# Understanding Mario: An Evaluation of Design Metrics For Platformers

Adam Summerville
University of California, Santa Cruz
1156 High St.
Santa Cruz, CA 95064
asummerv@ucsc.edu

Julian R. H. Mariño
Universidade de São Paulo
Av. Trabalhador Sancarlense, 400
São Carlos, SP 13566-590
julianmarino@usp.br

Sam Snodgrass
Drexel University
3141 Chestnut St.
Philadelphia, PA 19104
sps74@drexel.edu

Santiago Ontañón
Drexel University
3141 Chestnut St.
Philadelphia, PA 19104
so367@drexel.edu

Levi H. S. Lelis
Universidade Federal de Viçosa
Av. PH Rolfs, S/N
Viçosa, MG 36571-000
levi.lelis@ufv.br

## ABSTRACT

Evaluating the output of content generators is still one of the key open research challenges in Procedural Content Generation (PCG). This paper presents a collection of metrics for evaluating the quality of platform game levels, and analyzes how well these metrics are able to capture the human-perceived *difficulty*, *visual aesthetics* and *enjoyment* of these levels. We show empirically, in the context of *Infinite Mario Bros* (IMB), that some of the proposed metrics yield correlation values with human ratings that are near empirical upper bounds derived from a human inter-rater agreement study. We also show that a simple linear regression model using a subset of our metrics as input features is able to substantially outperform a previous approach that uses a neural network for predicting human-perceived difficulty, visual aesthetics, and enjoyment in IMB levels.

## CCS CONCEPTS

**•General and reference → Metrics; Design; •Human-centered computing → User studies; •Applied computing → Computer games;**

## KEYWORDS

Platformers, Design, Metrics, PCG

**ACM Reference format:**
Adam Summerville, Julian R. H. Mariño, Sam Snodgrass, Santiago Ontañón, and Levi H. S. Lelis. 2017. Understanding Mario:
An Evaluation of Design Metrics For Platformers. In *Proceedings of FDG'17, Hyannis, MA, USA, August 14-17, 2017,* 10 pages.
DOI: 10.1145/3102071.3102080


*FDG'17, Hyannis, MA, USA*

978-1-4503-5319-9/17/08…$15.00
DOI: 10.1145/3102071.3102080

## 1 INTRODUCTION

Evaluation is one of the key open research challenges in the field of procedural content generation (PCG). The amount of work on computational metrics to evaluate platform game maps [2, 20] significantly lags behind the amount of work concerning map generation [4, 5, 10, 15–17, 21–23], to name a few examples of PCG systems for generating levels of platform games.

Content evaluation schemes are necessary not only to verify the quality of the content generated by PCG systems, but also to guide search algorithms during the PCG process. Reis et al. [15] showed that a PCG system can employ human computation to evaluate content during the PCG process. However, such an approach becomes impractical if the number of evaluations required is large.

Mariño et al. [9] showed that computational metrics often used to evaluate the content generated by PCG systems are unable to capture the player's perceived visual aesthetics, difficulty, and enjoyment. In this paper we introduce and examine metrics based on simple rules that aim at better capturing the player's perceptions.

We evaluate the metrics introduced in this paper with levels of *Infinite Mario Bros* (IMB), a clone of *Super Mario Bros.* The IMB levels we use were rated by humans according to their perceived visual aesthetics, difficulty, and enjoyment. We compute the correlation coefficient between our metrics and the human ratings. One of the metrics introduced in this paper strongly correlates with difficulty (an impressive correlation of 0.72) and with enjoyment (0.42), while the best performing metric for visual aesthetics obtained a correlation of 0.23. These correlation values are near empirical upper bounds derived by an inter-user study with humans, which shows similar trends, i.e., independent humans tend to agree in terms of difficulty and enjoyment, but not in terms of visual aesthetics.

Guzdial et al. [6] presented a deep convolutional neural network (CNN) approach to automatically predict player ratings using the same dataset as we use in this paper. As a way of demonstrating the quality of the metrics we introduce, we also show that a simple linear regression model using a subset of our metrics as input features is able to substantially outperform the CNN approach introduced by Guzdial et al. in terms of predicting human-perceived visual aesthetics, difficulty, and enjoyment in IMB levels.

## 2 PROBLEM FORMULATION

The main research question addressed in this paper is whether it is possible to define computational metrics that can accurately predict the player's perceptions of game maps (e.g., whether a game map is difficult or not). Thus, given a level $M$ which was rated by humans, we seek to define metrics that can predict the human ratings of difficulty, visual aesthetics, and enjoyment of $M$. We verify the prediction accuracy of the metrics introduced in this paper in terms of correlation between the metric values and human ratings, and in terms of the mean absolute error computed from our model's predicted values and human ratings.

Our goal in this paper is not to develop computational metrics that will replace user studies in the evaluation of content generated by PCG algorithms, but to develop metrics that could be used as evaluation heuristics or fitness functions of PCG algorithms.

## 3 NOTATIONS AND DEFINITIONS

In this section we introduce the notation used to describe the computational metrics presented in this paper.

- We use the words level and map interchangeably.
- The game map is defined as a grid space $M$ with width, $w$, and height, $h$, where $M[x][y]$ is the grid cell in coordinates $x$ and $y$ of $M$. We define the bottom-left grid cell of $M$ as the origin of the grid (i.e., $x = 1$ and $y = 1$).
- Objects in the game map occupy grid cells. We define $empty(M[x][y]) = 1$ if grid cell $M[x][y]$ is empty (i.e., no objects occupy grid cell $M[x][y]$) and 0 otherwise.
- **G** is the set of objects (e.g., blocks and enemies) in $M$.
- **E** is the set of enemies in $M$.
- We also use the word tile to refer to a grid cell. We define the type of a tile $t$ according to the object occupying $t$. We consider 13 different types for *Super Mario Bros.* (SMB) in this paper: Solid, Enemy, Destructible Block, Question Mark Block With Coin, Question Mark Block With Power-up, Coin, Bullet Bill Shooter Top, Bullet Bill Shooter Column, Left Pipe, Right Pipe, Top Left Pipe, Top Right Pipe, and Empty.

Although most of the proposed metrics are generic, we use *Super Mario Bros.* as the testbed for our metrics in this paper.

## 4 RELATED WORK

Applying metrics to generated levels has been a common practice since Smith and Whitehead [19] introduced linearity and leniency. They used these metrics to describe the "expressive range" of their generator, i.e., what areas of the metric space did the generated levels cover. Horn et al. [8] extended these metrics with density [18] and pattern density [3] (the number of times certain mesopatterns appear in the level). Canossa and Smith [2] extended these metrics with a proposal for many more that attempt to address the complexities of properties of interest, such as aesthetics and difficulty.

There has been less work on mapping these metrics back to actual human affective responses. Pedersen et al. [13] predicted human responses using mostly features of the players' playtraces in addition to metrics related to the gaps in the levels (number of, width of, etc.). Summerville et al. [24] used playtrace metrics in addition to metrics related to the frequency of gaps, enemies, and rewards to predict players' responses. Most importantly for this work, we utilize the dataset used by Mariño et al. [9] who used some of the metrics used by Horn et al. [8] to predict the perceived difficulty, enjoyment, and visual aesthetics of generated levels.

### 4.1 Previous Computational Metrics

Here we describe some of the metrics introduced by previous work that are used in our experiments.

*Linearity:* The linearity of a level is computed by performing a linear regression on the center points of the platforms and mountains contained in the level [20]. The linearity is the average distance between the center point of platforms and mountains in each column of $M$ and the linear regression's line. The linearity values are first multiplied by $-1$ (so higher values indicate more linear levels) and then normalized into the range of $[0, 1]$.

*Leniency:* Leniency approximates how much challenge the player experiences while playing a level [20]. The leniency of a level is the sum of the lenience value $w(o)$ of all objects $o$ in $G$: $\sum_{o \in G} w(o)$, normalized by the width of $M$. We use the lenience values specified in previous works [9, 18]. That is, 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 [18]. The leniency values are first multiplied by $-1$ (so larger leniency values indicate more challenging levels) and then normalized into the range of $[0, 1]$.

*Density:* Some objects can occupy the same $x$-coordinate in $M$ (e.g., mountains in SMB can be 'stacked-up" together). The density of a level is the average number of mountains occupying the same $x$-coordinate in $M$ [18]. Density values are also normalized into the range of $[0, 1]$, where values closer to one indicate denser levels.

*Negative Space:* Negative Space is the percentage of the empty space that is reachable by the player [2]. Jumping in platform games such as SMB is the core way for players to navigate the vertical space. A higher Negative Space metric often means more "floating" platforms and mountains which tend to be more enjoyable and aesthetically pleasing than simply progressing along the ground.

*Other Metrics:* In addition to negative space, Canossa and Smith [2] introduced 19 other metrics, which we did not use in our study. Their metrics are categorized into: *aesthetic*, *difficulty*, *topology*, and *strategic* metrics. Aesthetic metrics cover aspects such as music and the visual palette. Difficulty metrics expand on leniency by categorizing the level's source of difficulty. Topology metrics look at the physical space of the level and measure relevant features. Strategy metrics focus on how a player will/must react in a level.

## 5 NOVEL COMPUTATIONAL METRICS

In this section we describe several novel computational metrics for evaluating game maps.

*Symmetry (S).* The notion of symmetry has been empirically shown to correlate with the visual aesthetics of graphical user interfaces [11] and images [1]. The model of symmetry we use is based on the work of Ngo et al. [12] and Mariño and Lelis [10]. In

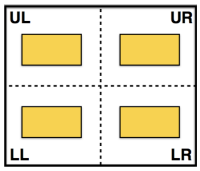



**Figure 1: Example of a symmetrical image.**

contrast with previous work, we are the first to use symmetry as a predictive metric of human annotated maps of a platform game.

The symmetry of a level is computed by dividing $M$ into four equal regions by a vertical and a horizontal separation line. The resulting regions are named Upper Left (UL), Upper Right (UR), Lower Left (LL), and Lower Right (LR). Let $X(LL)$ be the sum of the distances between the center of all objects in region LL and the vertical line; $Y(LL)$ be the sum of the distances between the center of all objects in LL and the horizontal line; and $A(LL)$ be the sum of the areas of all objects in LL. We define the symmetry value $S$ of a level $M$ in terms of functions $X(M)$, $Y(M)$ and $A(M)$, define below:

$$\begin{aligned} X(M) = &|X(UL) - X(UR)| + |X(LL) - X(LR)| \\ &+ |X(UL) - X(LL)| + |X(UR) - X(LR)| \\ &+ |X(UL) - X(LR)| + |X(UR) - X(LL)| \,. \end{aligned}$$

The value of $X(M)$ accounts for the "symmetrical" distance across the vertical line, across the horizontal line, and across the vertical and horizontal lines. The values of $Y(M)$ and $A(M)$ are defined analogously by using $Y$ and $A$-values instead of $X$-values. The $S$-value of a level is defined as as follows:

$$S(M) = X(M) + Y(M) + A(M) \,. \tag{1}$$

$S$ captures the intuitive notion of symmetry illustrated in Figure 1, where the yellow rectangles represent objects in the map. The map shown in Figure 1 has an $S$-value of zero, which means that the map is perfectly symmetrical according to $S$. The $S$-value of the map is zero because there are objects with exactly the same area in each region. Also, the objects in regions UL and LL are at the same distance from the vertical separation line as the objects in regions UR and LR; and the objects in regions LL and LR are at the same distance from the horizontal separation line as the objects in regions UL and UR.

*Balance (B).* According to Ngo et al. [11], the metric of balance measures whether the objects the player might find interesting, and thus attract their eyes, are well distributed in $M$. Here we assume that the "attractiveness" of an object is proportional to the object's distance to the horizontal separation line as well as the object's area. One could use richer schemes to define attractiveness (e.g., object color), which we intend to investigate in future work.

Balance is computed by dividing $M$ into two regions, Top (T) and Bottom (B), of equal size. In Figure 1 T is defined by the union of regions UL and UR, and bottom by the union of regions LL and LR. We define $G_T$ and $G_B$ as the set of objects in T and B, respectively. The Balance value of a map is computed in terms of function $W$, which is defined for the objects in region $T$ as follows:

$$W(G_T) = \sum_{o \in G_T} dy(o)A(o)$$

where $dy(o)$ is the distance between the center of the object $o$ and the horizontal separation line, and $A(o)$ is the area of $o$. $W(G_B)$ is defined analogously. The balance value of a map $M$ is then defined as the absolute difference between $W(G_T)$ and $W(G_B)$:

$$Balance(M) = |W(G_T) - W(G_B)|$$

*Reachability (R).* Reachability measures the proportion of elements placed in $M$ that are reachable by a player, i.e., that the player can directly interact with. Our hypothesis is that players rate poorly the visual aesthetics of maps that have fundamental flaws such as unreachable objects. The reachability is calculated as follows:

$$R(M) = \frac{n_{RC}}{n}$$

where $n_{RC}$ is the number of unreachable elements, and $n$ is the total number of objects in $M$. The value of $n_{RC}$ can be computed by applying domain-specific rules (e.g., in SMB Mario is unable to jump more than a given number of tiles).

*Decoration Frequency.* Levels are composed of many different objects, and the grid $M$ tends to be sparse, with most of the grid cells being empty. Some objects, such as the question-mark blocks, pipes, or enemies bring more visual variety to the level (i.e., they "decorate" the level), and as such we define the decoration metric as the number of decoration tiles over the size of the map:

$$DP(M) = \frac{\sum_{x=1}^{w} \sum_{y=1}^{h} \text{pretty}(M[x][y])}{w \times h}$$

where pretty($t$) is defined as being equal to 1 when $t$ is any of the following tile types: Pipe, Enemy, Destructible Block, Question Mark Block, or Bullet Bill Shooter Column and 0 otherwise.

*Tile Frequencies.* This metric is simply defined as the number of tiles of that type divided by the size of the map,

$$EP(M) = \frac{\sum_{x=1}^{w} \sum_{y=1}^{h} \text{Type}(M[x][y])}{w \times h} \,,$$

for each of 13 different types: Solid, Enemy, Destructible Block, Question Mark Block With Coin, Question Mark Block With Power-up, Coin, Bullet Bill Shooter Top, Bullet Bill Shooter Column, Left Pipe, Right Pipe, Top Left Pipe, Top Right Pipe, and Empty. We include these as they represent a base-line. More complex metrics tend to use these in different combinations and scalings (e.g., Leniency incorporates the number of enemies), but we wanted to see if the most simple metrics still held power.

*Tile Position Summary Statistics.* The distribution of object types in a level contains important information about the experience the player will encounter. For example, levels with more variance in the height of ground tiles will likely require the player to jump more. Levels with low variance on the $x$-coordinate where the enemies are placed will likely have a closely packed group of enemies. For each of the 13 tile types we get:

- $\mu_x$ and $\sigma_x$ - The mean and standard deviation $x$ position of that tile type.
- $\mu_y$ and $\sigma_y$ - The mean and standard deviation $y$ position of that tile type.

*Enemy Sparsity (ES).* In this metric we measure whether the enemies are grouped together or spread in the map. The enemy sparsity of map $M$ is computed as follows.

$$ES(M) = \frac{\sum_{e \in \mathbf{E}} |x(e) - \bar{x}|}{|\mathbf{E}|}.$$

Where $\mathbf{E}$ is the set of enemies in $M$, $x(e)$ is the $x$-position of enemy $e$ in $M$, $\bar{x}$ is the average $x$-position of all enemies in $\mathbf{E}$, and $|\mathbf{E}|$ is the total number of enemies in $M$.

Enemy Sparsity is similar to the metric $\mu_x$ for enemy tiles as they both measure the horizontal spread of enemies in the level. The difference between the two metrics is subtle: while the former computes the spread with the standard deviation formula, the latter uses the absolute differences between enemies and $\bar{x}$. We further discuss this subtle difference in Section 7.5.3.

*Tile Indicator.* Some of the above metrics only make sense if a given tile type is present in a level. For each tile type, this metric is defined as 1 if the tile type is present in the level and 0 if it is not.

*(Normalized) Number of Enemies.* In this metric we count the number of enemies in the level, as above, but we then normalize such that the highest number of enemies in a level (15 in our dataset) is 1 and the lowest (0 in our dataset) is 0.

*Path Length Percentage (Path %).* This metric is the proportion of the level that is taken up by a path from beginning to end (i.e., a sequence of grid cells from Mario's initial grid position to a grid position after the finish line of the level) found by an A* search [7]. We expect that the more obstacles that are in the level, the longer the required path. This is because the player will need to move around the grid to avoid the obstacles. The Path % metric of $M$ is computed by dividing the number of grid cells in the path found by A* for $M$ divided by the total number of tiles in $M$ ($w \times h$).

*Jump Count.* Using the same A* search we count the number of jumps required to complete the level. If all actions a player can issue (e.g., move, jump, etc.) have cost of one, an A* search will minimize the number of actions required to finish the level (i.e., for A* all actions are equally costly). However, for this metric we want to know the number of jumps required to finish the level, not the number of possible jumps, which can be very large. In fact, an A* search minimizing the number of actions could return a sequence of actions that includes jumps that are easily replaced by runs. In order to find a sequence of actions that includes jumps only when necessary, we make the jump action cost more than all other actions. In our A* implementation a jump action costs 2 while all other actions cost 1. This means that the sequence of actions encountered by A* to finish the level will run/walk if possible, and only jump when a gap/enemy/hill requires it.

*Summary of Metrics Introduced.* In total we introduce 85 metrics: Symmetry, Balance, Reachability, Decoration Frequency, Tile Frequency (13 metrics), Tile Position Summary Statistics (52 metrics, 13 for each of the following: $x_\mu$, $x_\sigma$, $y_\mu$, $y_\sigma$), Tile Indicator (13 metrics), Jump Count, Enemy Sparsity, and Path %.

# 6 DATASET

In our experiments we use the dataset described by Reis et al. [15].[1] Reis et al. used the Notch Level Generator (NLG),[2] to generate a library of 2,000 levels of size $20 \times 15$ (a typical *Super Mario Bros.* map is approximately 10 times longer than Reis et al.'s small levels). NLG receives as input a difficulty value $d$ for stochastically determining the number of enemies to be included in the map. Reis et al. used NLG to generate maps with different values of $d$ to ensure diversity in the dataset produced. These maps were made available online for evaluation, and volunteers played 1,437 distinct small levels and then provided 2,715 evaluations. The small levels were evaluated according to the volunteers' perceived visual aesthetics, enjoyment, and difficulty on a 7-point Likert scale. We use the median rating of a level if a level was evaluated by multiple volunteers. The evaluations were obtained in 125 different sessions of play. A session of play is defined by a volunteer entering the system, annotating a collection of small maps, and exiting the system. Since Reis et al. wanted to maximize the number of annotated small levels, in order to simplify the annotation process, they did not ask for the volunteer's identity nor their demographic information. The number of sessions of play offers a reasonable approximation of the number of volunteers who participated in their data collection.

Two independent volunteers agreed to contribute non-anonymously to Reis et al.'s data collection. The ratings provided by these two volunteers allow us to perform an inter-rater study in this paper (the two volunteers evaluated 453 levels in common). We use this subset of 453 evaluated levels to verify how well a volunteer is able predict the ratings of another independent volunteer. Also, one of these two volunteers evaluated 38 levels twice. We use these ratings to evaluate how the evaluations of a single person correlate with this person's own evaluations.

# 7 EMPIRICAL RESULTS

We treat the problem of predicting human ratings as a classification task which we tackle with a multinomial LASSO regression [25]. We use the metric values of a given level as the input features and the human ratings the values to be predicted. As a byproduct of its regression, LASSO also selects a subset of discriminative metrics for the multinomial regression task. Then, we compute the correlation of each selected metric with the human ratings.

We chose to use a multinomial regression instead of the more standard linear regression due to the nature of the ratings. While the ratings are Likert-like (i.e., they have a number associated with them and are not purely categorical responses such as "Poor" or "Great") we did not want to make any assumptions about the scaling (i.e., the difference between 1 and 4 might not be the same as the difference between 4 and 7).

## 7.1 Metrics Selection with LASSO

In this first experiment we perform a 10-fold cross-validation multinomial LASSO regression for each criterion: difficulty, visual aesthetics, and enjoyment. We chose multinomial LASSO for two reason **(1)** we believe the Likert style data should not be treated as interval (multinomial) **(2)** we wanted a regularization technique

[1] Available at http://www.dpi.ufv.br/~lelis/downloads/Mario-Dataset.zip
[2] The system is named after Markus "Notch" Persson.

that encouraged sparsity for variable selection (LASSO as opposed to ridge or elastic net regression). In addition to linearity, leniency, density, and negative space, we use all 85 metrics introduced in this paper in this experiment.

A multinomial LASSO regression minimizes the categorical cross entropy while limiting the absolute sum of the regression coefficients scaled by an input parameter $\lambda$. Many of the regression coefficients are set to zero due to the $\lambda$ limitation—LASSO performs feature selection during its regression procedure. A multinomial regression predicts a class, $k$, from a set of $K$ classes for each data point, $x$, by taking the class with the highest probability.

$$\Pr(k|x) = \frac{e^{\beta_k x}}{\sum_{l=1}^{K} e^{\beta_l x}} .$$

Our multinomial LASSO regression effectively predicts, for a given level $M$, which of the 7 "Likert classes" $M$ belongs to. Note that due to the fact that there is disagreement among the human raters it is impossible for the regression to achieve no error (e.g., Rater A gave a level a 3 while Rater B gave it a 5 means that the regression can get at most one of those correct).

Many of the metrics introduced in this paper encode similar information. For example, the percentages for tile types Left Pipe and Right Pipe are expected to provide similar information. Our goal with this experiment is to select a subset of discriminative metrics for the task of predicting each of the evaluation criteria.

We perform one LASSO regression for each of the three criteria and choose the maximal $\lambda$ parameter that was within one standard error of the minimal training error; the minimal training error is achieved by including all metrics. LASSO reduced from 89 metrics to only 12 metrics for difficulty, 16 metrics for visual aesthetics, and 14 metrics for enjoyment.

## 7.2 LASSO Prediction Results

The performance of our multinomial regressions can be seen in Table 1 in terms of *accuracy* and *mean absolute error* (MAE). The accuracy is computed as the percentage of levels classified correctly (assuming the ground truth for level $M$ is the median rated value for $M$), and MAE is the mean absolute difference between the predicted values and the ground truth values. Our model achieves an accuracy of 37.6% in difficulty, 33.1% in visual aesthetics, and 35.2% in enjoyment. A random classifier is expected to achieve an accuracy of ≈14.3% as our problem has 7 distinct classes.

The MAE values of our multinomial regression predictions vary from 1.16 (difficulty) to 1.29 (visual aesthetics), which means that our prediction model errs on average slightly more than one point in the 7-Likert scale. The 7-Likert scale is defined by the following points: 1 (strongly disagree), 2 (mostly disagree), 3 (somewhat disagree), 4 (neither agree nor disagree), 5 (somewhat agree), 6 (mostly agree), and 7 (strongly agree). By erring by slightly more than one point, it means that on average the prediction model could, for example, mostly agree that a given level is difficult while the human rater strongly agrees that the level is difficult.

More details about the classification results are provided in the confusion matrices shown in Figure 2. If our predictions were perfect, the squares in Figure 2 would be yellow across the diagonal. By observing the light-colored squares across the diagonal, we notice that difficulty is the easiest criterion to predict, followed by enjoyment, and then visual aesthetics. Most of the prediction errors for both visual aesthetics and enjoyment come from predicting a score of 5 when the actual rating is a 3, 4, 6, or 7 (see the row for score 5 in Figure 2 (a) and (c)). This is because the score of 5 is the most common score in the dataset, being approximately 50% more common than the next most common rating. As for difficulty, the most common prediction error comes from predicting a score of 7 when the actual rating is a 4, 5 or 6, and from predicting a score of 2 when the actual score is a 1, 3, and 4.

| Criterion | Accuracy | MAE |
|---|---|---|
| Difficulty | 37.6% | 1.16 |
| Visual Aesthetics | 33.1% | 1.29 |
| Enjoyment | 35.2% | 1.18 |

**Table 1: Percentage of correctly classified levels (accuracy) as well as the Mean Absolute Error (MAE) for the different metrics from the multinomial LASSO regressions. A more detailed analysis can be seen in Figure 2.**

| Criterion | LASSO MAE | Convolutional NN MAE |
|---|---|---|
| Difficulty | **0.66** | 0.92 |
| Aesthetics | **0.71** | 1.13 |
| Enjoyment | **0.68** | 1.04 |

**Table 2: MAE results of our linear regression using the metrics selected by the LASSO multinomial regression as input features (LASSO MAE) compared to the MAE results of Guzdial et al.'s [6] convolutional neural networks.**

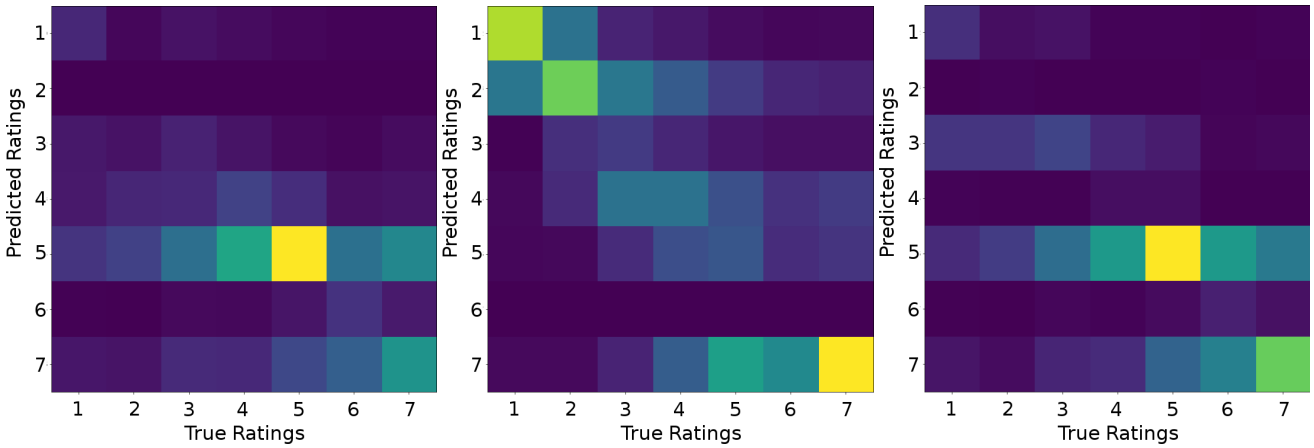
## 7.3 Previous Neural Network Model

We note that the results shown in Table 1 and Figure 2 are not directly comparable to the results of Guzdial et al. [6]. This is because, even if the original data set is the same, they are comparing to averaged ratings whereas we are comparing to the non-averaged ratings (e.g., Rater A rates a level with 1, Rater B rates it with a 3—we compare to both those points, while they merge it to a single rating of 2). As a point of comparison, we ran a linear regression with the metrics selected by the multinomial LASSO regression, the results of which can be see in Table 2.

We see that a small number of high quality metrics can substantially outperform more advanced neural network approaches like those of Guzdial et al. by 50%. For example, when predicting *difficulty*, our LASSO linear regression has a MAE of 0.66 (which means that the predicted value is, in average 0.66 points off the average score provided by humans in a 7-point Likert scale), while the error reported by Guzdial et al.'s approach was 0.92.

## 7.4 Metrics Selected by LASSO

The metrics selected for each regression can be seen in Table 3. Many of the metrics we introduce in this paper do not appear in the table because they either do not provide relevant information for the task of predicting human ratings in LASSO's model, or because the information they provide is made redundant by a more informative metric. It is possible that metrics not selected by the LASSO could be relevant to more complex models. Prior to the regression all metrics are scaled to have a mean of 0 and variance




(a) Aesthetics (b) Difficulty (c) Enjoyment

**Figure 2: Confusion matrices for the multinomial LASSO predictor, normalized by the number of ratings per category. Perfect prediction would be yellow along the diagonal. Difficulty is the easiest to predict with most of the errors coming from incorrectly predicting 2 and 7 too often. For both Aesthetics and Enjoyment the most common error source is predicting a 5. This comes from the fact that 5's are the most prevalent rating, being roughly 50% more common than the next most common rating, and over twice as most of the other ratings**

Difficulty

| **Metric** | **Weight** | $\rho$ |
|---|---|---|
| Number of Enemies | 1.00 | 0.72 |
| Enemy $\sigma_x$ | −0.07 | 0.50 |
| Enemy Indicator | −0.02 | 0.38 |
| Enemy $\sigma_y$ | 0.02 | 0.48 |
| Jump Count | −0.02 | −0.20 |
| Pipe Top $\mu_y$ | −0.01 | −0.20 |
| Enemy Sparsity | −0.01 | 0.27 |
| Bullet Bill $\sigma_y$ | < 0.01 | 0.01 |
| Path % | < 0.01 | −0.12 |
| Pipe $\sigma_x$ | < 0.01 | −0.20 |
| Pipe Top % | < 0.01 | −0.22 |
| Bullet % | < 0.01 | −0.02 |
| **Human Rater** | $\rho$ | |
| Same User | 0.75 | |
| Independent Users | 0.80 | |

Aesthetics

| Metric | Weight | $\rho$ |
|---|---|---|
| Power up $\mu_x$ | 1.00 | 0.23 |
| Reachability | −0.58 | −0.20 |
| Number of Enemies | 0.54 | 0.22 |
| Negative Space | 0.29 | 0.20 |
| Balance | 0.28 | 0.20 |
| Enemy $\mu_x$ | 0.27 | 0.17 |
| Enemy Sparsity | 0.22 | 0.16 |
| Power up $\mu_y$ | 0.13 | 0.23 |
| Enemy Indicator | 0.09 | 0.18 |
| Symmetry | 0.07 | 0.19 |
| Bullet Bill Column % | 0.05 | 0.06 |
| Pipe $\mu_x$ | 0.04 | −0.04 |
| Enemy $\sigma_y$ | 0.02 | 0.17 |
| Power up Indicator | 0.02 | 0.23 |
| Decoration % | 0.02 | 0.16 |
| Density | 0.01 | 0.13 |
| **Human Rater** | $\rho$ | |
| Same User | 0.55 | |
| Independent Users | 0.38 | |

Enjoyment

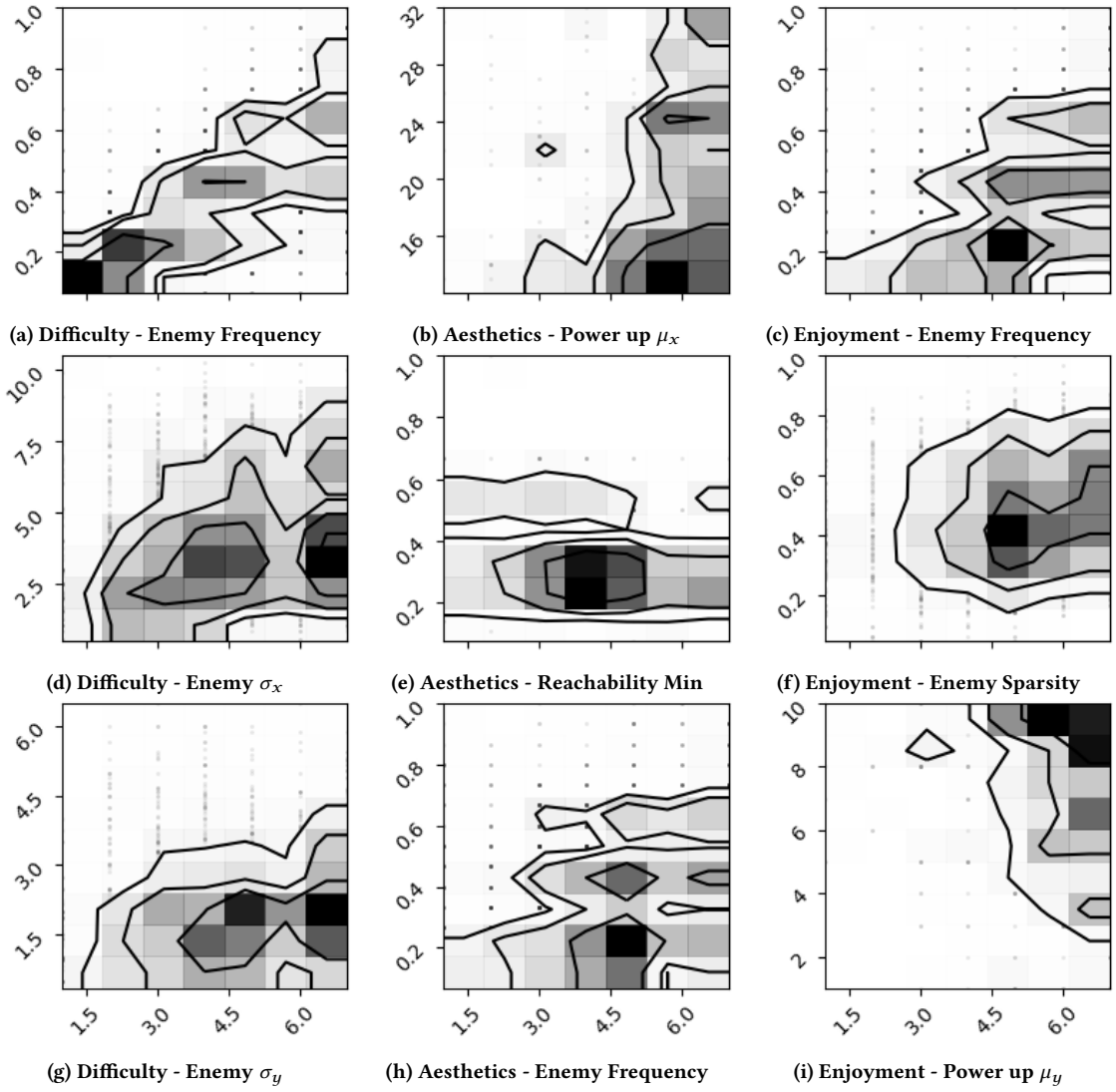
| Metric | Weight | $\rho$ |
|---|---|---|
| Number of Enemies | 1.00 | 0.42 |
| Enemy Sparsity | 0.24 | 0.27 |
| Power up Indicator | 0.22 | 0.25 |
| Power up $\mu_y$ | 0.20 | 0.25 |
| Power up $\mu_x$ | 0.16 | 0.25 |
| Negative Space | 0.15 | 0.26 |
| Symmetry | 0.12 | 0.27 |
| Enemy Indicator | 0.06 | 0.29 |
| Reachability | −0.05 | −0.12 |
| Enemy $\mu_y$ | 0.04 | 0.13 |
| Enemy $\sigma_y$ | 0.04 | 0.32 |
| Enemy $\mu_x$ | 0.03 | 0.24 |
| Coin $\mu_x$ | 0.03 | 0.16 |
| Bullet Column $\sigma_y$ | 0.01 | 0.06 |
| **Human Rater** | $\rho$ | |
| Same User | 0.64 | |
| Independent Users | 0.45 | |

**Table 3: The metrics selected by the regressions. The weights listed are scaled such that the maximum absolute value is 1.00. For each of the metrics, the Spearman rank coefficient is listed.**

of 1, so as to guarantee that the weights are on a similar scale. The weights listed in the table are scaled such that the weight with the highest absolute value is scaled to 1. Also listed are the non-parametric Spearman correlation coefficients, $\rho$, of each metric selected as relevant by LASSO's multinomial regression.

In addition to the correlation between the selected metrics and the human ratings, we also present in Table 3 the correlation of two independent volunteers ("Independent Users"), and the correlation between the ratings of a given volunteer ("Same User"). The correlations of Independent Users can be seen as an empirical upper bound on the expected correlation.

Figure 3 shows a closer look at the correlations between some of the individual metrics and the rated feature. Each plot in Figure 3 shows the human ratings in the x-axis and the metric values in the y-axis. Should a metric have a perfect positive correlation ($\rho = 1$) with one of the evaluated criterion, we would observe dark squares



(a) Difficulty - Enemy Frequency (b) Aesthetics - Power up $\mu_x$ (c) Enjoyment - Enemy Frequency

(d) Difficulty - Enemy $\sigma_x$ (e) Aesthetics - Reachability Min (f) Enjoyment - Enemy Sparsity

(g) Difficulty - Enemy $\sigma_y$ (h) Aesthetics - Enemy Frequency (i) Enjoyment - Power up $\mu_y$

**Figure 3: Sample density contours for the given metrics vs. the rated feature. The clearest trend can be seen in Difficulty-Enemy Frequency.**

across the secondary diagonal. Similarly, should a metric have a perfect negative correlation ($\rho = -1$) with one of the criterion, we would observe dark squares across the main diagonal. The clearest trends can be seen between difficulty and Number of Enemies, where a clear linear trend can be seen. Clear trends can also be seen between enjoyment and Number of Enemies and between difficulty and Enemy $\sigma_x$.

## 7.5 Discussion

The empirical upper bounds defined by the correlations between the ratings of two independent volunteers ("Independent Users") suggest that it is easier to predict difficulty, than enjoyment and visual aesthetics: the correlation between the ratings of independent users is 0.80 for difficulty, 0.45 for enjoyment, and 0.38 for visual aesthetics. This means that different people tend to agree more in terms of difficulty, than for enjoyment or visual aesthetics. A similar trend is observed in the correlation values obtained by the best performing metrics in each criterion: the best performing metric for both difficulty and enjoyment is Number of Enemies, with correlation values 0.72 and 0.42, respectively, and for visual aesthetics it is a set of metrics related to power ups: Power up $\mu_x$, $\mu_y$, and Indicator all with correlation values of 0.23.

Overall, the best performing metrics in each criterion are near the empirical upper bound given by the correlation of independent volunteers. Namely, Number of Enemies yields a correlation value nearly equal to the correlation value of the two independent users for enjoyment (0.42 for the former and 0.45 for the latter). The same Number of Enemies yields correlation value of 0.72 for difficulty, which is near the correlation value of 0.80 presented by the two independent users. Visual aesthetics is the only criterion for which

the difference between the correlation of the best performing metrics and the correlation of the two independent users is larger; the best performing metrics yield a correlation value of 0.23, while the ratings of the two independent users have a correlation of 0.38.

We conjecture that the correlation values for visual aesthetics tend to be smaller because visual aesthetics is perhaps the most subjective of the three criteria. Our conjecture is supported by the correlation value of only 0.55 for the visual aesthetics evaluations provided by the same person (see Same User in Table 3).

It is interesting to note that, in contrast with the other criteria, the correlation of the ratings of two independent volunteers is slightly higher than the correlation of the ratings given by same volunteer for difficulty—0.80 for the former and 0.75 for the latter. A possible explanation is that multiple scores given by the same volunteer for a fixed level are subject to ordering effects. That is, a level will likely be easier the second time a person plays that level.

*7.5.1 Difficulty.* The metric that obtained the highest correlation was Number Enemies, which simply counts the number of enemies in the level. The metric of leniency (not shown in Table 3 because it was not selected by LASSO) obtained a correlation of 0.53, which is much lower than the correlation obtained by the simpler Number of Enemies. Thus Number of Enemies is the current state-of-the-art single metric for predicting human-perceived difficulty in the dataset used in our experiments. However, this might just indicate that difficulty in levels generated by the Notch Level Generator used in our experiments comes primarily from adding enemies, and not by other factors such as gaps or platform configurations.

The number of enemies is the most important metric by over a factor of 10 when considering the regression weight but is not the only important metric. Moreover, notice that when interpreting the results presented in Table 3, we must have in mind that the correlation coefficient $\rho$ is calculated for each metric individually, but the regression weight results from applying LASSO to all metrics at the same time. So, a low regression weight might not mean that a metric is not relevant, but could also mean that LASSO found other metrics that represent similar information, and thus did not have to assign a higher weight to a given metric.

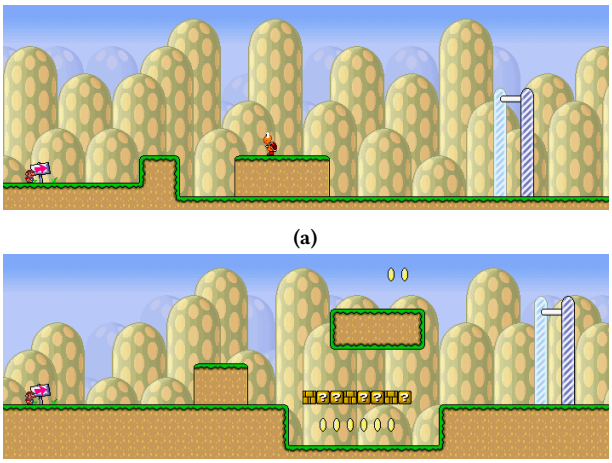
We see that the horizontal spread of the enemies (Enemy $\sigma_x$) has a negative impact on the difficulty (negative LASSO regression weight), meaning that humans tend to find levels with a larger spread of enemies to be easier than levels with a smaller spread of enemies. This is interesting because the horizontal spread has a positive correlation with difficulty. It is possible that this sign discrepancy between LASSO's weight and correlation value for Enemy $\sigma_x$ happens because the metric acts as a proxy for number of enemies when analyzed individually (correlation value). By contrast, when analyzed altogether with metrics that already account for the number of enemies (e.g., LASSO regression also accounting for Number of Enemies), Enemy $\sigma_x$ shows that humans tend to find easier to play levels in which the enemies are spread out. Intuitively, it makes sense that a larger spread is easier since the enemies will be spaced out, while a dense cluster will present a more difficult obstacle. Conversely, we see that the enemy vertical spread (metric Enemy $\sigma_y$) has a positive effect on the difficulty. Again, this makes sense as a large vertical spread of enemies tightly clustered horizontally will present a wall of enemies that is hard to navigate, whereas a tight cluster vertically can be avoided.

*7.5.2 Visual Aesthetics.* The metric of Symmetry is amongst the best performing metrics for visual aesthetics with respect to correlation values ($\rho = 0.19$). Note, however, that one might have expected a negative correlation between $S$-values and visual aesthetics. That is, symmetrical levels (small $S$-values) to be rated as visually pleasing by humans. However, we observed the opposite in our study: the positive correlation for Symmetry and visual aesthetics means that levels with larger $S$-values (less symmetrical levels) are rated as more visually pleasing. This result contrasts with the recent study performed by Mariño and Lelis [10], whose system builds symmetrical small maps of IMB. Mariño and Lelis' small maps were rated as visually pleasing by human subjects.

The explanation for this discrepancy is rooted at the number of objects in the levels: Symmetry might be working as a proxy for Negative Space. Intuitively, levels with fewer objects tend to have much smaller $S$-values than levels with a large number of objects. This is because with more objects the values of $X_M$, $Y_M$, and $A_M$ tend to be larger (see Equation 1 and the definitions of $X_M$, $Y_M$, and $A_M$). As an example, the $S$-value of an empty map is trivially zero. The positive correlation for Symmetry is explained by the fact that human subjects tend to attribute low visual aesthetics scores to levels with very few objects, and high scores to levels with more objects—in our study the symmetry metric is essentially measuring how much of the grid is filled with objects. Therefore, it is not surprising that Negative Space and Symmetry yield similar correlation values: 0.20 for the former and 0.19 for the latter. Mariño and Lelis [10] were able to create visually pleasing maps by minimizing the symmetry metric because they always compare $S$-values of levels with exactly the same set of objects.

Other metrics that show very strong correlations are related to the presence of power ups (e.g., Power Up $X\mu$ and Power Up $Y\mu$), perhaps because power ups are usually scarce and could appear amongst other more elaborate decorative tiles. The somewhat strong correlations between Negative Space and Reachability with visual aesthetics suggest that reachable vertical variety tends to be appreciated by the player. Interestingly, LASSO attributed small weights to both Power up Indicator (weight 0.02) and Power up Frequency (weight of 0.00, as the metric does not appear in Table 3) and much larger weights to Power up $\mu_x$ and Power up $\mu_y$. This difference in the weight values suggest that it is not the mere presence of power ups (Power up indicator) or the amount of power ups (Power up Frequency), but the positioning of the power ups that is most important for the aesthetics. The further to the right and higher the power ups (i.e., large values of both Power up $\mu_x$ and Power up $\mu_y$), the more aesthetically pleasing it was to the players. Perhaps players find it aesthetically pleasing to have a reward towards the end of the level that requires maneuvering to reach and dislike being handed a power up at the beginning.

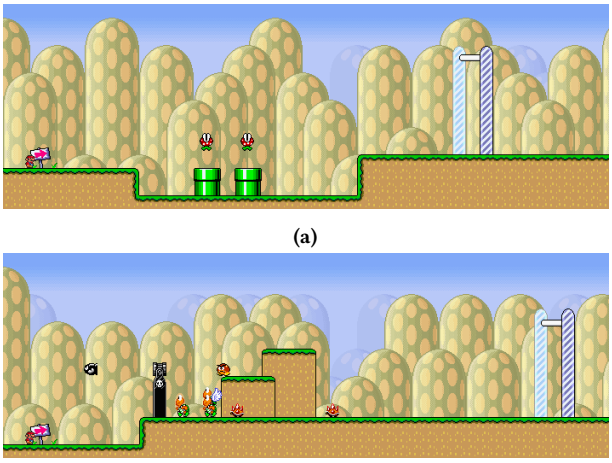
We also see Number of Enemies appears as a well performing metric speaking to the fact that players enjoy the variety that enemies bring. Interestingly, enemies are the only such type that have this effect. As discussed above, players find the presence of power ups pleasing, but the amount has a minimal effect. The amount of Bullet Bill Columns has an effect, but perhaps interestingly, not the



**(a)**

**(b)**

**Figure 4: Two levels with high misclassification error for the Difficulty rating. Both were classified as 1 by LASSO, but both have ratings of 7. Given that the players who rated them as 7's had no difficulty completing the level, we believe that it comes from a misunderstanding of the rating scale.**



**(a)**

**(b)**

**Figure 5: Two levels with high misclassification error for the Difficulty rating. Level (a) was classified as a 1 by LASSO but was given a rating of 6 by a player, while Level (b) was classified as a 7 by LASSO but was given a rating of 2 by a human.**

amount of Bullet Bill cannons. This means that the larger the column, the more visually pleasing, but that adding more cannons does not necessarily improve the human-perceived visual aesthetics.

*7.5.3 Enjoyment.* The metrics that correlated the most with enjoyment tend to be metrics related to elements in the game that the player can interact with (enemies, power ups, and coins). This is related to the Yerkes-Dodson law [27] demonstrated by Piselli et al. [14] in the context of video games. According to the Yerkes-Dodson law, enjoyment will be maximum for the right amount of challenge. The strong correlation between Number of Enemies and enjoyment ($\rho = 0.42$) suggests that the right amount of challenge for Reis et al.'s volunteers included a large number of enemies. As mentioned above, a wide spread of enemies vertically indicates more challenge for the player and the fact that it was one of the highest correlations ($\rho = 0.32$) reinforces that this challenge is enjoyable for players.

Interestingly, while the counting-based metric Number of Enemies is the most important factor for enjoyment (LASSO weight of 1.00 and $\rho$ of 0.42), positioning-based metrics also seem to be more important. The mean horizontal position for enemies (Enemy $\mu_x$), power ups (Power up $\mu_x$), and coins (Coin $\mu_x$) all have a positive impact on the players' enjoyment. This seems to indicate that players enjoy a brief amount of respite at the beginning and appreciate higher complexity towards the end of the level, which is a common level design tactic [26].

Additionally, metrics concerning the distribution of platforms such as Symmetry and Negative Space also had a high correlation. These metrics reflect the type of movement that players can execute through levels, again reinforcing that enjoyment is linked to how the player interacts with the level. Furthermore, Reachability has negative weight and $\rho$ values, suggesting that the player finds levels containing objects that they cannot interact with less enjoyable.

The metric with largest LASSO weight after Number of Enemies is Enemy Sparsity (weight of 0.24). Similar to Enemy $\sigma_x$, Enemy Sparsity also computes the spread of enemies in the level. The positive weight and $\rho$ values for Enemy Sparsity in enjoyment indicate that people tend to find levels in which the enemies are spread out to be more enjoyable. Although the difference between Enemy $\sigma_x$ and Enemy Sparsity is subtle (the former returns larger values than the latter for levels with enemies too far from the average enemy position), our results suggest that this subtle difference is important. That is, if we remove Enemy Sparsity from our pool of metrics, LASSO selects 24 instead of the 14 metrics shown in Table 3. This increase in the number of selected metrics suggests that one needs approximately 10 other metrics to make up for the lack of Enemy Sparsity.

## 7.6 Case Studies

We now turn our attention to a few of the level snippets that had the highest misclassification error for Difficulty. We look at difficulty since (1) it has the highest inter-rater reliability so we are more likely to able to make valid judgments and (2) the metrics and regressions both have the best predictive power for difficulty so disagreements are probably fundamental, and not a factor of noise.

In Figure 4 we see two levels that showcase the difficulty of our prediction task. Both of these levels were predicted to be a 1 in difficulty by LASSO, i.e., the easiest levels possible. However, human raters attributed a difficulty score of 7 to both of them, i.e., the hardest levels possible. At a glance, we can tell that these levels are indeed easy, with either a single, easily dodged enemy (a) or no possibility for death (b). In this case, we believe that there was a misunderstanding of the rating scale, i.e., the volunteers thought that 7 was easy and 1 was most difficult, or that the raters were not performing the task faithfully. In both cases, other raters rated the levels as extremely easy, giving them 1's or a 2, in the case of (a).

In Figure 5 we see two levels that were incorrectly classified for legitimate, interesting reasons. In the level shown in Figure 5.(a), we see two piranha plants in the middle of the screen. A patient player can bide their time, wait for the plants to return to the pipes and continue on their way. However, a novice player who is unaware of how the piranha plants behave could easily be in mid-jump over a pipe when the piranha plant emerges, catching them by surprise and killing them. In fact, the player who rated the level as a 6 in difficulty died on the level, so it is likely that they were caught by surprise. While from a purely count based view, the level is easy, hence why it was classified to be a 1, but it offers enough surprise that a novice player could find some difficulty with it. Our classification system incorporates no knowledge about player familiarity or skill which may not be representative.

In the level shown in Figure 5.(b), we see a dense cluster of enemies. At first glance, this appears to be an intimidating, skill intensive block for players. However, the goomba is about to fall off of the higher platform, leaving a clear path for a patient player who hops up to the bullet bill cannon and jumps over to the now clear path over the enemies. While one of the raters did die on this level, rating it a 5 instead of the 6 that we classified it as, the other passed it with no trouble bypassing the enemies altogether. Again, a novice or intermediate player is likely to have difficulty either through nervousness or a desire to kill all of the enemies (the player who died killed 7 enemies in total) that gets them in trouble, while the advanced player just ignores the enemies. Thus, a count-based metric can only find that a tight cluster of enemies is correlated with difficulty, even if there are clear paths through the level.

## 8 FUTURE WORK AND CONCLUSIONS

The strong correlation between several of the proposed metrics and human ratings, encourages us to investigate the use of some of these metrics to automatically adjust the difficulty of procedurally generated levels to match the player's skill, or to generate levels that maximize enjoyment or visual aesthetics.

In this paper we introduced several computational metrics that can potentially be used to guide the search process of PCG systems for creating platform game maps. We performed an experiment in which we treated the problem of predicting the player's perceived visual aesthetics, difficulty, and enjoyment as a classification task where our metrics were used as features and applied a feature selection approach to discover a subset of discriminative metrics. We then computed the correlation between the selected metrics with each of the evaluation criteria. One of the metrics presented an impressive correlation of 0.72 with difficulty and of 0.42 with enjoyment. The best performing metric for visual aesthetics obtained a correlation of 0.23. We derived an empirical upper bound for the correlation values by computing the correlation between the ratings of two independent volunteers on a subset of the levels tested. This inter-user study showed that humans also tend to agree more in terms of difficulty than in terms of enjoyment and visual aesthetics. Finally, we observed that the best performing metrics in each criterion were near the empirical upper bound (except for aesthetics, where the difference is larger).

As part of our future work, and informed by our results, we would like to consider a pool of level segments coming from different level generators, in order to have a more varied sample, as well as investigate additional sets of metrics.

## REFERENCES


[1] M. Bauerly and Y. Liu. 2006. Computational modeling and experimental investigation of effects of compositional elements on interface and design aesthetics. *International Journal of Human-Computer Studies* 64, 8 (2006), 670–682.

[2] Alessandro Canossa and Gillian Smith. 2015. Towards a Procedural Evaluation Technique: Metrics for Level Design. *Proceedings of FDG* (2015).

[3] Steve Dahlskog and Julian Togelius. 2013. Patterns as objectives for level generation. (2013).

[4] Steve Dahlskog, Julian Togelius, and Mark J. Nelson. 2014. Linear levels through n-grams. In *Proceedings of the 18th International Academic MindTrek Conference.*

[5] Matthew Guzdial and Mark O. Riedl. 2015. Toward Game Level Generation from Gameplay Videos. In *Proceedings of the FDG workshop on Procedural Content Generation in Games.*

[6] M. Guzdial, N. Sturtevant, and B. Li. 2016. Deep Static and Dynamic Level Analysis: A Study on Infinite Mario. In *Proceedings of the 3rd Experimental AI in Games Workshop.* 8.

[7] P. E. Hart, N. J. Nilsson, and B. Raphael. 1968. A formal basis for the heuristic determination of minimum cost paths. *IEEE Transactions on Systems Science and Cybernetics* SSC-4(2) (1968), 100–107.

[8] Britton Horn, Steve Dahlskog, Noor Shaker, Gillian Smith, and Julian Togelius. 2014. A comparative evaluation of procedural level generators in the mario ai framework. (2014).

[9] J. R. H Mariño, W. M. P. Reis, and L. H. S. Lelis. 2015. An Empirical Evaluation of Evaluation Metrics of Procedurally Generated Mario Levels. In *AAAI Conference on Artificial Intelligence and Interactive Digital Entertainment.*

[10] J. R. H. Mariño and L. H. S. Lelis. 2016. A Computational Model based on Symmetry for Generating Visually Pleasing Maps of Platform Games. In *Proceedings of the Conference on Artificial Intelligence and Interactive Digital Entertainment.*

[11] D. C. L. Ngo, A. Samsudin, and R. Abdullah. 2000. Aesthetic measures for assessing graphic screens. *J. Inf. Sci. Eng* 16, 1 (2000), 97–116.

[12] David Chek Ling Ngo, Lian Seng Teo, and John G Byrne. 2003. Modelling interface aesthetics. *Information Sciences* 152 (2003), 25–46.

[13] Christopher Pedersen, Julian Togelius, and Georgios N Yannakakis. 2010. Modeling player experience for content creation. *IEEE Transactions on Computational Intelligence and AI in Games* 2, 1 (2010), 54–67.

[14] Paolo Piselli, Mark Claypool, and James Doyle. 2009. Relating cognitive models of computer games to user evaluations of entertainment.. In *FDG*, Jim Whitehead and R. Michael Young (Eds.). ACM, 153–160.

[15] W. M. P. Reis, L. H. S. Lelis, and Y. Gal. 2015. Human Computation for Procedural Content Generation in Platform Games. In *Conference of Computational Intelligence and Games.* IEEE, 99–106.

[16] Santiago Londoño and Olana Missura. 2015. Graph Grammars for Super Mario Bros Levels. In *Proceedings of the Procedural Content Generation Workshop.*

[17] Noor Shaker and Moahamed Abou-Zleikha. 2014. Alone We Can Do So Little, Together We Can Do So Much: A Combinatorial Approach for Generating Game Content. In *Proceedings of AIIDE.*

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

[19] G. Smith, M. Treanor, J. Whitehead, M. Mateas, M. Treanor, J. March, and M. Cha. 2011. Launchpad: A Rhythm-Based Level Generation for 2D Platformers. *IEEE Transactions on Computing Intelligence and AI in Games* 3, 1 (2011), 1–16.

[20] Gillian Smith and Jim Whitehead. 2010. Analyzing the expressive range of a level generator. In *Proceedings of the 2010 Workshop on Procedural Content Generation in Games.* ACM, 4.

[21] Gillian Smith, Jim Whitehead, and Michael Mateas. 2010. Tanagra: A mixed-initiative level design tool. In *Proceedings of the Fifth International Conference on the Foundations of Digital Games.* ACM, 209–216.

[22] Sam Snodgrass and Santiago Ontanon. 2016. Learning to Generate Video Game Maps Using Markov Models. *IEEE TCIAIG* (2016).

[23] Adam Summerville and Michael Mateas. 2016. Super Mario as a String: Platformer Level Generation Via LSTMs. In *To Appear In Proceedings of the First International Conference of DiGRA and FDG.*

[24] Adam Summerville, Shweta Philip, and Michael Mateas. 2015. MCMCTS PCG 4 SMB: Monte Carlo Tree Search to Guide Platformer Level Generation. In *AAAI Conference on Artificial Intelligence and Interactive Digital Entertainment.*

[25] R. Tibshirani. 1994. Regression Shrinkage and Selection Via the Lasso. *Journal of the Royal Statistical Society, Series B* 58 (1994), 267–288.

[26] Christopher W Totten. 2014. *An architectural approach to level design.* CRC Press.

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