Procedural versus Human Level Generation: Two sides of the same coin?

Luiz Rodrigues[a,1,*], Robson Bonidia[b,1], Jacques Brancher[a]

[a] *Department of Computer Science, Londrina State University, Rodovia Celso Garcia Cid - Pr 445 Km 380, Londrina - PR, Brazil*
[b] *Bioinformatics Graduate Program, Federal University of Technology, Av. Alberto Carazzai, 1640, Cornélio Procópio - PR, Brazil*

Abstract

Game development often requires a multidisciplinary team, demands substantial time and budget, and leads to a limited number of game contents (e.g., levels). Procedural Content Generation (PCG) can remedy some of these problems, aiding with the automatic creation of content such as levels and graphics, in both the development and playing time. However, little research has been performed in terms of how PCG influences players, especially on Digital Math Games (DMG). This article addresses this problem by investigating the interactions of players with a DMG that uses PCG, investigating the hypothesis that interacting with this intervention can provide experiences as good as human-designed content. To accomplish this goal, an A/B test was performed wherein the only difference was that one version (*static*, N = 242) had human-designed levels, whereas the other (*dynamic*, N = 265) provided procedurally generated levels. To validate the approach, a two-sample experiment was designed in which each sample played a single version and, thereafter, self-reported their experiences through questionnaires. We contribute by showing how the participants' interactions with a DMG are reported in terms of (1) fun, (2) willingness to play the game again, and (3) curiosity, in addition to how they (4) describe their experiences. Our findings show that samples' experiences did not significantly differ on the four metrics, but did differ on in-game performance. We discuss possible factors that might have influenced players' experiences, in terms of the participants' performances and their demographic attributes, and how our findings contribute to human interaction with computers.

Keywords: Procedural Content Generation, Player Experience, User Experience, A/B test, Digital Math Game, Serious Game

*Corresponding author
Email address: luiz_rodrigues17@hotmail.com (Luiz Rodrigues)
[1] Present address: Institute of Mathematics and Computer Science, University of São Paulo, Av. Trab. São Carlense, 400, São Carlos - SP, Brazil

1. Introduction

Many students perceive math as difficult, do not like it and consider the subject to be displeasing (Biswas et al., 2001). Digital Math Games (DMG) might be used to remedy this problem, improving students' math learning (McLaren et al., 2017), while increasing their positive attitudes toward the subject (Ke, 2008), and it can even reduce the users' anxiety while increasing their engagement (Kiili and Ketamo, 2017). Additionally, rather than conventional paper and pencil exercises, computer-based practice is preferred by them (Yurdabakan and Uzunkavak, 2012), which can also aid in solving the mentioned problems. Based on this context, the usage of this game's type is fundamental, which is demonstrated by the significant attention that DMG have been receiving by the academia (Kiili et al., 2018; de Carvalho et al., 2016; Ibarra et al., 2016; Cheng et al., 2015). However, even for general purpose games, its development is still a slow and costly task, which commonly requires several designers, artists, and developers (Amato and Moscato, 2017; Hendrikx et al., 2013).

An alternative that might tackle these problems is Procedural Content Generation (PCG) (Hendrikx et al., 2013; Carli et al., 2011; Togelius et al., 2011a). It has shown to be a reliable tool that can provide diversified, automatically generated outputs, which can be controlled through generation parameters (Horn et al., 2014), and it has great potential for educational games (Hooshyar et al., 2018; Horn et al., 2016). It has been mainly used in games to automate, aid in creativity and speed up the creation of various types of content (Korn et al., 2017; Moghadam and Rafsanjani, 2017; Smith and Whitehead, 2010) such as vegetation, rivers, terrains, networks, scenarios, levels, non-player characters behavior, and control games' difficulty level (Hendrikx et al., 2013). Furthermore, it is a powerful technique to tackle another problem that is faced in the aforementioned context: the fact that technologies must provide positive experiences; otherwise, it is unlikely that players will interact or accept it, especially children (Bauckhage et al., 2012; Sim and Horton, 2012). To tackle this problem, PCG might be used as a way to constantly provide players with new, unseen content and therefore promote positive outcomes (Korn et al., 2017; Rodrigues et al., 2017; Horn et al., 2014; Togelius et al., 2011b). This context demonstrates the value of this technique to enhance game development, as well as how it can increase the amount of content available to a game without over-charging developers. However, what is the real impact of PCG on players has received little attention from the academic community (Korn et al., 2017).

Therefore, this work will expand on the literature through the investigation of this gap. Within the DMG context, this research will investigate the influences of procedurally generated levels on players' feelings, using an A/B test. Thereby, according to the baseline of a game version that contains human-authored content, we demonstrate how procedural level generation influences players in a DMG. Thus, we contribute by presenting an empirical analysis of the effects that a computational intervention (PCG), which improves game development, has on users' perception of their interaction with a game, demonstrating how their experience is expressed in terms of psychological aspects.

Thus, this research is valuable to professionals who want to employ similar interventions, showcasing how it impacts users' perceptions.

Hereafter, we refer to the game version using human-designed levels as the *static* version, and the other, which uses procedurally generated levels, as the *dynamic* version. Additionally, in the scope of this work, we consider playing a game level to be gameplay, while playing a set of levels is considered to be a game session. Hence, each level finished by a player (winning or losing it) originates gameplay. Considering this context and based on this research's goal, when comparing the experiences of players of procedurally generated versus human-designed levels, we assume the following:

Hypothesis 1 (H1): Players' fun levels do not differ.

Hypothesis 2 (H2): Players' willingness to play the game again - *returnance* - does not differ.

Hypothesis 3 (H3): Players' curiosity levels do not differ.

Hypothesis 4 (H4): Players' descriptions of their experiences do not differ.

The remainder of this article presents background on PCG in Section 2.1, related work in Section 2.2, the research method in Section 3, analysis and results in Section 4, a discussion of its findings in Section 5, and our final considerations in Section 6.

2. Related Work

First, this section introduces a brief background of what is PCG, what is content in the context of a game, a taxonomy of terms used to specify a PCG system, and the methods used to evaluate these systems. Then, it reviews research studies that performed A/B comparisons of procedurally generated content versus human-designed content in terms of players' perspectives, which is the main concern of this article.

2.1. What is PCG?

PCG (Togelius et al., 2011a) is the algorithmic creation of outputs that are good enough according to some criterion of the context in which they will be used (Togelius et al., 2012). This definition might be clarified with the distinction between *necessary* and *optional* content/output presented in Togelius et al. (2011b). The authors argue that both terms are highly dependent on the applied context. Every *necessary* content must be correct, providing the minimal requirements to accomplish the context's goal. In contrast, *optional* is content that the player might avoid and is allowed to be unusable and/or unreasonable. For example, in the generation of a maze, the minimal requirement to progress in it, and therefore, a *necessary* content, is the existence of at least one path through the initial point to the exit. On the other hand, *optional* content might be the insertion of enemies or resources that aid the maze's traverse, in which they could be faced/used or not during the gameplay. In a role-playing game, we consider the creation of its mission (sequence of actions). A minimal requirement to a player's progress might be the possibility of collecting the stage's key

to complete this mission. If it is available when needed, then the content should be considered to be *necessary*. In contrast, *optional* content would be the creation of a weapon that does nothing, or hazards that are intended to make the player's path difficult. Therefore, the generated content in these examples, regardless of whether they are *necessary* or *optional*, should be considered to be good enough. This aspect is true because they are in agreement with their context of application and objective. Those examples were based on games, which is the main application of PCG, wherein it has been used to create different types of content (Hendrikx et al., 2013).

2.1.1. Game Contents

Content is a key term when PCG is discussed, given that it might be used to refer to multiple elements contained in a game. Examples are levels (Shaker et al., 2012) (e.g., stages from the Super Mario Bros game); race tracks (Cardamone et al., 2011); missions from a role-playing game, which might be a sequence of objectives to be accomplished (Karavolos et al., 2015); game progression, which might be viewed as the sequence of game levels presented to the player (Butler et al., 2015); and music (Scirea et al., 2014).

This article approaches the use of PCG to generate game scenarios, which according to Hendrikx et al. (2013) is defined as follows: describe how and in which order the game events will occur. Puzzles, storyboards, story and the concept of a level (playable game space where the player seeks some objectives) are examples of game scenarios. This is one of the most popular types of content used in PCG for games and nearly all genres can benefit from its usage (Hendrikx et al., 2013). Consequently, it was selected to be the matter of research in this study.

2.1.2. Taxonomy

The way that a generator is used, what type of content it produces and which type of interaction it requires by the designer/developer might be defined according to the notions presented in Togelius et al. (2011b) and Carli et al. (2011). Although the taxonomy in Togelius et al. (2011b) was originally designed for search-based PCG, it is also suitable for almost any type of PCG. Next, key definitions are presented.

- As previously mentioned, a *necessary* content must always be correct and is required for the completion of a level. An *optional* content could be avoided or discarded and might be incorrect. Depending on which of these fits the content under generation, whether it is good enough will differ.

- Content might be created with a *constructive* algorithm, wherein its results are obtained in a single sequence of steps. Thus, the method is required to guarantee that its outputs will be at least good enough during their construction. In contrast, a *Generate-and-Test* (GaT) method features two phases: generating and testing, as the name suggests. They

are commonly put together in a loop until some generated instance passes the testing criterion, which is dependent on the application context.

- It is possible for a PCG method to be *adaptive* if it considers players' behaviors and/or profiles to create the outputs. This approach is unusual in commercial games; most of them use a *generic* method that does not take players into account.

- *Online* generation is the case when the content is created at runtime, while the game is running, as stated in (Togelius et al., 2011b). It enables the adaptation of outputs and the creation of endless gameplay, but it requires speed, predictable runtime and, often, predictable quality. In contrast, *offline* generation is when the content's creation is accomplished before the game starts or on the game's development. It can be used to aid designers in creativity or permit the use of methods that are infeasible for real-time execution.

- If a generator receives the same set of parameters and creates the same output, it is *deterministic*. In contrast, if the method is *stochastic*, this guarantee is inexistent. While one provides reproducible results, the other can be used to achieve diversity.

- An *assisted* technique requires significant human intervention during its setup. In contrast, if a simple interaction such as just setting up a few parameters is needed, the technique is considered to be *non-assisted*.

  In this article, we used a PCG system that creates *necessary* content *online*, in a *generic* way, using a *constructive* method that is *non-assisted* and provides *stochastic* outputs.

2.2. How is PCG used and evaluated in Games?

Fundamentally, there are two perspectives that might be adopted for PCG evaluation. One is focused on the algorithm's capabilities, which is commonly performed through the analysis of the *expressive range*, the perspective on which most research relies (Moghadam and Rafsanjani, 2017; Valls-Vargas et al., 2017; Dahlskog et al., 2014; Horn et al., 2014; Linden et al., 2013). In contrast, the others are concerned with how player *experience* the algorithm's outputs, and their impact, which must be captured through the players' interactions with the content. Next, both are briefly presented.

The analysis of an algorithm's expressive range might be summarized in three main steps. First, the evaluation metrics must be defined, which might be a level's linearity or a difficulty score given by an artificial agent. These will be used to assess a large set of content (e.g., 10000 levels) generated through the algorithm under evaluation. Finally, the assessment results should be analyzed through plots, such as heat maps and histograms, to visualize which is the generator's expressive range according to the selected metrics (Smith and Whitehead, 2010). While the aforementioned approach is reliable for investigating how well a method is according to computational metrics, it is insufficient for replacing

user-based studies (Mariño et al., 2015). This leaves a need for the second perspective of PCG evaluation, investigating how the content is experienced based on studies with real users. Metrics (or measures) to accomplish this goal might be questionnaires, observational experiments, facial reactions, voice recordings or physiological responses, such as the heartbeat intensity (Yannakakis and Togelius, 2011). Using these measures, it is possible to evaluate the PCG algorithm through its content, according to the Players' Experiences (PX)[2], subjectively and objectively. Although experiments wherein users interact with procedurally generated content capture in which fashion players perceive the content, they do not provide concerns about the PCG's impact. They show how players perceived the generator outputs; however, how their perception would be if that content was human-authored remains unknown. To actually identify PCG's impact on players, using the same game with and without the generator, in an A/B test fashion, is the most feasible procedure (Korn et al., 2017).

In sum, PCG might be evaluated through its content's expressive range, focusing on the algorithm's capabilities, through user-based studies, investigating PX according to their interaction with the application using it or through A/B comparisons to identify the PCG's impact. Selecting the best approach will depend on what are the study's goals and what questions are expected to be answered. Given the goal of this article, we employed the A/B approach.

2.3. How does Player Experience compare between PCG and Human Design?

To the best of our knowledge, there are three studies that evaluated the impact of PCG on PX, which were found through a non-systematic process based on searching popular scientific databases (e.g. Scopus and Google Scholar) and snow-balling. Next, they are surveyed and compared. Then, literature gaps and contributions of this article are highlighted.

Butler et al. (2015) introduced a game progress system and evaluated it through a serious game according to in-game measures. They addressed an A/B research methodology based on a two-sample analysis, comparing both the time and number of levels played according to data from 2377 players. A DMG, which approaches fractions as the serious subject, Refraction, was used as the testbed. As it was online and collected data from an uncontrolled environment, players' demographics were not available due to the nature of the environment wherein their testbed was hosted. The analyzed PCG system was responsible for creating game levels, which mainly focused on creating a progression based on the solutions of the math problems (fractions). In their results, the authors found a small significant difference in the number of played levels. With respect to the total time that each version was played, the difference was insignificant in spite of the sample size. However, the PCG-based approach was played approximately 92% of the time compared with the human-designed, which shows

[2]In the scope of this article, we refer to PX as how players' interaction with the game are experienced, following Yannakakis et al. (2013).

that their solution is capable of engaging players for similar amounts of time in comparison to the human-authored approach.

Connor et al. (2017) performed similar research, investigating PCG impacts according to players' self-reports through an abstract game. The abstract game was selected with the aim of mitigating any bias that the game design could insert into the A/B test. A player immersion questionnaire was the measure to capture the self-reports, allowing them to capture players' explicit opinions about a game version, where the two-sample design was also adopted. Twenty players participated in this research, where both samples were composed of adults, between 18 and 35 years, that completed the questionnaire (30 questions) after playing one of the two versions only. They used a generator that created levels independently, using a *generate-and-test* method, unlike the approach of Butler et al. (2015), which was focused on the game progression, and they used a constructive method. In terms of their results, Connor et al. (2017) found a significant difference between these versions in favor of the human-designed content when considering their total immersion. However, when analyzing each questionnaire's answer, the authors found that this difference was significant only in less than 17% (5/30) of the questions.

Last, Korn et al. (2017) evaluated the use of a procedural generation system, based on the players' self-reports, through a documentary game. Their generation system was in charge of creating the game's reefs, an optional game element, unlike the aforementioned studies, which generated the necessary content (i.e., levels). They also used a constructive method for the generation, comparing it to a human-designed approach, according to the feedback of 41 subjects. These subjects were adults, who played both game versions (10-15 minutes each) and responded to a questionnaire after playing each one, in contrast to previous approaches. Players indicated their experiences based on the visual aspects and preference for one version or another, favoring the automatically generated content in both perspectives. In addition, they found that older players were more likely to favor the procedurally generated reefs. Their findings show that PCG can provide games with more than money-saving, impacting the PX and that a game environment's change is an advantage.

A summary and comparison of related works are shown in Table 1. From the table, we summarize the main limitations of this field as follows. Generally, little research has been performed in terms of A/B comparison of PCG systems versus human-designed content. Nevertheless, when the problem of using a DMG was approached, the evaluation did not address any aspect related to the math subject, or to players' affects explicitly, and did not provide evidence regarding whether or not the sample characteristics were related to the subject. Furthermore, relatively small samples were used when performing analysis that captured a player's explicit opinions regarding their experiences. This article will cover these aspects and expand the literature by: i) using a game focused on arithmetic operations; ii) adopting measures are related to encouraging the educational subject training (i.e., curiosity) (Wouters et al., 2011) and to the general entertainment purpose of a game (i.e., fun and *returnance* (Read and MacFarlane, 2006)); iii) comparing groups' demographics to identify

whether our sample is related to the testbed's educational subject and whether differences in the sample could represent a threat to our research; iv) gathering a significant amount of data compared to related work and; v) using a PCG system setup similar to related works. Thus, whereas we tackle the literature gaps, we still adopt a similar approach for the purpose of advancing this field of research with similar studies.

Table 1: Related works comparison. Connor et al. (2017) used a non-educational game, therefore, educational subject is not applicable. Korn et al. (2017) employed a within-subject design, therefore, comparing groups' demographics is not applicable.

Attribute	Butler et al.	Connor et al.	Korn et al.
Year	2015	2017	2017
Educational Subject	Fractions	N.A.	History
Measures	Time and number of levels played	Immersion	Contents' visual aspects
Compared groups' demographics?	No	Yes	N.A.
Sample N	2377	20	41
Generate content	Levels	Levels	Reefs
Generated content is:	Necessary	Necessary	Optional
Generation method	Constructive	Generate-and-test	Constructive

N.A. = Not applicable.

3. Method

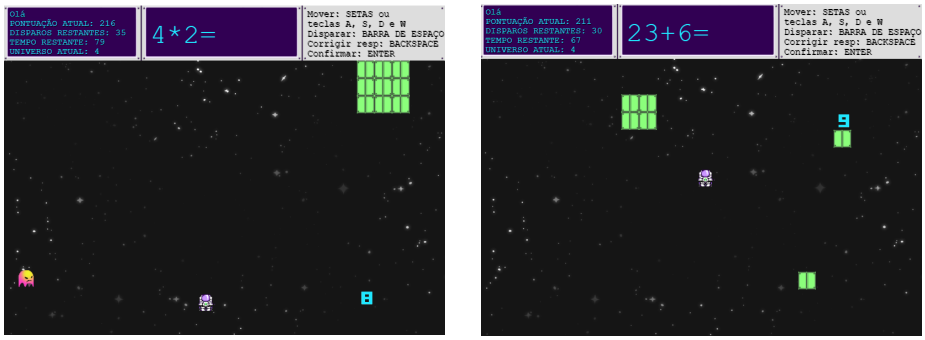
This section describes our method in terms of design, material, participants, measures, procedure, and data analysis.

3.1. Design

We employed a between-subject design with random assignment and two-levels: *experimental*, that is, playing a game version that contained procedurally generated levels (*dynamic* version; experimental group), and *control*, playing the version that features human-designed levels (*static* version; control group). To assess PX, we adopted a postinteraction intervention, considering that this approach was used in most related research (Connor et al., 2017; Butler et al., 2015), as well as the selected design (Korn et al., 2017; Butler et al., 2015).

3.2. Material

The DMG SpaceMath[3] was used as material to perform the study related to this article. It is a game that encourages its players to practice the four basic arithmetic operations, wherein at each level, the players must solve a different math challenge. Correctly solving it leads the player to the next level; otherwise, the player goes back to the beginning. Therefore, the levels represent a key aspect of the game, wherein players must explore them to solve the challenge and, thus, progress into the game. Figure 1 shows two screenshots of the game, which demonstrates two different levels in which the player must explore and solve the math challenge in order to advance. By exploring the level, we mean finding the number hidden below the boxes, and by solving, we refer to collecting the numbers (e.g., Figure 1a bottom-right) that form the math problem's answer (eight and 29 on Figures 1a and 1b, respectively).



(a) **(b)**

Figure 1: Screenshots of two levels of SpaceMath.

Moreover, it was developed for research goals, and hence, it contains two playing modes that differ in a single aspect: the way that its levels were generated, which fulfills the needs of this study. A game developer generated the levels of the *static* version, who graduated in Technology in Digital Games and had three years of experience in the field at the time of designing them. A total of 20 levels were available in this version and were designed to increase the game's difficulty in a linear way. Thereby, as players increase their winning streak, the levels that were designed to be harder are presented. On the other hand, the *dynamic* version features a PCG system that uses a constructive method to create its levels. Following the idea of the *static* version, the *dynamic* version maintains a parameter that is used by the level generation algorithm, allowing it to also increase the levels' difficulty according to a player's winning streak.

Therefore, there are two key differences between these versions. One is that while the *dynamic* version can provide pseudoinfinite levels that are algorithmi-

[3]spacemath.rpbtecnologia.com.br

cally created, the *static* one has a limited set of 20 levels created by a human developer that is experienced in the field. This fact leads to another dissimilarity: whereas players from the *dynamic* version always interact with newly generated content, players from the other version will face already seen levels once they lose and must restart from the first level. The context of the *static* version, which requires the player to replay some content, is common in several games, while constantly providing new content is a feature provided by the use of PCG. However, to what extent one approach or another is better from a player's perspective will be answered by this research.

3.3. Participants

Five hundred and seven players participated in this research ($N_{dynamic}$ = 265, N_{static} = 242). The testbed game is available online and, to perform this research, it was disclosed to specific audiences. We contacted fellow teachers and professors, presented our research's goal and procedure, and asked them to help us by applying the testbed game to collect data to perform our study. Additionally, it was disclosed through some email lists and social network groups, presenting the testbed game and always mentioning that it was part of a research. As a result, data collection was performed in four institutions with the supervision of an institution's supervisor and a researcher, in addition to the players who were achieved through emails and social networks and who agreed to collaborate. Nevertheless, most participants came from institutional applications (over 70%). Note that all players were aware that the game was part of a research study and had the option to not participate from the beginning or to quit at any time. None of the associated institutions required ethics approval for this study.

Due to the two-sample research design, it is necessary to investigate whether these populations differ from each other. This aspect is important to avoid biases from players with different background characteristics since a key goal of this work is to compare their experiences. At the first step, their demographics distribution was assessed independently, through the Shapiro-Wilk test, which provided significant evidence that these attributes do not follow a normal distribution. Therefore, we examined groups' differences through non-parametric hypothesis tests. The usual alpha level of 0.05 was adopted for all hypothesis tests, including the normality assessment.

We compared the numeric attributes (i.e., the age and weekly playing hours) of both groups through the Mann-Whitney U test (hereafter, U test). It showed that in terms of the groups' age, they had an insignificant difference (U = 31860, p = 0.9005), where the average age was 14.1 years (SD = 5.7) for the control and 14.5 (SD = 6.5) for the experimental groups. In terms of the amount of the participants' playing time during a week, the difference between groups was also insignificant (U = 32943, p = 0.5931), with an average of 14.1 hours (SD = 25.6) for the control and 13.3 hours (SD = 26.4) for the experimental groups. In relation to the categorical demographics, Table 2 presents their distributions across classes. This table shows these classes in terms of males (M) and females (F) for gender and, concerning gamers and the ones with internet access through

a computer at home, it shows the number of players that were/had (Y) and were/had not (N). Additionally, it presents the results of the Chi-Squared (χ^2) homogeneity test, which evaluated whether the distribution is the same for both samples, according to each attribute. The significance would be denoted by an asterisk; however, there were no significant differences between groups, as seen in the table.

Table 2: Comparison of samples' categorical attributes.

Attribute	**Control**	**Experimental**	χ^2 **(df)**
Gender	F=79, M=163	F=106, M=159	2.64 (1)
Gamers	N=122, Y=120	N=139, Y=126	0.14 (1)
Has Net	N=27, Y=215	N=23, Y=242	0.62 (1)

* $p < 0.05$

Therefore, we can conclude that there is no significant difference between the samples under analysis in terms of the attributes evaluated. Hence, it is expected that the subjects' background characteristics will not bias this research's results. Furthermore, we highlight that we consider the DMG of this article feasible to the approximately 14 years old sample. This is because basic arithmetic operations is a math fundamental topic that provides background for almost all other topics, such as quadratic equations, functions, calculus, and so on. Thereby, the game is suitable for players of varied ages.

3.4. Measures

Four factors were captured to measure PX: experienced fun; *returnance*; curiosity; and experience description. This selection was based on previous studies that already used and validated these factors. Two of them are based on the widely used Fun-Toolkit (Read and Macfarlane, 2002) and the others are inspired by its use for rapid assessment (Moser et al., 2012). Hence, our questionnaire not only evaluates enjoyment (fun) and endurability (i.e., *returnance*) but also curiosity, which is valuable to educational systems (Wouters et al., 2011), in addition to a description of the experience, which is related to enjoyment, however, assesses it in a deeper way. Next, the four factors are explained, and the way that they were captured by the questionnaire is described. The full questionnaire is available as a supplementary material *link of the article to be added*.

3.4.1. Experienced Fun

Experienced Fun was captured using the Smileyometer from the Fun-Toolkit (Read and Macfarlane, 2002). It provides a simple and intuitive way for players to indicate this factor. It was encoded as a rating, on a five-point scale that ranges from 1 to 5, where higher values indicate more fun.

3.4.2. Returnance

Returnance identifies the players' willingness to play the game again. In other words, it asks users to indicate if they would play the game again, choosing between yes (5), maybe (3) or no (1). This questionnaire's section was based on another tool of Fun-Toolkit, the Again Again Table (Read and Macfarlane, 2002).

3.4.3. Curiosity

Curiosity was adapted from the questionnaire used in (Wouters et al., 2011). It was captured through the following statements, which were also encoded as ratings on a five-point scale that ranged from 1 (completely disagree) to 5 (completely agree):

- **C1:** The game motivated me to learn more about math;
- **C2:** I wanted to continue playing because I wanted to see more about the game levels;
- **C3:** Playing the game raised questions about the game levels;
- **C4:** I was curious about the next event in the game;
- **C5:** I sought explanations for what I encountered in the game;
- **C6:** Playing the game raised questions regarding math;
- **C7:** I wanted to continue playing because I wanted to know more about math.

3.4.4. Experience Description

Experience Description allows us to investigate PX in a deeper way than the Smileyometer. It is based on predefined opposed attributes to have a semantic balance. This approach was inspired by its usage in (Moser et al., 2012). The following attributes were captured in a boolean-based way, wherein players could indicate as many attributes as they wanted: simple - difficult; great - childish; fun - boring; exciting - tiring; and intuitive - confusing.

3.4.5. Playing Performance

The players' performances play an important role in their experience. Therefore, it can provide valuable insights concerning self-reports from different game versions. These data are automatically captured, as the game is played, storing metrics such as the score, time and wins per level, for each player.

3.5. Procedure

Basically, the research procedure might be defined in four steps, as can be seen in Figure 2. The first was to introduce the testbed game, which describes how it works and how to play it, which was performed by the researcher or the institution's supervisor. Thereafter, players had to register into the game, providing some demographic characteristics. Third, they played a total of 20 levels of a specific game version, which took an average of 9.08 and 8.35 minutes for control and experimental groups, respectively. The number of levels to be played (20) was selected to guarantee equal conditions to the two groups of players, since there were only 20 human-designed levels. Last, the players completed the measures questionnaire, which was the same for all of them. Afterward, the players could continue to play the game if they wanted to.

Figure 2: Procedure performed in this research. The figure shows the steps that were followed, which indicates who performed each one of them.

3.6. Data analysis

The Kruskal-Wallis test was used to analyze **H1** to **H3**, similar to Connor et al. (2017), since they also used a two-sample design and Likert-scale measures. Unlike the measures from **H1** to **H3**, **H4** captured the participants' feedback in a true or false way. Additionally, they could select zero or multiple attributes. Thereby, **H4** was analyzed through the χ^2 homogeneity test, to investigate whether the completion distribution of both groups differs or not. Therefore, to support our hypothesis that these measures do not differ between game versions, the tests must yield p-values greater than the 0.05 alpha, thus providing insufficient evidence to reject the null hypothesis that significant differences exist. On the other hand, for more exploratory analyses, such as assessing individual characteristics' impact on PX, a less conservative alpha level of 0.1 was adopted.

4. Results

This section presents results from the analysis of this study's hypotheses, beginning with the overall findings. Subsequently, further analyses were conducted in terms of the participants' performances and the differences in the responses within subsamples.

4.1. Overall analyses

Figure 3 shows the statistics of the participants' questionnaire completion from both conditions. It presents a boxplot with data from the following measures: fun, *returnance*, and curiosity (seven questions and aggregated). From this figure, one can assess multiple statistics of players' answers, such as minimum, maximum, and median value. Also, it is possible to assess experience levels' deviation from the median based on the size of the boxes, which allows identifying each player experience factor variation based on the interquartile, that is, the difference between third and first quartile.

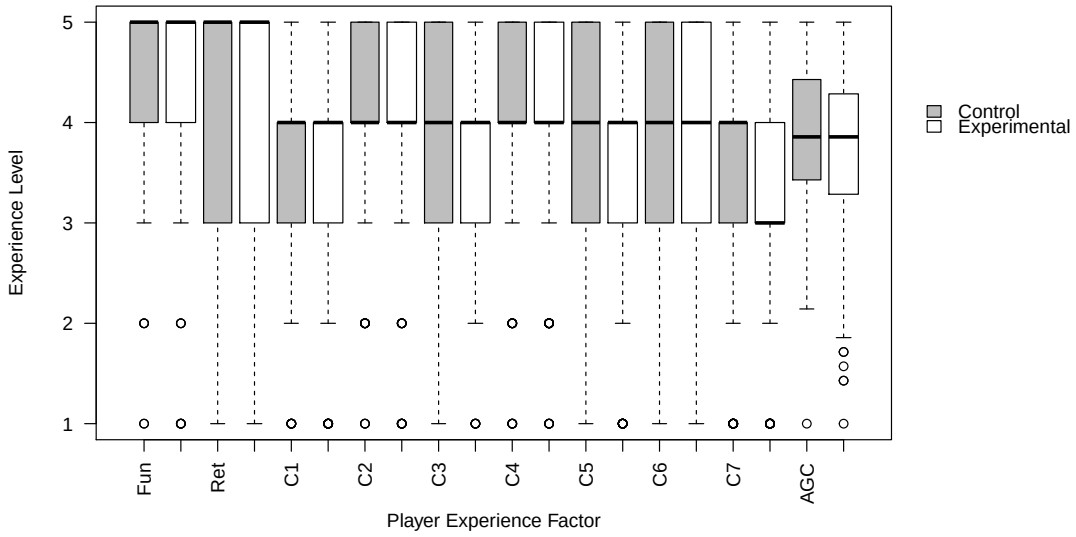


Figure 3: Boxplot of the participants' completions from each group for fun, *returnance* and curiosity.

As seen, for all factors, the distribution of responses from both groups was similar, and their medians were mostly approximately four (agreement). To provide evidence about their similarity, Table 3 shows the mean rank and standard deviation of each PX measure (PXM), along with the results of the Kruskal-Wallis tests. For C1 to C7, p-values were corrected using the Bonferroni approach (multiplying p-values by the number of comparisons; i.e., seven). Tests' results failed to provide evidence (all p-values > 0.05) groups' opinions were significantly different using the standard 0.05 alpha, which can be seen as strong evidence groups' do not differ considering the samples size. Thereby, these experiments support **H1**, **H2**, and **H3**. Therefore, we can conclude that the *dynamic* game version promoted experiences as well as the *static* version in the three measures analyzed here, according to the players' opinions, with 95%

confidence.

Table 3: Comparison of the groups' experiences. Data represented as Mean rank (SD). P-values from C1 to C7 were corrected using the Bonferroni approach. All p-values were > 0.05.

PXM	**Control**	**Experimental**	χ^2**(1)**
Fun	4.45 (0.78)	4.43 (0.87)	0.10
RET	4.40 (1.04)	4.25 (1.19)	1.37
AGC	3.90 (0.68)	3.78 (0.78)	2.95*
C1	3.79 (0.98)	3.64 (1.04)	2.59
C2	4.23 (0.84)	4.16 (0.87)	1.03
C3	3.89 (0.90)	3.82 (0.92)	0.79
C4	4.16 (0.89)	4.04 (0.89)	3.17
C5	3.81 (1.04)	3.56 (1.12)	6.58*
C6	3.99 (0.98)	3.85 (1.08)	1.59
C7	3.45 (1.20)	3.41 (1.22)	0.16

AGC = aggregated curiosity; * $p < 0.1$

Nevertheless, Table 3 shows groups' curiosity was marginally different. That is, for the curiosity measure, the Kruskal-Wallis tests yielded results that would be significant if a more exploratory approach had been adopted (i.e., an alpha level of 0.1). These marginally significant results were found for aggregated curiosity (AGC), possibly originating from the single item that also yielded a marginally significant result after correcting the p-values: C5 (*I sought explanations for what I encountered in the game*). The descriptive analysis suggests that players from the experimental group were less curious compared to the control's players (e.g., Figure 3). This suggests that despite the similar experiences between groups, there might be some factors that played a role and lead to this curiosity difference. The following section explores factors that possibly influenced in this *exploratory-significant* differences.

Moreover, in terms of **H4**, Figure 4 shows the distributions of the participants' experience description. It displays the five attribute pairs in the order of their opposites (e.g., great - tiring, simple - boring), which allows us to see that positive attributes, mainly great and fun, were selected substantially more than the others for both versions. Besides, the variability between groups can be seen by comparing the percentage of selection from one group to another, which are visually similar. We confirmed this suspect through the χ^2 homogeneity test ($\chi^2 = 7.218$, df = 9, p = 0.614). Therefore, players described their experiences mostly using positive attributes, where the differences in the distributions of the selected attributes were insignificant between groups, which corroborates with **H4**.

4.2. Difference analyses

These analyses have the aim of providing insights into the participants' characteristics (e.g., gender and age) that could influence their general opinions.

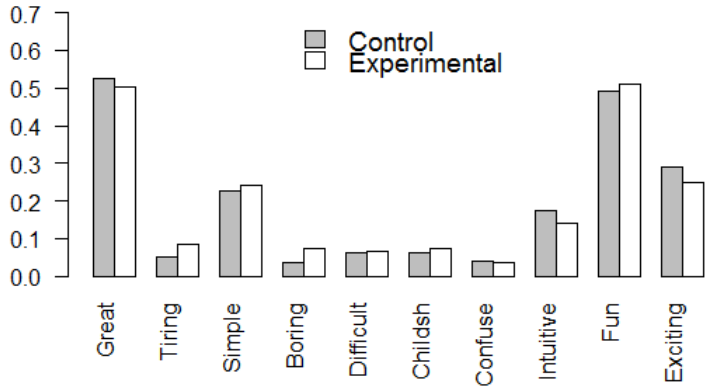


Figure 4: Bar plot of PX description. It shows the percentage that each attribute was selected between all selections from all players.

4.2.1. Performance

The performance of players from each version was compared to investigate whether some differences could lead to the difference in their self-reports. From the captured in-game measures, the following performance metrics were extracted and evaluated: average score, the highest level achieved, win rate, the maximum score achieved, and total time spent until reaching the questionnaire (playing 20 levels). Note that these data are related to the participants' gameplays before completing the questionnaire only. To analyze it, the U test was adopted since these data were measured in a similar way as in Butler et al. (2015).

The comparison of the groups' performance can be seen in Table 4. It shows the average and standard deviation for the aforementioned metrics and the statistical results of the U test. It shows that players from the control group performed significantly better than players from the experimental group and took more time to reach the questionnaire. Since answering the questionnaire is based on the number of played levels, the small win rate shows that players from the experimental group lost more and, consequently, spent less time playing game levels. A possible explanation for that difference is that those in the control group could play repeated levels, unlike those in the experimental group. Analyzing the relationship between the ratio of the number of repeated levels and players' average score, we found an insignificant relationship ($p = 0.27$, $r = 0.07$), which suggests those factors were not correlated. That is, playing more repeated levels did not affect participants' average scores, possibly because all levels present different math problems (see Section 3). Therefore, based on in-game metrics, we can conclude that the *dynamic* version provided players with more challenging experiences and that their performance was not related to playing repeated levels.

Table 4: Comparison of groups' performances. Data represented as Mean (SD).

Metric	Control	Experimental	U test
Avg Score	54.74 (5.37)	53.30 (5.51)	37394.50*
Highest Level	8.82 (2.60)	7.54 (3.00)	41988.00*
Win Rate	0.88 (0.06)	0.85 (0.08)	42162.50*
Maximum Score	536.43 (155.33)	465.45 (175.62)	41259.50*
Total Time	544.58 (175.82)	501.16 (151.26)	36587.50*

Time in seconds; AGC = aggregated curiosity; * $p < 0.05$.

4.2.2. Demographics

This study's sample is composed of a varied set of participants with distinct characteristics. Hence, we compared the control and experimental groups in terms of subsamples, to identify whether differences between them exist (e.g., control males vs experimental males) and, possibly, explain the marginally significant difference in curiosity.

The results of the comparison of self-reports according to each subsample of players' gender, being a gamer, and having the internet is presented next. As seen in Table 5, no significant difference was found between male players from the two groups, neither between females of both groups. Gamer players had no difference in their experiences, while non-gamers significantly differed in aggregated curiosity, as demonstrated by Table 6. Subjects with internet access through a computer at home did not significantly differ, neither those without it (Table 7). Hence, we have evidence that considering oneself a gamer or not had a marginally significant ($p < 0.1$) influence on players' curiosity, whereas gender and whether a player has access to the internet at home through a computer had not.

Table 5: Comparison of groups' experiences for subsamples according to gender. Data represented as Mean (SD). P-values from C1 to C7 were corrected using the Bonferroni approach. All p-values were > 0.05.

	Males			Females		
PXM	**Control**	**Experimental**	χ^2**(1)**	**Control**	**Experimental**	χ^2**(1)**
Fun	4.41 (0.80)	4.38 (0.93)	0.09	4.53 (0.73)	4.50 (0.77)	0.00
RET	4.37 (1.06)	4.30 (1.17)	0.16	4.44 (1.01)	4.19 (1.23)	1.88
AGC	3.94 (0.69)	3.81 (0.78)	1.41	3.82 (0.67)	3.74 (0.77)	1.17
C1	3.81 (1.02)	3.67 (1.07)	2.69	3.75 (0.91)	3.68 (1.00)	0.20
C2	4.24 (0.82)	4.19 (0.89)	0.14	4.20 (0.88)	4.10 (0.84)	1.18
C3	3.95 (0.87)	3.84 (0.87)	1.27	3.76 (0.95)	3.77 (0.99)	0.02
C4	4.20 (0.88)	4.06 (0.94)	1.86	4.09 (0.89)	4.01 (0.81)	0.83
C5	3.81 (1.10)	3.65 (1.11)	1.95	3.81 (0.92)	3.43 (1.13)	5.35
C6	4.08 (0.96)	3.84 (1.12)	3.20	3.80 (0.99)	3.86 (1.02)	0.33
C7	3.51 (1.24)	3.48 (1.20)	0.07	3.33 (1.12)	3.30 (1.25)	0.03

AGC = aggregated curiosity; * $p < 0.1$

Table 6: Comparison of groups' experiences for subsamples according to being a gamer or not. Data represented as Mean (SD). P-values from C1 to C7 were corrected using the Bonferroni approach. All p-values were > 0.05.

	Gamers			**Non-gamers**		
PXM	**Control**	**Experimental**	$\chi^2(1)$	**Control**	**Experimental**	$\chi^2(1)$
Fun	4.55 (0.74)	4.47 (0.94)	0.09	4.34 (0.80)	4.40 (0.80)	0.53
RET	4.53 (0.96)	4.25 (1.21)	3.57	4.26 (1.10)	4.25 (1.19)	0.03
AGC	3.96 (0.75)	3.86 (0.87)	0.34	3.84 (0.60)	3.71 (0.67)	3.40*
C1	3.84 (1.06)	3.75 (1.10)	0.49	3.74 (0.90)	3.55 (0.99)	2.57
C2	4.20 (0.92)	4.17 (1.00)	0.01	4.25 (0.75)	4.14 (0.74)	1.90
C3	3.88 (0.98)	3.90 (0.94)	0.01	3.90 (0.82)	3.74 (0.90)	1.83
C4	4.26 (0.93)	4.08 (0.99)	2.79	4.07 (0.83)	4.01 (0.79)	0.64
C5	3.88 (1.04)	3.67 (1.19)	1.32	3.75 (1.04)	3.47 (1.05)	5.77
C6	4.07 (0.93)	3.87 (1.21)	0.56	3.91 (1.02)	3.83 (0.95)	0.86
C7	3.63 (1.24)	3.60 (1.26)	0.04	3.27 (1.14)	3.25 (1.16)	0.10

AGC = aggregated curiosity; * $p < 0.1$

Second, we investigated how the relationship from numerical demographic variables differs between game versions. Since weekly playing hours had no significant correlation to any PX measure, only correlations from age are presented in Table 8. It shows, for each group (G), Kendall's correlation coefficient from age to the nine self-reported PX measures. Mainly, this correlation test was selected over Pearson's test due to the data's measure, which is neither interval nor ratio, and over Spearman's due to the number of tied ranks (Statistics, 2018b,a). In this context, the farther the correlation is from 0, the more the dependent variable (PX measure) is affected by the independent variable (demographics). Thus, we can see that the experimental group (E) was less affected by age than the control (C). Additionally, the table shows that for players from the experimental group, age had an insignificant correlation to C5, unlike for players from the control, despite the degree of correlation being almost the same. Thereby, while the age of the players from the *static* version impacted their experience more than those from the *dynamic* version, the participants' weekly playing hours did not affect either.

5. Discussion

This section discusses our findings in terms of whether they support our hypotheses, the rationales for the results achieved, and the limitations and issues that represent threats to the validity of our study. Overall, our findings support three of the four experimental predictions that were assumed in this research. These results are in line with previous research in this field, wherein PCG has demonstrated to provide experiences that are almost as good as human-designed levels (Connor et al., 2017; Butler et al., 2015). In our experiments, nevertheless, we found marginally significant evidence that the curiosity of players differed between the control and experimental groups, possibly due to an also marginally

Table 7: Comparison of groups' experiences for subsamples according to having a computer with internet access at home or not. Data represented as Mean (SD). P-values from C1 to C7 were corrected using the Bonferroni approach. All p-values were > 0.05.

	Has it			**Has not**		
PXM	**Control**	**Experimental**	χ^2**(1)**	**Control**	**Experimental**	χ^2**(1)**
Fun	4.45 (0.75)	4.43 (0.86)	0.16	4.44 (1.01)	4.44 (0.99)	0.01
RET	4.39 (1.04)	4.26 (1.18)	0.96	4.48 (1.05)	4.22 (1.31)	0.51
AGC	3.90 (0.66)	3.79 (0.77)	2.14	3.89 (0.87)	3.66 (0.81)	1.01
C1	3.79 (0.96)	3.63 (1.04)	2.76	3.78 (1.16)	3.78 (1.13)	0.00
C2	4.20 (0.83)	4.16 (0.86)	0.22	4.44 (0.85)	4.09 (1.00)	2.38
C3	3.90 (0.87)	3.83 (0.89)	0.75	3.78 (1.09)	3.65 (1.15)	0.09
C4	4.18 (0.85)	4.05 (0.88)	2.59	4.04 (1.16)	3.91 (1.00)	0.70
C5	3.82 (1.02)	3.60 (1.11)	4.62	3.74 (1.20)	3.17 (1.23)	2.98
C6	3.99 (0.98)	3.86 (1.05)	1.59	3.97 (0.98)	3.78 (1.31)	0.04
C7	3.45 (1.20)	3.43 (1.21)	0.04	3.48 (1.22)	3.22 (1.31)	0.56

AGC = aggregated curiosity; * $p < 0.1$

Table 8: Correlation from age to PX measures for each group.

G	**Fun**	**RET**	**AGC**	**C1**	**C2**	**C3**	**C4**	**C5**	**C6**	**C7**
C	-0.25*	-0.26*	-0.26*	-0.30*	-0.21*	-0.24*	-0.25*	-0.11*	-0.20*	-0.27*
E	-0.16*	-0.20*	-0.20*	-0.23*	-0.16*	-0.23*	-0.15*	-0.07	-0.18*	-0.19*

AGC = aggregated curiosity; * $p < 0.1$

significant in one of the seven curiosity items. Furthermore, our experiments provided insights concerning factors that might have influenced this difference, which are discussed next.

5.1. Insights on Experience Difference

Here, we discuss two new hypotheses, from different perspectives, that possibly contribute to groups' marginally significant different in curiosity.

One perspective is the participants' performance, which plays an important role in PX and was significantly different between groups. The participants of the experimental group performed worse than those of the control group. The fact that control group players played repeated levels could explain this difference, as playing repeated levels could lead them to achieve higher scores; however, our findings indicate the ratio between the number of repeated levels played and the average score was negligible. Then, we argue that those in the experimental group had to make more effort while playing, which might have led them to have less time to feel curious about the game, since they were more focused on trying to advance in the game to yield better performances. On the other hand, the participants from the control group possibly had to make less effort while playing, especially because they would play repeated levels, hence, the second time they already knew where to find the numbers. Therefore, these players potentially could explore the game more than those of the other group, exercising their curiosity more than the others.

Thus, the hypothesis that players feel lass curious when they are more challenged arises. Our claim is based on the fact that players from both groups had similar profiles (see Section 3.3). Thereby, it is expected that all participants have similar levels of gaming experience and that the performance difference emerged from the game levels. However, we are not aware of their previous experience with specific game genres. For example, those who have more experience in the testbed's genre had an easy life, whereas the participants experienced in other genres had more difficulty than the average. Hence, despite the similar characteristics, other background factors possibly played a role in the participants' experiences. Therefore, we sought rationales of this difference from the second perspective, demographic attributes.

In terms of this second perspective, we found that comparing the answers of each version, the difference in aggregated curiosity was significant only for a specific group of participants: those who consider themselves as gamers. That is, when considering only gamer players, the marginal difference in aggregated curiosity hold, but when considering other subsamples (e.g., only non-gamer or females) no even marginal differences were found. Additionally, investigating the relationship of the participants' ages to their experiences, we found that for those of the control group, all correlations were higher than for the participants of the experimental group. Thereby, the hypothesis that the aforementioned demographic attribute affected the influences of PCG on the participants' experiences also emerges.

Furthermore, for demographics attributes such as gender and having internet access at home, significant results would have been found as well if the Bonferroni correction had not been applied. For instance, groups' differences considering only males were not found even without applying the correction. On the other hand, groups' differences for females would be found if the correction was not applied. Similarly, no difference would be found for those without internet access at home, but would for those with it. This is important to note because the Bonferroni correction is considered conservative and there is an open debate on which correction to apply, especially because samples' size was decreased for subsample comparisons, which decreases statistical power.

Since demographic attributes' differences between groups are insignificant, our findings provide evidence that the PX provided by PCG was sensible to the participants' characteristics (being a gamer and age), as well as opens the question on whether other attributes, such as gender and having internet access at home, are relevant. A significant effort from the game design community has been made to deliver games that provide experiences equally good for males and females. Our results contribute in this vein suggesting that the *static* version delivered experiences that were not different from the *dynamic* version for both males and females. Moreover, the *dynamic* version did not differ from the *static* for gamers, which suggests that PCG matched the human-designer for experienced/skillful players.

In contrast, for non-gamers, it was evident that a difference between versions was perceived, which might emerge because they are less experienced/skillful and thus faced more difficulty on the *dynamic* version. This finding corrobo-

rates with the indication that players who face more challenges feel less curious, given the weaker playing background of non-gamers compared to gamers. Considering the participants with internet access at home through a computer or not, the curiosity significant difference did not hold. Lastly, age had small interactions with the experiences of the participants from both groups; however, those were higher for players in the control group in all measures. Moreover, the only measure that was not significantly correlated to age was C5 from the experimental group. Thereby, suggesting the *static* version provided experiences that were less dependent on the participant's age.

In summary, these further analyses provided valuable insights concerning factors that influence participants' curiosity. Based on these perspectives, we raised hypotheses about factors that could have led to the groups' marginally significant differences in aggregated curiosity and C5. Although we made an effort to present rationales about why such factors possibly affected PX, further research is required to confirm those hypotheses.

5.2. Threats to Validity and Recommendations for Addressing Them

There are some threats that emerged from our experiments, such as human bias and participants' performances. There is a critique with regard to using a human for levels development because, if those are of poor quality, a bias might be inserted into the results (e.g., favoring the PCG version) (Horn et al., 2016). One way to remedy this problem is to use more than one *static* version, which would allow us to identify whether one is better than another and, then, to use the best one as a baseline comparison. Another alternative is to also use a completely random generator, which would enable the comparison of whether both human- and algorithmic-authored content excels random contents (Butler et al., 2015). In this way, human biases would be mitigated, which would improve the reliability of the study's findings. Although human-authored content can represent a bias, we argue that the high self-reported experiences remedied this threat. As shown, most participants reported positive experiences, and hence, our findings suggest that developers could take advantage of PCG benefits without jeopardizing their games' outcomes.

On the other hand, the average scores close to the upper end of the five-point Likert scale raises attention to whether differences between conditions were masked by ceiling effects. That is, it might be that because there was no higher score to choose, differences between groups' experiences could not be found. One point is that the five-point Likert scale is commonly used, and we implemented it according to previous related studies, which is expected to reduce such threats. Another point is that both groups reported medium and low experiences (although not much), which suggests that if participants had not considered their experiences to match the top of the scale, they would not choose it. Hence, we believe that ceiling effects do not represent a significant threat to our findings.

The performance of players from different groups was significantly different. Given the similarity of the participants' profiles, we argue that the use of a PCG algorithm had the most impact on it, either by providing harder-to-play levels

or by preventing players from playing repeated levels. On one hand, in spite of playing the same game with a single difference (level generation process), experiencing different levels of challenge might have affected the PX. Aiming to mitigate this effect, we analyzed the players' performances and discussed the possible implications. Another way to remedy this effect is to perform a pilot study, to balance the difficulty of both versions, which could be achieved by adapting the parameter of the PCG algorithm. Consequently, similar research wherein participants yield even performance could answer one of the hypotheses previously mentioned in our discussion. On the other hand, we consider that playing repeated levels is not a threat as it is common for games not using PCG to provide repeated content.

Additionally, in this research, levels were generated through a simple and straightforward technique, the *constructive* method. It has been shown that more complex approaches (e.g., search-based (Togelius et al., 2011b)) are preferred compared to both constructive (Khalifa et al., 2016) and randomly created content (Scirea et al., 2018). Therefore, performing similar research wherein human-created content is compared to those techniques is expected to show that PCG can overcome the human-authored content. However, search-based PCG, for example, requires a fitness function development, which is the most complex task of using it (Togelius et al., 2011b). Successfully employing it will mainly depend on designing this function to provide game levels according to both the users' and the developers' expectations. Moreover, there is a trade-off in terms of speed, whereas these are GaT algorithms that tend to be more costly than constructive approaches. In the case of using PCG online, such as in this study and in the reviewed research, time constraints represent a relevant limitation. Hence, employing search-based PCG has the potential to improve PX, even though it is harder to develop and computationally more expensive.

6. Conclusions

The research presented in this article investigated players' interactions with a DMG. The goal was to identify whether a computational intervention that improves game development, creating game levels through PCG, could lead to PX that are as good as the ones led by human-designed levels. Our hypotheses were based on the assumption that the experience of players from one intervention would not differ from the experience of players from the other, considering four types of measures (i.e., fun, *returnance*, curiosity, and description of experience), and thus, there were four hypotheses. Hence, according to the feedback of 507 participants, we performed an A/B test based on a between-subject design, wherein 242 participants played on the human-designed levels (control group) and 265 interacted with the procedurally generated levels (experimental group).

We highlight that the main findings of our experiments show that there were no significant differences between the control and experimental groups for the four metrics. Nevertheless, curiosity presented a marginally significant difference. Based on further investigations, we also found that players from

the control group had better performances than participants from the experimental group. Additionally, performing the same comparisons but considering subsamples of the participants (e.g., males from control versus males from experimental, gamers from control versus gamers from experimental), the results suggested players demographics' characteristics influenced the impact of PCG on PX.

In summary, this article's findings support our four hypotheses, while they raise attention to the fact that groups' curiosity was marginally different. Hence, the main contribution of this article is the empirical analysis of how players report their experiences when interacting with two versions of the same game, wherein the single difference is to use a technique that improves the game development by automatically generating levels. Additionally, we contribute by showing that it is possible to use PCG to improve the game's development process and still promote experiences that are almost as good as the experiences from human-authored content, from the perspective of the players' feelings. We believe that these results can generalize to similar games, considering that PCG can increase their replay value by constantly providing new content as well as having its outputs controlled to achieve outcomes as expected by the designers/developers. Thereby, developers can benefit from the advantages of PCG in development while providing experiences that might be almost equivalent to human-designed content.

Although some research that concerns the influences of PCG on players' interactions has been conducted, this specific field is yet emerging and demands more studies to further validate ours and similar findings. This aspect is one strand that we recommend to be tackled in future research. In addition, we suggest the investigation of the hypotheses that emerged from the deeper analysis of our results, as mentioned in our discussion. The goal would be to analyze whether players having dissimilar performances is, indeed, a factor that impacts on their curiosity and whether it extends to other PX measures (e.g., fun and *returnance*). Similarly, we cal for further research in terms of how players' demographic attributes affect their experience as well. These studies would contribute not only to human interaction with computers but also to more specific fields such as player modeling, which could then rely on the findings to deliver personalized games that provide players with tailored experiences.

Declarations of Interest

None.

Acknowledgments

L. Rodrigues was supported by Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES) - Finance Code 001. R. Bonidia was supported by Federal University of Technology - Paraná (UTFPR - Grant: April/2018).

References

Amato, F., Moscato, F., March 2017. Formal procedural content generation in games driven by social analyses. In: 2017 31st International Conference on Advanced Information Networking and Applications Workshops (WAINA). pp. 674–679.

Bauckhage, C., Kersting, K., Sifa, R., Thurau, C., Drachen, A., Canossa, A., Sept 2012. How players lose interest in playing a game: An empirical study based on distributions of total playing times. In: 2012 IEEE Conference on Computational Intelligence and Games (CIG). pp. 139–146.

Biswas, G., Katzlberger, T., Bransford, J., Schwartz, D., TAGV, 2001. Extending intelligent learning environments with teachable agents to enhance learning. In: Artificial Intelligence in Education, J.D. Moore et al. (Eds.) IOS. Press, pp. 389–397.

Butler, E., Andersen, E., Smith, A. M., Gulwani, S., Popović, Z., 2015. Automatic game progression design through analysis of solution features. In: Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems. CHI '15. ACM, New York, NY, USA, pp. 2407–2416.
URL http://doi.acm.org/10.1145/2702123.2702330

Cardamone, L., Loiacono, D., Lanzi, P. L., 2011. Interactive evolution for the procedural generation of tracks in a high-end racing game. In: Proceedings of the 13th Annual Conference on Genetic and Evolutionary Computation. GECCO '11. ACM, New York, NY, USA, pp. 395–402.
URL http://doi.acm.org/10.1145/2001576.2001631

Carli, D., Bevilacqua, F., Pozzer, C., d'Ornellas, M., Nov 2011. A survey of procedural content generation techniques suitable to game development. In: SBGAMES 2011. pp. 26–35.

Cheng, H. N. H., Lin, Y. J., Wang, M., Chan, T. W., July 2015. Math detective: Digital game-based mathematical error dectection, correction and explanation. In: 2015 IEEE 15th International Conference on Advanced Learning Technologies. pp. 122–126.

Connor, A. M., Greig, T. J., Kruse, J., Dec 2017. Evaluating the impact of procedurally generated content on game immersion. The Computer Games Journal 6 (4), 209–225.
URL https://doi.org/10.1007/s40869-017-0043-6

Dahlskog, S., Togelius, J., Nelson, M. J., 2014. Linear levels through n-grams. In: Proceedings of the 18th International Academic MindTrek Conference: Media Business, Management, Content & Services. AcademicMindTrek '14. ACM, New York, NY, USA, pp. 200–206.
URL http://doi.acm.org/10.1145/2676467.2676506

de Carvalho, M. F., Gasparini, I., da Silva Hounsell, M., 2016. Digital games for math literacy: A systematic literature mapping on brazilian publications. In: Rocha, Á., Correia, A. M., Adeli, H., Reis, L. P., Mendonça Teixeira, M. (Eds.), New Advances in Information Systems and Technologies. Springer International Publishing, Cham, pp. 245–254.

Hendrikx, M., Meijer, S., Van Der Velden, J., Iosup, A., feb 2013. Procedural content generation for games: A survey. ACM Trans. Multimedia Comput. Commun. Appl. 9 (1), 1:1–1:22.

Hooshyar, D., Yousefi, M., Wang, M., Lim, H., 2018. A data-driven procedural-content-generation approach for educational games. Journal of Computer Assisted Learning 34 (6), 731–739.
URL https://onlinelibrary.wiley.com/doi/abs/10.1111/jcal.12280

Horn, B., Clark, C., Strom, O., Chao, H., Stahl, A. J., Harteveld, C., Smith, G., 2016. Design insights into the creation and evaluation of a computer science educational game. In: Proceedings of the 47th ACM Technical Symposium on Computing Science Education. SIGCSE '16. ACM, New York, NY, USA, pp. 576–581.
URL http://doi.acm.org/10.1145/2839509.2844656

Horn, B., Dahlskog, S., Shaker, N., Smith, G., Togelius, J., 2014. A comparative evaluation of procedural level generators in the mario ai framework. In: Foundations of Digital Games 2014. pp. 1–8.
URL http://www.fdg2014.org/

Ibarra, M. J., Soto, W., Ataucusi, P., Ataucusi, E., Oct 2016. Mathfraction: Educational serious game for students motivation for math learning. In: 2016 XI Latin American Conference on Learning Objects and Technology (LACLO). pp. 1–9.

Karavolos, D., Bouwer, A., Bidarra, R., jun 2015. Mixed-initiative design of game levels: integrating mission and space into level generation. In: FDG 2015. p. 8.
URL http://graphics.tudelft.nl/Publications-new/2015/KBB15

Ke, F., 2008. A case study of computer gaming for math: Engaged learning from gameplay? Computers & Education 51 (4), 1609 – 1620.
URL http://www.sciencedirect.com/science/article/pii/S0360131508000523

Khalifa, A., Liebana, D. P., Lucas, S. M., Togelius, J., 2016. General video game level generation. In: 2016 GECCO. pp. 253–259.

Kiili, K., Ketamo, H., 2017. Evaluating cognitive and affective outcomes of a digital game-based math test. IEEE Transactions on Learning Technologies PP (99), 1–1.

Kiili, K., Moeller, K., Ninaus, M., 2018. Evaluating the effectiveness of a game-based rational number training - in-game metrics as learning indicators. Computers & Education 120, 13 – 28.
URL http://www.sciencedirect.com/science/article/pii/S0360131518300125

Korn, O., Blatz, M., Rees, A., Schaal, J., Schwind, V., Görlich, D., Apr. 2017. Procedural content generation for game props? a study on the effects on user experience. Comput. Entertain. 15 (2), 1:1–1:15.
URL http://doi.acm.org/10.1145/2974026

Linden, R. v. d., Lopes, R., Bidarra, R., oct 2013. Designing procedurally generated levels. In: IDPv2 2013 - Workshop on Artificial Intelligence in the Game Design Process. AAAI, AAAI Press, AAAI Press, Palo Alto, CA, pp. 41–47, iSBN 978-1-57735-635-6.

Mariño, J. R. H., Reis, W. M. P., Lelis, L. H. S., 2015. An empirical evaluation of evaluation metrics of procedurally generated mario levels. In: Proceedings of the Eleventh AAAI. pp. 44–50.

McLaren, B. M., Adams, D., Mayer, R. E., Forlizzi, J., 2017. A computer-based game that promotes mathematics learning more than a conventional approach. IJGBL 7, 36–56.

Moghadam, A. B., Rafsanjani, M. K., March 2017. A genetic approach in procedural content generation for platformer games level creation. In: 2017 2nd Conference on Swarm Intelligence and Evolutionary Computation (CSIEC). pp. 141–146.

Moser, C., Fuchsberger, V., Tscheligi, M., 2012. Rapid assessment of game experiences in public settings. In: Proceedings of the 4th International Conference on Fun and Games. FnG '12. ACM, New York, NY, USA, pp. 73–82.
URL http://doi.acm.org/10.1145/2367616.2367625

Read, J., Macfarlane, S., 2002. Endurability, engagement and expectations: Measuring children's fun. In: Interaction Design and Children. Shaker Publishing, pp. 1–23.

Read, J. C., MacFarlane, S., 2006. Using the fun toolkit and other survey methods to gather opinions in child computer interaction. In: Proceedings of the 2006 Conference on Interaction Design and Children. IDC '06. ACM, New York, NY, USA, pp. 81–88.
URL http://doi.acm.org/10.1145/1139073.1139096

Rodrigues, L., Bonidia, R. P., Brancher, J. D., 10 2017. A math educational computer game using procedural content generation. In: SBIE 2017. pp. 756–765.
URL http://www.brie.org/pub/index.php/sbie/article/view/7604/5400

Scirea, M., Cheong, Y.-G., Nelson, M. J., Bae, B.-C., 2014. Evaluating musical foreshadowing of videogame narrative experiences. In: Proceedings of the 9th Audio Mostly: A Conference on Interaction With Sound. AM '14. ACM, New York, NY, USA, pp. 8:1–8:7.
URL http://doi.acm.org/10.1145/2636879.2636889

Scirea, M., Eklund, P., Togelius, J., Risi, S., 2018. Evolving in-game mood-expressive music with metacompose. In: Proceedings of Audio Mostly (AM 2018). pp. 1–8.

Shaker, N., Yannakakis, G. N., Togelius, J., Nicolau, M., O'Neill, M., 2012. Evolving personalized content for super mario bros using grammatical evolution. In: Proceedings of the Eighth AAAI Conference on Artificial Intelligence and Interactive Digital Entertainment. AIIDE'12. AAAI Press, pp. 75–80.
URL http://dl.acm.org/citation.cfm?id=3014629.3014643

Sim, G., Horton, M., 2012. Investigating children's opinions of games: Fun toolkit vs. this or that. In: Proceedings of the 11th International Conference on Interaction Design and Children. IDC '12. ACM, New York, NY, USA, pp. 70–77.
URL http://doi.acm.org/10.1145/2307096.2307105

Smith, G., Whitehead, J., 2010. Analyzing the expressive range of a level generator. In: Proceedings of the 2010 Workshop on Procedural Content Generation in Games. PCGames '10. ACM, New York, NY, USA, pp. 4:1–4:7.
URL http://doi.acm.org/10.1145/1814256.1814260

Statistics, L., oct 2018a. Kendall's tau-b using spss statistics.
URL statistics.laerd.com/spss-tutorials/kendalls-tau-b-using-spss-statistics.php

Statistics, L., oct 2018b. Pearson's product-moment correlation using spss statistics.
URL statistics.laerd.com/spss-tutorials/pearsons-product-moment-correlation-using-spss-statistics.php

Togelius, J., Justinussen, T., Hartzen, A., 2012. Compositional procedural content generation. In: Proceedings of the The Third Workshop on PCG in Games. PCG'12. ACM, New York, NY, USA, pp. 16:1–16:4.
URL http://doi.acm.org/10.1145/2538528.2538541

Togelius, J., Kastbjerg, E., Schedl, D., Yannakakis, G. N., 2011a. What is procedural content generation?: Mario on the borderline. In: Proceedings of the 2Nd International Workshop on Procedural Content Generation in Games. PCGames '11. pp. 3:1–3:6.
URL http://doi.acm.org/10.1145/2000919.2000922

Togelius, J., Yannakakis, G., Stanley, K., Browne, C., Sept 2011b. Search-based procedural content generation: A taxonomy and survey. Computational Intelligence and AI in Games, IEEE Transactions on 3 (3), 172–186.

Valls-Vargas, J., Zhu, J., Ontañón, S., 2017. Graph grammar-based controllable generation of puzzles for a learning game about parallel programming. In: Proceedings of the 12th International Conference on the Foundations of Digital Games. FDG '17. ACM, New York, NY, USA, pp. 7:1–7:10.
URL http://doi.acm.org/10.1145/3102071.3102079

Wouters, P., van Oostendorp, H., Boonekamp, R., van der Spek, E., 2011. The role of game discourse analysis and curiosity in creating engaging and effective serious games by implementing a back story and foreshadowing. Interacting with Computers 23 (4), 329 – 336, cognitive Ergonomics for Situated Human-Automation Collaboration.
URL http://www.sciencedirect.com/science/article/pii/S0953543811000415

Yannakakis, G. N., Spronck, P., Loiacono, D., André, E., 2013. Player modeling. In: Dagstuhl Follow-Ups. Vol. 6. Schloss Dagstuhl-Leibniz-Zentrum fuer Informatik, p. 45.

Yannakakis, G. N., Togelius, J., July 2011. Experience-driven procedural content generation. IEEE Transactions on Affective Computing 2 (3), 147–161.

Yurdabakan, I., Uzunkavak, C., 07 2012. Primary school students' attitudes towards computer based testing and assessment in turkey. Turkish Online Journal of Distance Education 13, 177–188.