impacted player behaviour if participants had played the game over a longer time. It is possible that, in a longer study, participants in the Base condition would stop playing earlier for a lack of variety, while those with supplemental mechanics would continue. However, the data does

not seem to indicate this at present, as the most common motivator commented on by participants was collecting coins and finishing levels.

8 IMPLICATIONS

Surprisingly, the addition of noticeable game mechanics made very little difference to the participants' measured behaviour, game experience and perception. Does it make sense then, for developers of serious games, to add mechanics to supplement a core gameplay loop that engages players in the intended behaviour? Nothing in the data suggests that adding mechanics improves players' game experience or alters their behaviour. If anything, it seems that additional mechanics can, in fact, *detract* players from progressing in the aspect of the game that is supposed to be beneficial when they play for the same amount of time. Considering mechanics are time-consuming to develop and create, there is little reason to include them based on the results of this study. While there are many other possible mechanics that were not tested here, the data suggests that creating a well-rounded game experience around a single core mechanic could be the preferred approach.

However, examining the data also gives an indication of how additional mechanics might meaningfully contribute to a serious game's design. Instead of being used to increase the entertainment value of the serious game, additional mechanics can instead enforce its purpose. If additional mechanics are added to the game's core design with this intent, informing the player of the game's benefits or purpose, this in turn can influence their behaviour and game experience. In the case of *Pocket Odyssey*, for example, an aesthetic theme and narrative relating to people experiencing memory issues (rather than a sea-faring theme) could prove more effective in enforcing the purpose of the game. There are, however, many ways such mechanics could be implemented. This experiment does not form enough of a basis to formulate guidelines to this effect, nor is the data gathered conclusive in supporting this theory. Connecting the findings of this study (and, ideally, the ones that will follow it), to existing design methods would be a worthwhile endeavour.

Pocket Odyssey was created with intentions of providing a beneficial task (related to memory and navigation) in the form of a game that would motivate users to engage in this task. As discussed before, it is not yet proven that it can actually provide this effect but, assuming that it can, its measure of success is in whether people played it, how much they played it, and whether they would continue to play it. In this way, at least for the duration of the study, the game is a success, and an evaluation of the game as it was meant to be used (i.e., the Narcho_Self condition) would have shown it to be so. Some participants could have commented on the supplemental mechanics not being particularly meaningful to them or the quality of the narrative but, overall, the game design would have been presented successful. Based on this, the developers might suggest that including the tested mechanics is recommended for other, similar projects. While it is not necessarily wrong to do so, the findings in this paper would have gone undetected had that been the only test carried out and such suggestions would not present the full picture.

Adding choices to a narrative, something quite expected with the interactive nature of games, is a time-consuming process. Writing engaging narratives is a difficult task. Creating artwork for collectables (e.g., decorating choices) requires resources. While all these aspects were created with the best intentions, they did little to (1) improve engagement and (2) to emphasise the game's purpose. Many aspects of *Pocket Odyssey* can be critiqued — and the presented study shows clear directions in which it can be improved — but the rationale behind its design is not unique. Though most clearly exemplified in the trend of gamification, where game elements are 'added' to a non-gaming task or context, many serious games are designed around the notion of 'fun' parts to make the 'serious' aspects more engaging. When evaluating serious games as a whole, even if they may individually be successful in meeting their targets, how such 'fun' parts add to or detract from the experience and the game's purpose may go unnoticed.

The main contribution from this paper does not necessarily lie in its individual findings, although the authors argue for continued empirical studies into game mechanics and other elements of a game's design, both to further investigate what was found here and what is yet to be discovered. The primary takeaway, however, should be for readers to reconsider how serious games are evaluated and, possibly, designed. Games are complex systems with many interacting elements — creating and evaluating them is no simple task. During gameplay, a player's attention shifts between various aspects of the game's design. This process is difficult enough to capture and understand in entertainment games, but serious games add another layer of complexity — namely, whether those aspects actually help the game achieve its purpose. While some aspects can simply be for 'fun' or to add to the game's aesthetics, elements that take up significant portions of time (and attention) should ideally be designed to contribute to the serious game's overarching goals. Where possible, such elements should be evaluated for their contribution, or at least for how they interact with the rest of the game. Developers and researchers should consider on a per project basis which aspects of a game are in need of closer examination, as the presented methodology spreads participants over multiple conditions, increasing time and resources required for attaining usable results. However, what this article shows, is that the effort may be worthwhile.

9 CONCLUSION

This study investigated the effects of game mechanics found in commercial casual games when implemented in a serious game for cognitive training. Additionally, it examined the effects of different serious game purposes and awareness of said purpose on perceived value, player behaviour, and game experience.

While it may intuitively seem that games require a certain diversity and complexity to their mechanics to be engaging, statistical evidence suggests that adding supplemental game mechanics does not necessarily impact player behaviour or game experience. This, in turn, suggests that serious game designers have some leeway in their decision of how much different content to create, as players are not immediately reacting positively or negatively to this aspect of game development. On the other hand, awareness of the game's purpose improved players' perception of the game. Improved perception, in turn, led to increased game experience and players progressing further in the game. Based on the results of this study, simply providing supplemental game mechanics on their own do not improve engagement with the serious game. However, if additional mechanics are used to enforce the game's purpose, they might.

This study examined a limited number of mechanics and others may provide different results. Similarly, the type of serious game used in this study is one in which players are encouraged to engage with a specific task. The results should therefore not be generalised to all kinds of serious games. In educational games that aim to explain complex subjects, for example, a completely different design approach may be needed. However, the game tested here provides a solid approximation of existing serious games that revolve around encouraging a specific task, for example in healthcare and for training purposes. Future work should focus on repeating this experiment with other game mechanics, and examine how these related to serious game with another purpose.

The primary contribution of this paper is to show that, contrary to serious game validation approaches, it is beneficial to examine specific game mechanics for how they affect player behaviour and the game experience. The effects of mechanics that are evaluated as part of a serious game's entire design, may in fact not add any measurable effect to its success. At worst, they may even detract from it. Secondly, the findings on purpose awareness and perceived value suggest these factors are more important than so far has been discussed. It may be beneficial to control for this effect in validation studies, in which participants are aware they are part of a research study. Additionally, it opens up possibilities for game mechanics to be specifically used to enforce the

game's purpose and context. This, in turn, may actually lead to increased engagement. In addition to these conclusions, this study's results may form the basis for future work due to the presented statistical findings, its data, and methodology. Finally, the game *Pocket Odyssey* may be used in future studies as well.

ACKNOWLEDGMENTS

The authors would like to thank all participants for their time, energy, and dedication.

REFERENCES

[1] G5 Entertainment AB. 2019. Match Town Makeover: Match 3. [iOS].
[2] Andreas Alexiou and Michaéla C Schippers. 2018. Digital game elements, user experience and learning: A conceptual framework. Education and Information Technologies 23, 6 (2018), 2545–2567.
[3] Sylvester Arnab, Theodore Lim, Maira B Carvalho, Francesco Bellotti, Sara De Freitas, Sandy Louchart, Neil Suttie, Riccardo Berta, and Alessandro De Gloria. 2015. Mapping learning and game mechanics for serious games analysis. British Journal of Educational Technology 46, 2 (2015), 391–411.
[4] Atari. 2020. RollerCoaster Tycoon Story. [iOS].
[5] Diego Ávila-Pesántez, Luis A Rivera, and Mayra S Alban. 2017. Approaches for serious game design: A systematic literature review. The ASEE Computers in Education (CoED) Journal 8, 3 (2017).
[6] Kat Bailey. 2020. Ghost of Tsushima Pretty Much Sums Up This Generation. https://www.usgamer.net/articles/ghost-of-tsushima-pretty-much-sums-up-this-generation. Accessed: 2021-02-12.
[7] Julia Claudia Binder, Jacqueline Zöllig, Anne Eschen, Susan Mérillat, Christina Röcke, Sarah Schoch, Lutz Jäncke, and Mike Martin. 2015. Multi-domain training in healthy old age: Hotel Plastisse as an iPad-based serious game to systematically compare multi-domain and single-domain training. Frontiers in aging neuroscience 7 (2015), 137.
[8] Jeanne H Brockmyer, Christine M Fox, Kathleen A Curtiss, Evan McBroom, Kimberly M Burkhart, and Jacquelyn N Pidruzny. 2009. The development of the Game Engagement Questionnaire: A measure of engagement in video game-playing. Journal of experimental social psychology 45, 4 (2009), 624–634.
[9] Benjamin Byl, Matthias Sünchsen, and Michael Teistler. 2018. A serious virtual reality game to train spatial cognition for medical ultrasound imaging. In 2018 IEEE 6th International Conference on Serious Games and Applications for Health (SeGAH). IEEE, 1–4.
[10] Kelly Caine. 2016. Local standards for sample size at CHI. In Proceedings of the 2016 CHI conference on human factors in computing systems. 981–992.
[11] Gordon Calleja. 2011. In-game: From immersion to incorporation. mit Press.
[12] Rachel Ivy Clarke, Jin Ha Lee, and Neils Clark. 2017. Why video game genres fail: A classificatory analysis. Games and Culture 12, 5 (2017), 445–465.
[13] Thomas M. Connolly, Elizabeth A. Boyle, Ewan MacArthur, Thomas Hainey, and James M. Boyle. 2012. A systematic literature review of empirical evidence on computer games and serious games. Computers & Education 59, 2 (2012), 661–686. https://doi.org/10.1016/j.compedu.2012.03.004
[14] Ann DeSmet, Dimitri Van Ryckeghem, Sofie Compernolle, Tom Baranowski, Debbe Thompson, Geert Crombez, Karolien Poels, Wendy Van Lippevelde, Sara Bastiaensens, Katrien Van Cleemput, et al. 2014. A meta-analysis of serious digital games for healthy lifestyle promotion. Preventive medicine 69 (2014), 95–107.
[15] Katharina Emmerich and Mareike Bockholt. 2016. Serious games evaluation: processes, models, and concepts. In Entertainment Computing and Serious Games. Springer, 265–283.
[16] Bruno Ferreira and Paulo Menezes. 2020. An Adaptive Virtual Reality-Based Serious Game for Therapeutic Rehabilitation. (2020).
[17] PopCap Games. 2001. Bejeweled. [Windows].
[18] Andrew Gelman, Jennifer Hill, and Masanao Yajima. 2012. Why we (usually) don't have to worry about multiple comparisons. Journal of Research on Educational Effectiveness 5, 2 (2012), 189–211.
[19] Kathrin Maria Gerling, Frank Paul Schulte, Jan Smeddinck, and Maic Masuch. 2012. Game design for older adults: effects of age-related changes on structural elements of digital games. In International Conference on Entertainment Computing. Springer, 235–242.
[20] Borja Gil, Iván Cantador, and Andrzej Marczewski. 2015. Validating gamification mechanics and player types in an e-learning environment. In Design for Teaching and Learning in a Networked World. Springer, 568–572.
[21] Marcello A Gómez-Maureira, Michelle Westerlaken, Dirk P Janssen, Stefano Gualeni, and Licia Calvi. 2014. Improving level design through game user research: A comparison of methodologies. Entertainment Computing 5, 4 (2014), 463–473.

[22] Christian Karl Grund. 2015. How games and game elements facilitate learning and motivation: A literature review. INFORMATIK 2015 (2015).
[23] Juho Hamari and Lauri Keronen. 2017. Why do people play games? A meta-analysis. International Journal of Information Management 37, 3 (2017), 125–141. https://doi.org/10.1016/j.ijinfomgt.2017.01.006
[24] Casper Harteveld. 2011. Triadic game design: Balancing reality, meaning and play. Springer Science & Business Media.
[25] Khe Foon Hew, Biyun Huang, Kai Wah Samuel Chu, and Dickson KW Chiu. 2016. Engaging Asian students through game mechanics: Findings from two experiment studies. Computers & Education 92 (2016), 221–236.
[26] Geoffrey Hookham and Keith Nesbitt. 2019. A systematic review of the definition and measurement of engagement in serious games. In Proceedings of the Australasian Computer Science Week Multiconference. 1–10.
[27] JASP Team. 2020. JASP (Version 0.14.1)[Computer software]. https://jasp-stats.org/
[28] Harold Jeffreys. 1961. Theory of probability (3rd ed.). Oxford University Press, Oxford.
[29] Fares Kayali, Naemi Luckner, Peter Purgathofer, Katta Spiel, and Geraldine Fitzpatrick. 2018. Design considerations towards long-term engagement in games for health. In Proceedings of the 13th international conference on the foundations of digital games. 1–8.
[30] Isabelle Kniestedt, Stephan Lukosch, and Frances Brazier. 2018. User-centered design of an online mobile game suite to affect well-being of older adults. In International Conference on Entertainment Computing. Springer, 355–361.
[31] Petros Lameras, Sylvester Arnab, Ian Dunwell, Craig Stewart, Samantha Clarke, and Panagiotis Petridis. 2017. Essential features of serious games design in higher education: Linking learning attributes to game mechanics. British journal of educational technology 48, 4 (2017), 972–994.
[32] Effie L-C Law, Florian Brühlmann, and Elisa D Mekler. 2018. Systematic review and validation of the game experience questionnaire (geq)-implications for citation and reporting practice. In Proceedings of the 2018 Annual Symposium on Computer-Human Interaction in Play. 257–270.
[33] Andreas Lieberoth. 2015. Shallow gamification: Testing psychological effects of framing an activity as a game. Games and Culture 10, 3 (2015), 229–248.
[34] Minhua Ma, Andreas Oikonomou, and Lakhmi C Jain. 2011. Serious games and edutainment applications. Vol. 504. Springer.
[35] Christopher Madge, Richard Bartle, Jon Chamberlain, Udo Kruschwitz, and Massimo Poesio. 2019. Incremental game mechanics applied to text annotation. In Proceedings of the Annual Symposium on Computer-Human Interaction in Play. 545–558.
[36] Maarten Marsman and Eric-Jan Wagenmakers. 2017. Bayesian benefits with JASP. European Journal of Developmental Psychology 14, 5 (2017), 545–555.
[37] Anjum Matin, Mardel Maduro, Rogerio de Leon Pereira, and Olivier Tremblay-Savard. 2020. Effect of Timer, Top Score and Leaderboard on Performance and Motivation in a Human Computing Game. In International Conference on the Foundations of Digital Games. 1–10.
[38] Niamh A Merriman, Eugenie Roudaia, Matteo Romagnoli, Ivan Orvieto, and Fiona N Newell. 2018. Acceptability of a custom-designed game, CityQuest, aimed at improving balance confidence and spatial cognition in fall-prone and healthy older adults. Behaviour & Information Technology 37, 6 (2018), 538–557.
[39] David R Michael and Sandra L Chen. 2005. Serious games: Games that educate, train, and inform. Muska & Lipman/Premier-Trade.
[40] Lennart E Nacke and Christoph Sebastian Deterding. 2017. The maturing of gamification research. Computers in Human Behaviour (2017), 450–454.
[41] HH Nap, YAW De Kort, and WA IJsselsteijn. 2009. Senior gamers: preferences, motivations and needs. Gerontechnology 8, 4 (2009), 247–262.
[42] Heather L O'Brien and Elaine G Toms. 2008. What is user engagement? A conceptual framework for defining user engagement with technology. Journal of the American society for Information Science and Technology 59, 6 (2008), 938–955.
[43] Anthony O'Hagan. 2008. The Bayesian approach to statistics. Handbook of probability: Theory and applications (2008), 85–100.
[44] The pandas development team. 2020. pandas-dev/pandas: Pandas. https://doi.org/10.5281/zenodo.3509134
[45] Avinash Parnandi and Ricardo Gutierrez-Osuna. 2015. A comparative study of game mechanics and control laws for an adaptive physiological game. Journal on Multimodal User Interfaces 9, 1 (2015), 31–42.
[46] Mikki H Phan, Joseph R Keebler, and Barbara S Chaparro. 2016. The development and validation of the game user experience satisfaction scale (GUESS). Human factors 58, 8 (2016), 1217–1247.
[47] Playrix. 2016. Gardenscapes. [iOS].
[48] Christian Roth. 2016. Experiencing interactive storytelling. Ph.D. Dissertation.

[49] Jeffrey N Rouder, Richard D Morey, Paul L Speckman, and Jordan M Province. 2012. Default Bayes factors for ANOVA designs. Journal of Mathematical Psychology 56, 5 (2012), 356–374.
[50] Jesse Schell. 2008. The Art of Game Design: A book of lenses. CRC press.
[51] Miguel Sicart. 2008. Defining game mechanics. Game Studies 8, 2 (2008).
[52] Arvid Sjölander and Stijn Vansteelandt. 2019. Frequentist versus Bayesian approaches to multiple testing. European Journal of Epidemiology 34, 9 (May 2019), 809–821. https://doi.org/10.1007/s10654-019-00517-2
[53] Sharon T Steinemann, Elisa D Mekler, and Klaus Opwis. 2015. Increasing donating behavior through a game for change: The role of interactivity and appreciation. In Proceedings of the 2015 annual symposium on computer-human interaction in play. 319–329.
[54] Gustavo F Tondello, Karina Arrambide, Giovanni Ribeiro, Andrew Jian-lan Cen, and Lennart E Nacke. 2019. "I don't fit into a single type": A Trait Model and Scale of Game Playing Preferences. In IFIP Conference on Human-Computer Interaction. Springer, 375–395.
[55] Vanessa Vallejo, Patric Wyss, Luca Rampa, Andrei V Mitache, René M Müri, Urs P Mosimann, and Tobias Nef. 2017. Evaluation of a novel Serious Game based assessment tool for patients with Alzheimer's disease. PLoS One 12, 5 (2017), e0175999.
[56] Katinka van der Kooij, Evert Hoogendoorn, Renske Spijkerman, and VT Visch. 2015. Validation of games for behavioral change: connecting the playful and serious. International Journal of Serious Games 2, 3 (2015), 63–75.
[57] Johnny Van Doorn, Alexander Ly, Maarten Marsman, and Eric-Jan Wagenmakers. 2018. Bayesian inference for Kendall's rank correlation coefficient. The American Statistician 72, 4 (2018), 303–308.
[58] Juan A Vargas, Lilia García-Mundo, Marcela Genero, and Mario Piattini. 2014. A systematic mapping study on serious game quality. In Proceedings of the 18th International Conference on Evaluation and Assessment in Software Engineering. 1–10.
[59] Georgios N Yannakakis and Ana Paiva. 2014. Emotion in games. Handbook on affective computing 2014 (2014), 459–471.
[60] Amri Yusoff, Richard Crowder, and Lester Gilbert. 2010. Validation of serious games attributes using the technology acceptance model. In 2010 Second International Conference on Games and Virtual Worlds for Serious Applications. IEEE, 45–51.

Received February 2021 ; revised June 2021 ; accepted July 2021