An Empirical Evaluation of Evaluation Metrics of Procedurally Generated Mario Levels

Julian R. H. Mariño and **Willian M. P. Reis** and **Levi H. S. Lelis**
Departamento de Informática
Universidade Federal de Viçosa
Viçosa, Minas Gerais, Brazil

Abstract

There are several approaches in the literature for automatically generating Infinite Mario Bros levels. The evaluation of such approaches is often performed solely with computational metrics such as leniency and linearity. While these metrics are important for an initial exploratory evaluation of the content generated, it is not clear whether they are able to capture the player's perception of the content generated. In this paper we evaluate several of the commonly used computational metrics. Namely, we perform a systematic user study with procedural content generation systems and compare the insights gained from our user study with those gained from analyzing the computational metric values. The results of our experiment suggest that current computational metrics should not be used in lieu of user studies for evaluating content generated by computer programs.

Introduction

Automatic generation of good-quality content is a long-term goal in Artificial Intelligence (AI), where content could mean levels of a computer game, stories, sport commentaries, and others. The research area of automatic content generation is known as Procedural Content Generation (PCG), and in this paper we refer to systems which automatically generate content as PCG systems. We focus the present work on the problem of automatically generating levels of the game of Infinite Mario Bros (IMB), a variant of Super Mario Bros (SMB). IMB has recently received a lot of attention from AI researchers—see Togelius et al. (2011) for a review. The reason for its popularity amongst AI researchers is that IMB is an excellent testbed for PCG systems: the game is simple enough to allow researchers to quickly experiment with novel PCG approaches and yet quite entertaining.

There are various approaches in the literature for automatically generating IMB levels. The evaluation of such approaches is often performed solely with computational metrics such as *leniency* and *linearity* (Smith and Whitehead 2010; Horn et al. 2014). While these metrics are important for performing initial exploratory evaluations of the levels generated, it is not clear whether they are able to capture the player's perception of the content generated. A focus in many PCG research projects is to know whether the content generated has good quality from the player's perspective, and the literature lacks a systematic evaluation of the computational metrics used for evaluating PCG systems.

The contributions of this paper are empirical. Namely, we perform a systematic user study with IMB PCG systems and compare the insights gained from our study with those gained from analyzing a set of commonly used computational metrics. As an example of the results we present in this paper, all computational metrics used in our experiment rated the levels generated by two PCG systems very similarly, while subjects in our user study found that the levels generated by one of the systems were significantly more enjoyable to play than the levels generated by the other system. As another example, the computational metric of leniency, which was designed to approximate the difficulty of a given level, only weakly correlated with the difficulty rated by the subjects in our study. Perhaps the most important conclusion one can draw from our experiments is that although the computational metrics can be valuable for an initial exploratory study of the content generated by PCG systems and for verifying the diversity of levels a PCG system can generate, these metrics should not replace user studies for analyzing the player's perception of the content generated.

The current paper is of interest to the AI community because our results provide valuable insight on how to design experiments for evaluating PCG systems and potentially on how to develop novel computational metrics for guiding the AI search process of procedurally generating IMB levels.

This paper is organized as follows. First, we provide a literature review of evaluation strategies of PCG systems for Mario games. Following, we explain the computational metrics evaluated in our experiment. Finally, we describe our experiment and discuss the results obtained.

Review of Evaluation Strategies

In this section we review strategies used by others for evaluating IMB and SMB PCG systems. The literature in evaluation strategies for games in general is much broader than the review we present in this paper; we focus on Mario games.

We group the works in our review as follows: works using user studies to evaluate PCG systems (user studies for evaluation), works using user studies not to evaluate PCG systems, but to collect data to learn predictive models (user

studies for data collection), works using computational metrics and/or artificial agents to evaluate PCG systems (computational evaluation), and works that evaluate PCG systems with self critique or some other sort of evaluation.

User Studies for Evaluation

The Mario AI Competition (Togelius et al. 2013) offers a user study for comparing different PCG systems. In contrast with the competition, which is interested in ranking different PCG systems, in this paper we perform a user study to evaluate computational metrics.

Shaker et al. (2010) describe a system for generating adaptive player-tailored IMB levels. Their system directly asks questions to the players about their preferences. Shaker et al.'s main experiment is carried out with artificial agents, but an experiment with human subjects compares the proposed adaptive approach with a non-adaptive one. Dahlskog and Togelius (2013) also present a user study comparing different levels of SMB. Bakkes et al. (2014) describe a system for balancing game challenging in IMB levels; their system is also evaluated with human subjects.

User Studies for Data Collection

Pedersen et al. (2009) presented a system for modeling player experience based on empirical data collected in a user study. They were aiming at learning statistical models for predicting, given an IMB level $\mathbf{L}$, the challenge $\mathbf{L}$ will offer to the player, and how much enjoyment and frustration the player will have while playing $\mathbf{L}$. Similarly, in different works, Shaker et al. (2011; 2012; 2013) showed how to extract features to learn predictive models of the player's experience in IMB. Pederson et al.'s and Shaker et al.'s long-term goal is to use these models to guide the search for good-quality player-tailored IMB levels. By contrast, in this paper we are not interested in models for guiding the process of generating good-quality IMB levels, we are interested in comparing different strategies for evaluating PCG systems.

Computational Evaluation

Smith and Whitehead (2010) and Horn et al. (2014) introduced several computational metrics for evaluating what the authors called the expressivity of PCG systems—we describe some of these metrics below. Their metrics were used in several works as a form of evaluating PCG systems.

For example, Smith et al. (2010) presented Tanagra, a system for generating levels of 2D-platform games which was evaluated solely with computational metrics. Later, Smith et al. (2011) presented Launchpad, a system that uses rhythm groups to generate levels of platform games. Launchpad was also evaluated with computational metrics. Shaker et al. (2012a) use a grammar to concisely encode design constraints for evolving IMB levels. Shaker et al.'s system is also solely evaluated with computational metrics similar to those used by Smith et al. to evaluate Tanagra and Launchpad. In another work Shaker et al. (2012b) evaluate the personalized content generated by a grammar-based PCG system with artificial agents. In a recent work Shaker and Abou-Zleikha (2014) used non-negative matrix factorization to generate levels based on patterns learned from levels generated by other systems; their method is also evaluated with computational metrics.

Dahlskog et al. (2014) use n-grams created from original levels of SMB to generate novel levels of the game. In two other works Dahlskog and Togelius (2014a; 2014b) present systems which use patterns to generate levels of the game of SMB. All these works were evaluated with the computational metrics introduced by Smith and Whitehead (2010).

Sorenson et al. (2011) presented a system which uses the idea of rhythm groups introduced by Smith et al. (2008) to define a computational model of player enjoyment to evolve levels of IMB. This model is also used to evaluate the resulting levels. Sorenson et al. also evaluate their system in terms of the results of the Mario AI Competition.

Other Evaluation Strategies

Some PCG systems are evaluated neither with user studies nor with computational metrics. For example, the Occupancy-Regulated Extension (ORE) PCG system is evaluated by the authors themselves with an analysis of the levels generated (Mawhorter and Mateas 2010). Kerssemakers et al. (2012) also presents a self critique of the proposed approach in addition to an empirical running time analysis of the system. The seminal work of Compton and Mateas (2006) on content generation for platform games does not present evaluations of the proposed approach.

Recently, Canossa and Smith (2015) introduced several metrics based on design theory for evaluating IMB levels. Their metrics were obtained through discussions with design students. However, in contrast with the computational metrics introduced by Smith and Whitehead (2010) and Horn et al. (2014), some of Canossa and Smith's metrics are not formal enough to be implemented as a computer procedure.

Most of the user studies in the PCG literature of Mario Bros games have been performed to collect data to learn predictive models of the player's perception of the game. Exceptions include the Mario AI competition, and the works of Shaker et al. (2010), Dahlskog and Togelius (2013), and Bakkes et al. (2014). A large number of PCG systems have been solely evaluated with computational metrics similar to the ones introduced by Smith and Whitehead (2010). It is not clear whether such metrics provide insights about the player's perception of the generated levels. Our work is the first to systematically compare the computational metrics with the human's perception of the generated levels.

The PCG Problem for Infinite Mario Bros

In this paper we are interested in the problem of evaluating the content generated by PCG systems for the game of IMB. The levels of IMB are grid spaces containing a set of objects such as platforms, mountains and shooting cannons. Every object is associated with a location on the grid (x and y coordinates) and some of the objects such as mountains can have different heights and widths.

Let $\mathbf{L} = \{o_1, o_2, \cdots, o_n\}$ be a level of IMB where $o_1, o_2, \cdots, o_n$ are the n objects composing the level. The PCG problem for IMB is to choose the set of objects in $\mathbf{L}$

as well as the objects' x and y coordinates. For some of the objects such as pits and mountains the PCG system also needs to define their height and width values. In this paper we assume that the goal in PCG for IMB is to generate levels which are both visually appealing and enjoyable to play.

Computational Metrics

In this section we describe the computational metrics used in our experiment: *linearity* and *leniency* introduced by Smith and Whitehead (2010), *density*, and *Compression Distance* introduced by Shaker et al. (2012a). Similarly to previous works, to ease the presentation of the results, we normalize all metrics to the $[0, 1]$ interval. Normalization is performed by accounting for the levels generated by all systems evaluated. Thus, the metric values we present in this paper are not directly comparable to the values presented in other works as the normalized values depend on the systems evaluated. We note that the normalization we perform does not affect the results of our experiment.

Linearity The linearity of level $\mathbf{L}$ is computed by performing a linear regression on the center points of each platform and mountain contained in $\mathbf{L}$. The linearity of $\mathbf{L}$ is the average distance between the center points and the linear regression's line. Normalized values closer to one indicate more linear levels. The linearity of a PCG system ρ is the average normalized linearity of the levels ρ generates.

Linearity measures the changes in height (y-coordinate) the player experiences while going through the level. Smith et al. (2011) pose the linearity metric as a visual aesthetics metric, which is reasonable since levels with different linearity values are expected to look different from one another.

Leniency Leniency measures how much challenge the player is likely face while playing the level. The leniency of level $\mathbf{L}$ is the sum of the lenience value $w(o)$ of all objects o in $\mathbf{L}$, defined as $\sum_{o \in \mathbf{L}} w(o)$. We use the lenience values specified by Shaker et al. (2012a). Namely, power-up items have a weight of 1, cannons, flower tubes, and gaps of -0.5, and enemies of -1. We subtract the average gap width of the level from the resulting sum as defined by Shaker et al. Leniency is meant to approximate the difficulty of the levels. Normalized values closer to one indicate more lenient levels. The leniency of a PCG system ρ is the average normalized leniency of the levels ρ generates.

Density Mountains can occupy the same x-coordinate on the grid defining a IMB level by being "stacked-up" together. The density of $\mathbf{L}$ is the average number of mountains occupying the same x-coordinate on the grid. Intuitively, a level with high density could have different layouts and challenges than a level with low density. Normalized values closer to one indicate denser levels. The density of a PCG system ρ is the average normalized density of the levels ρ generates.

Compression Distance The Compression Distance (CD) measures the structural dissimilarity of a pair of levels. CD is computed as follows. First, we convert the pair of levels $\mathbf{L}$ and $\mathbf{L}'$ into two sequences of integers $\mathbf{S}$ and $\mathbf{S}'$, respectively. Each integer in $\mathbf{S}$ represents one of the following in $\mathbf{L}$: (i) an increase or decrease in the platform's height, (ii) the existence or the nonexistence of enemies and items, and (iii) the beginning or ending of a gap. The conversion of $\mathbf{L}$ into $\mathbf{S}$ is done by traversing the level's grid from left to right and for each x-value on the grid we insert the appropriate integer into the converted sequence (e.g., the integer 1 in position 10 could represent an enemy at x-coordinate 10 of the level). Intuitively, if sequences $\mathbf{S}$ and $\mathbf{S}'$ are very different, then one would expect $\mathbf{L}$ and $\mathbf{L}'$ to be structurally different.

The CD value of a PCG system ρ is the average normalized compression metric (Li et al. 2004) of $\mathbf{S}$ and $\mathbf{S}'$ for pairs of levels $\mathbf{L}$ and $\mathbf{L}'$ ρ generates. Normalized values closer to one indicate that the PCG system is able to generate levels with a larger structural variety.

Evaluating Evaluation Metrics

We now evaluate the computational metrics described above. First we describe the methodology of our experiment. Then, we present the results of the computational metrics, followed by the results of the user study. Finally, we discuss the insights gained from each evaluation.

Methodology

We now describe our experimental methodology.

Systems Tested We used four different IMB PCG systems in our experiments: Notch Level Generator (NLG), Human-Computation Tension Arc-Based (HCTA) level generator with a random tension arc (HCTA+R) and with a parabolic tension arc (HCTA+P) (Reis, Lelis, and Gal 2015), and Occupancy-Regulated Extension (ORE) generator (Mawhorter and Mateas 2010).

The NLG system receives as input a difficulty value d for stochastically determining the number of enemies and challenges to be placed in the level. The levels NLG generates tend to be harder for larger values of d. NLG starts with an empty level grid and adds objects to the grid according to the value of d. HCTA+R and HCTA+P are variants of NLG. The HCTA systems work by having human subjects rating a set of small levels generated by NLG. Then, HCTA combines the small levels into a regular-sized IMB level according to the human-rated difficulty of the small levels. HCTA+P combines the small levels into a regular level in a way that the difficulty of the resulting level follows a parabolic curve: difficulty increases as the player progresses into the level until reaching its largest value, then difficulty decreases until the end of the level. HCTA+R combines the small levels in a way that the difficulty is random (but still respecting a user-specified upper bound) throughout the level. See Reis et al. (2015) for details on HCTA.

We chose NLG, HCTA+R, HCTA+P, and ORE for two reasons. First, the computational metrics will tend to give similar scores to the levels HCTA+R and HCTA+P generate for they both use similar strategies for level generation. Yet, levels generated by the HCTA systems could still be rated differently by the participants in the user study. Second, ORE generates levels which are structurally different from the ones the other systems generate, allowing us to verify whether the user study is able to capture nuances which

are likely to be captured by the computational metrics. Ideally we would use more systems in our experiment, but the time required for each participant to complete the experiment could be prohibitively long should we required them to play extra levels.

Participants Our within-subject experiment had 37 participants: 32 males and 5 females with an average age of 23.95 and standard deviation of 4.48. Each participant played one level generated by each system, resulting in the evaluation of 37 levels of each PCG system. The experiment was carried out online: our system was made available in the Internet and our experiment advertised in different mailing lists. Participation was anonymous and volunteered.

Evaluated Metrics In the user study the systems are evaluated according to the following criteria: enjoyment, visual aesthetics, and difficulty. Each participant was asked to answer how much they agreed or disagreed, in a 7-likert scale, with the following sentences: "This level is enjoyable to play"; "this level has good visual aesthetics"; "this level is difficult". A score of 1 for enjoyment and visual aesthetics means that the participant strongly agrees that the level played is enjoyable and has excellent visual aesthetics; a score of 1 for difficulty means that the participant strongly agrees that the level is difficult.

We compute the computational metric values only for the levels evaluated in our user study: 148 levels in total (37 levels for each of the four systems). This is to allow a fair comparison of the insights gained from the computational metrics with those gained from the user study.[1] The normalization of the computational metrics to the $[0, 1]$ interval was made by considering all 148 levels used in our experiment.

Experimental Design In the beginning of the experiment the subjects filled a questionnaire informing their age, and their skills in the game of Mario Bros. Subjects were instructed about the controls of the game before playing a practice level. The practice level is important so that the participants get acquainted with the controls of the game. NLG was used to generate the practice levels. Only after playing the practice level that the participants evaluated the levels generated by the PCG systems. Each participant played one level generated by each of the four PCG systems. After playing each level the participants gave scores according to the criteria described above in a 7-likert scale. In addition to the scores, the participants had the option to enter comments informing us of technical issues they might have had during the experiment. Since all participants played one level generated by each system, we used a balanced Latin square design to counteract ordering effects. The tested levels were generated during the experiment by the evaluated systems, we did not pre-select a set of levels to be tested.

In order to have a fair comparison of the levels generated by different systems we had all systems generating levels of the same size: 160×15. We chose this size because we did not want the experiment to be too long. In total each participant played 5 levels (1 practice level and 4 other levels for evaluation), and using larger levels could be tiring for the participants. Finally, to ensure a fair comparison of the different approaches, we tuned the systems to generate levels with similar difficulty. This was done by manually setting the d-values of NLG, HCTA+P, and HCTA+R so that the three systems generated levels which we thought to be of difficulty similar to the ones generated by ORE.

Data Cleaning The data of participants who did not finish playing all 5 levels (1 practice level plus 4 levels to be evaluated) is not included in the results. We also removed the data of one participant who had never played the game of Mario before. By examining the logs of the experiment we noticed that this participant was not able to get too far into the game and thus not able to properly evaluate the levels. The number of 37 participants was obtained after cleaning the data.

Computational Metric Results

We start by presenting the computational metric results. Although the computational metrics are systematic and do not represent a source of variance in our experiment, all PCG systems are stochastic and insert variance in the results. Moreover, as we explained above, the number of levels considered in this experiment is somewhat limited (37 levels for each system). Therefore, we present statistical tests for the computational metric results. Table 1 shows the average value and standard deviation for each metric and PCG system. Different letters in a given row of the table indicate that the two means are significantly different.

We now explain how the statistical significance was computed for the results in Table 1. First, we ran Shapiro-Wilk tests for each metric and verified that the leniency, density, and CD values were unlikely to be normally distributed. Thus, repeated-measures ANOVA was used only for linearity, and the test indicated statistically significant results ($p < .001$). The non-parametric Friedman test was applied to remaining metrics and indicated statistically significant results for leniency and density ($p < .001$), the differences in CD were not significant. Pairwise comparisons with Tukey tests for linearity showed that the only averages that are not significantly different are those of HCTA+P and HCTA+R, all other differences are significant ($p < .001$). Pairwise comparisons with Wilcoxon signed-rank tests for leniency and density showed that the averages that are not significantly different are those of the HCTA+P and the HCTA+R systems for both leniency and density, and ORE and NLG for leniency; all other results are statistically significant ($p < .001$).

We highlight the following observations from Table 1.

1. HCTA+P and HCTA+R generate similar levels as both systems scored similarly in all four metrics tested.
2. The average leniency value of the HCTA systems are much lower than ORE and NLG, indicating that the levels generated by HCTA are more difficult than those generated by the other two systems.
3. The HCTA approaches generate levels with nearly equal linearity averages, ORE generates highly non-linear lev-

[1] We also computed the computational metrics for a larger number of levels and observed results similar to the ones we report in this paper.

	HCTA+P	HCTA+R	ORE	NLG
Leniency	0.45 ± 0.10^a	0.48 ± 0.18^a	0.71 ± 0.11^b	0.77 ± 0.15^b
Linearity	0.52 ± 0.17^a	0.52 ± 0.15^a	0.33 ± 0.14^c	0.83 ± 0.07^b
Density	0.74 ± 0.13^a	0.73 ± 0.12^a	0.17 ± 0.15^c	0.49 ± 0.09^b
CD	0.61 ± 0.02^a	0.61 ± 0.02^a	0.60 ± 0.02^a	0.56 ± 0.02^a

Table 1: Computational metric results. Larger values of leniency, linearity, and density indicate are more lenient, linear, and dense levels; larger values of CD indicate that the PCG system is able to generate a larger variety of structurally different levels. Different letters in the same row indicate statistically significant results.

els, and NLG generates highly linear levels. The linearity results suggest that the HCTA approaches generate levels with similar visual aesthetics while NLG and ORE generate levels which are visually different than the levels generated by the other systems.

4. The density averages follow a pattern similar to linearity's: the HCTA approaches have very similar values while NLG and ORE differ from the other systems. The density results indicate that the HCTA approaches often use the pattern of superposing mountains while ORE rarely uses such a pattern. Similarly to linearity, the difference in the density average values show that the levels generated by ORE are visually different than the levels generated by other systems.

5. The difference on the average values of CD is minimal, indicating that all systems generate levels with similar structural diversity.

User Studies Results

We now present the user study results. The mean results and standard deviations are shown in Table 2. Different letters in a given row indicate that the two means are significantly different. Shapiro-Wilk tests showed that our data is unlikely to be normally distributed ($p < .0001$ for all criteria). Thus, we used the non-parametric Friedman test which showed a significant difference on enjoyment ($p < .05$) and on visual aesthetics ($p < .05$) across different systems; there was no significant difference for difficulty.

Next, we use Wilcoxon signed-rank tests to perform pairwise comparisons of the results obtained by the evaluated systems. We present the effect size of the comparisons (r-values) in addition to p-values. HCTA+P generates levels which are significantly more enjoyable to play than the levels HCTA+R generates ($p < .05$, $r = 0.21$) and the levels that ORE generates ($p < .001$, $r = 0.35$). The levels HCTA+R generates are significantly more enjoyable to play than the ones ORE generates ($p < .05$, $r = 0.26$). Finally, the levels NLG generates are significantly more enjoyable to play than the ones ORE generates ($p < .05$, $r = 0.24$).

Pairwise comparisons on visual aesthetics (Wilcoxon signed-rank test) showed that HCTA+P generates levels with significantly better visual aesthetics than the levels ORE generates ($p < .01$, $r = 0.27$) and than the levels NLG generates ($p < .05$, $r = 0.24$). HCTA+R generates levels with significantly better visual aesthetics than the levels ORE generates ($p < .01$, $r = 0.37$).

All pairwise comparisons reported as statistical significant have effect sizes around the medium size mark of 0.3, indicating substantial differences among the levels generated by the different systems.

We highlight the following observations from Table 2.

1. The system that generates the most enjoyable levels is HCTA+P. The difference between enjoyment of HCTA+P and HCTA+R is significant and substantial. That is, HCTA+P yielded an average score of 2.24 which is close to 2 (score marked by participants who *agreed* that the level played is enjoyable). By contrast, HCTA+R yielded an average score of 2.70 which is close to 3 (score marked by participants who *somewhat agreed* that the level played is enjoyable).

2. The HCTA approaches generated the levels with best visual aesthetics, followed by NLG and then ORE. In particular, HCTA+P generates levels with significantly better visual aesthetics than NLG and ORE.

3. There is little difference amongst the difficulty scores of the systems, indicating that the evaluated systems generate levels with similar difficulty.

Next, we discuss the strengths and weaknesses of the user study evaluation and of the computational evaluation by comparing the conclusions drawn from the two evaluations.

Strengths of the User Study Evaluation

We organize the discussion of the strengths of the user study evaluation by the evaluated criteria.

Enjoyment The user study shows a significant and substantial difference between the average enjoyment score of the levels generated by HCTA+P and by HCTA+R, while the computational evaluation yielded nearly the same score for both systems in all metrics. Spearman correlation tests between enjoyment and the computational metrics yielded coefficients close to zero, indicating that none of the metrics correlated with enjoyment.

Enjoyment is perhaps the most important evaluation criterion for PCG systems as we are interested in generating content which users find enjoyable to play. The computational metrics used in our experiment were not able to estimate the player's enjoyment. This result is not surprising. First, none of the computational metrics used in the literature were designed for measuring enjoyment. Second, enjoyment is difficult to measure without accounting for human input as it depends on various factors such as cultural background.

	HCTA+P	HCTA+R	ORE	NLG
Enjoyment	2.24 ± 1.75^{a}	2.70 ± 1.91^{b}	3.35 ± 2.04^{c}	2.62 ± 2.00^{ab}
Visual Aesthetics	2.32 ± 1.65^{a}	2.38 ± 1.64^{ab}	3.43 ± 2.21^{c}	2.92 ± 1.93^{b}
Difficulty	3.46 ± 1.76^{a}	3.38 ± 1.72^{a}	3.27 ± 1.90^{a}	3.84 ± 2.35^{a}

Table 2: User study results. Lower values of enjoyment and visual aesthetics indicate levels which are more enjoyable to play and have better visual aesthetics; lower values of Difficulty indicate levels which participants found more challenging to play. Different letters in the same row indicate statistically significant results.

Our user study required the participants to answer questions after playing each level. Another promising way of receiving human input for evaluating PCG systems is by analyzing facial expressions of the players (Shaker and Shaker 2014; Tan, Bakkes, and Pisan 2014).

Visual Aesthetics Both linearity and density indicated that HCTA+P and HCTA+R would generate levels with similar visual aesthetics, while ORE and NLG would generate levels with different visual aesthetics. The user study indicates that the HCTA approaches have nearly the same score for visual aesthetics, while ORE and NLG have higher values (indicating worse visual aesthetics). While the computational metrics indicated levels with different visual aesthetics, the metrics are not able to distinguish good from bad aesthetics. By contrast, through the user study we are able to rank the systems with respect to the visual quality of the levels generated. A Spearman's test shows that linearity weakly correlates with visual aesthetics (coefficient of 0.18 and $p < .05$); none of the other metrics correlates with visual aesthetics.

Difficulty While there is a large difference in the leniency values of the systems tested, according to the user study, there is little or no difference in the difficulty rated by the participants. Difficulty is also an important criteria for evaluating PCG systems as it is closely related to enjoyment. That is, it is known that the Yerkes-Dodson law (Yerkes and Dodson 1908) applies to computer games in the sense that enjoyment will be maximum somewhere in between the largest and the smallest difficulty (Piselli, Claypool, and Doyle 2009). Thus, when comparing different PCG systems the difficulty of the levels generated should be controlled to yield a fair comparison of the systems. It is hard to automatically measure difficulty because difficulty depends on factors such as the disposition of objects on the level. For example, there could be a level full of enemies and challenges (non-lenient level according to the metric) but with an easy path for Mario to follow and win the game—in such cases leniency will be misleading. Although our leniency results were somewhat misleading, we observed a weak but significant correlation between leniency and difficulty—the Spearman's coefficient was of 0.199 with $p < .05$.

The weak correlations observed between the computational metrics and the human-evaluated criteria of visual aesthetics and difficulty indicate that there is hope that future research will develop novel computational metrics to automatically (without asking the user) estimate the visual aesthetics and difficulty of IMB levels.

Strengths of the Computational Evaluation

Although our experiment showed that the computational metrics can be misleading, this kind of automatic evaluation also has its strengths. In contrast with the user study, one can easily achieve statistical significance by computing the metrics for a large number of levels. Moreover, we found the metrics to be very easy to implement. Taken together, these features make the computational metrics an easy and cheap way to perform an initial exploratory evaluation of the content generated by PCG systems.

The computational metrics can be particularly useful for gaining insight on evaluation criteria which are hard to test in user studies. For example, if one wants to verify the diversity of levels generated by a given PCG system in a user study, then the participants would have to play several levels generated by the same system and then inform the diversity of levels played. If the subjects had to play several levels of each system, then the experiment would likely be too long to be practical. One could then use the CD metric—or another similar metric such as edit distance (Smith et al. 2011)—to gain insight on the structural diversity of levels generated.

Conclusions

In this paper we tested several computational metrics used to evaluate levels generated by PCG systems for the game of IMB. We conducted a user study for evaluating four PCG systems according to the following criteria: enjoyment, visual aesthetics, and difficulty. Then, we compared the results obtained in the user study with those obtained by using computational metrics. Our evaluation showed that the computational metrics (i) are not able to accurately estimate enjoyment, (ii) provides limited information about the visual aesthetics of the levels, and (iii) can be misleading with respect to the player's perceived difficulty. Yet, the computational metrics can provide important information by measuring features which are hard to be measured in user studies.

Perhaps the most important conclusion drawn from our experiment is that a well-designed user study cannot be replaced by the current computational metrics. However, we believe that the computational metrics are suitable to be used during the design process, for quick and easy exploratory evaluations of the PCG system.

Acknowledgements

This research was supported by CAPES, CNPq, FAPEMIG, and the program Science Without Borders. We thank Rob Holte, Noam Tractinsky, Kobi Gal, and the anonymous reviewers for their invaluable feedback on this research.

References

A, C., and Smith, G. 2015. Towards a procedural evaluation technique: Metrics for level design. In *FDG*. ACM.

Bakkes, S.; Whiteson, S.; Li, G.; Visniuc, G. V.; Charitos, E.; Heijne, N.; and Swellengrebel, A. 2014. Challenge balancing for personalised game spaces. In *Games Media Entertainment*, 1–8. IEEE Press.

Compton, K., and Mateas, M. 2006. Procedural level design for platform games. In *Conference on Artificial Intelligence and Interactive Digital Entertainment*, 109–111. AAAI Press.

Dahlskog, S., and Togelius, J. 2013. Patterns as objectives for level generation. In *Proceedings of the Workshop on Design Patterns in Games at FDG*.

Dahlskog, S., and Togelius, J. 2014a. A multi-level level generator. In *IEEE Conference on Computational Intelligence and Games*, 1–8.

Dahlskog, S., and Togelius, J. 2014b. Procedural content generation using patterns as objectives. In *Proceedings of the European Conference Applications of Evolutionary Computation*, 325–336.

Dahlskog, S.; Togelius, J.; and Nelson, M. J. 2014. Linear levels through n-grams. In *Proceedings of the International Academic MindTrek Conference*.

Horn, B.; Dahlskog, S.; Shaker, N.; Smith, G.; and Togelius, J. 2014. A comparative evaluation of level generators in the Mario AI framework. In *FDG*. ACM.

Kerssemakers, M.; Tuxen, J.; Togelius, J.; and Yannakakis, G. N. 2012. A procedural procedural level generator generator. In *Conference of Comp. Intell. and Games*, 335–341. IEEE.

Li, M.; Chen, X.; Li, X.; Ma, B.; and Vitnyi, P. M. B. 2004. The similarity metric. *IEEE Transactions on Information Theory* 50(12):3250 – 3264.

Mawhorter, P. A., and Mateas, M. 2010. Procedural level generation using occupancy-regulated extension. In *Conference of Comp. Intell. and Games*, 351–358. IEEE.

Pedersen, C.; Togelius, J.; and Yannakakis, G. N. 2009. Modeling player experience in Super Mario Bros. In *Conference on Computational Intelligence and Games*, 132–139. IEEE Press.

Piselli, P.; Claypool, M.; and Doyle, J. 2009. Relating cognitive models of computer games to user evaluations of entertainment. In Whitehead, J., and Young, R. M., eds., *FDG*, 153–160. ACM.

Reis, W. M. P.; Lelis, L. H. S.; and Gal, Y. 2015. Human computation for procedural content generation in platform games. In *Conference of Comp. Intell. and Games*. IEEE.

Shaker, N., and Abou-Zleikha, M. 2014. Alone we can do so little, together we can do so much: A combinatorial approach for generating game content. In *Conference on Artificial Intelligence and Interactive Digital Entertainment*, 167–173. AAAI Press.

Shaker, N., and Shaker, M. 2014. Towards understanding the nonverbal signatures of engagement in Super Mario Bros. In *Proceedings of the Conference on User Modeling, Adaptation, and Personalization*, 423–434.

Shaker, N.; Nicolau, M.; Yannakakis, G. N.; Togelius, J.; and O'Neill, M. 2012a. Evolving levels for Super Mario Bros using grammatical evolution. In *Conference of Comp. Intell. and Games*, 304–311. IEEE.

Shaker, N.; Yannakakis, G. N.; Togelius, J.; Nicolau, M.; and O'Neill, M. 2012b. Evolving personalized content for Super Mario Bros using grammatical evolution. In *Conference on Artificial Intelligence and Interactive Digital Entertainment*, 75–80. AAAI Press.

Shaker, N.; Yannakakis, G. N.; and Togelius, J. 2010. Towards automatic personalized content generation for platform games. In *Conference on Artificial Intelligence and Interactive Digital Entertainment*, 63–68. AAAI Press.

Shaker, N.; Yannakakis, G. N.; and Togelius, J. 2011. Feature analysis for modeling game content quality. In *Conference of Comp. Intell. and Games*, 126–133. IEEE.

Shaker, N.; Yannakakis, G. N.; and Togelius, J. 2012. Digging deeper into platform game level design: Session size and sequential features. In *EvoApplications*, volume 7248 of *Lecture Notes in Computer Science*. Springer.

Shaker, N.; Yannakakis, G.; and Togelius, J. 2013. Crowdsourcing the aesthetics of platform games. *IEEE Transactions on Computational Intelligence and AI in Games* 5(3):276–290.

Smith, G., and Whitehead, J. 2010. Analyzing the expressive range of a level generator. In *Proceedings of the Workshop on Procedural Content Generation in Games*, 1–7. ACM.

Smith, G.; Treanor, M.; Whitehead, J.; Mateas, M.; Treanor, M.; March, J.; and Cha, M. 2011. Launchpad: A rhythm-based level generation for 2d platformers. *IEEE Transactions on Computing Intelligence and AI in Games* 3(1):1–16.

Smith, G.; Cha, M.; and Whitehead, J. 2008. A framework for analysis of 2d platformer levels. In *ACM SIGGRAPH Symposium on Video Games*, 75–80. ACM.

Smith, G.; Whitehead, J.; and Mateas, M. 2010. Tanagra: a mixed-initiative level design tool. In *FDG*, 209–216. ACM.

Sorenson, N.; Pasquier, P.; and DiPaola, S. 2011. A generic approach to challenge modeling for the procedural creation of video game levels. *IEEE Transactions on Computing Intelligence and AI in Games* 3(3):229–244.

Tan, C. T.; Bakkes, S.; and Pisan, Y. 2014. Inferring player experiences using facial expressions analysis. In *Proceedings of the Conference on Interactive Entertainment*, 7:1–7:8. ACM.

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

Togelius, J.; Shaker, N.; Karakovskiy, S.; and Yannakakis, G. N. 2013. The Mario AI championship 2009-2012. *AI Magazine* 34(3):89–92.

Yerkes, R. M., and Dodson, J. D. 1908. The relation of strength of stimulus to rapidity of habit formation. *Journal of Comparative Neurology and Psychology* 18:459–482.