The Impact of Procedural Level Generation on Players' Experiences and In-game Behavior

Luiz Rodrigues and Jacques Brancher (advisor)
Department of Computer Science
Londrina State University
Londrina, Brazil
luiz_rodrigues17@hotmail.com, jacques@uel.br

***Abstract*—Game development is an expensive process which commonly requires a multidisciplinary team. Procedural content generation can remedy some problems of this process, aiding on the creation of different types of contents (e.g. levels). However, few studies have been done in terms of how it influences players, especially on digital math games. This work tackles this problem by investigating how procedural level generation influences players of an introduced digital math game. Besides identifying these influences, we validated the introduced game and analyzed both the relationship between fun, willingness to play the game again (*i.e., returnance*), and curiosity as well as the impact of demographic and in-game data on player experience and performance. A two-sample experiment was designed where participants played a game version with (*dynamic*) or without PCG (*static*) in which in-game (N = 724) and questionnaire (N = 506) data were gathered and empirically analyzed. Mainly, the results demonstrate the experiences of players from the *dynamic* version were similar to those of the *static* in all but one question, while being more difficult and providing equivalent engagement. Furthermore, the findings show: the game is fun and arises players' curiosity and *returnance*, players' curiosity has a strong correlation to fun and *returnance*, and demographics and in-game performance impact players' experiences. Our results are valuable to developers and designers by showing the impact of procedural level generation on players experiences, as well as how and which factors might play a role in their experiences.**

***Keywords*-procedural content generation; player experience; game development; A/B test; educational game.**

I. Introduction

Some students perceive math as a difficult subject, do not like it, and consider it displeasing [1], which might be related to the ease of access to interactive technology of nowadays that, consequently, leads to a lack of interest in the traditional way of teaching [2]. Digital Math Games (DMG) might be used to address it, improving aspects such as students learning [3], positive attitudes towards the subject [4], and engagement [5]. Despite that, the development process of DMG is a slow and costly task, even for the broader category of general purpose games, which commonly requires several designers, artists, and developers [6].

An alternative to tackle these problems is the Procedural Content Generation (PCG) [7], a reliable tool to provide diversified, automatically generated outputs, which can be controlled through generation parameters [8], while automating, aiding in creativity, and speeding up the creation of various types of game contents [6]. In spite of that, few studies have applied it on educational games [9], [10]. Additionally, a limitation of PCG literature is that most studies focus on algorithms capacities [11], failing to demonstrate the true impact of automating content generation from players' perspective [12]. This is important because technologies must provide positive experiences, especially for children; otherwise, players are unlikely to interact or accept it [13], [14] and might have their learning experience harmed [15]. To address the challenge of using PCG to improve educational games development whilst providing players with positive experiences, as well as analyze PCG's impact on players, this work introduced a DMG featuring two PCG algorithms and validated it with 724 players.

We performed an A/B test [16] to identify the influences of Procedural Level Generation (PLG) on players, comparing the game version using it (*dynamic*) to a game version featuring expert-designed levels (*static*), thus demonstrating whether using PCG is able to provide experiences as good as those provided by human-designed contents in terms of six Player Experience (PX) metrics. The findings show the only difference was that players of the *static* version sought more explanations for what they encountered in the game, whereas all other metrics differences were statistically insignificant. Thereby, demonstrating PCG was able to provide experiences nearly equivalent to expert-designed contents. Furthermore, the results demonstrate the introduced game led to positive experiences, PX metrics were highly correlated, and demographics attributes impacted on both PX and performance.

II. Background and Related Works

PCG refers to creating contents automatically without or with limited human intervention [7]. Mainly, there are two perspectives that might be adopted to evaluate it. One is focused on the algorithm's capabilities, commonly performed through the analysis of the expressive range [11]. However, that approach is insufficient to replace user-based studies

[17], leading to the other perspective, which concerns how the algorithm's outputs are experienced, investigating PX according to their interaction with the application using PCG [18], or through A/B comparisons to identify PCG's impact [19]. Hence, the only approach that reveals the impacts of PCG usage is the A/B test method [12], which was the main goal of this work.

However, few studies have addressed the impacts of PCG from the players perspective. In Butler *et al.* [20], a framework to create game progressions via PCG was introduced and validated by applying it in the DMG *Refraction*. The authors compared it to the game's original version and found the version using their method was almost played as much as the original. In Korn *et al.* [12], game reefs were procedurally generated and compared to those generated by designers. The findings demonstrated that users evaluated significantly better the reefs created through the PCG method. In Connor *et al.* [19], the impact of PCG on players' immersion was analyzed comparing levels automatically and manually generated. Players' reports demonstrated PCG led to smaller immersion in two out of 30 aspects of immersion.

From those, only two addressed the impacts of level generation [20], [19], and a single study used an educational game as the testbed [20]. Additionally, neither of those research investigated PX in terms of both opinions and behaviors, as well as did not involve a heterogeneous sample from the perspective of subjects characteristics, which would increase the generalization of their findings [21]. Furthermore, the reduced sample size [19] and the lack of evidence concerning the groups of players compared [20] also threats related works [21]. Considering this context, this work differs from those of the literature by (i) analyzing the impacts of PLG in an educational game, according to both players' opinions (n = 506) and in-game behavior (n = 724), (ii) based on a heterogeneous sample (iii) of substantial size compared to other studies, which (iv) features similar characteristics (statistically insignificant differences) between sub-samples.

III. Materials and Method

In summary, this study's main goal was to answer the following question: *Do the opinions and in-game behavior of players are influenced by PCG-created levels compared to those created by a human?* The hypothesis was that no difference would be found, considering the PCG algorithm would provide levels as good as those manually designed although its simplicity. To measure possible differences, both players' opinions and in-game behavior were analyzed. To enable the comparison, we performed an A/B test comparing two versions of the same game in which one featured levels generated through PCG (*dynamic*) and the other contained levels created by a game developer (*static*). A two-sample design was adopted, following similar research [19], [20], in which players were randomly assigned to the *static* or to the *dynamic* version, hence, featuring the control or the experimental group, respectively. Thereby, enabling the comparison of both samples to identify possible differences. Data collection was performed in face-to-face applications in four institutions (over 70% of the collected data) and *in the wild* (players reached via emails and social networks). The procedure was as follows: (i) introducing the game and the research itself; (ii) players registering into the game and completing the demographics questionnaire; (iii) players playing exactly 20 game levels; and (iv) participants completing a PX questionnaire. Additionally, in-game data log was constantly stored after each level was played. Note that players were not aware of our hypothesis or that the game had two different versions. Figure 1 summarizes both the setup and the procedure mentioned.

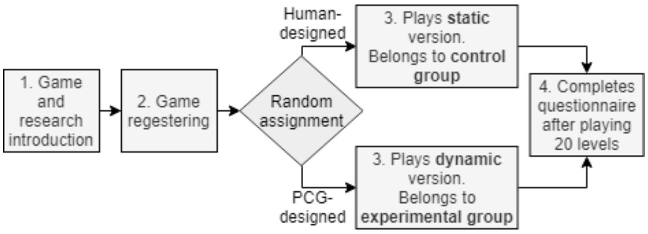
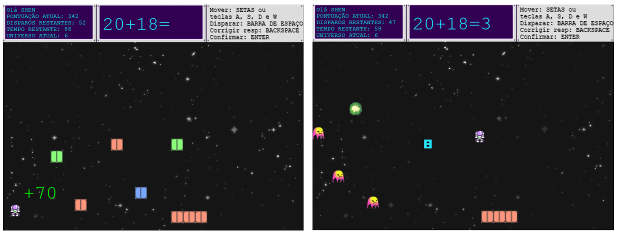

1. Game and research introduction
2. Game regestering
Random assignment
Human-designed
PCG-designed
3. Plays static version. Belongs to control group
3. Plays dynamic version. Belongs to experimental group
4. Completes questionnaire after playing 20 levels

Figure 1. Study's method and participants groups.

A. Testbed Game

To enable this research, we developed *SpaceMath*[1] [22], [23], [24], a DMG that fosters the practice of basic mathematical operations (summation, subtraction, multiplication, and division) and uses PCG to create both its levels and math puzzles. In this game, players control an astronaut towards exploring multiples parallel universes (levels) and solving math puzzles, as shown in Figure 2. Solving math puzzles is the pedagogical aspect of the game, in which players have to solve arithmetic operations by collecting the numbers that form the correct answer (38 in the figure's case). As the game provides endless gameplay, players repeatedly solve several problems, which helps them in learning by repetition. For information concerning the generation process of both the *dynamic* and *static* versions, see [25].



(a) Initial Arrangement. (b) After partial exploration.

Figure 2. Interface of the testbed game developed in this work.

[1] Available online at http://spacemath.rpbtecnologia.com.br

B. Measures

To compare game versions, players reported their opinions through an adapted version of a questionnaire for rapid assessment [26] after playing 20 levels, which captured PX in terms of fun, *returnance*, curiosity (composed of seven statements, referred to as C1, C2, up to C7) [22], and experience description. The threshold was 20 because this is the number of human-created levels available, hence, guaranteeing players of both versions completed the questionnaire after playing the same number of levels. Additionally, in-game data were captured throughout the process as well, which enabled the analysis of players' performance and in-game behavior. For further explanations, see [25].

C. Data Analysis Process

First, we found both groups' demographics were insignificantly different, preventing threats that could emerge from comparing data of players with different characteristics [21]. Second, we compared the experiences (opinions) of participants of both groups ($N_{control}$ = 242; $N_{experimental}$ = 265), as well as their in-game behavior ($N_{control}$ = 355; $N_{experimental}$ = 369), through Kruskal-Wallis and Mann-Whitney hypothesis tests, respectively. Then, we compared the correlations from fun and *returnance* to curiosity and which attributes impact on PX and performance (N = 507), through Kendall's correlation tests and Chi-squared association tests. For justifications and further information, we refer to [25].

IV. Results

The difference between the experiences from each version was insignificant in all self-reported factors but one, the *I sought explanations for what I encountered in the game* (C5) statement. Further investigating this concern, we found that this difference was significant only for: females, gamers, and those with internet access through a computer at home. Also, the influence of age was strongest on those who played the *static* version, whilst the remainder relied more on their affinity with math. Considering in-game data, the differences in players' retainment (*i.e.*, playing 20 levels or more) and engagement (*i.e.*, number of played levels) were insignificant, whereas their performances were mostly significantly different, wherein the most impacting factor was having internet access at home from a computer. Thus, game versions differed in only one of the nine self-reported factors assessed and in players' performance, wherein demographic attributes showed insights from where these differences emerged. Figure 3 demonstrates the distribution of players' self-reported experiences and Table I summarizes their behavior.

Furthermore, the results provided evidence that, considering both groups, players' self-reports of fun and *returnance* are significantly correlated to their average curiosity as well as to its factors separately, with a degree that ranges from moderate to strong. Additionally, our analyses demonstrated that some attributes (*e.g.*, players' school stage and age) have small to moderate negative significant correlations to PX, in contrast to others (*e.g.*, players' affinity to math), which have a small but significant positive correlation to fun, *returnance*, and curiosity. On the other hand, whilst curiosity is associated with genre, being a gamer and having internet access through a computer at home, *returnance* is not, while fun depends on being a gamer only. Lastly, we found players' performance metrics have small significant negative correlations to their experience, with the exception of average shots per level.

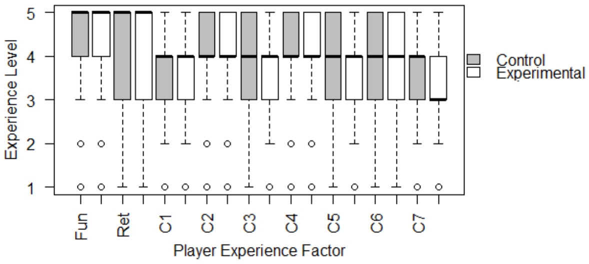


Figure 3. Boxplot of PX from participants of both versions.

Table I
Groups' performance. Data represented as Mean (SD).

Metric	Control	Experimental	U test
Avg Score	54.741 (5.366)	53.304 (5.507)	37394*
Highest Level	8.822 (2.602)	7.540 (3.002)	41988*
Wins Rate	0.884 (0.064)	0.846 (0.080)	42162*
Max. Score	536.430 (155.328)	465.453 (175.621)	41259*
Total Time	544.583 (175.822)	501.158 (151.263)	36587*

* $p < 0.05$

V. Main Contributions

This dissertation contributes to the fields of Human-Computer Interaction, in terms of the impacts of PLG, in-game performance, and demographic data on PX; and Computers & Education, introducing, validating, and showcasing the impacts of using a technique to improve the development of a DMG. Consequently, providing valuable contributions to the fields of Games and Entertainment as well. In summary, the contributions are: (i) A DMG that encourages its players to practice math and provide them with pseudo-infinite game levels and arithmetic problems; (ii) empirical evidence that, besides providing players with positive experiences, this game arises their curiosity; (iii) to demonstrate that using PCG-created game levels promoted experiences equivalent to human-designed levels in all but one PX factor; (iv) to reveal demographic characteristics associated with PX as well as how in-game performance is correlated with their experiences; (v) to confirm that the difficulty of the *dynamic* game version can be adjusted through the level generation parameter; and (vi) to provide evidence that players experienced fun and *returnance* are correlated to their curiosity. These contributions generated a series of scientific publications, and a registered software (registered at the Brazilian National Institute of Industrial Property - INPI). Each of these contributions is detailed in the following external link, for the sake of space-saving: http://bit.ly/CTD-SBGames-Rodrigues2019.

VI. Concluding Remarks

This study analyzed the influences of PLG based on both players' opinions and in-game behavior through a DMG that we introduced to enable the identification of those influences on an educational game. The main findings are that both human- and PCG-created levels led to indifferent in-game behavior and to PX that differed in a single aspect, and that demographics, in-game behavior, and curiosity are correlated to PX. We highlight that there are some threats and limitations that were not mentioned here due to reduced space, in which we refer the interested reader to [25], as well as there is no evidence that our findings will generalize to other games and contexts, concerns that demand further research to verification. As main future works, we suggest performing similar research to ground PCG's impacts, also evaluating the impacts of PCG on learning rather than on PX, and exploiting our findings to derive models of PX.

Acknowledgements

This research was financed in part by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil.

References

[1] G. Biswas, T. Katzlberger, J. Bransford, and D. Schwartz, "Extending intelligent learning environments with teachable agents to enhance learning, 2001.

[2] C. Madeira, L. Câmara, I. Beserra, and R. Tavares, "Mathmare: um jogo de plataforma envolvendo desafios matemáticos do ensino médio," in *Proceedings of the SBGames*, 2015, in portuguese.

[3] B. M. McLaren, D. Adams, R. E. Mayer, and J. Forlizzi, "A computer-based game that promotes mathematics learning more than a conventional approach," *International Journal of Game-Based Learning*, vol. 7, 2017.

[4] F. Ke, "A case study of computer gaming for math: Engaged learning from gameplay?" *Computers & Education*, vol. 51, no. 4, 2008.

[5] K. Kiili and H. Ketamo, "Evaluating cognitive and affective outcomes of a digital game-based math test," *IEEE Transactions on Learning Technologies*, vol. PP, no. 99, 2017.

[6] M. Hendrikx, S. Meijer, J. Van Der Velden, and A. Iosup, "Procedural content generation for games: A survey," *ACM Trans. Multimedia Comput. Commun. Appl.*, vol. 9, no. 1, 2013.

[7] J. Togelius, E. Kastbjerg, D. C. Schedl, and G. N. Yannakakis, "What is procedural content generation?: Mario on the borderline," in *Proceedings of the 2nd PCGames*, 2011.

[8] B. Horn, S. Dahlskog, N. Shaker, G. Smith, and J. Togelius, "A comparative evaluation of procedural level generators in the mario ai framework," in *Proceedings of the 14th FDG*, 2014.

[9] Y. Dong and T. Barnes, "Evaluation of a template-based puzzle generator for an educational programming game," in *Proceedings of the 12th FDG*, 2017.

[10] L. Rodrigues, R. P. Bonidia, and J. D. Brancher, "A math educational computer game using procedural content generation," in *Proceedings of the SBIE*, 2017.

[11] G. Smith and J. Whitehead, "Analyzing the expressive range of a level generator," ser. Proceedings of the 1st PCGames, 2010.

[12] O. Korn, M. Blatz, A. Rees, J. Schaal, V. Schwind, and D. Görlich, "Procedural content generation for game props? a study on the effects on user experience," *Comput. Entertain.*, vol. 15, no. 2, 2017.

[13] C. Bauckhage, K. Kersting, R. Sifa, C. Thurau, A. Drachen, and A. Canossa, "How players lose interest in playing a game: An empirical study based on distributions of total playing times," in *IEEE CIG*, 2012.

[14] G. Sim and M. Horton, "Investigating children's opinions of games: Fun toolkit vs. this or that," in *Proceedings of the 11th IDC*, 2012

[15] R. Paiva, I. I. Bittencourt, T. Tenório, P. Jaques, and S. Isotani, "What do students do on-line? modeling students' interactions to improve their learning experience," *Computers in Human Behavior*, vol. 64, 2016.

[16] H. Desurvire and M. S. El-Nasr, "Methods for game user research: Studying player behavior to enhance game design," *IEEE Computer Graphics and Applications*, vol. 33, no. 4, 2013.

[17] J. R. H. Mariño, W. M. P. Reis, and L. H. S. Lelis, "An empirical evaluation of evaluation metrics of procedurally generated mario levels," in *Proceedings of the 11th AIIDE*, 2015.

[18] G. N. Yannakakis and J. Togelius, "Experience-driven procedural content generation," *IEEE Transactions on Affective Computing*, vol. 2, no. 3, 2011.

[19] A. M. Connor, T. J. Greig, and J. Kruse, "Evaluating the impact of procedurally generated content on game immersion," *The Computer Games Journal*, vol. 6, no. 4, 2017.

[20] E. Butler, E. Andersen, A. M. Smith, S. Gulwani, and Z. Popović, "Automatic game progression design through analysis of solution features," in *Proceedings of the 33rd CHI*, 2015.

[21] C. Wohlin, P. Runeson, M. Hst, M. C. Ohlsson, B. Regnell, and A. Wessln, *Experimentation in Software Engineering*. Springer Publishing Company, Incorporated, 2012.

[22] L. Rodrigues and J. D. Brancher, "Playing an educational game featuring procedural content generation: Which attributes impact players' curiosity?" *RENOTE*, 2019, in press.

[23] L. Rodrigues and J. D. Brancher, "Improving players' profiles clustering from game data through feature extraction," in *Proceedings of SBGames 2018 - Computing Track*, 2018.

[24] W. Oliveira, L. Rodrigues, A. M. Toda, P. T. Palomino, S. Isotani, "Automatic Game Experience Identification in Educational Games," in *Proceedings of the SBIE*, 2019.

[25] L. Rodrigues, "Assessing the influences of procedural level generation through a digital math game: An empirical analysis," Master's thesis, Londrina State University, 2018.

[26] C. Moser, V. Fuchsberger, and M. Tscheligi, "Rapid assessment of game experiences in public settings," in *Proceedings of the 4th the FnG*, 2012