What Makes a Level Hard in Super Mario Maker 2?

Carlo A. Furia
Software Institute – USI Università della Svizzera italiana
Lugano, Switzerland
bugcounting.net

Andrea Mocci
Software Institute – USI Università della Svizzera italiana
Lugano, Switzerland
andrea.mocci@usi.ch

***Abstract*—Games like Super Mario Maker 2 (SMM2) lower the barrier for casual users to become level *designers*. In this paper, we set out to analyze a vast amount of data about SMM2 user-written levels, in order to understand what factors affect a level's *difficulty* as experienced by other users. To this end, we perform two kinds of analyses: one based on regression models and one using natural language processing techniques. The main results shed light on which level characteristics (e.g., its style, popularity, timing) and which topics and sentiments have a consistent association with easier or harder levels. While none of our findings are startling, they help distill some key differences between easy and hard SMM2 levels, which, in turn, can pave the way for a better understanding of end-user level design.**

***Index Terms*—platformer, difficulty, sentiment analysis, Super Mario Maker 2, end-user level design**

I. Introduction

Super Mario Maker 2 (SMM2) is a popular game for the Nintendo Switch that sold over eight million copies since its release in 2019.[1] The game's key innovation is combining a 2D platformer with a level editor: SMM2 users can just play the levels provided with the game, but can also create their own levels and share them with other users by uploading them to Nintendo's servers. Thus, every SMM2 user can be both a player and a *maker*; and we can say that SMM2 makers engage in a form of *end-user programming*.

Starting with its predecessor for the Wii U, a vibrant community of passionate gamers and content creators coalesced around the game. As part of their efforts, a large dump of SMM2 level data has been collected, reverse engineered, and made publicly available [22]. Broadly speaking, our goal is analyzing these SMM2 data with techniques commonly used in empirical software engineering research, in order to shed some light on this peculiar form of end-user programming.

In this paper, we focus on a fundamental question: *what denotes the difficulty of a level in* SMM2*?* Even casual players are well aware of the broad range of level difficulties one encounters in SMM2—from facile "little Timmy" levels to unforgiving kaizos.[2] Finding out what characteristics of a level are associated with its difficulty is thus a fundamental way of understanding how SMM2 users harness the game's features to create a broad variety of challenges and experiences. Our analysis considers both intrinsic level characteristics (e.g., its visual style) and community engagement (e.g., how many likes a level received).

After an overview of the data, we tackle the question in two ways. First, we perform a regression analysis of the level data to determine which variables contribute the most to a level's clear rate (a fundamental measure of its difficulty). Second, we use NLP machine learning techniques to perform a topic classification of the levels' titles, descriptions, and user comments, which is the basis for a qualitative analysis of how certain topics and sentiments are linked to a level's difficulty. These two analyses are complementary in the data they use and the insights they provide.

Data availability. For reproducibility, the detailed results and the analysis scripts are available in a replication package.[3]

II. Data Overview

The dump of SMM2 data we analyzed covers a whopping 26 609 725 levels—a comprehensive snapshot taken in February 2022 [22]. Let us give an overview of the level and user data, which we analyze in Sec. III.

A. Level Data

The level data include 36 variables; after culling undocumented and redundant variables, as well as others that are unsuitable for a regression analysis (e.g., text such as a level's title or user comments), we ended up with a selection of 22 variables, which characterize each level along different dimensions. Variables are of two main types: numeric and nominal (possibly ordinal); Tab. Ia and Tab. Ib overview several variables of each kind. Independent of their type, we group and color variables into categories (style, plays, timing, and difficulty) according to what aspect of a level's design they pertain to. Rather than tediously going through all variables systematically, let's illustrate those that are most relevant for Sec. III's analysis, while explaining how levels are made and played in SMM2.

When they create a level in the game's editor, SMM2 users can choose a *style* among four classic Nintendo games: SMB1 (Super Mario Brothers 1), SMB3 (Super Mario Brothers 3), SMW (Super Mario World), NSMBU (New Super Mario Brothers U), and SM3DW (Super Mario 3D World). A level's *theme* denotes the styling of its graphical elements to resemble environments such as a *castle*, a winter landscape with *snow*,

Work partially supported by SNF grant 200021-207919 (LastMile).

[1] Source: https://en.wikipedia.org/wiki/Super_Mario_Maker_2

[2] An overview of SMM2's lingo, which we'll occasionally use in the paper: https://supermariomaker2.fandom.com/wiki/Terminology

[3] https://dx.doi.org/10.6084/m9.figshare.28525223

	plays							timing			user				
	attempts	boos	clear rate	clears	comments	likes	players	autoscroll: speed	timer	world record	cleared	first clears	maker points	records	uploaded
min	0	0	0.00	0	0	0	0	0	10	−0	0	0	0	0	0
25%	22	0	0.10	4	0	0	10	0	300	10	80	1	780	0	10
50%	54	1	0.25	10	0	1	21	0	300	21	219	4	1 599	2	25
75%	127	2	0.50	28	1	4	48	0	300	44	534	11	2 441	7	52
99%	1 431	15	1.00	247	13	49	325	0	500	258	4 502	215	7 687	268	100
99.99%	142 163	576	1.00	15 557	906	4 590	18 171	2	500	2 628	57 813	7 803	22 633	10 674	100
max	67 895 905	37 886	1.00	1 804 430	120 515	564 890	1 302 862	2	500	6 000	513 124	67 102	29 100	105 379	100

(a) An overview of the main *numeric* variables in SMM2's data. Variables are grouped according to whether they refer to a level's plays statistics, timing, or the user who created it. For each variable, the table reports the minimum, maximum as well as the 25%, 50%, 75%, 99%, and 99.99% percentiles in the data.

difficulty				style					theme										version					
easy	normal	expert	super expert	SMB1	SMB3	SMW	NSMBU	SM3DW	airship	castle	desert	forest	ghost	overworld	sky	snow	underground	water	1.0.0	1.0.1	1.1.0	2.0.0	3.0.0	3.0.1
33	44	16	8	14	8	17	33	28	6	15	7	8	5	31	10	9	7	2	0	24	5	15	14	42

(b) An overview of the main *nominal* variables in SMM2's data. Variables are colored according to whether they refer to a level's style or to its difficulty rating. For each discrete value of each variable, the table reports the *percentage* of levels with that value.

TABLE I: An overview of the level and user data in SMM2.

a *forest*, etc. Throughout its life, SMM2 went through several *version* numbers, which affect some of the features available in the editor. Tab. Ib shows that, in the dataset we analyzed, the most popular game *style* is NSMBU, but all styles are fairly used; the most popular *theme* is *overworld* (possibly simply because it is the default for each style); and the game *version* with more levels is 3.0.1 (probably just because it was the most recent when the data was collected). Every level also has a *timer* (the number of seconds a player has to clear it), and may feature autoscroll with different speeds, as well as a clear condition (e.g., "do not take damage"). Tab. Ia shows that the median level[4] has a timer of 300 seconds (the default in the editor) and no autoscroll. Before a user is allowed to upload a level, they must clear it to show that it can be beaten. The system records the number of attempts (*upload attempts*) and the time (*upload time*) taken by the maker to clear a level before uploading it, as well as the timestamp of when the upload finally took place, and the maker's user id.

Once a level is uploaded, any SMM2 user can play it. For each level, the system keeps track of the total number of users who played the level (*players*) the total number of *attempts*, and of successful *clears*; the id of the first user who cleared the level and of the current world record holder, as well the *world record* itself (excluding the level's creator). Users can express their appreciation of a level with *likes* or *boos*, or by leaving *comments*: the corresponding variables record the total numbers of each. According to Tab. Ia, there is a huge spread in these metrics across levels, and a limited number of most popular levels generate massive amounts of plays, likes, and comments. More interesting, the median level generates modest, yet non-negligible, engagement (it was cleared 10 times, and played 54 times by 21 users, who left 1 like, 1 boo, and 0 comments); and only a small minority of levels never gets played.[5]

Most relevant for this paper, two variables characterize a level's difficulty. The *clear rate* is simply the ratio *clears*/*attempts*. In addition, SMM2 automatically assigns an ordinal *difficulty* rating with four levels: *easy*, *normal*, *expert*, and *super expert*. Tab. I shows that the median level is rated *normal*, and has a clear rate of 0.25 (i.e., 25%). The formula used by Nintendo to assign difficulty ratings is not known; however, it is likely based on the clear rate (although it's not *exclusively* based on it, since two levels with the same clear rate may get different difficulty ratings) and is dynamic (i.e., it may change over time). Intuitively, the clear rate is a more precise assessment of a level's difficulty, given that it is more fine grained. However, it's not a perfect measure: the clear rate does not distinguish between popular (which have been played by myriad players with all kinds of skills) and unpopular levels (whose clear rate may simply reflect the skills of the small, self-selected group of players who attempted it); and, ultimately, "difficulty" is partly subjective (e.g., depending on a player's familiarity with the mechanisms used by a game). Nevertheless, *clear rate* remains the best proxy for a level's difficulty in the metadata; hence, we'll rely on it in our analysis.

B. User Data

The user data include 28 variables; as for the level data, we distilled these down to 19 variables, which we grouped into the user category. Each user has a unique identifier, and is associated a country and a *region* (roughly corresponding to continents) of activity.

[4] With a little abuse of terminology, "median level" denotes a hypothetical level all of whose variables are at the level of the median.

[5] In no small part thanks to community initiatives such as Team 0%.

The system stores a few key metrics of user activity as a *player*: the total number of levels they *played* and successfully *cleared*, as well as the number of *attempted* plays, and how many were *deaths*;[6] Other variables record a user's *first clears* (the number of levels that they cleared before any other player) and world *records* (the number of levels that they cleared faster than any other player). Then, a few variables record the user's activity in specific game modes: their high *score* in the endless challenge (where a user plays random levels of a certain difficulty until they run out of lives) for each difficulty; and their *versus rating* in multiplayer versus (where four users race to reach the end of a level).

Finally, there are a few key metrics of a user's activity as a *maker*: the system assigns *maker points*, which reward a maker's activity and achievements; the number of levels a user *uploaded*,[7] and whether they allow other users to leave comments (variable *comments?*); and the total number of likes the maker's levels received (variable *maker likes*).

Just like for level data, Tab. Ia shows that there is a large spread in user data as well, and that most users have a non-trivial activity record. The median user cleared 219 levels, and claimed first clear on 4 and world record on 2. Remarkably, an ample majority of users are also *makers*: the median user collected 1 599 maker points by uploading 25 levels.

III. Regression Analysis of the Data

In order to quantitatively analyze which factors affect a level's difficulty, we fit on the level and user data generalized linear regression models of the general form:

$$\begin{aligned} \textit{clear rate}_i \sim\ & \mathsf{Normal}(\mu_i, \sigma) \\ \log(\mu_i) =\ & \alpha + \sum_{v \in \mathcal{V}} \beta_v \mathsf{v}_i \end{aligned} \tag{1}$$

In all models, variable *clear rate* is used as outcome, and other variables in $\mathcal{V}$ are used as predictors; the logarithmic link function ensures that the predicted *clear rate* is always a nonnegative value.

Fitting even a simple model such as (1) on SMM2's massive dataset is challenging even with plenty of memory, CPUs, and disk space; this restricted the techniques and models that we could use: *i*) classic frequentist fitting algorithms instead of more robust, yet computationally demanding Bayesian simulation-based techniques [11]; *ii*) a normal distribution as likelihood, instead of more precise choices (such as a beta distribution, which constrains the outcome to be a value over $[0, 1]$, like *clear rate*) that use less efficient fitting algorithms.

A. Choosing Linear Predictors

A key choice is which variables to include as predictors in (1). Again for practical performance reasons, we have to exclude a few level and user variables: a level's id, upload timestamp, clear conditions, and ids of the user who first cleared the level and of the one who holds the world record; and a user's id and country. All these variables are nominal with a very large number of levels; since an ℓ-level nominal variable is modeled as $\ell - 1$ binary indicator variables, such variable would effectively blow up the number of variables in $\mathcal{V}$, thus rendering (1) intractable.

Apart from this restriction, we include as many variables as possible, since we are only interested in *associations* and *prediction* (as opposed to causal relations, which we'll briefly analyze separately later). This angle also justifies including variables that correlate closely (i.e., they exhibit multicollinearity), as long as they do not introduce convergence problems. Based on these principles, we consider two models m_ℓ and m_u. Model m_ℓ includes as predictors in $\mathcal{V}_\ell$ all 15 viable level variables: *difficulty*,[8] *clears*, *attempts*, *likes*, *boos*, *players*, *world record*, *upload time*, *comments*, *timer*, *autoscroll*, *style*, *theme*, *version*, *upload attempts*. Model m_u tries to include, on top of the level data, the user data about each level's maker. Extending m_ℓ with all viable user variables incurs convergence problems; as a workaround, we excluded the three level variables with the smallest effect in m_ℓ (*players*, *upload time*, *timer*) and added all 16 viable user variables in $\mathcal{U}$: *played*, *cleared*, *attempted*, *deaths*, *maker points*, the high *score* in endless for each difficulty, the *versus rating*, *first clears*, *world records*, *uploaded* levels, *region*, *comments?*, and *maker likes*. In all, m_u uses the 28 predictors in $\mathcal{V}_u = \mathcal{V}_\ell \setminus \{\textit{players}, \textit{upload time}, \textit{timer}\} \cup \mathcal{U}$.

B. Regression Analysis of All Data

After fitting models m_ℓ and m_u on the SMM2 data, we can interpret its coefficients β as strength of association between a predictor variable and the clear rate. For a set of n variables $\mathsf{v}^1, \ldots, \mathsf{v}^n$, a datapoint $\vec{p}$ is an n-tuple $\langle p^1, \ldots, p^n \rangle$ of values for each variable. Consider two datapoints $\vec{p}, \vec{q}$ for the same set of n variables such that $p^i = q^i$ for all components i, except $p^k = q^k + x$. According to model (1), the ratio of the expected value of the clear rate of a level with data $\vec{p}$ over the expected value of the clear rate of a level with data $\vec{q}$ is thus:

$$\frac{\mu^p}{\mu^q} = \frac{\exp(\alpha + \beta_1 p^1 + \cdots)}{\exp(\alpha + \beta_1 q^1 + \cdots)} = \exp(\beta_k(p^k - q^k)) = \exp^x(\beta_k) \tag{2}$$

This leads to the following interpretation for numeric and for nominal variables:

numeric: the expected ratio of change of *clear rate* associated with a *unit* change of a numeric variable v is $\exp(\beta_\mathsf{v})$.

nominal: since a nominal variable v with n levels $x_1, \ldots, x_n$ is modeled as $n-1$ binary variables $\mathsf{v}_2, \ldots, \mathsf{v}_n$, the ratio of change of *clear rate* associated with a change of v from the baseline[9] level x_1 to level $x_{k>1}$ is $\exp(\beta_{\mathsf{v}_k})$.

In the following, we will actually analyze, for each variable v, the value $\delta(\mathsf{v}) = \exp(\beta_\mathsf{v}) - 1$. Since $\exp(\beta_\mathsf{v})$ is a ratio

[6] Thus, *played* and *cleared* refer to *levels*, whereas *attempted* and *deaths* refer to individual plays of any level.

[7] At any given time, a user can share 100 levels maximum; however, one may delete some levels and replace them with new uploads. Variable *uploaded* records the current number of published levels, and hence it is capped at 100.

[8] Interestingly, omitting *difficulty* from the model incurs convergence problems and leads to a fit with poor predictive capabilities.

[9] The baseline levels of nominal variables in our models are: *easy* for *difficulty*, *airship* for *theme*, *1.0.0* for *version*, *Asia* for *region*.

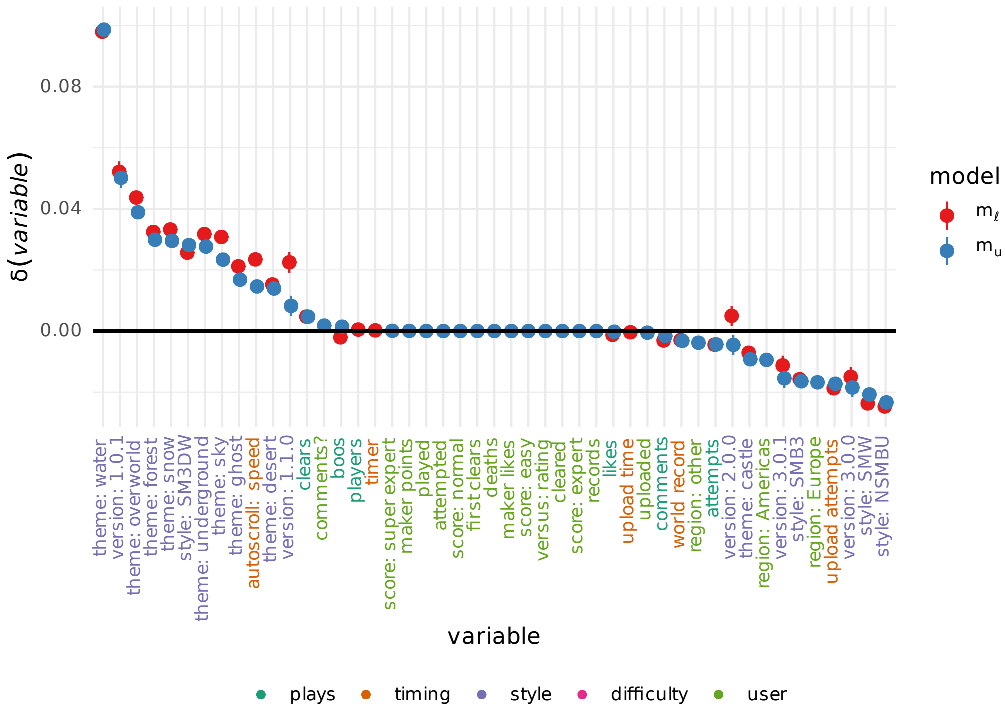

Fig. 1: For each regression variable v, the value of $\exp(\beta_v) - 1$, where β_v is v's coefficient in model m_ℓ and in model m_u.

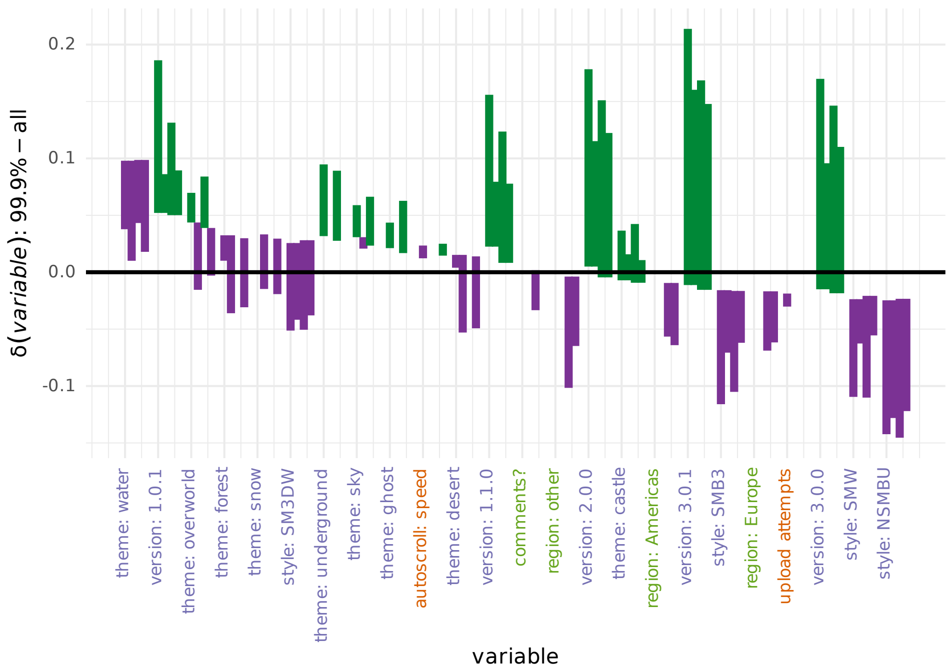

Fig. 2: Each bar is the difference $\tau^d_m(v)$ of $\delta(v)$ fitted on the 99.9% most popular levels and fitted on all level data. Bars ▌ represent positive differences and bars ▌ negative ones. For each v, there are (up to) four bars: $\tau^m_\ell(v)$, $\tau^p_\ell(v)$, $\tau^m_u(v)$, $\tau^p_u(v)$.

of means, $\delta(v)$ can be readily interpreted as the fractional increase (if positive) or decrease (if negative) of *clear rate* associated with a unit change in that variable.

Let's start with variable *difficulty*: in model m_ℓ, $\delta(\textit{normal}) = -0.54$, $\delta(\textit{expert}) = -0.84$, $\delta(\textit{super expert}) = -0.98$. This means that, on average, a *normal* level has a clear rate that is 54% lower than an *easy* level; an *expert* level's clear rate is 84% lower; and a *super expert* level's clear rate is 98% lower. This is reasonable, as it is unusual to find a super expert level with a clear rate that is much higher than 2%.

Given that *difficulty* is based on the clear rate, it is unsurprising that the corresponding δs have a disproportionate association with the outcome. The rest of the analysis, in particular Fig. 1, focuses on the δs for all *other* predictors in models m_ℓ and m_u.[10]

Most *user* variables have a negligible association with the outcome; in fact, model m_u has worse (higher) AIC score [16] than m_ℓ despite including many more variables, which suggests that m_ℓ outperforms m_u in predictive power. The only *user* variables with a noticeable association ($|\delta| > 10^{-3}$) with the outcome are *region* and *comments?*. Thus, makers based in Asia tend to build easier levels than makers in other regions; and makers who allow others to leave comments on their levels tend to design easier levels.

Among variables related to timing, the *timer* has negligible impact—probably because most levels stick to the default timer. The *autoscroll speed* is instead associated with higher clear rate; this is probably just a result of most levels not using autoscroll, and hence does not provide any clear insight. In contrast, the *upload time*, *world record*, and *upload attempts* are all associated with lower clear rates: naturally, if a level took a long time to clear it is usually harder. Variables *comments* and *attempts* are also all associated with lower clear rates, which would seem to indicate that levels that generate a lot of plays tend to be harder. Then, variables *likes* and *boos* do not have a consistent association one way or another; whereas variable *clears* is associated with easier levels, as these would easily pile up a lot of clears.

The impact of variables of group style is mixed. Fig. 1 suggests the following ranking of game *style*, from easier to harder: SM3DW, SMB1, SMB3, SMW, NSMBU. Each style from SMB1 to NSMBU extends the game with new mechanics (e.g., holding objects, spin jumps, wall jumps), which provide additional ways of building challenging levels. However, SM3DW also offers its own distinct game mechanics, which is hard to square with it being the style with the easier levels on average.[11] Newer *version*s of SMM2 are associated with harder levels; in this case too, this may reflect the additional items and features that have been added with every major version, as well as a more mature, skilled community of makers. Most *theme*s are associated with easier levels than those with the baseline *airship* theme; the exception is the *castle* theme, which is the style associated with the hardest levels. We might speculate that the features available in these themes (e.g., a *castle*'s lava floor or an *airship*'s bobbing camera) may support introducing additional challenges. Somewhat surprisingly, theme *water* has the strongest association with easier levels—despite the perception that water levels can be quite challenging due to the mechanics of swimming.

C. Regression Analysis of Popular Levels

Fig. 1 shows that the predictions of models m_ℓ and m_u are largely consistent (except possibly for certain *version*s). Does the relative impact of each variable on the clear rate change if we focus on a smaller set of *popular* levels? To answer this question, we introduce two datasets: D^p includes all 26 610 levels whose number of *players* is above the 99.9% percentile; and D^m all 26 678 levels whose maker's *maker points* are above the 99.9% percentile. For each variable v, $\tau^d_m(v)$ denotes the difference between $\delta(v)$ in model m_m fitted on D^d and the same model fitted on all data.

For several variables v, $\tau^d_m(v)$ is smaller than 0.01, and hence negligible. Fig. 2 plots the non-negligible values of $\tau^d_m(v)$. For several *theme*s and for the *SM3DW* style the difference crosses the zero line, which means that we see opposite trends in popular vs. all levels. In particular, popular SM3DW levels tend to be harder than popular SMB1 levels, which corroborates the conjecture that the variety of mechanics available in SM3DW support experienced makers to build harder levels. A similar observation holds for certain themes (*water*, *forest*, *snow*, *underground*, and *desert*) that tend to be harder to beat in popular levels than in all levels. Conversely, popular levels using theme *castle* tend to be easier; and the changes are inconsistent for the other themes. For most variables, however, the difference $\tau^d_m(v)$ does not cross the zero line; the most popular levels often present the same kind of associations between v and the level's clear rate, but in stronger form. For example, the difficulty ranking of styles *SMB3<SMW<NSMBU* still roughly holds for popular levels, but each has an overall stronger association with the clear rate.

For clarity, Fig. 2 omits variables related to *difficulty*, which range over quite different values than all other variables. While the strong association between difficulty rating and clear rate remains for *difficulty* variables, τ^m_ℓ and τ^m_u are usually negative, whereas τ^p_ℓ and τ^p_u are positive. Thus, within the same difficulty rating, popular makers tend to produce harder levels, but the most played levels tend to be a bit easier.

IV. NLP Analysis

We leverage NLP (natural language processing) techniques to analyze how level titles and descriptions (Sec. IV-A), and comments (Sec. IV-B) relate to a level's *difficulty*.

A. Titles and Descriptions

The title and the description of a level represent the creator's highlights of the level's content. To study how this correlates with the level's *difficulty*, we leverage BERTopic, a topic modeling approach [12]. BERTopic extracts latent topics from a collection of documents; in our case, each level is a document consisting of the concatenation of the level's title and description. The base BERTopic algorithm uses pre-trained transformer-based language models to build document embeddings, which are clustered by similarity to derive latent topics. Then, BERTopic derives topic representations according to their class-based term frequency-inverse document frequency (TF-IDF). We configured BERTopic to use words and bigrams as topics; in practice, a topic is a collection of words that tend to occur together in several documents. In the following, "topic-word" refers to any of the words that characterize a topic.

[10] This does not mean that we remove *difficulty* from our models, but simply that we zoom in on the other predictors.

[11] On the other hand, trivial "refreshing" levels abound in SM3DW, which might account for part of the difference.

Consistently with the game's worldwide popularity, SMM2 titles and descriptions are written in many different languages. We focus on English text, which is the most widely used language. Since the SMM2 dataset has no information about language, we use the LINGUA language detector [12], obtaining a total of 10 689 031 levels with title and descriptions in English.

To understand which topics characterize each difficulty, we calculate the relative frequency of a topic t for each difficulty d (easy, normal, expert, super expert) as the fraction of levels with difficulty d whose title/description matches topic t. Based on this metric, we call topic t a *characterizing topics* for difficulty d if t's frequency in levels of difficulty d is higher than its frequency in levels of other difficulties. Of the total 179 topics extracted with BERTOPIC, 52 are characterizing for *super expert*, 20 for *expert*, 37 for *normal*, and 70 for *easy*. Tab. II shows the top-5 characterizing topics for the *super expert* and *easy* difficulties.

topic		frequency per difficulty			
#	top-4 topic words	easy	normal	expert	super expert
#15	[speed, speedrun, run, seconds]	1.90%	2.28%	3.74%	4.45%
#13	[jump, jumps, jumping, long]	0.93%	1.59%	2.55%	3.15%
#32	[level, lol, sorry, little]	0.22%	0.49%	1.56%	2.80%
#56	[practice, tricks, jumps, basic]	0.07%	0.15%	0.61%	2.37%
#25	[ride, spin, run, victory]	0.36%	0.59%	1.07%	1.45%
#0	[maker, super, bros, 11]	9.51%	7.57%	5.90%	4.75%
#5	[level, level easy, easy level]	3.86%	3.46%	3.57%	3.69%
#30	[world, 11, 12, 13]	2.36%	1.81%	0.98%	0.65%
#16	[hard, easy, try, impossible]	1.89%	1.78%	1.77%	1.61%
#40	[water, dangerous, madness, life]	1.85%	1.01%	0.73%	0.57%

TABLE II: Top-5 characterizing topics, by frequency, in titles and descriptions of *super expert* (top) and *easy* (bottom) levels.

Topics characterizing super expert levels. The most frequent topic-word characterizing *super expert* levels is **speed**. Overall, **speed** is a topic-word in 4 topics, of which 3 are characterizing for *super expert*. Similarly, **jump** is a topic-word in 14 topics, of which 13 are characterizing for *super expert*. All the 4 topics containing **spin** as topic-word, and all the 5 topics containing **practice**, are also characterizing for *super expert*. These probably indicate levels that let players practice and master a variety of SMM2's challenging game mechanics; in fact, these topics also include topic-words such as **jump**, **spin**, **fly**, and **tricks**.

Topics characterizing easy levels. The most frequent topic-words characterizing *easy* levels are very generic terms like **maker** and **world**, which probably just means that easy levels tend to have generic titles and descriptions. Also self-explanatory is the second-most occurring topic-word: **easy**; among the 14 topics that contains it, 7 are characterizing for the *easy* difficulty. One seemingly unexpected topic-word is the number **11**; this is actually due to the way BERTOPIC encodes topic-words by removing non-alphanumeric characters. Thus, **11** actually stands for **1–1**, which is the canonical way of referring to the first level of a multi-level world and, in particular, to the iconic *World 1–1* in the original Super Mario Bros. Replicating such classic level in SMM2, often with unique twists, is its own sub-genre;[13] our findings indicate that this is particularly popular with easy level—although 1–1 variants are found at all difficulties. Other topic-words consisting of two-digit numbers (e.g., 12, 13) have a similar explanation: they stand for *world–level* identifiers (e.g., 1–2, 1–3). These are either replicas of original Super Mario Bros. levels, or identify levels that belong to *Super Worlds*.[14] Some of the remaining topics characterizing easy levels seem to contradict the levels' difficulty ranking, as they include topic-words such as **hard**, **impossible**, **dangerous**, and **madness**. A possible explanation is the makers' taste for deliberately misleading "trolling"[15] titles, whereupon trivial levels are marked "100% impossible" and very hard ones are titled "easy".

B. Comments

While titles and descriptions express the maker's point of view on their levels, comments allow any players to voice their opinion. We first look at the *sentiment* of level comments, and then at their *topics*. As in Sec. IV-A, we only consider comments written in English, as detected by LINGUA. To have a meaningful set of comments for each level, we only consider levels with at least 6 comments (3% of all levels), corresponding to a total of 100 148 levels and 1 996 367 comments. Let $C(\ell)$ denote the set of comments of a level ℓ.

Sentiment. We analyze the sentiments of the comments by using Barbieri et al.'s [6] transformer-based pipeline. For a comment c, the model estimates its sentiment $S(c)$ as a triple $\langle S_-(c), S_0(c), S_+(c)\rangle$ of scores representing the fraction of *negative* $-$, *neutral* 0, and *positive* $+$ sentiment, where $S_i(c) \in [0, 1]$ and $\sum_i S_i(c) = 1$, $i \in \{-, 0, +\} \equiv K$. Consider the following derived metrics, summarized in Tab. III:

- The *average sentiment* $S_i(\ell)$ of ℓ is the mean of $S_i(c)$ over all comments c of level ℓ.
- The *dominant sentiment* $D(c) \in K$ is the sentiment with the highest score in $S(c)$: $D(c) = \operatorname{argmax}_i S_i(c)$.
- $C_i(\ell) \subseteq C(\ell)$ is the set of all comments of ℓ whose dominant sentiment is $i \in K$.
- $D_i(\ell)$ is the fraction $|C_i(\ell)|/|C(\ell)|$ of ℓ's comments whose dominant sentiment is $i \in K$.
- Similarly, $\overline{S_i(X)}$ and $\overline{D_i(X)}$ denote the mean of $S_i(\ell)$ and $D_i(\ell)$ for all levels of difficulty in the set X of difficulties.

A trend visible in Tab. III is that *super expert* level comments display a higher negative average sentiment than other difficulties; this holds both for the S_- and the D_- metrics, that is whether we aggregate by average or by dominant sentiment. To quantify this observation, we compute Cliff's delta—a non-parametric effect size, suitable to quantify how often the values in one set are larger than the values in another, independent set,

[12]See https://github.com/pemistahl/lingua-py.

[13]See for example https://www.youtube.com/watch?v=PhyG0s9tJaM

[14]See https://supermariomaker2.fandom.com/wiki/Super_Worlds

[15]https://docs.google.com/document/d/13ZoqeblLs45HuEfTtsOrq6X0LAuEnA8nB721_doxE38

difficulty x	$d[S_i^x]$ −	$d[S_i^x]$ +	$d[D_i^x]$ −	$d[D_i^x]$ +	$\overline{S_i(\{x\})}$ −	$\overline{S_i(\{x\})}$ +	$\overline{D_i(\{x\})}$ −	$\overline{D_i(\{x\})}$ +
easy	−0.03	−0.09	−0.04	−0.09	0.22	0.38	0.20	0.36
normal	−0.09	0.08	−0.09	0.07	0.21	0.42	0.19	0.41
expert	0.03	0.06	0.04	0.06	0.23	0.42	0.21	0.41
super expert	0.24	−0.05	0.25	−0.03	0.27	0.38	0.26	0.37

TABLE III: An overview of the sentiments in level comments. The table shows the Cliff's delta of the relationship between a difficulty and all other difficulties, as well as the mean of the average sentiment $S_i(\ell)$ and of the fraction of dominant sentiment $D_i(\ell)$ over all levels ℓ of each difficulty x.

without assumptions about their underlying distributions. Let $d[S_i^x]$ denote Cliff's delta of the differences between $\overline{S_i(\{x\})}$ and $\overline{S_i(\{y \neq x\})}$, i.e., between the average S_i for all levels of difficulty x and the average S_i for all other levels; similarly, $d[D_i^x]$ denotes Cliff's delta of the corresponding difference of metric D_i. Tab. III shows that $d[S_-^{\text{super expert}}] = 0.24$ and $d[D_-^{\text{super expert}}] = 0.25$, Such effect sizes are usually considered *small* but not negligible [18];[16] in contrast, $d[S_i^x]$ and $d[D_i^x]$ are *negligible* for all other difficulties x other than *super expert*, and in all ratings corresponding to positive sentiment, which supports our observation that players of super expert levels tend to express more negative sentiments.

Topics. We analyzed English user comments by applying BERTopic as in Sec. IV-A, aggregating comment topic occurrences by *level*; thus, a topic t has occurrence frequency $\tau\%$ for a certain difficulty x if $\tau\%$ of the x-difficulty levels include at least one comment with topic t. Tab. IV lists the top-5 characterizing topics for the *super expert* and *easy* difficulties.

#	top-4 topic words	easy	normal	expert	super expert
#0	[level, great level, good level]	39.30%	45.46%	48.88%	55.66%
#5	[hard, easy, beat, challenge]	9.82%	12.16%	17.64%	22.24%
#10	[jump, spin, jumping, jumped]	7.01%	7.92%	11.10%	13.81%
#50	[shell, jump, throw, double]	1.49%	2.62%	4.14%	8.85%
#38	[troll, bully, mad, lad]	3.78%	3.82%	4.63%	8.81%
#1	[mario, toad, bros, maker]	24.11%	21.49%	16.37%	12.42%
#2	[nintendo, switch, minecraft]	16.95%	14.37%	14.29%	14.90%
#3	[course, great course, fun course]	16.15%	15.00%	13.79%	11.74%
#4	[awesome, good job, thanks]	16.12%	14.31%	12.58%	9.44%
#8	[brain, eyes, pain, dad]	12.42%	9.18%	9.39%	9.80%

TABLE IV: Top-5 characterizing topics, by frequency, in comments of *super expert* (top) and *easy* (bottom) levels.

Comment topics characterizing super expert levels. The most frequently occurring characterizing topic in super expert levels seems to denote appreciation (e.g., **great/good** level); indeed, the average + sentiment for comments with this topic is 0.67, which is comparatively high. In contrast, topic 38 indicates dislike, and the average − sentiment for comments with this topic is 0.46, which is also above average. This finding complements the previous observation about negative sentiments in super expert comments: while users are frustrated by trolly or gratuitously hard levels, they arguably appreciate when the difficulty is fair and determines a challenging but satisfying level. The other characterizing topics of super expert level comments are in line with Sec. IV-A's analysis, showing frequent mention of topic-words that suggest challenging game mechanic—such as **shell jumps** and **double**/triple jumps.

Comment topics characterizing easy levels. Topics #3 and #4 in Tab. IV seem to indicate generic user appreciation of a level; in fact the average + sentiment for comments with these topics are 0.62 and 0.86 respectively—both clearly above average. In contrast, topic #8 leans towards negative; its average negative sentiment is 0.34, clearly above the average negative sentiment of *easy* levels. Topic-words such as **brain**, **eyes**, and **pain** suggest that the most common criticism of easy levels targets those that make an unnecessary, excessive usage of flashy effects,[17] which can significantly deteriorate the user experience. Finally, topics #1 and #2 seem generically positive (or possibly neutral) comments. However, topic-word **Minecraft** stands out, which may indicate attempts at translating some of Minecraft's features as a SMM2 level; the stronger association with easy levels might indicate that such attempts tend to be uninteresting from the perspective of platforming level design.

V. Threats to Validity

The operationalization of the concept of level *difficulty* is a fundamental threat to *construct* validity. While *clear rate* and *difficulty* are reasonable proxies, they may fail to capture other aspects of a level's difficulty, including a user's *perception*; analyzing these aspects would require a different kind of study.

Since we analyzed observational data, our statistical analyses merely report *associations* rather than genuine causal relations—which is a fundamental threat to *internal* validity. This paper's replication package includes a structural analysis of causal relations, which seems to be consistent with the main model presented in the paper; while a reliable causal analysis requires controlled experiments, even purely predictive models can give interesting insights. BERTopic was trained on general text, which may not adequately cover SMM2's vernacular; fine-tuning an NLP model on the texts commonly used in comments may thus improve its effectiveness for our analysis.

Our NLP analysis was limited to English descriptions and comments, which may restrict the *external* validity of our results (i.e., their generalizability). Extending our analysis to other linguistic groups (e.g., Japanese, which is the second most used language in SMM2) belongs to future work.

VI. Related Work

Difficulty is a fundamental characteristic of videogame experience [7], which has been investigated in disparate ways [2], [4], [5]. For *platformers* specifically, a very recent work introduced a model to automatically evaluate their difficulty based on a level's structure [10]; applying this model to SMM2 levels is an interesting future work direction. The *Platformer Experience Dataset (PED)* [15] corpus collects various kinds

[16]Null-hypothesis statistical testing is uninformative in this case (possibly in general [8]), since the sheer amount of data leads to minuscule p-values.

[17]https://supermariomaker2.fandom.com/wiki/Sound_Effects

of data that capture the experience of Super Mario Bros. players with rankings, game content recordings, and visual recordings of players. Another approach [19] measures difficulty in platformers through a machine learning model trained on both performance telemetry data and emotional data estimated from electrodermal activity. All these approaches focus on precisely measuring the difficulty of individual levels and player runs. In contrast, our paper mainly analyzes *metadata*, such as a level's style or description; by focusing on such data, we have access to large amounts of user information, including measures, such as the clear rate, that are normally not available for single-player games. Our approach is also justified by the specific game we targeted: in SMM2, (some) players are also level designers, which adds an interesting dimension to the analysis. While metadata is arguably more coarse-grained than, say, a level's detailed structure, our work demonstrated that it can provide complementary, broad-stroke insights.

A very different angle to analyze games is as models of computation [1], [3], [20]. For example, the undecidability [13] and the computational complexity [3], [9] of classic Nintendo games, in particular Mario games, has been painstakingly studied. NLP sentiment analysis has been applied to artifacts related to videogames, including reviews [24], chats [23], and player speech [21]. Other work focused on specific (sub)genres, like Virtual Reality [14] or "Souls-like" games [17]; their findings include that player negative emotions can also be influenced by cultural elements.

VII. Conclusions

This paper analyzed a treasure trove of data about SMM2 levels and users, with the goal of studying the factors that are associated with a level's difficulty. The main findings include: *i*) Levels that generate a lot of plays tend to be harder. *ii*) While most variables characterizing makers have negligible associations with level difficulty, makers with more experience tend to design easier levels (possibly because they are better designers). *iii*) Usually, older game styles (e.g., Super Mario Bros. 1) tend to be associated with easier levels than newer game styles (e.g., Super Mario World)—especially in popular levels. *iv*) Descriptions of the harder levels often refer to challenging game mechanics, whereas easy levels' descriptions may be bland or generic. *v*) Comments expressing negative sentiments are more common in harder levels. *vi*) However, several frequently occurring comment topics in the same levels express appreciation—when the difficulty is justified by a high-quality design.

References

[1] Zachary Abel and Della Hendrickson. Baba Is Universal. In *Conference on Fun with Algorithms (FUN)*, pages 1:1–1:15, 2024.

[2] Justin T Alexander, John Sear, and Andreas Oikonomou. An investigation of the effects of game difficulty on player enjoyment. *Entertainment computing*, 4(1):53–62, 2013.

[3] Greg Aloupis, Erik D. Demaine, Alan Guo, and Giovanni Viglietta. Classic Nintendo games are (computationally) hard. *Theoretical Computer Science*, 586:135–160, 2015.

[4] Maria-Virginia Aponte, Guillaume Levieux, and Stéphane Natkin. Scaling the level of difficulty in single player video games. In *Proc. of ICEC 2009*, pages 24–35. Springer, 2009.

[5] Maria-Virginia Aponte, Guillaume Levieux, and Stéphane Natkin. Difficulty in videogames: an experimental validation of a formal definition. In *Advances in computer entertainment technology*, pages 1–8, 2011.

[6] Francesco Barbieri, Jose Camacho-Collados, Luis Espinosa Anke, and Leonardo Neves. TweetEval: Unified benchmark and comparative evaluation for tweet classification. In *Findings of the ACL: EMNLP*, pages 1644–1650, 2020.

[7] Jenova Chen. Flow in games (and everything else). *Commun. ACM*, 50(4):31–34, April 2007.

[8] Jacob Cohen. The earth is round ($p < .05$). *American Psychologist*, 49(12):997–1003, 1994.

[9] Erik D. Demaine, Giovanni Viglietta, and Aaron Williams. Super Mario Bros. is Harder/Easier Than We Thought. In *Conference on Fun with Algorithms (FUN)*, pages 13:1–13:14, 2016.

[10] Yannick Francillette, Hugo Tremblay, Bruno Bouchard, and Bob-Antoine Menelas. A comprehensive model of automated evaluation of difficulty in platformer games. *ACM Games*, 3(1), January 2025.

[11] Carlo A. Furia, Richard Torkar, and Robert Feldt. Applying Bayesian analysis guidelines to empirical software engineering data. *ACM TOSEM*, 31(3):40:1–40:38, 2022.

[12] Maarten Grootendorst. Bertopic: Neural topic modeling with a class-based tf-idf procedure. *arXiv preprint arXiv:2203.05794*, 2022.

[13] MIT Hardness Group, Hayashi Ani, Erik D. Demaine, Holden Hall, Ricardo Ruiz, and Naveen Venkat. You Can't Solve These Super Mario Bros. Levels: Undecidable Mario Games. In *Conference on Fun with Algorithms (FUN)*, pages 22:1–22:20.

[14] Tibor Guzsvinecz and Judit Szűcs. Investigation of sentiments in virtual reality game reviews. In *IEEE International Conference on Cognitive Infocommunications*, pages 000011–000016, 2024.

[15] Kostas Karpouzis, Georgios N. Yannakakis, Noor Shaker, and Stylianos Asteriadis. The platformer experience dataset. In *Affective Computing and Intelligent Interaction (ACII)*, pages 712–718, 2015.

[16] Richard McElreath. *Statistical rethinking: A Bayesian course with examples in R and Stan*. CRC press, Florida, USA, 2 edition, 2020.

[17] Sicheng Pan, Gary J. W. Xu, Kun Guo, Seop Hyeong Park, and Hongliang Ding. Cultural insights in souls-like games: Analyzing player behaviors, perspectives, and emotions across a multicultural context. *IEEE Transactions on Games*, 16(4):758–769, 2024.

[18] J. Romano, J.D. Kromrey, J. Coraggio, and J. Skowronek. Appropriate statistics for ordinal level data. In *Annual meeting of the Florida Association of Institutional Research*, pages 1–3, 2006.

[19] Marcos P. C. Rosa, Eduardo A. dos Santos, Iago L. R. de Moraes, Tiago B. P. e Silva, Mauricio M. Sarmet, Carla D. Castanho, and Ricardo P. Jacobi. Dynamic difficulty adjustment using performance and affective data in a platform game. In *Proc. of HCI International 2021*, page 367–386, Berlin, Heidelberg, 2021. Springer-Verlag.

[20] Matthew Stephenson, Jochen Renz, and Xiaoyu Ge. The computational complexity of angry birds. In *Proceedings of IJCAI 2020*, 2021.

[21] Philipp Sykownik, Felix Born, and Maic Masuch. Can you hear the player experience? A pipeline for automated sentiment analysis of player speech. In *2019 IEEE Conference on Games*, pages 1–4.

[22] TheGreatRambler. Mario Maker 2 datasets. https://tgrcode.com/posts/mario_maker_2_datasets, sep 2022.

[23] Joseph J Thompson, Betty HM Leung, Mark R Blair, and Maite Taboada. Sentiment analysis of player chat messaging in the video game StarCraft 2: Extending a lexicon-based model. *Knowledge-Based Systems*, 137:149–162, 2017.

[24] Markos Viggiato, Dayi Lin, Abram Hindle, and Cor-Paul Bezemer. What causes wrong sentiment classifications of game reviews? *IEEE Transactions on Games*, 14(3):350–363, 2022.