Procedural Level Generation to Improve a Digital Math Game's Development: Does it Impact Player Experience?

Luiz Rodrigues[1] **and Jacques Brancher**[1] **(advisor)**

[1]Department of Computer Science - Londrina State University - Londrina, Brazil

luiz_rodrigues17@hotmail.com, jacques@uel.br

Abstract. *Procedural content generation can improve the game development process, however, few studies evaluated 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. Additionally, we validated the game and analyzed both the relationship between fun, willingness to play the game again (i.e., returnance), and curiosity, and 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. 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. The findings also 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, showing the impact of procedural level generation on players, and how and which factors might play a role in their experiences.*

1. Introduction

Some students perceive math as a difficult subject, do not like it, and consider it displeasing [Biswas et al. 2001], 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 [Madeira et al. 2015]. Digital Math Games (DMG) might be used to address it, improving aspects such as students learning [McLaren et al. 2017], positive attitudes towards the subject [Ke 2008], and engagement [Kiili and Ketamo 2017]. Despite that, the development process of DMG is a slow and costly task, even for general purpose games, which commonly requires several designers, artists, and developers [Hendrikx et al. 2013].

An alternative to tackle these problems is the Procedural Content Generation (PCG) [Shaker et al. 2016], a reliable tool to provide diversified, automatically generated outputs that can be controlled through generation parameters [Horn et al. 2014], while automating, aiding in creativity, and speeding up the creation of various types of game contents [Hendrikx et al. 2013]. In spite of that, few studies have applied it on educational games [Dong and Barnes 2017, Rodrigues et al. 2017]. Additionally, a limitation of PCG literature is that most studies focus on algorithms capacities [Smith and Whitehead 2010],

DOI: 10.5753/cbie.wcbie.2019.1032

failing to demonstrate the true impact of automating content generation from players' perspective [Korn et al. 2017]. This is important because technologies must provide positive experiences, especially for children; otherwise, players are unlikely to interact or accept it [Bauckhage et al. 2012, Sim and Horton 2012] and might have their learning experience harmed [Paiva et al. 2016]. 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 [Desurvire and El-Nasr 2013] 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*), aiming to demonstrate 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.

2. Background and Related Works

PCG refers to creating contents automatically without or with limited human intervention [Shaker et al. 2016]. 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 [Smith and Whitehead 2010]. However, that approach is insufficient to replace user-based studies [Mariño et al. 2015], 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 [Shaker et al. 2016], or through A/B comparisons to identify PCG's impact [Connor et al. 2017]. Hence, the only approach that reveals the impacts of PCG usage is the A/B test method [Korn et al. 2017], 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. 2015], 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 played almost as much as the original. In [Korn et al. 2017], 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. 2017], 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. This context demonstrates the literature on PCG's impact has mixed results, which shows the necessity of further research.

From those studies, only two addressed the impacts of level generation [Butler et al. 2015, Connor et al. 2017], and a single study used an educational game as

the testbed [Butler et al. 2015]. 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 [Wohlin et al. 2012]. Furthermore, the reduced sample size [Connor et al. 2017] and the lack of evidence concerning the groups of players compared [Butler et al. 2015] also threats related works [Wohlin et al. 2012]. 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.

3. Materials and Method

Given our goal of demonstrating PCG's impact, we mainly sought 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 [Connor et al. 2017, Butler et al. 2015], 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. Figure 1 summarizes both the setup and the procedure mentioned.

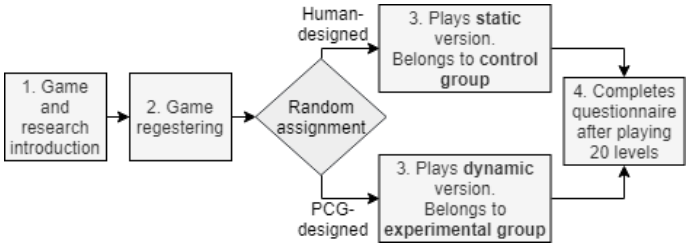


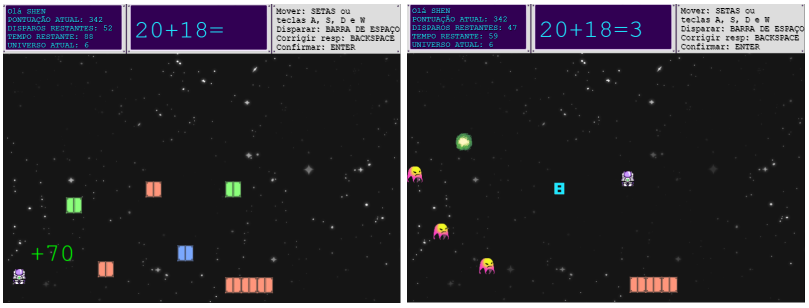
Figure 1. Study's method and participants groups.

3.1. Testbed Game

To enable this research, we developed *SpaceMath*[1] [Rodrigues and Brancher 2019, Oliveira et al. 2019, Rodrigues and Brancher 2018], a DMG that fosters the practice of

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

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. Arithmetic problems are procedurally generated, following a template-based approach which guarantees that all problems have exactly two numbers and that all solutions are positive integer numbers. Game levels are procedurally generated through a straight-forward constructive algorithm [Rodrigues et al. 2017], which behaves in the same way regardless of players' characteristics, and that allows the generation process to increase their difficulty level as players' win-streak increases. Moreover, a set of 20 game levels were manually designed by a experienced game developer to provide a baseline comparison, which were arranged in a way that, as players' win-streak increases, those levels difficulty increase accordingly, similarly to the PCG method, to promote gameplays similar to each other.



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

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

3.2. Measures

We compared groups based on six PX metrics, considering their opinions (four) and in-game behavior (two). Players reported their opinions through an adapted version of the questionnaire for rapid assessment by [Moser et al. 2012] after playing 20 levels, which captured PX in terms of curiosity (composed of seven statements, referred to as C1, C2, up to C7) [Rodrigues and Brancher 2019], fun, *returnance*, 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. The four opinion-based metrics were captured as follows. **Fun** was captured using the Smileyometer from the Fun-Toolkit [Read et al. 2009]. It provides a simple and intuitive way to players indicate this factor. It was encoded as a rating, on a five-point scale ranging from 1 to 5, where higher values indicate more fun. **Description of Experience** investigates PX in a deeper way than the Smileyometer. It is based on predefined opposed attributes in order to have a semantic balance. This approach was inspired by its usage in [Moser et al. 2012]. Players could select none, one, or many of the following attributes: simple - difficult; great - childish; fun - boring; exciting - tiring; and intuitive - confusing. **Returnance** identifies 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 from the Fun-Toolkit, the Again Again Table [Read et al. 2009]. **Curiosity** was adapted from the questionnaire used in [Wouters et al. 2011]. It was captured through the following questions, that were encoded as ratings in a five-point scale ranging from 1 (completely disagree) to 5 (completely agree):

- The game motivated me to learn more about math;
- I wanted to continue playing because I wanted to see more about the game levels;
- Playing the game raised questions about the game levels;
- I was curious about the next event in the game;
- I sought explanations for what I encountered in the game;
- Playing the game raised questions regarding math;
- I wanted to continue playing because I wanted to know more about math.

Additionally, in-game data were captured throughout the process as well, which enabled the analysis of players' performance and in-game behavior. These data are: Average score per level; maximum summed score achieved; average of shots fired per level; average of time spent to complete each level in seconds; total time spent playing the game in seconds; largest sequence of wins achieved; total of played levels; and total of wins in all levels. Based on these, besides players' performance, we measured their in-game behavior in terms of **retention** (i.e., played 20 levels or more, thus, were retained until answering the questionnaire) and **engagement** (i.e., how many levels each group played, thus, to what extent they were engaged in continuing playing the game), similar to [Butler et al. 2015].

3.3. Data Analysis Process

First, we found both groups' demographics were insignificantly different, preventing threats that could emerge from comparing data from players with different characteristics [Wohlin et al. 2012]. Second, we compared the opinions of participants of both groups ($N_{control}$ = 242; $N_{experimental}$ = 265) through Kruskal-Wallis and, then, their in-game behavior ($N_{control}$ = 355; $N_{experimental}$ = 369) using Chi-squared homogeneity and Mann-Whitney hypothesis tests. Third, 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. All analyses were performed based on a 95% confidence level, using hypothesis tests suitable for data types, selected following similar research available in the literature (e.g., [Connor et al. 2017, Butler et al. 2015]) after assessing data normality via the Shapiro-Wilk test, when necessary.

4. Results

The difference between the experiences from each version was insignificant in all PX metrics but one question of the curiosity metric, the *I sought explanations for what I encountered in the game* (C5) statement (see Figures 3 and 4 and Table 1). Further investigating this concern, we found this difference was significant only for: females, gamers, and those with internet access through a computer at home. Also, the age's influence on PX 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 and engagement were insignificant, whereas their performances were mostly significantly

different (see Table 2). Thus, game versions differed in only one of the nine self-reported question assessed and in players' performance, wherein demographic attributes showed insights from where these differences emerged, whereas in-game behavior did not show significant differences.

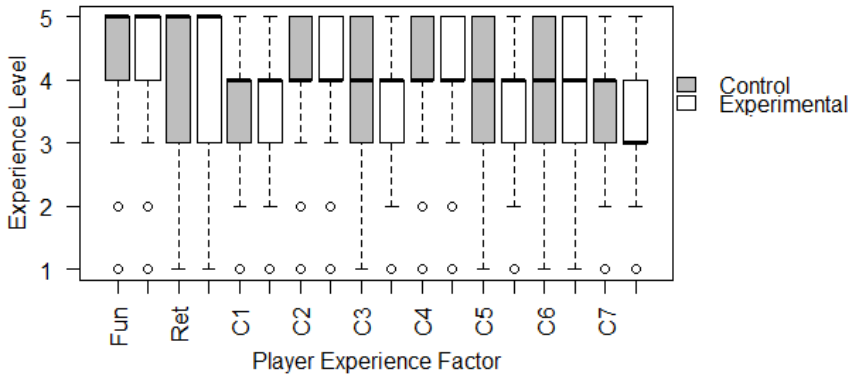


Figure 3. Boxplot of fun, *returnance*, and curiosity from participants of both versions.

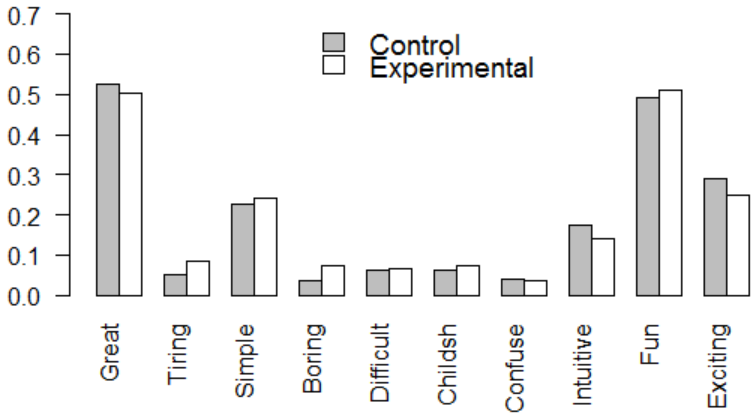


Figure 4. Barplot of the experience description from players of both versions.

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 (see Figure 5). Additionally, our analyses demonstrated 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 (see Table 3). On the other hand, whilst curiosity is associated with genre, being a gamer and having internet access through a computer at home, *returnance* is associated to none, while fun depends on being a gamer only. In addition, we found players' performance metrics have small significant negative correlations to their experience, with the exception of average shots per level (see Table 4). Moreover, we found win-streak had a strong negative correlation to average score, showing the PLG increased the levels' difficulty as expected.

Table 1. Comparison of groups experience. Data represented as Mean (SD).

PXF	Control	Experimental	KW-χ^2
Fun	4.446 (0.778)	4.430 (0.868)	0.104
RET	4.397 (1.039)	4.253 (1.194)	1.373
C1	3.789 (0.982)	3.642 (1.043)	2.593
C2	4.227 (0.836)	4.155 (0.872)	1.028
C3	3.888 (0.897)	3.815 (0.917)	0.786
C4	4.161 (0.885)	4.042 (0.889)	3.166
C5	3.810 (1.041)	3.562 (1.123)	6.583*
C6	3.988 (0.979)	3.849 (1.077)	1.585
C7	3.450 (1.201)	3.411 (1.219)	0.161

* $p < 0.05$; KW-χ^2 = Kruskal-Walis statistic

Table 2. 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*

* $p < 0.05$; U test = Mann-Whitney statistic

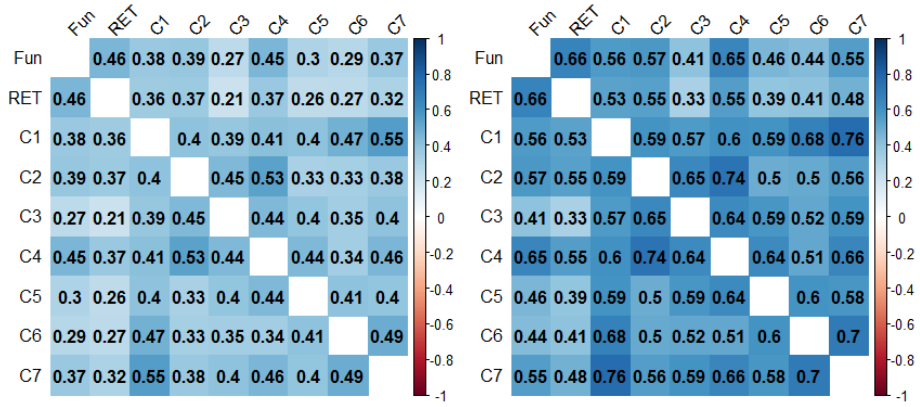
5. Main Contributions

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 and Education, introducing, validating, and showcasing the impacts of using a technique to improve the development of a DMG. In summary, the contributions are: (i) A DMG that encourages its players to practice math and provides 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 aspect of one PX metric; (iv) to reveal demographic characteristics associated with PX as well as how in-game performance is correlated with their experiences; (v) to

Table 3. Correlation degree from demographics to PX factors.

Attribute	Fun	*Returnance*	Curiosity
Age	-0.199 (-0.307)*	-0.226 (-0.347)*	-0.226 (-0.347)*
School Stage	-0.157 (-0.244)*	-0.123 (-0.192)*	-0.186 (-0.288)*
Weekly Playing Hours	-0.010 (-0.015)	-0.062 (-0.098)	-0.018 (-0.029)
School Type	-0.006 (-0.010)	-0.028 (-0.044)	0.021 (0.033)
Likes Math	0.181 (0.280)*	0.173 (0.269)*	0.215 (0.331)*
Knows Math	0.131 (0.204)*	0.069 (0.108)	0.095 (0.148)*

*$p < 0.05$



(a) Kendall's coefficient.

(b) Kendall's coefficient converted to Pearson's.

Figure 5. Degree of correlation between PX factors.

Table 4. Correlation degree from performance to PX factors.

Metric	Fun	*Returnance*	Curiosity
Average Score	-0.144 (-0.225)*	-0.164 (-0.255)*	-0.175 (-0.271)*
Average Time	0.090 (0.141)*	0.099 (0.154)*	0.098 (0.154)*
Average Shots	-0.071 (-0.111)*	-0.082 (-0.128)*	-0.060 (-0.094)
Highest Level	-0.085 (-0.133)*	-0.129 (-0.201)*	-0.073 (-0.114)*
Wins Rate	-0.089 (-0.140)*	-0.126 (-0.197)*	-0.121 (-0.189)*
Maximum Score	-0.112 (-0.176)*	-0.139 (-0.217)*	-0.102 (-0.160)*
Total Time	0.094 (0.147)*	0.106 (0.165)*	0.105 (0.164)*

*p $<$ 0.05

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-CBIE-Rodrigues2019`

6. Concluding Remarks

This study analyzed the influences of PCG 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. As main future works, we suggest performing similar research to ground PCG's impacts, also evaluating the impacts of PCG on learning gains, and exploiting our findings to derive models of PX aiming to design personalization mechanisms.

Acknowledgements

This research was financed in part by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES) - Finance Code 001.

References

Bauckhage, C., Kersting, K., Sifa, R., Thurau, C., Drachen, A., and Canossa, A. (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)*, pages 139–146.

Biswas, G., Katzlberger, T., Bransford, J., and Schwartz, D. (2001). Extending intelligent learning environments with teachable agents to enhance learning. *Artificial Intelligence in Education*, pages 389–397.

Butler, E., Andersen, E., Smith, A. M., Gulwani, S., and 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, pages 2407–2416, New York, NY, USA. ACM.

Connor, A. M., Greig, T. J., and Kruse, J. (2017). Evaluating the impact of procedurally generated content on game immersion. *The Computer Games Journal*, 6(4):209–225.

Desurvire, H. and El-Nasr, M. S. (2013). Methods for game user research: Studying player behavior to enhance game design. *IEEE Computer Graphics and Applications*, 33(4):82–87.

Dong, Y. and Barnes, T. (2017). Evaluation of a template-based puzzle generator for an educational programming game. In *Proceedings of the 12th International Conference on the Foundations of Digital Games*, FDG '17, pages 40:1–40:4, New York, NY, USA. ACM.

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

Horn, B., Dahlskog, S., Shaker, N., Smith, G., and Togelius, J. (2014). A comparative evaluation of procedural level generators in the mario ai framework. In *Foundations of Digital Games 2014*.

Ke, F. (2008). A case study of computer gaming for math: Engaged learning from game-play? *Computers & Education*, 51(4):1609 – 1620.

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

Korn, O., Blatz, M., Rees, A., Schaal, J., Schwind, V., and Görlich, D. (2017). Procedural content generation for game props? a study on the effects on user experience. *Comput. Entertain.*, 15(2):1:1–1:15.

Madeira, C., Câmara, L., Beserra, I., and Tavares, R. (2015). Mathmare: um jogo de plataforma envolvendo desafios matemáticos do ensino médio. In *Proceedings of the Brazilian Symposium on Computer Games and Digital Entertainment (SBGames 2015)*. In portuguese.

Mariño, J. R. H., Reis, W. M. P., and Lelis, L. H. S. (2015). An empirical evaluation of evaluation metrics of procedurally generated mario levels. In *Proceedings of the Eleventh Artificial Intelligence and Interactive Digital Entertainment*, pages 44–50.

McLaren, B. M., Adams, D., Mayer, R. E., and Forlizzi, J. (2017). A computer-based game that promotes mathematics learning more than a conventional approach. *International Journal of Game-Based Learning (IJGBL)*, 7:36–56.

Moser, C., Fuchsberger, V., and Tscheligi, M. (2012). Rapid assessment of game experiences in public settings. In *Proceedings of the 4th International Conference on Fun and Games*, FnG '12, pages 73–82, New York, NY, USA. ACM.

Oliveira, W., Rodrigues, L., Toda, A., Palomino, P., and Isotani, S. (2019). Automatic game experience identification in educational games. In *Proceedings of the Brazilian Symposium on Computers in Education (SBIE 2019)*.

Paiva, R., Bittencourt, I. I., Tenório, T., Jaques, P., and Isotani, S. (2016). What do students do on-line? modeling students' interactions to improve their learning experience. *Computers in Human Behavior*, 64:769–781.

Read, J., MacFarlane, S., and Casey, C. (2009). Endurability, engagement and expectations: Measuring children's fun. *Interaction Design and Children*.

Rodrigues, L., Bonidia, R. P., and Brancher, J. D. (2017). A math educational computer game using procedural content generation. In *Proceedings of the Brazilian Symposium on Computers in Education (SBIE 2017)*.

Rodrigues, L. and Brancher, J. D. (2018). Improving players' profiles clustering from game data through feature extraction. In *Proceedings of the Brazilian Symposium on Computer Games and Digital Entertainment (SBGames 2018) - Computing Track*.

Rodrigues, L. and Brancher, J. D. (2019). Playing an educational game featuring procedural content generation: Which attributes impact players' curiosity? *Revista Novas Tecnologias (RENOTE)*.

Shaker, N., Togelius, J., and Nelson, M. J. (2016). *Procedural Content Generation in Games: A Textbook and an Overview of Current Research*. Springer.

Sim, G. and 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, pages 70–77, New York, NY, USA. ACM.

Smith, G. and Whitehead, J. (2010). Analyzing the expressive range of a level generator. In *Proceedings of the 1st International Workshop on Procedural Content Generation in Games (PCGames '10)*.

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

Wouters, P., van Oostendorp, H., Boonekamp, R., and 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.