An Examination of the Hidden Judging Criteria in the Generative Design in Minecraft Competition

SUBMISSION DATE / POSTED DATE

11-05-2023 / 12-05-2023

CITATION

Hervé, Jean-Baptiste; Salge, Christoph; Warpefelt, Henrik (2023): An Examination of the Hidden Judging Criteria in the Generative Design in Minecraft Competition. TechRxiv. Preprint. https://doi.org/10.36227/techrxiv.22698484.v1

DOI

10.36227/techrxiv.22698484.v1

An Examination of the Hidden Judging Criteria in the Generative Design in Minecraft Competition

Jean-Baptiste Hervé, Christoph Salge, and Henrik Warpefelt

***Abstract*—Game content has long been created using procedural generation. However, many of these systems are currently designed in an ad-hoc manner, and there is a lack of knowledge around the design criteria that lead to generators producing the most successful results. In this study, we conduct a qualitative examination of the comments left by judges for the 2018–2020 *Generative Design in Minecraft* competition. Using abductive thematic analysis, we identify the core design criteria that contribute to a generator that creates "good" content – here defined and interesting or engaging. By performing this study, we have identified that the core design criteria are usability of the settlement environment, the thematic coherence within the settlement, and an anchoring in real-world simulacra are the main factors that create an interesting settlement.**

***Index Terms*—Procedural Content Generation, Games, Minecraft, Human Factors**

I. Introduction

In this paper we perform a methodical and qualitative examination of the comments made by judges of the 2018, '19 and '20 Generative Design in Minecraft Competition (GDMC) [16, 18]. The GDMC is about creating a procedural generator that can generate a "good" or "interesting" settlement for any given map. It was designed to foster interest in Procedural Content Generation (PCG) and adaptive computational creativity and co-creativity. From a computational creativity standpoint the core challenges are the adaptation to unknown content (i.e. the input map), and the vaguely defined quality criteria. Consequently, the competition relies on human expert judges to evaluate the generated settlement quantitatively through scores according to certain scales, as well as qualitatively as written comments.

The instructions given to judges both talk in general terms about using their best judgment to determine the quality of each settlement, and to rate the settlement in four categories, each of which are introduced by a short description, and a list of explicitly illustrative not exhaustive criteria. This immediately raised the question: do those criteria capture the overall quality assessment of the judges, or are there further elements that are currently not covered? In 2021 one of the judges of the competition mentioned the Discord server of the GDMC competition[1] that they scored all settlements based on the scoring guide and their favorite settlement generator was not the one with the most points. By looking at the written, optional comments over the years, we hope to identify, what might be the missing or hidden criteria that caused this discrepancy. Another relevant complaint participants have made occasionally is that they feel it is evident from the evaluations that some judges do not actually play Minecraft. This indicates that there might be both an element of experience and actual interaction, or a lack thereof, that can be perceived in the judgment text.

In this paper we are examining the written feedback made by the human judges that accompanies their numerical score of the generated settlements in the competition. Both the scores and comments provided by the judges are made publicly available by the competition organizers [2].

The primary aim of this paper is to gain a better understanding of criteria the judges of the competition *actually* use to ascertain the quality of a Minecraft settlement. In doing so, we will support work towards the development of formalized and automatic metrics for procedural content generation. Although there exists a range of metrics for this purpose, few of them have been explicitly compared to human qualitative judgment, or have shown to be good predictors of human quality assessment, particularly in the domain of Minecraft settlements. Many existing metrics focus more on measuring what the generator can express rather than how this content is received and interpreted by players. Additionally, we currently know of no works that have studied human self-reported quality judgments for any kind of procedural content generation in games. While this work is focused on Minecraft settlements, we still hope that this analysis can shed some light on how humans evaluate PCG in general.

II. Background

In this section, we will start by introducing the game Minecraft and the GDMC competition. We will then present theory critical to understanding Procedural Content Generation, as well as how players interpret and evaluate PCG artifacts.

A. Minecraft and the GMDC competition

Minecraft [12] is a voxel based game developed by Mojang Studio, where the players progress in an open world made out of blocks. These blocks represent different materials, such as wood, rock, or coal. Players can destroy blocks, place them in any position within the world, or even combine them through crafting mechanics in order to create new types of block or item. Minecraft is mostly known for its open-endedness. Even if the game offers a main objective, which include visiting alternative dimensions and fighting a dragon, it is mostly used

Manuscript received April 13, 2023

[1]https://discord.gg/MtYJfsUnVN

[2]https://gendesignmc.engineering.nyu.edu

Fig. 1. An panoramic view of a Minecraft settlement generated for the competition

as a sandbox game. Many players use the block mechanic to terraform the game world, create structures such as houses, castles or cities, and create their own game's rules. Since the art style and setting of Minecraft are very generic, the game affords free creation of almost any kind of artifact, with only the player's imagination setting the limits.

The GDMC is a yearly competition in which teams submit a settlement generator [15]. These generators work by adding and removing Minecraft blocks, the same way a player would do. All the submitted generators are then tested on 3 maps with a fixed size of 256x256 blocks, which are selected by the organizers [18]. An example of these settlements can be see in Figure 1 on page 2. All the generated settlements are then sent to the jury. The jury includes experts in various field, such as Artificial Intelligence (AI), Game Design or urbanism. Each judge scores the settlements between 0 to 10 points, in each of the following categories : *Adaptability, Functionality, Narrative, and Aesthetic*. *Adaptability* is how well the settlement is suited for its location - how well it adapts to the terrain, both on a large and small scale. *Functionality* is about what affordances the settlement provides, both to the Minecraft player and the simulated villagers. It covers various aspects, such as food, production, navigability, security, etc. *Narrative* reflects how well the settlement *itself* tells an evocative story about its own history, and who its inhabitants are (There is a separate bonus challenge about also adding a written PCG text that tells the story of the settlement [17]). Finally, *Aesthetic* is a rating of the overall look of the settlements. In the competition, the rating of each category is computed for each generator by averaging (mean) across all judge's scores. The rating works in the following way : a grade of 5 means that the result looks human made, a 6-9 correspond to what we would expect from an expert human, and finally a 10 would be attributed to a "superhuman performance". After they have examined each generator's performance on each of the three maps included in the competitions, the judges provide a score for each of the four categories.

In addition to these ratings, judges provide qualitative written feedback for each generated settlements. The comments left by judges are wide-ranging, and describe the feelings evoked by the settlement, the perceived quality of the generated artifacts, and the ways in which the generated content does or does not fulfill their expectations. The comments also serve as a connection between the real-world understanding of PCG artifacts. However, in order to provide context for our evaluation of these comments we must first describe some theory of PCG.

B. Procedural Content Generation

Procedural Content Generation (PCG) is the creation of content through algorithmic mean. In computer science, it usually refers to a software, a *generator*, which produces content. A single output from a generator is also commonly referred as an *artifact*. PCG techniques are used in particular in the video games industry, where generators are used to produce gameplay elements (levels, items, etc), aesthetic elements (trees, buildings, characters, etc) or even narrative elements (quests, lore, dialogues, etc). It has been used in a wide range of games, with different genre and ambition. It can be used in many different ways, from a production tool to a core

element of the game itself. The procedural generation of game content is a complex process, and as such there exists a need to describe some of the issues inherent in the nature of PCG.

C. Possibility Space

For a given type of artifact we intend to generate, the *Possibility Space* [6] represent the range of any artifact we can think of. If we take the GDMC as an example, the Possibility Space of an entry would be any combination of 256x256x256 Minecraft blocks, matching the dimensions within the competition occurs. In the same time, a generator might produces artifacts in only a portion of our Possibility Space, that we will refer as our *Generative Space* [6]. The generative space is a space contained within our possibility space, and part of the work of the creation of a generator is actually designing its Generative Space. A GDMC competitor wants to prune the Possibility Space, firstly to remove anything that is not a settlement, but also to give its own style to their generator. The generative space could be constrained for instance to contains only medieval looking village or modern city. It is also important to point out such space is also limited by technical elements. Translating one's vision into a software is not a simple task, and therefore the Generative Space can be impacted by the inability from the generator to design certain aspect of the artifact, or even bugs.

A common flaw in procedural content generation is the repetitiveness of the artifacts, which leads to less interest into the generated content. This effect has been defined as the *Kaleidoscope Effect* [2], which occurred as the player begins to visualize the Generative Space of a generator and its boundaries. Once the Generative Space is fully identified, players can guess the nature of the next artifacts.

But even the more limited Generative Space is often to complex to represent - hence *Expressive Range* [19] Analysis is commonly employed. The Expressive Range is similar in nature to the Generative Space, but is defined by human selected dimensions. The analysis of the Expressive Range of a generator can be useful in order to understand its behavior and how the artifacts are spread among the dimensions. Therefore, the usefulness of an Expressive Range depends mostly on the relevance of the dimensions by which it is defined. Usually, dimensions used are automatically computed metrics applied on the whole artifact. They do not necessarily need to capture something associated with quality, but there is often an underlying assumption that higger values in certain dimensions is preferred. More importantly thought, is that the metrics capture meaningful differences, so artifacts lying in different areas of the expressive range appear different to the relevant players. From the idea of similarity and difference among generated artifact, the concepts of *Perceptual differentiation* and *Perceptual uniqueness* were introduced [4]. *Perceptual differentiation* is the feeling that an artifact is different in some way from the previous one, while *Perceptual uniqueness* occurs when a single artifact is distinguishable and has its own character. Those two definitions rely on the individual perception and feeling of a player, and therefore are hard to capture through computational means. While a generator can create a large amount of technically different and unique artifacts, their differences might not be judged relevant for a human, and it is less likely that one of them end up being perceived as unique. The PCG research community has developed tools aiming at automatically compute these distinctions. The most established one is the Expressive Range Analysis (ERA) [19], which is used in numerous scientific publication as an illustration of a generator's capabilities. There is even a will to make this analytical tool as accessible as possible, beyond the scientific community [5]. But this method has received critics and efforts among the field are being made to improve it [20]. Obviously ERA is not the single analytical tool available and used, but ultimately it is usually a matter of representing the distribution of artifacts along the chosen dimensions. Eventually Perceptual differentiation remains difficult to evaluate in an consistent way and automated evaluation of Perceptual uniqueness is still a challenge that has to be tackled.

Within the field of PCG, a truly qualitative evaluation of these artifacts is still challenging. Several experiments have already been conducted with the intent to critically examine some of the commonly used metrics and their relevance [10, 8]. Although part of the tool set is pertinent in evaluation scenarios, it is also clear that we are missing player-driven evaluation methods. We believe that new metrics, built with the goal of capturing human perception, could be helpful for tackling the issues related to PCG listed previously.

D. Player Interpretation

The following sections present a theoretical framework that aims to describe the nuances of player-driven evaluation of PCG artifacts. It should be noted that many of these theories refer to the "user". In this paper we replace the word "user" with "player", which for our purposes is functionally identical but thematically more correct. Although the term *experiencer* may technically be more apt, the use of "player" more closely matches our intended interpretation.

At its core, the player-driven evaluation of PCG artifacts is centered around the player's understanding of what they are being shown by the game. As described by Warpefelt [21], the player observes a collection of details, or what Warpefelt calls *indicators* presented by the game, interprets them, and forms expectations on the game. However, the player interpretation and forming of expectation is a complex process, which is influence by how the player is situated (when and where they are playing), their previous experiences with this or other games, and what expectations have been set before the player begins their play session (for example by advertising or reviews). These factors together work as a lens through which the player forms an interpretation of the gaming experience. From this interpretation arises the player's experience.

Within the context of this paper, we are focusing on two of the main parts of the underlying theories that describe the player's situation: how the environmental storytelling [7] of the game sets expectations and provides indicators [21], and how the player's previous experience can be described in terms of the human-computer interaction of *character* [9].

1) Character: The concept of *character* was first introduced by Janlert & Stolterman [9] in 1997. Through the evaluation of character, we are able to discern the nature of objects that we observe – and to understand how these objects are different from one another. More formally, character encapsulates how we can understand what objects are, and discern the difference between subclasses of objects, say a sports car versus an SUV. Character is composed of a number of *characteristics* that help the viewer of an artifact understand its nature. Through the evaluation of character, we are able to understand how we can use an object (which affordances it provides) and what we can expect from an object.

2) Mechanisms of player interpretation: Warpefelt [21] has incorporated the concept of characteristics into the indicator theory, where each indicator not only feeds into the usability aspects (signifing of affordances) but also the narrative aspects (indices for storytelling [7, 14]) and the setting of expectations (through characteristics). Together, these factors allow us to deconstruct how the player interprets the game using a bottom-up approach focused on the examination of how the interpretation of details feed into the player's holistic understanding of the game. This approach is especially useful for PCG artifacts, since their creation is by nature detail-oriented.

3) Player interpretation and the game environment: Players also interpret the environment in order to recreate the context and the narrative environment in which they evolve. The ruins of a castle and gigantic skyscrapers both lead to different conclusions regarding the time period, the region, or even the series of event that occurred in that place. This storytelling strategy is commonly used in level design, and has been defined as *indexical storytelling* [7]. Behind this term is the concept of stories told through traces, or *indices* as defined by Peirce [14], which the player connects in order to recreate the context of a place or the past events. However, the player has to be able to interpret the indices and their connections [13], referencing their repertoire of character [9]. Furthermore, the player themselves can contribute to the environmental storytelling, as their own actions can leave traces. Thus, the game's story is not not only the one intended by the designers, but also one created and influenced by the player. In essence, this is the mechanic by which the player iteratively refines and evolves the materials used in the construction of their alterbiography [1].

Nitsche [13] exposes how architectural features and their characteristics (in the parlance of Janlert & Stolterman [9]) have an evocative effect that impacts human interactions. These evocations are however defined through several parameters, like the distance from the features or the environment for instance. But more importantly, they are defined by the past experience and the culture of the player. Thus, one of the key points of examination for the player's interpretation of a game would be to examine the composition of the game world.

III. Methods and materials

In this study, we used an abductive thematic analysis methodology (combining both inductive and deductive coding) to analyze free-text comments provided by judges for the 2018, 2019, and 2020 editions of GDMC. We initially performed inductive coding on the judge comments for 2018, and then used as a base for the deductive coding of the comments in the 2019 and 2020 instances of the GDMC. For each year, the researchers reached consensus around the codes, and this consensus-decided collection of codes was then carried on to the next year. Once we had coded all three years, we elicited themes from the compounded list of codes. All authors played an equal role in coding and theme elicitation.

This study also provides two unique perspectives of evaluation, in contrast to looking at qualitative evaluations of fully human-designed game content in general. Firstly, the judges all knew that the content was algorithmically generated, and it is interesting to see how and if this influences their judgments. Secondly, the diversity of different generators might help to illuminate quality criteria that only become apparent by contrasting the approaches of different generators, both within a given year, and over the course of the three years. Together, these two methodological and data attributes should help provide a greater understanding into the problem described by this paper.

A. Respondents and data

In the theme descriptions below, the judging group is referred to as "the respondents". In total across the 3 editions of the GMDC, we had 17 unique respondents evaluating 21 different settlements over 3 years. Split across the different years, 2018 had 9 judges evaluating 4 settlements, 2019 had 11 judges evaluating 6 settlements, and 2020 had 9 judges evaluating 11 settlements. In order to preserve some level of anonymity for the respondents, we have chosen to not include a demographic breakdown of the respondent group in this paper. Furthermore, the small size of the respondent group makes between-groups analysis based on demographic data largely meaningless.

IV. Themes

The data analysis resulted in a total of 15 themes, composed of 8 main level themes, and 7 sub-themes. All themes, complete with overview descriptions, can be seen in Table I. The detailed specifications can be found in the following sections.

A. Navigation

This theme describes the ease of navigation within the settlement. Overall, the respondents quickly identified visual indications for navigation, for example signposts. Visual landmarks, for example large buildings, also played a key role in how visitors oriented themselves within the settlement and game world, echoing theories introduced by Nitsche [13] and Fernández-Vara [7].

As can be expected, roads also played a key role in how the respondents reasoned about navigation within the settlement. However, the design of the roads within the settlement need to be of sufficient quality if they are to contribute to the navigational ability of players. Roads need to match the

TABLE I
THEMES ELICITED FOR THIS STUDY

Theme	Overview description
Navigation	Ease of navigation in the settlement
Quality of Life (For the player)	The assumed quality of life for inhabitants of the settlement
Environmental Narrative	How narrative is conveyed to players using the environment of the settlement. This has four sub-themes: *Environmental narrative Dissonance*, *Cultural/Trope Evocation*, *Individual Exterior Quality*, and *Individual Exterior Quality*
Environmental Narrative Dissonance	When the environmental narrative is incongruent with what is expected by the respondent. This is a sub-theme to *Environmental Narrative*.
Cultural/Trope Evocation	When the generator uses cultural or trope shorthand to invoke a certain feeling in the player. This is a sub-theme to *Environmental Narrative*.
Individual Exterior Quality	The quality of the exteriors of buildings. This is a sub-theme of *Environmental Narrative*.
Individual Interior Quality	The quality of the interiors of buildings. This is a sub-theme to *Environmental Narrative*.
Settlement Composition	Covers how the settlement and its layout is adapted to where it is situated. This has a sub-theme: *Interconnectedness*
Interconnectedness	Describes the mental model of the settlement and how it is constructed. This is a sub-theme to *Settlement Composition*
Real-World Evocation	Covers how the settlement evokes the real world. Distinct from cultural evocation in that it focuses on functional elements rather than experiential elements.
Affect on Player	Describes how interesting and standout features capture the attention of the player
Various Varieties	Describes the ways in which affects the player experience. This has two sub-themes: *Generator Style* and *Situated Adaptivity*
Generator Style	Describes how the design decisions for each generator influences the interpretation, and the importance of coherence in generator style. This is a sub-theme to *Various Varieties*
Situated Adaptivity	Covers the delimitations of how variety can be expressed by a generator. This is a sub-theme to *Various Varieties*
Gameplay Elements	Describes how domain-specific understanding of Minecraft as a game influenced the interpretation of the generated settlements.

terrain, connect to other roads, and generally be perceived as sensible. Overall, respondents transfer a large degree of real-world expectations onto roads found in the settlement. This indicates some degree of transfer between the game and real world in terms of repertoire of character [9].

Lastly, the difficulty of traversing the landscape affected to what degree the respondents considered the landscape to be navigable. Broken terrain, for example cliffs, or traversal aids, such as bridges, both contributed to the perceived level of difficulty of navigation. This is indicative of the face that indices [13, 21] inform the evaluation of the game world in terms of affordances – which is in accordance with how theories described by McGrenere & Ho [11] and Warpefelt [21, 22].

B. Quality of life

When evaluating a settlement, the respondents included notions of livability into their evaluation. This included the navigability of the settlement (see the *Navigation* theme in Section IV-A) but also to what extent the settlement provided food or services to the inhabitants. Furthermore, the quality of the buildings and the level of safety afforded by protective measures such as walls and lighting against monsters and natural hazards (see the *Gameplay Elements* theme in Section IV-H for a further discussion on this). As with the *Navigation* (see Section IV-A) the findings here echo the theory of repertoire of character described by Janlert & Stolterman [9].

C. Environmental Narrative

The theme *Environmental Narrative* encompasses a large number of codes, all describing various aspects of how narrative is conveyed through the game environment in the Minecraft world. In the data, respondents discussed how environmental narrative arises from many sources, including the architecture and layout of cities, the various functions provided by buildings (for example farms), how distinct building types and landmarks all work together to provide a narrative to the player, without any traditional storytelling, allowing them to suspend disbelief. This is essentially a manifestation of indexical storytelling [7], where the game world provides indices for the player to latch onto to build their alterbiography [1], i.e. the story of how the specific player's experienced play session. This process has previously been described by Warpefelt [21] and involves how narrative storytelling helps convey the affordances [11] of the game world to the player.

The respondents also reacted particularly strongly to incomplete environmental narratives – where the game world seemingly is setting up the fundamentals for some kind of narrative but by and large fails to follow through. Essentially, there is an interesting hook but no resolution. This can be particularly problematic when the generator uses a single hook to make the space seem alive, as described by respondent 2020:1 in terms of a 2020 submission that made extensive use of a monorail running through the settlement:

> *The whole approach of using the monorail of course fell down on the isolated island where there wasnt* (sic) *opportunity to build the actual settlement.* – 2020:1

When the core element of the environmental narrative takes up so much space that the settlement itself cannot be instantiated in the game world, this is obviously problematic in terms of creating a believable settlement, and not just a monorail stop. As such, that particular generator's over-reliance on a single feature led it to not be feasible for constrained spaces. It should be noted that this theme operates on the gestalt of the

artifact interpreted holistically, rather than individual details, in contrast to how Warpefelt [21] describes this phenomenon. However, this phenomenon is inextricably linked to the details as created by the generator. As seen in the monorail example above, the holistic interpretation of the artifact is still critically dependent on the details.

1) Environmental Narrative Dissonance: The subtheme *Environmental Narrative Dissonance* covers the parts of the experience where the generated artifact portrays an environmental narrative that is incongruent with what is expected by the respondent. This causes a mismatch of expectation and delivered content, thus breaking immersion and disrupting the suspension of disbelief. Like its parent theme, this operates on the gestalt of the artifact, but is inextricably linked to the generated details of the artifact. The problems associated with this these are often strongly correlated with the kaleidoscope effect described by Cardona-Rivera [2] and the concept of expressive range described by Smith & Whitehead [19], i.e. that the player is starting to see the limits of what the generator can express, and that the content is starting to be perceived as repetitive and predictable.

2) Cultural/Trope Evocation: The subtheme *Cultural/Trope Evocation* is related to the instances when the generator uses cultural or trope shorthand to create connections to existing bundles of expectations in the player. Examples of this are generators using torii gates to invoke a "Japanese" feeling, or the use of natural materials and medieval architecture to invoke a fantasy feeling.

3) Individual Exterior Quality: The subtheme *Individual Exterior Quality* describes the occurrences where the judges drew conclusions based on the exterior of the buildings. This phenomenon is related to several different types of details found on the exterior of a building, including the general aesthetic of the building, the evocation of certain tropes (as described by the subtheme *Cultural/Trope Evocation* in Section IV-C2), the materials from which the building is constructed, and the accessibility of the building.

Furthermore, the exterior of buildings need to fulfill Compton's principle of *perceptual differentiation* [4]. Buildings need to be both different and alike – if they are too similar, they end up blending together, and if they are too different the cohesive look of the settlement is lost. Thus, there exists a Goldilocks range where the perceptual differentiation of buildings is just right. We have not been able to elicit the exact parameters of this zone in our study.

It should also be noted that individual elements in the settlement's makeup can have a strong impact on the overall perception of the settlement. "Signature buildings", that have a unique look and are strongly evocative, tend to leave better impressions and more long-lasting memories. This is coherent with Compton's principle of *perceptual uniqueness*, and is exemplified in many settlements from 2019, which prominently featured signature buildings such as windmills and a monorail.

However, the immersive effect of buildings is also fragile. Single discordant details can have a strong negative impact on the interpretation of, and favorable disposition towards, a settlement. If buildings are placed in ways that are seemingly unrealistic (for example half hanging off a cliff with a giant foundation) this can have a deleterious effect on the acceptance of the settlement. To a large extent, this evaluation seems to be done based on the "realism" of the building – i.e. if it is possible to build such a structure in the real world. This is indicative of the judges transferring and applying their repertoire of character [9] as a set of expectations on the settlement.

Overall, the external evaluation of the settlement seems to be largely holistic. Settlements that were favorably rated often provided a fond of buildings that were varied enough to be perceptually different, but without being so varied that the settlement seems haphazardly constructed. Furthermore, they had a few perceptually unique signature buildings that provided anchoring points within the settlements. The character of buildings was also evaluated within the context of the other buildings present in the settlement.

4) Individual Interior Quality: The *Individual Interior Quality* sub-theme describes the occurrences where judges interpreted a building's interior. It should be noted that very few buildings had developed interiors in the first years of the competition, and those comments mostly dealt with a lack of furniture inside the building. Once these were starting to become prevalent in later years, judges started commenting on the quality of the furnishing.

A core factor in evaluation was also incoherence in building interiors. Rooms that were improperly scaled, or floors that were unreachable, were rated particularly poorly by judges. This reinforces the idea that judges bring with them their repertoire of character into the game, as mentioned in the *Individual Exterior Quality* sub-theme (see Section IV-C3).

One item of particular interest is that the interiors of buildings seem to have been evaluated more or less separately from the exteriors of buildings. Indoor environments were judged one-by-one, and were seemingly not impacted by the overall, holistic, evaluation of the settlement. This seems to suggest that there is a second layer of analysis done for interior.

D. Settlement Composition

This theme covers how the settlement is fitted to terrain and adapted to local features, laid out, and sized. Overall, what we are evaluating here are the functional components of the evoked character [9] of the settlement.

Terrain fitting is evaluated based on how well the settlement follows the features of the terrain, how it handles terrain features like rivers or small islands, and how the quality of how the terrain has been altered in order to make the settlement fit in the desired space. The key success factor is finding a balance between a real-looking settlement and one where the world has not been entirely bulldozed to make space for the buildings. Success in terms of settlement composition is also related to how well the materials used to build the settlement match the materials found in the nearby areas and biomes, and how well these materials are integrated into the building designs. An example of good matching would be log houses in a heavily wooded colder area, or Adobe-style housing in the desert. Although these are not necessarily the

only type of building that fits with such terrain, they do evoke a certain related character that we as players would expect to see in such a climate. By extension, this also connects to the *Cultural/Trope Evocation* theme (see Section IV-C2 on page 6) This is also echoed the evaluations by the respondents, here from Respondent 2019:1:

> *[B]iome-variants of varied buildings are clustered around habitable areas, with a particularly thorough road system and farms with varied crops* (sic) – Respondent 2019:1

As we can see in this quite, there is also a notion that the layout and positioning of the settlement is evaluated based on "soft" ideas of reasonableness - i.e. a notion of what looks like a "real" place. Ideally buildings should be placed in a way that makes sense to the player as they bring in their real-world based understanding into the evaluation process, essentially fulfilling the evoked character [9] of the settlement. As explained by Respondent 2019:2:

> *Great design work here - the watchtowers are fantastic, as is their placement [...]. Lovely winding paths and farm arrangements, great sense of place.* – Respondent 2019:2

Respondent 2019:2 picks out a core feature in one of the judged maps for 2019, which contained central markers in the form of watchtowers. However, they also pick up on the winding paths and farms, both features that are classically associated with pastoral landscapes, thus providing a strong sense of place. This highly interconnected presentation of tropes provides the respondent with a strong sense that this is a "real" place.

However, settlement composition can also have a negative effect on the player experience. Cardinal sins in this area includes overlapping buildings or placing them in a way that is physically impossible in the real world (such as hanging off a ledge with no support structure). Sizing is also a concern among respondents, and it is critical that the buildings be sized in a way that is perceived as plausible – for example a straw hut several hundred meters on the side is seen as less believable. Finally, the pathing of the settlement needs to be appropriate for the presented thematics – again connecting to the Cultural/Trope Evocation theme.

1) Interconnectedness: The *Interconnectedness* theme operates more on the mental model of the settlement than the settlement itself. It exists adjacent to the navigational affordances of the pathing in the generated settlement, but instead represents the player's mental model of how the pathing fits together, and how it evokes and/or reinforces the character of the settlement. In addition, the theme covers the understanding of how one settlement may be composed of multiple smaller components, for example a city split into districts or a collection of villages.

A negative impact of poor interconnectedness can be exemplified by this quote from Respondent 2019:3:

> *The houses are spread out throughout the world, with no easy access paths between them, and some houses are too close together and hard to access, making it not very functional* – Respondent 2019:3

Here the respondent expresses that the lack of paths, and by extension connections, between elements of the settlement make it difficult to navigate the space. Conversely, a level of interconnectedness between parts of the settlement can be beneficial, as seen in the following quotes:

> *Nice architecture and design choices - I also think the pathing was quite good. I like how understated some aspects of it were - the houses were small, nestled in between trees, with nice dirt paths. Really felt like a small forest village* – Respondent 2020:2

> *The paths around the housing is - for me - one of the strongest aspects of this, given it helps me understand the larger structure of the settlement. Plus the lamps that help signpost the path* – Respondent 2019:4

As we can see in the quotes from these three respondents, the pathing of settlements is a core part of how we understand how they are connected. Note, however, that the pathing is not the same as interconnectedness. Instead, it acts as a kind of facilitator for the formation of a mental map of the settlement.

E. Real-World Evocation

The real-world evocation theme is related to the ways in which the settlement evokes the real world. This is similar but distinct from the cultural evocation mentioned in the theme cultural / trope evocation and instead focuses on the details of the generator's output rather than the gestalt of the settlement as a whole.

Respondents raised interest not only in the real world alike aesthetic, but also on the feeling it produces. Any clues that lead to thinking actual people are or could be living in the settlements are perceived as positive additions. The presence of furniture within the building, sources of food, place of work and other proof of human activities. Also, settlements suggest human and legal organization, like boundaries around houses or farms, different neighborhoods and so on. These evocations can also be linked to cultural elements, in the architecture and the settlement layout. But if some elements contribute to the 'real world' feeling, others degrade it. In particular any elements that are either unbelievable, or unrealistic. It could be something impossible in our reality, such as a floating building, or a bad measurement of real-world characteristics, like an over-flatten terrain.

We can deduce that real world evocation in the context of Minecraft settlement assists the player in understanding the sense of the settlement, and how to interact with its surroundings.

F. Effect on the Player

In some cases, generated settlements presented especially interesting features that stand out from the rest of the generated content, for example strong architectural attractions like windmills or towers (see Figure 2 on page 8). In these cases, the respondents have described these features as being awe-striking, and providing a visceral sense of wonder and a call to adventure . As exemplified by a response from Respondent 2019:5:

Fig. 2. An example of strong architectural attractions in a generated town

> *Wow! This entry was very different. The skyscrapers are outstanding: they look right, furniture, floor colours, etc - and stairs ,and light fittings!* – Respondent 2019:5

Simply put, these generated features act as a focal point for the player's attention, and provide an impactful gaming experience. In some cases, the features also act as a clarion call to adventure within the generated world. The respondents (who were judging the generated artifacts for a competition) sometimes described generated features as leading them off from their core mission of judging and made them explore the world a bit more than they otherwise would. As explained by Respondent 2019:6:

> *Loved the stone structures and mossy rock - almost like ruins in some places? They were so interesting I actually dug them up to see if there was anything hidden!* – Respondent 2019:6

As described by the respondent, the ruin-like stone structures acted as a call to adventure, where they felt a need to explore more. Essentially, these structures act as what Fernández-Vara calls wieners - essentially a feature that draws the attention of the player and acts as an attractant to the area in which it is located. As described by Fernández-Vara, this is a concept imported from theme park design [7].

G. Various Varieties

Several of the judges made comments related to the overall concept of variety. This is noteworthy since variety was not one of the judging criteria. In PCG, variety is often seen as synonymous with the expressive range [19] of the generator, i.e. the range of different artifacts the generator can produce. The comments made by the judges related to the GDMC context can help us decompose this concept further, thus aiding in our understanding of how variety impacts PCG artifacts in this specific case.

In general, was mentioned as something positive and desirable. A large portion of the comments related to variety related to the lack thereof. Some comments were ambiguous and could refer to the differences between the overall artifacts (the three different settlements made for three different maps), but several comments from judges did speak specifically about the variety between houses or elements of the settlements.

This illustrates that Minecraft settlements are composite artifacts, larger artifacts composed of several similar sub-units. In technical terms, this means that there is an easy separation between hierarchies, and a generator could generate an overall composite structure, while another generator then fills in the sub-units. Alternatively, one of those two levels might be human-made, such as the houses in this case, which are often templates, which are then arranged by a generator. This idea of a composite artefact is not unique to the GDMC challenge. Other creative artefacts, such as books, pieces of music, or maps, can be seen and generated as composites of small units, and faces similar challenges. Again, the variety between the sub-units is remarked on positively (quote).

There is a competing quality criterion though, codified as cohesion. The sub-units, usually houses in our case, should be similar in a way that makes them believable belong to the same overall settlement. This is in contrast to the desire to

have them be different. A good generator here seems to be able to strike a balance between those two drives.

1) Generator Style: Addressing this conflict between cohesion and variety raises the question if there are certain elements that should be varied, while others should be kept constant. The concept of generator style is discussed extensively by respondents, and there are several codes that form a sub-theme of variety around this topic.

This sub-theme encapsulates the possible dimensions of the buildings into three categories, related to how they value variety in a specific dimension:

1) Those that should always be varied between sub units to make for believable variety
2) Those that correspond to a positive quality, and should be set towards a target value (or maximized)
3) Those that relate to a stylistic choice, i.e. should be chosen once, and then adhered to for all buildings in the settlement, to create a sense of cohesion.

For individual respondents several codes could be sorted into these categories. However, it was difficult to sort the various codes for stylistic dimensions in a way that is consistent across responses. What some might see as a quality criterion, is seen as a stylistic choice by others. As such, we have introduced this theme to encapsulate the multi-faceted nature of generator style and how it impacts a complex generated artifact like a Minecraft settlement.

2) Situated Adaptivity: Another sub-theme related to variety is adaptivity, something that was specifically promoted as a criterion by the judging guide. It is about the ability of the generator to react appropriately to different input maps, and should be evident by the fit of the final settlement artifact into the provided map. This is in opportunity to introduce variety by building on the variety of the provided map prompts, but is more technically challenging, as it requires more than just randomness - directed adaption.

As it is a criterion for the challenge, it presence is generally commented on in a positive way (quotes), and the codes related to this sub-theme are closely related to the illustrative comments on the judging guide. The two biggest focus are the adaption to biome materials, and how well buildings are places in the landscape, i.e. the adaption to the height map.

H. Gameplay Elements

Minecraft is not just an editor to generate maps and settlements with 3d blocks, but also a game, and as such, artifacts within Minecraft can and are seen as game play elements. We see this reflected in the comments of the judges, with several of them relating to the theme of game play elements.

Common and very Minecraft-specific comments talk about the presence of monsters and how the settlements fail to keep them from spawning, or how the settlements do not offer protection from monsters. This effect might have arisen from one of the judging criteria, a subject which we will discuss during the conclusion of this paper.

There are also several comments talking about the navigability, or lack thereof, or the settlements. This is less Minecraft specific, but still evaluates the artifact as more than just an observable creative piece, and more as something to be interacted with.

Several respondents also talk about food, but it is unclear here if this is commented on from a perspective of a player trying to obtain food, or as a comment on the environmental narrative.

Another code that is related to this theme is light - which plays a surprising role in many different evaluation elements, as it is both relevant for game play, such as monster spawn prevention, but also allows the player to see, and also contributes to the mood and aesthetic impression of the settlement.

However, we saw a large variance in how often gameplay related codes were expressed by different respondents. We suspect that this may arise from a difference in how experienced the judges were with Minecraft, which may have informed expectations.

V. Conclusions, Discussion, and Future Work

The core finding of this study is that the unifying evaluative theme for all judges seems to be the Suspension of Disbelief [3]. More precisely, the judgment of an individual human seems in large part determined by how the judges' expectations are fulfilled with by the generator. Unfortunately, for any attempt to produce more computational metrics, those expectations are influenced by a large range of hard to quantify factors. For a start, there are cultural and biographical factors. Expectations seem to also shift from year to year - were in the first year biome-based material replacement was seen as innovative, it was criticized as boring in year four. There was also some indication that judges took into account that this was made by an algorithm - reporting that the output was great for PCG, but others harshly judged entries against human standards. Comparison to real world places were sometimes seen as positive, but also led to the identification of glaring flaws, that other judges yet identified as clever conceits to game logic and gaming conventions. Overall, there was a lot of language indicating that judges had been both positively surprised and negatively disappointed, with some of the same elements being inconsistently labeled as both good and bad. Overall, the comments demonstrated that the players have a lot of different experiences, and hence expectations, we would need to consider, if we wanted to model their judgment. Furthermore, we also saw that they paid attention to, or at least reported on, very different aspects of the artifact. Taken together, the plethora of different factors make the case for a very multifaceted analysis of generative artifact. In turn, this makes it even harder to conceive of a computational metric that evaluates a complex PCG artifact, such as a Minecraft settlement, holistically in a similar way to a human observer.

Human judges seem to address multifaceted analysis problem in part by conceiving of Minecraft settlements as composite artifacts, and judging both their parts, and their composition, on multiple levels - as evidence by the distinction between interior and exterior environments, or the theme of interconnectedness – which mostly seems to operate on conceptual rather than concrete parts of the settlement. The nature of a Minecraft Settlement as a composite artifact thus

had some interesting consequences. Primarily, it allowed for some part of the artifact to set expectation for the rest. Within this scope there then seems to be some trade off between the positive novelty of interesting variation, balanced with providing an expected consistency in style and tone, echoing Compton's theory of perceptual uniqueness and differentiation [4] as well as Janlert and Stolterman's concept of character [9]. It is unclear what exactly the sweet spot between novelty and consistency is here thought. Judges comments indicate that there are some dimensions, such as style, were variation is less desired then in others. Given that many PCG artifacts, such as longer text, music, game levels can be seen as composites, we posit that determining one way towards evaluation of complex, composite artifacts would be automatically determine both the good and bad dimensions for expected variety, and the right trade off between novelty and consistency. Warpefelt's indicator theory [21] is useful as a qualitative descriptor for these concepts, but would need further development to be useful as a quantitative measure, and even so will likely be highly situational. Compositionality also raises the question on how we deal with the order of experiences when we have a human judge evaluate an artifact that cannot be holistically perceived in one moment - yet offers no pre-defined perception order (unlike a book). One interesting methodological option here would be to actually measure order effects by selectively exposing participants or players to different part of an artifact, and take existing order effects as a positive sign of the existence of expectation setting. The inherent non-linear nature of in-game narrative, especially when it comes to environmental narrative, makes the player experience of generated artifacts difficult to analyze using traditional tools for narrative inquiry, and we find that there is a need for a theory allowing for more quantifiable analysis of these non-linear narratives. Some initial tools exist, for example Warpefelt's indicator theory [21] or Fernández-Vara's indices [7]. However, as mentioned above these tools will need specific adaptation to generative settlements in Minecraft.

Finally, we were intrigued by the judges reporting on the affect some of the generation had on them. The idea that brilliant art is capable of moving us, both emotionally and to action, is not new. However, we now have empirical evidence that even the slightly flawed art produced by generators can cause these effects, which suggests that the threshold for when this effect is induced may be lower, and the induction of the effect may not be reserved simply for *brilliant* art. Furthermore, the direct report by the players that a specific component of the generated settlement caused them to reconsider their goals for a given interaction shows the importance of actually interacting and experiencing the artifact, rather than just observing it. It also shows that singular stand-out features, again realizing Compton's theory, can have a large impact on the player experience. This might also suggest a new way of evaluating PCG artifacts. We could, for example, measure how much players will deviate a player from a chosen path, or moderate the players behavior, simply by virtue of subverting or fulfilling their expectations using strong indicators or indices within the game environment.

In summary, analyzing the judges comments from the GMDC competition has provided interesting insights into how players interpret generative artifacts. Although there are several theories that help us deconstruct and understand these experiences, there still exists a strong need for a more in-depth understanding of how we evaluate generative game artifacts – especially complex composite artifacts like settlements.

References

[1] Gordon Calleja. *In-game: From immersion to incorporation*. MIT Press, 2011.

[2] Rogelio Enrique Cardona-Rivera. "Cognitively-grounded procedural content generation". In: *Workshops at the Thirty-First AAAI Conference on Artificial Intelligence*. 2017.

[3] Samuel Taylor Coleridge. *Biographia Literaria*. Retrieved May 16, 2022 from Project Gutenberg http://www.gutenberg.org/ebooks/6081. 1817.

[4] Kate Compton. "So you want to build a generator". In: *Published online: https://galaxykate0.tumblr.com/post/139774965871/so-you-want-to-build-a-generator* (2016).

[5] Michael Cook, Jeremy Gow, and Simon Colton. "Danesh: Helping bridge the gap between procedural generators and their output". In: (2016).

[6] Mike Cook. "Tutorial: Generative & Possibility Space". In: *Published online: https://www.possibilityspace.org/tutorial-generative-possibility-space/index.html* (2019).

[7] Clara Fernández-Vara. "Game Spaces Speak Volumes: Indexical Storytelling". In: *DiGRA '11 - Proceedings of the 2011 DiGRA International Conference: Think Design Play*. DiGRA/Utrecht School of the Arts, 2011.

[8] Jean-Baptiste Hervé and Christoph Salge. "Comparing PCG metrics with Human Evaluation in Minecraft Settlement Generation". In: ().

[9] Lars-Erik Janlert and Erik Stolterman. "The character of things". In: *Design Studies* 18.3 (1997), pp. 297–314.

[10] Julian Mariño, Willian Reis, and Levi Lelis. "An empirical evaluation of evaluation metrics of procedurally generated Mario levels". In: *Proceedings of the AAAI Conference on Artificial Intelligence and Interactive Digital Entertainment*. Vol. 11. 1. 2015.

[11] Joanna McGrenere and Wayne Ho. "Affordances: Clarifying and evolving a concept". In: *Proceedings of Graphics Interface 2000*. 2000, pp. 179–186.

[12] Mojang Studios. *Minecraft*. Video game. 2011.

[13] Michael Nitsche. *Video game spaces: image, play, and structure in 3D worlds*. MIT Press, 2008.

[14] Charles Sanders Peirce. *The essential Peirce: selected philosophical writings*. Vol. 2. Indiana University Press, 1992.

[15] Christoph Salge et al. "Generative design in minecraft (GDMC) settlement generation competition". In: *Proceedings of the 13th International Conference on the Foundations of Digital Games*. 2018, pp. 1–10.

[16] Christoph Salge et al. “Generative Design in Minecraft (GDMC): Settlement Generation Competition”. In: *Proceedings of the 13th International Conference on the Foundations of Digital Games*. FDG ’18. Malmö, Sweden: ACM, 2018, 49:1–49:10. ISBN: 978-1-4503-6571-0. DOI: 10.1145/3235765.3235814. URL: http://doi.acm.org/10.1145/3235765.3235814.

[17] Christoph Salge et al. “Generative Design in Minecraft: Chronicle Challenge”. In: *Proceedings of the 10th International Conference on Computational Creativity*. 2019.

[18] Christoph Salge et al. “The AI Settlement Generation Challenge in Minecraft”. In: *KI-Künstliche Intelligenz* 34.1 (2020), pp. 19–31.

[19] Gillian Smith and Jim Whitehead. “Analyzing the expressive range of a level generator”. In: *Proceedings of the 2010 Workshop on Procedural Content Generation in Games*. 2010, pp. 1–7.

[20] Adam Summerville. “Expanding expressive range: Evaluation methodologies for procedural content generation”. In: *Fourteenth artificial intelligence and interactive digital entertainment conference*. 2018.

[21] Henrik Warpefelt. “Micro-level examination of games using Indicator Analysis”. In: *FDG ’20: International Conference on the Foundations of Digital Games*. ACM, Sept. 2020, pp. 1–9. DOI: https://doi.org/10.1145/3402942.3402980.

[22] Henrik Warpefelt. “The case for naive and low-fidelity narrative generation”. In: *Games and Narrative: Theory and Practice*. Ed. by Barbaros Bostan. Springer, 2022.