Towards a Procedural Evaluation Technique: Metrics for Level Design

Alessandro Canossa Gillian Smith

Northeastern University
Playable Innovative Technologies Lab
360 Huntington Ave, 100 ME
Boston, MA 02115

{a.canossa, gi.smith}@neu.edu

ABSTRACT

Existing approaches to characterizing and evaluating level designs involve either the application of theory-based language to qualitatively describe the level's structure, or empirical evaluation of how players experience the levels. In this paper, we propose a method for evaluation that bridges these two approaches: theory-based, quantitative metrics that measure the qualities of levels. The metrics are drawn from a design activity with student designers creating levels for both a 2D and 3D platformer. The results from this activity are 20 new metrics that aim to shed light on the aesthetic and topological properties of level design, and the ways they induce players to use tactics and experience difficulty. The existence of quantitative metrics for evaluating and analyzing levels offers the opportunity for computational implementation. This could provide valuable, rapid feedback to designers during their design process, as well as have repercussions for new evaluation techniques for procedural content generation.

Categories and Subject Descriptors

H.5.m [**Information Interfaces and Presentation**]: Miscellaneous.

General Terms

Design, Human Factors, Experimentation

Keywords

Game Design, Level Design, Procedural Content Generation, Design Patterns, Metrics

1. INTRODUCTION

How to usefully describe, analyze, and evaluate levels is an open problem in level design. Level design involves manipulating the content and composition of a game space in order to direct player experience. Relatively small changes in content—even as simple as a change in color palette—can lead to large differences in the way a player experiences a level. The variety of stimuli that can potentially affect player experience is very broad: from aural cues to slight variation in topology or luminosity of the environment. Because of the multi-modal nature of these stimuli, evaluating game levels is not a trivial task. In this article, we argue that evaluation is both a reflection on the player experience and a critique upon the design itself, including considerations of necessary conditions.

Understanding how level qualities map onto player experience is important in many avenues of level design research and practice. Designers need to be able to communicate a desired aesthetic and hypothesize about ways to reach it. Game analysis research requires a vocabulary for describing and reasoning about levels. Procedural content generation and automated design tools need ways to automatically and objectively evaluate the content being produced by a system.

Existing methods for understanding level design include the theory-based, qualitative approach of design patterns [3, 6, 11, 23] and player-based, empirical evaluation methods from game user research [4, 14, 26]. Design patterns support a common vocabulary for designers and researchers to describe and analyze levels, but are typically not sufficiently formal for use in automated analysis, and also typically focus on level elements, rather than a level as a whole. Game user research involves empirical, often quantitative, analysis of large-scale data from existing players. This provides a wealth of information about a level design, but only after it has been deployed and has been played. What is needed is a method for evaluating level designs that bridges the gap: one that provides rapid feedback on a level, is grounded in game design theory, and can be used without needing to gather data from playtesting.

Within procedural content generation, a method for evaluating the quality of content generators is to analyze and rate the content they produce using a set of consistent metrics [10, 24]. In this paper, we propose generalizing this evaluation method to become a means for analyzing human-created levels, and suggest new metrics that are grounded in game design theory and have been gathered through discussion, level construction, and critique with student designers. The metrics have been distilled from an analysis of levels created for a Super Mario World clone[1] (a common test domain for PCG research [13]) and the game Portal 2 [29]. Several of the metrics we propose are generalizable across the two domains, while others are unique to a particular game.

Proceedings of the 10th International Conference on the Foundations of Digital Games (FDG 2015), June 22-25, 2015, Pacific Grove, CA, USA.
ISBN 978-0-9913982-4-9. Copyright held by author(s).

[1] Super Mario Flash 2, retrieved from: http://www.pouetpu-games.com/index.php?section=2&game_id=2&w=640&h=480

The primary contribution of this paper is a set of design considerations that are formalizable as metrics, that can be used to rapidly evaluate a level by mapping features of the content and composition to a desired player experience. As a secondary contribution, we also present a method for gathering new metrics via communication with level designers.

2. RELATED WORK

Two existing approaches for evaluating and analyzing levels are design pattern analysis and empirical study of players and player behavior. Design patterns are common elements that are found across many different games; they are descriptive and qualitative, used to describe the components of a game and, optionally, how the presence of those components (and relationship with other components) impacts player experience.

Existing work in design patterns for level design typically focuses on analyzing the structure and composition of the game space, often at a fine level of granularity [3, 11, 23]. For example, Smith et al.'s patterns for level design in role-playing games [23] include patterns describing common room layouts and ways that treasure can be hidden in a level. These patterns are useful for providing a common vocabulary for describing game levels, and have even been used more formally for automatically generating levels [6, 22]. However, as a tool for analyzing levels, they are highly descriptive, and typically afford only stating that a level uses or does not use a particular pattern. It is harder to use design patterns to directly measure or compare entire levels; design patterns are more focused on detailed description than automated analysis. In this work, we are aiming to build a vocabulary of metrics that can be used to rate and compare entire levels (rather than individual pieces) consistently and rapidly.

Game User Research (GUR) is a discipline concerned with studying players' response to games. GUR's main objective is to investigate interactions between players and games and the surrounding context of play, providing actionable assessments of player experience and its adherence to a designer's intent. Its methods range from user testing to physiological measurements and can be loosely categorized as affective, behavioral or cognitive. Affective evaluation attempts to gauge the emotional responses of players, i.e. what they feel; behavioral evaluation looks at quantifiable measures of players' actions, i.e. what players are doing; and cognitive evaluation tries to infer opinion and attitudes, i.e. what players are thinking [1].

Traditionally, GUR has evaluated game levels using a number of methods. The methods are generally subdivided according to three axes: qualitative versus quantitative; what people say versus what people do; and natural versus artificial context of use [18]. Here are some of the most used methods employed to evaluate player experience arising from interaction with levels.

Think-aloud is a method used to gather data in usability testing that requires participants to speak out loud as they are performing a set of specified tasks, for example playing a defined portion of the game [27].

Heuristics are design guidelines which serve as a useful evaluation tool to assess additional properties of the game experience such as narrative values, strategy and challenge or skill development [7, 8].

Heat maps are graphical representations of data using colors to indicate the level of activity, For example, a heatmap could indicate the number of player deaths in a multiplayer map of a given game [25, 15, 4].

Time-spent reports are used to examine how much time playtesters spent on different types of activity in a game level [31].

Force-Directed Network Graphs are becoming more and more common. Modeling games as state machines is a fairly established strategy to provide game-agnostic visualization of the games' possibility spaces and enumerate the strategies players enact while playing. In order to understand cause and effect relationships, it is essential to examine multiple variables and their interrelations simultaneously while accounting for temporal progression [2, 30, 5].

Current work in defining metrics for evaluating levels exists in the procedural content generation literature, specifically focused on evaluating generators. Horn et al. summarize several metrics for evaluating platformer levels produced by a range of different automated designers [10]. These metrics are highly formal, with a computational implementation, and for each level produce a value between 0 and 1 rating how well it meets the characteristics described by the metric. The metrics we present in this paper are not as heavily formalized as those used in evaluating PCG systems, but have the potential to be formalized in future work.

2.1 A Review of Current Level Metrics

The current metrics used to evaluate levels exist only for 2D platforming levels, specifically those created in the Mario AI framework [13]. These metrics are grounded in design theory, rather than through interviews with level designers or extracted through a critical design process. In this section, we provide an overview of existing metrics and the design rationale used in creating them [10].

Linearity. Linearity is a metric that measures the extent to which the walkable surface in the 2D level fits to a line. It is a means for measuring the vertical "flow" of a level as the player progresses through it -- a low linearity level will involve a lot of rolling hills, while a high linearity level has the player making relatively few changes in the y direction. This metric was chosen to be one that is relevant to how the player experiences a level in terms of their movement through it.

Leniency. Leniency is an approximation to the difficulty of a level, based only on components in the level without consideration for individual player preferences and skills. Each element in the level is assigned a leniency score, based on the seriousness of the harm the element would cause if the player were to fail to negotiate it correctly. For instance, a platform has a leniency of 1, as it cannot harm the player at all. A gap has a leniency of 0, because if the player falls through it they die. An enemy has a leniency between these two, since the player might not die, but rather lose a health increment or undergo a similar change in state.

Density. Density measures the average amount of content that lies in vertical slices of a level. For instance, a vertical slice that contains three platform tiles and an enemy is denser than the same sized vertical slice with only one platform tile. This metric gives an approximation of the number of paths that are available through a level, as well as the amount of visual clutter in the level.

Pattern Variation. Based upon an analysis of design patterns in existing Mario games, the pattern variation metric measures how many different kinds of typical Mario-like level constructs exist in a single level. This metric, as well as pattern density (below), is intended to measure how similar a generated level is to an original, human-designed Mario level.

Pattern Density. Pattern density measures how much of a level can be explained by Mario design patterns. A level that has many patterns in it is thought to be more similar to an original Mario level than one that does not have a dense concentration of patterns.

Other metrics for level evaluation are designed more for evaluating the work of an automated (or human) designer. For example, compression distance [10] and edit distance [22] evaluate the similarity between pairs of levels, rather than acting as metrics for individual levels. This paper focuses more on defining metrics that can be used to evaluate individual levels that have been created by either a human or automated designer; however, there is interesting future work in determining metrics that can help characterize a space of levels in a design-relevant manner. For example, metrics for evaluating level pairs could help a designer understand the coherence of a family of levels being designed for an individual game, or help game scholars evaluate the coherence and style of levels created by an individual designer across multiple games.

3. METHOD

For this study we recruited 28 students from the class "Level Design and Architecture" at Northeastern University in Fall 2014 and Spring 2015. The activity was structured as a series of class exercises intended to augment the students' comprehension of the design process, while also serving as a research activity for the authors. Students were first presented with the circumplex model of affect [20, 21]. The model describes affective space as a circumplex charted over a Cartesian space where one axis represents arousal (high and low) while the other represents valence (pleasant and unpleasant). The students practiced by designing game environments that attempted to trigger defined emotional states, utilizing a very simple multiplayer level [32]. Subsequently they were introduced to the editor *Super Mario Flash 2* and asked to design levels targeting a specific affective value. The motivations for the exercise included focusing designer intent towards a specific goal without restricting excessively the expressive potential and helping the students get started very quickly. Furthermore, the exercise parameters insured a maximum of variety among the levels.

Twenty-eight levels were produced, distributed evenly among the 4 quadrants individuated by the circumplex model (high arousal and pleasant; high arousal and unpleasant; low arousal and pleasant; low arousal and unpleasant). After discussing the article "A Comparative Evaluation of Procedural Level Generators in the Mario AI Framework" by Horn et al. [10], students were asked to use the original level design metrics identified to describe the levels generated. Each of the levels generated were described by the students in terms of how much or how little they instantiated the existing metrics (Table 1) using a 5 points Likert scale (low, medium/low, medium, medium/high, high). Although the existing metrics proved somewhat useful to differentiate among the levels generated, it became evident that the students felt adding new metrics was needed, in order to capture the differences and nuances between all the levels. A comprehensive list of the new metrics generated by the students is given in section 4.

	High-Arousal Pleasant	Low-Arousal Unpleasant	High-Arousal Unpleasant	Low-Arousal Pleasant
Leniency	high	high	low	high
Linearity	-	medium/ high	-	high
Density	-	low	high	low
Pattern Variation	medium/ low	high	medium/ low	low
Pattern Density	-	low	-	low

Table 1. Four categories of levels (High Arousal, Pleasant; Low Arousal, Unpleasant; High Arousal, Unpleasant; Low Arousal, Pleasant) described according to the perceived intensity of each metric.

The next assignment saw the participants generate levels for the game Portal 2 and try to evaluate whether any of the newly identified metrics could be used to describe the design space in the Portal editor. The results are presented in section 5.

4. NEW PROPOSED METRICS FOR EVALUATING LEVEL DESIGNS

Following the method described in Section 4, we have identified 20 new metrics. This section details the identified level properties and proposes a means for their quantitative evaluation. The metrics have been divided into four categories based on the aspect of level design they are describing: aesthetic choices made by a designer, methods for approximating level difficulty, topological features, and means for scaffolding particular player tactics.

4.1 Aesthetic Choices

Aesthetic metrics are related to choices made by the designer in terms of visual composition, color, texture, and sound. It is important to note that we are referring here only to aesthetics in terms of what Niedenthall describes as "sensory phenomena" [16]. Designers tended to rate these visual and auditory metrics as highly important to evaluating a level design, yet none of these considerations have been previously touched upon by other metrics for evaluating levels. This is perhaps because of the domain in which metrics have been confined--researchers in procedural content generation have historically not concerned themselves so much with the visual or auditory experience of a level, focusing instead on the underlying level structure and player experience from a mechanics standpoint. That is partly due to the fact that aesthetics are difficult to quantify and partly because pattern researchers are somewhat inclined to favor structural and systems analysis, rather than visual appearance. Furthermore, most of the work in this area has relied on the Mario AI framework; a unified and limited tileset and sound effects does not afford much experimentation for these aspects of aesthetics.

Music. The soundtrack chosen for a game is an important aspect that accompanies the entire game experience. Properties of the soundtrack include rhythm, intervals, dynamics (volume) and pitch (frequency), as well as the number of overlaid tracks. A potential way to measure and quantify the music used in a level is to adopt music analysis techniques such as those used in music recommendation systems [12] or automated rhythm games.

Sound Effects. Similar to music choice, the sound effects used in a game have a strong influence on the overall mood of a level.

Some aspects of the sound effects can be measured in the same way as music: the average length of the effect, the dynamic range of the effect in relation to the volume of the music, and the average pitch are all applicable. There is also the consideration of how often the sound effects will play, which is a function both of the number of sound-producing elements in a level and of the designer's choice as to what elements and activities the sound effect should be tied to.

Texture. The only set of level features identified by every group of designers were those related to texture and color palette. Texture has many elements, each of which can be measured in different ways. The first is by averaging different elements of color across an entire level. The average luminosity of a level can explain if a level is overall composed of light or dark elements. The average warmth or coolness of colors can suggest the mood. The standard deviation from averaging the colors across the entire level can be a pointer to how much variation there is in the color palette, an indicator of consistency in the visual design. The second method for assessing texture is to extract and analyze the color palette itself, looking at the type of palette (e.g. analogous or split complementary) as well as drawing from research in color psychology to determine how the color choices map to sentiment and mood [17].

Finally, there are many features of a level palette that are useful to examine in terms of salience. A quantitative way to measure salience is to count the number of salient points in a level. There are many different kinds of salience that can be examined. Motion salience occurs when there is a moving element, such as an animated sprite, among a field of static elements, or an element that is moving differently (e.g. direction) than those surrounding it. Pattern salience occurs when the pattern for a set of elements is significantly different from surrounding elements. Both motion and pattern salience could be objectively analyzed by manually classifying different tiles in a tileset before the analysis is performed. Salient colors can be extracted by converting the image to greyscale (using the luminosity of each color as the lightness) and finding areas that stand out from each other on the greyscale map. Saliency of elements due to their size is another feature that can be analyzed--for example, if a level contains mostly similarly sized enemy entities, except for one that is larger. Saliency is a way of indicating that certain elements of a level are important: quantifying the number and frequency of salient elements is a useful metric for understanding whether the designer has included focal areas in a level.

4.2 Proxies for Level Difficulty

Difficulty is an inherently subjective quality: what one player finds difficult, another might find easy, and vice versa. However, it is possible to examine several features of a level that may contribute towards its overall difficulty. The following level features were all identified as proxies for interpreting the difficulty of an individual level.

Leniency Differentiation. A level that derives all its challenge from environmental hazards looks and feels different to play than one that predominantly includes enemies or NPCs, especially aggressive ones. With leniency differentiation, the aim is to understand the ratio of environmental hazards to NPC hazards.

Failure States. The existing leniency metric measures only the average leniency of the different failure states in a level; however, this provides only a partial view of the overall leniency. Further investigation into failure states includes examining the variety of ways in which players can fail a level (e.g. only failing by falling down gaps, vs. failing against a variety of different elements) and the total number of failure states in the level.

Threat Level. This level quality is the extent to which the player is likely to feel threatened during the game. A game in which there are large clusters of enemies marching toward the player has a higher threat level than one in which there are the same number of enemies but they are scattered across the level. A proxy for this emotional response is to count the density and frequency of clusters of enemies.

4.3 Structure and Topology

The topology of a level describes the geometry and the spatial relationships of a structure. It is a useful concept in evaluating level designs as it allows assessing the properties of space that are preserved through non-Euclidean transformations.

Negative space is a measure of the empty space in a level that is potentially traversable by players (by jumping or falling), versus the empty space that is simply out of reach.

Verticality is the general trend of a game space, and can be easily summarized as an upward or downward vector between the possible beginning and ending points. It is different from linearity as it does not just account for the distance between the walkable path and an hypothetical line (or a plane in a game with three dimensions), but it also suggests an averaged vector describing whether the walkable path is going upwards, downwards or in any other direction.

Rhythm refers to actions initiated by players by pressing buttons, jumping for example. We considered two types of rhythms: compulsory and optional. Jumping to avoid a chasm is a compulsory button press, while jumping to reach a coin is optional. While the first is easily derived, the second requires assessing the potential stimuli that could induce a player action.

Relative size refers to the size of the game space relative to the of the player character. This measure can serve to create and alleviate tension, as suggested by Totten [28]. Relative size is a continuum but can be split in three categories. Narrow spaces are small enclosed spaces where the occupant feels confined and unable to move. These spaces create a sense of vulnerability in the player's inability to properly defend themselves. Intimate spaces are neither confining nor overly large; their defining characteristic is that everything in the space is immediately accessible and within reach to the player. Prospect space describes a spatial condition that is wide open, within which the occupant is exposed to potential enemies and often on a lower ground.

Path properties refer to properties of the navigable paths enforced by designers. Assuming more than one navigable path, there are two interesting properties: the number of intersections that allow moving from one path to another, and the proportion of the shared spaces where two or more paths coincide.

Section consistency is a measure of topological variation of game levels. All the topological metrics listed above are sampled as averages for a whole level. Section consistency entails defining arbitrary sections for a given level and evaluating whether the averages for negative space, verticality, rhythm, relative size or path properties remain constant.

Length is the objective measure of the extent of a game space.

Completion time is an estimate of the time required to complete the level, as measured by averaging the player population and not the shortest possible time necessary to finish a level.

4.4 Tactics and Strategies

This group of metrics estimates the potential affordances that a level offers in terms of planning actions to achieve goals.

Trial and error assesses whether a level presents players with a challenge and allows them to try various strategies without being punished for failure.

Reasonability is a property of the challenges presented to players. It gauges whether the success state or goal of a challenge is clearly communicated or visible to players so that they can reason and logically compose plans to achieve said goal without resorting to trial and error.

Situational awareness describes the distance between the presentation of a challenge and its resolution, and whether players are required to take action without complete information. An example of low situational awareness is if players are required to jump from one platform and land on another that initially lies offscreen, thus requiring players to take a leap of faith without being able to infer the outcome of their actions.

Number of solutions refers to the number of ways each level challenge can be overcome.

Power-ups is a complex metric that describes the properties of player-enhancers disseminated throughout the level. Power-ups are defined as elements that have an immediate effect on the mechanics, for example modifying the length of jumps. Frequency is the number of power-ups available in the level. Density is the relative distribution of power-ups along a level. Accessibility describes the distance and the effort that players must make to obtain the power-up.

Rewards is a descriptive metric that accounts for the properties of rewards, such as the coins in Super Mario. As with power-ups, frequency, density and accessibility are also applicable to this metric but there is an additional property afforded by the much higher frequency of rewards compared to the more scarce power-ups: “breadcrumbing”. This property describes whether rewards are used to guide player navigation by creating a visible alternate path that deviates from the original path.

5. DISCUSSION

The patterns described above were derived primarily through the design and analysis of Super Mario World-style levels, and secondarily by examining Portal 2 levels. Table 2 shows the applicability of the metrics between the two games. The vast majority are applicable to both. For example, in both Super Mario World and Portal 2, it is relevant to consider the length of the level, or the frequency of elements that are tied to particular styles of sound effects. However, the specific considerations used for the patterns would differ across games: counting up the number of potential failure states in a Portal 2 level is a different and more nuanced task than that of counting up the failure states in Super Mario World. The exact implementation of each metric will differ across games based on their mechanics.

As seen in table 2 most metrics are portable from Super Mario to Portal 2 but there are some metrics that are endemic to the first game and are not applicable to the second, such as aesthetic metrics based on palette variation and palette type. This is because the technology used to implement Mario levels relies on multiple equivalent tilesets, while Portal relies on a library of assets that do not have alternative textures or skins. The other set of metrics that are not portable consist of strategic metrics based on power-ups and rewards; this is because the game Portal 2 does not possess any mechanics that is categorizable as a power-up or a reward.

Aesthetic Metrics	**Super Mario**	**Portal 2**
Music	x	x
Sound Effects	x	x
Texture	x	x
-Luminosity	x	x
-Warm/Cool	x	x
-Palette Variation	x	-
-Palette Type	x	-
-Color Psychology	x	x
-Motion Salience	x	x
-Pattern Salience	x	x
-Color Salience	x	x
-Size Salience	x	x

Difficulty Metrics	**Super Mario**	**Portal 2**
Leniency Differentiation	x	x
Failure States	x	x
Threat Level	x	x

Topology Metrics	**Super Mario**	**Portal 2**
Negative Space	x	x
Verticality	x	x
Rhythm	x	x
Relative Size	x	x
Path Properties	x	x
Section Consistency	x	x
Length	x	x
Completion Time	x	x

Strategic Metrics	**Super Mario**	**Portal 2**
Trial and error	x	x
Reasonability	x	x
Situational Awareness	x	x
Number of Solutions	x	x
Power-ups	x	-
-Frequency	x	-
-Density	x	-
-Accessibility	x	-
Rewards	x	-
-Frequency	x	-
-Density	x	-
-Accessibility	x	-
-Breadcrumbing	x	-

Table 2. A comparison of the metrics applicable to Super Mario and Portal 2

It is important to note that while each of these metrics is applicable to a game, some of them cannot be applied when others are in use. For example, the leniency metric requires that the length of a level be held constant and equal to all other levels that it is being compared to, as it is normalized by level length. Thus the level length metric would not reveal any useful information in this scenario. A full computational implementation of each metric is required to better understand the conflicts between the metrics and potential ways to address them.

The classification and clustering of metrics was performed after working with the student designers to define each metric individually. This higher-level structure for metrics itself forms a useful vocabulary for describing different aspects of level design, and we anticipate expanding upon this vocabulary as more metrics are defined.

6. CONCLUSION AND FUTURE WORK

In this paper, we have described a novel approach for considering the evaluation and analysis of level designs, using an analytical vocabulary that spans multiple games. The quantitative approach to evaluation will be able to fill a gap in level design research, supporting rapid feedback to designers and the automated evaluation of PCG systems.

With this new approach for evaluation, there are several new avenues of research unlocked.

Several of the metrics described in this paper have revealed new design considerations for procedural level generators, identifying aspects that may be desirable to control for and reason about explicitly in the content generator. Aesthetics are rarely considered in procedural level design, with the focus instead being largely on the overall structure of levels, and the assumption that any tileset placed on top of that structure would be equally valid. However, an entire classification of metrics for level design that human designers have identified as an important feature is that of aesthetic choices in terms of both music and art assets. This research has reinforced the need for PCG researchers to begin taking seriously the problem of procedural art direction.

A computational implementation of these metrics to enable automated level evaluation is a next step that we are actively working on. The ability to perform automated level analysis has the potential for us to learn more about the aesthetic styles and design preferences of many human designers, as well as automated designers. Each metric score for a given level forms part of its “stylistic fingerprint”; by identifying these fingerprints quantitatively, it becomes possible to cluster levels together to find larger-scale patterns across levels, both within the same game and across different ones.

In order to envision this future in which we can compare levels between games, it is important to research the applicability of these patterns to games outside of the platforming genre. How these metrics translate to levels for physics-based games like Angry Birds [19] or scenarios for strategy games like Civilization IV [9] is an open question. There is much work to be done pushing the boundaries of this evaluation method, both in terms of how much in levels metrics can explain, as well as how valid it is to compare levels from completely different genres.

To expand and validate the metrics library, an area of future research would be to apply our method for finding metrics to different groups of designers, including prominent level designers on completed games. It is also possible to “validate” certain metrics by comparing the scores they provide on individual levels to human responses when playing them. For instance, a high leniency metric for a level should correspond to a one where players do not typically struggle or die frequently.

In creating this library of metrics, we have identified a new vocabulary for the quantitative analysis and qualitative description of level designs. Finding a shared vocabulary for level design, for the purposes of both facilitating communication between designers and teaching level design to students, is an important problem to solve. With these metrics and associated categorization, we hope to have provided such a vocabulary, and look forward to extending it in our future research.

7. ACKNOWLEDGMENTS

We wish to thank the students in the level design classes in which this activity was run, for their imagination and flexibility in considering this new approach to level design analysis. We also acknowledge Andrew Xu, Jingyu Yao, and Lindsey Kennard for their assistance with running discussion in the class, and for fruitful discussions about the computational implementation of some of the metrics, which helped further define them.

8. REFERENCES

[1] Amaya, G. et al. 2008. Game User Research (GUR): Our Experience with and Evolution of Four Methods. In 'Game Usability: Advancing the Player Experience', edited by Katherine Isbister, Noah Schaffer. Morgan Kaufman publishers

[2] Andersen, E., Liu, Y.-E., Apter, E., Boucher-Genesse, F., and Popović, Z. Gameplay analysis through state projection. *Proceedings of the Fifth International Conference on the Foundations of Digital Games*, ACM (2010), 1–8.

[3] Bjork, S. and Holopainen, J. 2004. Patterns in Game Design (Game Development Series). Charles River Media.

[4] Canossa, A. et al. 2011. Arrrgghh!!!: Blending Quantitative and Qualitative Methods to Detect Player Frustration. Proceedings of the 6th International Conference on Foundations of Digital Games (New York, NY, USA, 2011), 61–68.

[5] Canossa, A. Seif El-Nasr, M., Truong, H. 2014. Beyond Visualization: Democratizing Access to Game Analytics Through Interactive Sense-making. In CHI PLAY Game User Research Workshop (2014)

[6] Dahlskog, S. and Togelius, J. 2012. Patterns and Procedural Content Generation. Proceedings of the Workshop on Design Patterns in Games (DPG 2012), co-located with the Foundations of Digital Games 2012 conference (Raleigh, NC, May 2012).

[7] Desurvire, H., Caplan, M., and Toth, J.A. Using Heuristics to Evaluate the Playability of Games. *CHI ’04 Extended Abstracts on Human Factors in Computing Systems*, (2004).

[8] Desurvire, H. and Wiberg, C. Game Usability Heuristics (PLAY) for Evaluating and Designing Better Games: The Next Iteration. In *Online Communities and Social Computing*. Springer Berlin Heidelberg, 2009, 557–566.

[9] Firaxis Games. Civilization IV (PC Game). 2005

[10] Horn, B. et al. 2014. A Comparative Evaluation of Procedural Level Generators in the Mario AI Framework. Proceedings of the Foundations of Digital Games 2014 (Fort Lauderdale, FL, Apr. 2014).

[11] Hullett, K. and Whitehead, J. 2010. Design Patterns in FPS Levels. Proceedings of the 2010 International Conference on the Foundations of Digital Games (FDG 2010) (Monterey, CA, Jun. 2010).

[12] John, J. 2006. Pandora and the music genome project. Scientific Computing World. 23, 10 (2006), 40–41.

[13] Karakovskiy, S. and Togelius, J. 2012. The Mario AI Benchmark and Competitions. IEEE Transactions on Computational Intelligence in AI and Games. 4, 1 (Mar. 2012), 55–67.

[14] Kim, J.H. et al. 2008. Tracking Real-time User Experience (TRUE): A Comprehensive Instrumentation Solution for Complex Systems. Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (New York, NY, USA, 2008), 443–452.

[15] Kim, J.H., Gunn, D.V Schuh, E., Phillips, B., Pagulayan, R., and Wixon, D. 2008. Tracking real-time user experience (TRUE): a comprehensive instrumentation solution for complex systems. In Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '08). ACM, New York, NY, USA, 443-452.

[16] Niedenthal, S. What we talk about when we talk about game aesthetics. *Proceedings of the 2009 Digital Games Research Association Conference*, (2009).

[17] Ou, L.-C. et al. 2004. A study of colour emotion and colour preference. Part I: Colour emotions for single colours. Color research and application. 29, 3 (2004), 232–240.

[18] Rohrer, C., (2014) When to Use Which User-Experience Research Methods, white paper, Nielsen Norman Group

[19] Rovio Entertainment. Angry Birds (iOS). 2009.

[20] Russell, J.A., Barrett, L.F.. (1999) Core Affect, Prototypical Emotional Episodes, and Other Things Called Emotion: Dissecting the Elephant. Journal of Personality and Social Psychology 1999. Vol. 76. No. 5.805-819

[21] Russell J,A. A circumplex model of affect. Journal of Personality and Social Psychology. 1980;39:1161–1178.

[22] Smith, G. et al. 2011. Launchpad: A Rhythm-Based Level Generator for 2D Platformers. IEEE Transactions on Computational Intelligence in AI and Games. 3, 1 (Mar. 2011).

[23] Smith, G. et al. 2011. Situating Quests: Design Patterns for Quest and Level Design in Role-Playing Games. Proceedings of the 2011 International Conference on Interactive Digital Storytelling (Vancouver, BC, Canada, Dec. 2011).

[24] Smith, G. and Whitehead, J. 2010. Analyzing the Expressive Range of a Level Generator. Proceedings of the Workshop on Procedural Content Generation in Games, co-located with FDG 2010 (Monterey, CA, Jun. 2010).

[25] Swain, C. 2008. Master Metrics: The Science Behind the Art of Game Design. In 'Game Usability: Advancing the Player Experience', edited by Katherine Isbister, Noah Schaffer. Morgan Kaufman publishers

[26] Tan, C.T. et al. 2014. Inferring Player Experiences Using Facial Expressions Analysis. Proceedings of the 2014 Conference on Interactive Entertainment (Dec. 2014), 1–8.

[27] Tan, C.T., Leong, T.W., and Shen, S. Combining Think-aloud and Physiological Data to Understand Video Game Experiences. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, ACM (2014), 381–390.

[28] Totten, C. 2014. An Architectural Approach to Level Design. A K Peters/CRC Press

[29] Valve Software 2011. Portal 2 [PC Game].

[30] Wallner, G. and Kriglstein, S. 2012. A spatiotemporal visualization approach for the analysis of gameplay data. In Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '12). ACM, New York, NY, USA, 1115-1124.

[31] Zoeller, G. 2013. Game Development Telemetry in Production. In Game Analytics Maximizing the Value of Player Data. Edited by Seif El-Nasr, M., Drachen, A., Canossa, A. Springer Verlag

[32] Ubisoft 2008. Far Cry 2 [PC Game]