Exploration through video games

Gómez Maureira, M.A.

7 Empirical Evaluation of Level Design Patterns

This chapter describes the empirical evaluation of level design patterns for spatial exploration. To facilitate this evaluation, the research game *Shinobi Valley* was designed, developed, and piloted (Chapter 6) based on hypothesized design patterns (Chapter 5), following observations from games that elicited curiosity from players (Chapter 4).

> The research question that guides the work in this chapter is:
> **How do design patterns for exploration influence player behavior and experience?**

Although designers have an intuitive sense that curiosity is an important factor of game design (Schell 2008; Costikyan 2013; Klimmt 2003), how it can be purposefully elicited is not obvious and has not been studied empirically. This is unfortunate, as a more evidence-based understanding of what design features elicit the desire to explore would provide a stronger foundation for the research of player experience. It would also benefit the practice of game design and the development of engaging procedural environments.

The study presented in this chapter aims to perform fundamental work in filling this research gap. It assesses the impact of four level design patterns that are integrated through twelve individual implementations (three for each pattern). The four implemented patterns are:

- Overcoming *Extreme Points* (EXP) such as mountain peaks or other hard-to-reach structures

- Resolving *Visual Obstructions* (OBS) in the environment to discover what they might hide
- *Out-of-Place Elements* (OBS) that appear to not 'fit' into the environment
- *Understanding Spatial Connections* (SPC) between areas in the game environment

The expectation is that an environment purposefully designed to stimulate exploration causes players to behave differently and regard the environment more positively.

In the study, exploration is measured by the combination of in-game actions (i.e., game metrics) and players' accompanying emotional investment (through a post-game questionnaire and self-reported emotion words during gameplay). The gameplay is divided into a period of free exploration and one where participants must wait before completing the game.

This study aims to evaluate the following hypotheses:

- **H1a:** Level design patterns elicit more exploratory behavior from players
  H1b: Presence of level design patterns positively affects the emotional experience of the game
- **H2:** Having an explicit goal reduces exploratory behavior
- **H3:** Players with a higher predisposition for curiosity engage in more exploratory behavior

The study of level design patterns is the primary focus of the presented study and is what motivates the two H1 hypotheses. However, some aspects implicit in a video game can also influence exploration. For example, many games have a stated goal (a "main quest") that guides player movement. Furthermore, game environments do not simply consist of neutral topologies but have a distinct visual aesthetic to make them look enticing. To examine whether such extraneous factors impact player's curiosity for exploration, they are implemented as separate testing conditions. Another potential impact that is part of the evaluation is whether compensation for participation in the study influences exploratory behavior. Compensation is common in academic studies, and researchers of future studies building upon this work may consider offering it to participants. It also represents an incentive for completing the game in a short amount of time, thus increasing the curiosity that design patterns would need to elicit to motivate exploration. In this study

Overall, the study follows a *between-subjects 2x2x2x2 factorial design* with fixed factors: the presence of patterns, goal statement, environment aesthetic, and compensation.

An online experiment was conducted with 254 participants who were randomly sorted into different condition groups. Data collection consisted of questionnaires (both in- and post-game) and game metrics (e.g., distances from path and destination over time, play duration, player position and camera rotation, and instances of going out of bounds, i.e., jumping into a chasm). In addition to direct measures, interpreted emotion ratings were gathered through self-reported emotion words gathered during gameplay. These words were matched to affective components (valence, arousal, and dominance) based on the Glasgow Norms corpus (Scott et al. 2019).

The study results provide evidence for H1 and H2, but not for H3. Section 7.4 discusses the interpretation of results. The contributions of this study lie in providing an initial empirical study into level design patterns for spatial exploration. Results show that level design patterns impact exploratory behavior and that other factors further influence their effects. An explicit goal severely reduces exploratory behavior until that goal is fulfilled and the game becomes more open-ended. Receiving monetary compensation reduces exploration, but patterns motivate players to explore and perceive the experience more positively than when they play without them.

These findings can inform future game design considerations and should also be considered for the design of further studies in this area (e.g., variables to include, data analysis, and whether or not to offer compensation). Overall, the study aimed to be a nuanced, practical example of the complexity of studying video games experimentally and analytically, and provides a foundation for future work in the study of design patterns, curiosity, and player experience.

The following sections describe the design of the experiment, followed by the procedure and detailed results. The discussion section details the findings and explains how the results were interpreted. Finally, the chapter concludes with limitations and an overall conclusion of the study.

Chapter Publications

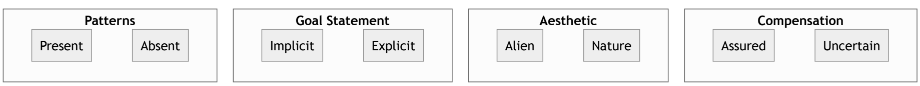
Work presented in this chapter has been published in this peer-reviewed venue:

- **Proceedings of the ACM on Human-Computer Interaction (PACMHCI Journal, Volume 5, Issue CHI Play) – 2021**
  "Level Design Patterns That Invoke Curiosity-Driven Exploration: An Empirical Study Across Multiple Conditions" (M. A. Gómez-Maureira et al. 2021)

7.1 Experiment Design

This section describes the structure and design of the study. Given that it is the second part of a two-part study, it involves aspects that were already discussed in the previous chapter. Especially the research game, *Shinobi Valley* (described in Chapter 6), remains a fundamental component of the experiment design in the second part of the study. While the game will not be described again here, relevant game features are briefly summarized where necessary as a reminder. Design considerations of the game are described with a focus on the role they play for participants, i.e., for the "serious" purpose that the game fulfills as a research artifact rather than its role as a simulated entertainment game.

The experiment for this study is designed to take place online and follows a between-subjects 2x2x2x2 factorial design. It tests the impact of four factors, each of which consists of two conditions. The fixed factors are (1) the presence of patterns, (2) the goal statement, (3) the environment aesthetic, and (3) compensation. Participants are randomly sorted into a condition group and given one of two options in each fixed factor. Participants play "their" randomized version of *Shinobi Valley*, during which data is collected through game metrics and a periodic in-game survey. Afterward, participants answer a post-game survey.



The following sub-sections describe the game variables that are modified between conditions, details regarding randomization of variables, participant sampling, experiment measures, and finally, the experiment setup.

7.1.1 Independent Game Variables

To examine the research questions, the design of *Shinobi Valley* varies in three aspects: the presence of patterns, the presence of a goal statement, and aesthetics. Participants are randomly assigned a combination of these variables. While the controls and general game progression remain unchanged by these variables, the game environment differs depending on which condition (i.e., the combination of variable states) a participant is assigned to. Variables are tested pair-wise, meaning that individual variables always overlap with others. Each participant plays only one possible combination of variables.

The individual game variables are described in the following sub-sections. In addition to game variables, another experiment variable is based on whether participants received financial compensation. This variable is further discussed in section 7.1.3.

7.1.1.1 Level Design Patterns for Spatial Exploration

A crucial goal of the experiment is to investigate whether the implemented design patterns successfully elicit curiosity in participants. To test this, participants take the experiment either with such *patterns present* or with *patterns absent*.

The *patterns present* condition has been the focus of previous design descriptions (see Chapter 6) but is summarized here briefly. Overall, there are 12 *pattern instantiation regions* (PIRs) that include a design pattern to motivate exploratory behavior. These regions are made up of four different kinds of patterns with three implementations each. The regions are distributed across the game environment along a primary path.

In the *pattern absent* condition, the 12 regions of interest are not present in the environment. In some cases, this slightly reduces the amount of explorable space, such as lacking mountains that were designed to be climbable and spatial connections that could be explored. Visual obstructions such as dense forests and the fog zone are removed and replaced with vegetation that mimics the distribution across the rest of the environment. The same is true for out-of-place elements that were meant to attract attention.

It is important to note that the game environment was created and developed with patterns in mind. Initially, the experiment only provided information on which design patterns are most successful in invoking exploratory behavior. The absence of patterns was implemented after they had already been implemented. While care has been taken to ensure that both conditions appear complete and consistent, participants will naturally find more opportunities for exploration in the presence of such patterns. At the same time, participants playing the game in the absence of patterns condition might still exhibit exploratory behavior. This is because the absence of patterns means the *absence of intentionally designed patterns* for this experiment.

Both conditions include areas that may very well attract the attention of participants but are not considered regions of interest, even if they otherwise exhibit traits of some design patterns. Especially *Extreme Points* patterns arise almost automatically depending on the surrounding environment. What differs is the extent and the intentionality of such patterns.

7.1.1.2 Goal Statement

In the experiment, participants need to follow an S-shaped path in the environment to its end to progress and complete the game session. The goal statement condition pair separates participants into playing a version in which they are given this task *explicitly* and a version in which the goal is only *implicitly* evident due to the existence of a path.

In the *explicit goal statement* condition, participants start their game session with two subsequent text messages before they gain control of the player character. The messages state the following (numbers indicate separate message boxes in sequence):

> (1) “This is you! You are a monkey ninja on a journey to meet your master.”
> (2) “Your master awaits your arrival at the end of this path.”

After the messages have been confirmed, participants are shown a sequence of three slow camera pans from predefined points along the path (figure 7.1). Each of the three pans lasts five seconds and fades over black to the next pan. The sequence of camera pans indicates how to get to the end. The last camera pan shows the ninja master characters sitting on a large stone in the distance, giving participants a visual preview of their destination. In addition to showing the path, the camera pans also present glimpses of the regions of interest meant to elicit curiosity for exploring the surrounding environment. The path preview thus acts as a goal statement for participants and a visual stimulus for what could be encountered if they leave the path.

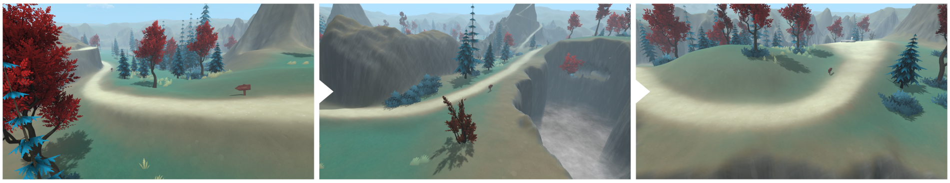


Figure 7.1: Camera shot sequence at the beginning of the game, following the primary path. The signposts shown in the sequence are only present in the explicit goal condition.

After the third camera pan, the camera shows the player character again, and the game proceeds to explain the control scheme to participants. After completing the tutorial, participants receive the following message before starting the game:

> "Now go! Your master awaits."

Already during the path preview, participants can see a series of wooden signs next to the primary path. The signs point towards the location of the master and, thus, to the apparent target destination. In addition to reminding players of the goal statement, the signs provide navigation support, so they do not get lost.

Overall, the *explicit goal statement* consists of three messages, a visual preview of the primary path lasting 15 seconds, and nine wooden signs pointing to the path destination. Consequently, these elements are absent in the *implicit goal* condition.

In the *implicit goal* condition, participants start directly with the game tutorial and receive the following message before starting the game:

> "You can now control your character."

During the game, participants need to orient themselves using the surrounding environment. At the same time, the primary path remains a distinctive visual (and topological) feature. *As such, neither condition is entirely free from any goal statement.* The *implicit goal* condition still presents participants with a goal, but one that is more uncertain and open to interpretation.

7.1.1.3 Visual Aesthetic

To draw valuable conclusions from the experiment, the expressed behavior of participants should be due to the impact of PIRs in the environment. The design of such regions is based on design patterns that are, for the most part, independent of how a game looks. At the same time, creating a game environment that is entirely neutral in its aesthetic appearance is not possible. For this reason, participants play *Shinobi Valley* in one of two visual aesthetics: the *nature aesthetic* condition and the *alien aesthetic* condition. The aesthetic aspects of each are further described in 6.

This control condition pair is not expected to impact exploration behavior. None of the game parameters and mechanics are affected by the game's visual aesthetic. The placement of vegetation is identical for both conditions, and care was taken to keep the dimensions of individual bushes and trees as close as possible.

7.1.2 Game Randomization

Each participant in the experiment plays one possible game version, consisting of a combination of the independent variables described in the previous sub-sections. Whenever a participant begins the experiment, they are assigned a combination of the three game variables (i.e., patterns, goal, aesthetic). Each variable has two possible options. The goal statement and aesthetic variables have a 50% chance for either option.

The pattern variable has a *70% chance of playing with level design patterns present* versus a 30% chance of playing in the absence of patterns. Although this creates an imbalance in the sample sizes between the different conditions, it allows for collecting more data on participants interacting with the (individual) level design patterns and subsequent analysis of those interactions. Considering that the analysis of individual patterns is a fundamental goal in this study, this was considered an acceptable trade-off to maximize available resources (i.e., time and budget for financial compensation). Implications of this decision are discussed as part of the study limitations in section 7.5.

Apart from the independent game variables, participants are randomly assigned one of two possible play directions. Randomizing the starting position in this manner is meant to counteract the impact of uneven distributions of individual PIRs and the potential order effects in encountering them. Since the play direction is not hypothesized to impact the overall efficacy of patterns, play direction is included as a nuisance variable rather than a fixed experiment factor.

In summary, the following game parameters are randomized for each participant:

- Pattern present (70% chance) or absent (30%)
- Goal statement implicit (50%) or explicit (50%)
- Nature aesthetic (50%) or alien aesthetic (50%)
- A→B (50%) or B→A (50%) play direction (treated as nuisance variable)

7.1.3 Sampling

Participants in the study were recruited through a combination of snowball sampling and crowd-sourcing. Snowball sampling included reaching participants through social media, the University environment, and word-of-mouth. Crowd-sourcing took place on three recruitment platforms: *Mechanical Turk* (Amazon 2021), *Prolific* (Prolific 2021), and *SurveyCircle* (SurveyCircle 2021). *Mechanical Turk* (*MTurk*) and *Prolific* participants received *assured monetary compensation* of 3.00 EUR each for completing the study. All other participants could opt into a lottery for *uncertain compensation* of three 20.00 EUR vouchers. The only requirement for participating in the study was that participants have basic English comprehension and that their computer can run the game smoothly. As described in section 7.2, the experiment logs game performance and terminates if it falls below a minimum threshold. From *Prolific*, only female participants could participate; a decision made to counter-balance a lean towards male participants up to that point in the data collection.

Players are hypothesized to be intrinsically motivated to engage with games due to the experiences that they offer (e.g., enjoyment from overcoming challenges, interacting with the game world, social interaction; see R. M. Ryan, Rigby, and Przybylski (2006)). However, it is possible that playing a game for research purposes influences how players engage with it, especially when they are compensated for their time. While some participants may be inclined to finish as fast as possible to maximize their gains, others might feel obligated to do well and earn their reward. In this study, the involvement of an extrinsic reward may affect intrinsic motivation and, possibly, the behavior that is being studied (Eisenberger, Pierce, and Cameron 1999). Given that the effect of an extrinsic reward has not yet been empirically tested for how it influences exploration in video games, *whether players did or did not receive assured monetary compensation is included as a fourth independent experiment variable.*

7.1.4 Measures

The *Shinobi Valley* game experiment primarily collects quantitative measures that are expressions of player behavior and assessments of their personality concerning curiosity. Qualitative measures are taken in the form of free text input. While quantitative measures can appear to deliver precise results, it is important to keep in mind that they are proxy measures for what the *Shinobi Valley* experiment seeks to investigate: curiosity as an emotional impact of design interventions. Quantitative data in the experiment should be understood as sources for constructing arguments for when and why design elements in the game impact participants. Qualitative data is used similarly; as supporting material for formulating arguments. Where it differs is in what material it sources. While quantitative data describes the expressed behavior of participants, qualitative measures are taken to construct an understanding of the motivational and emotional aspects that accompany that behavior.

The following subsections describe the individual measures that are taken in the study.

7.1.4.1 Game Metrics

The game logs player parameters at a repeated interval several times per second. Each log line includes position, camera rotation, avatar movement velocity, the closest distance to the primary path, and distances to start and destination points. Apart from repeated logs, the jumping and running of the player character are logged at the time of input as timestamped events. Timestamps are also logged for arrival and leaving of PIRs, arrival and leaving of start and destination areas, instances of resetting the player character (e.g., jumping into the chasm or when getting stuck in the level geometry for longer than 2 seconds), as well as triggering and completion of the in-game survey.

Additional metrics are generated from the aforementioned measures, primarily through spatio-temporal data created by participants in combination with the predefined locations of PIRs. These include measures for PIR visit counts, PIR stay duration, and spatial entropy of player movement; specifically, Altieri's entropy (Altieri, Cocchi, and Roli 2019), which captures the impact of localized clustering in addition to the overall heterogeneity of spatial data.

The following game metrics are captured at a regular interval of 5 measures per second (5 Hz; i.e., every 200ms):

- **Position of the player character (and position delta)**
  Measured in X, Y, Z position coordinates from the center of the environment (i.e., the virtual world origin point), and as position delta since the last position log. Positions are measured with an accuracy of three digits after the decimal point, with each unit being the rough virtual equivalent of a meter. The position accuracy is thus tracked at the equivalent of 1 mm in the real world. For comparison, the playable area in the game measures 256 x 256 in-game units (equivalent to meters), and the player character can run at a speed of 4 units per second.
- **Distance from the primary path**
  The path distance is tracked using the same measurement unit and accuracy as the player character's position, but with respect to the shortest distance to the primary path. This measure is based on a simplified bezier curve that roughly follows the middle of the visual representation of the primary path in the environment. The path distance expresses the shortest distance between the player character and the primary path. Given that the path's visual representation varies, a path distance of 1 unit (i.e., meter) may be just about off the path in some areas while still on the path in others.
- **Camera rotation (and rotation delta)**
  Measured in quaternions, delta rotations in degrees (pitch and yaw), and absolute delta in degrees. Both delta measures indicate the rotation since the last log entry. In contrast to the movement speed of the player character, the maximum possible delta rotation can differ between participants. This is because the camera rotation directly corresponds to mouse tracking sensitivity and the sensitivity setting that participants choose at the beginning of the experiment.
- **Player character state**
  Two player character states are tracked as binary (i.e., "true" or "false") parameters: whether the player character is currently jumping (i.e., airborne and in an

upwards trajectory) and whether the player character is currently running (as opposed to moving at regular walking speed).

Additional game metrics are captured when specific events occur. In contrast to metrics tracked at an interval, *event-based metrics* are timestamped at the moment of the event itself and are independent of a fixed measure frequency. The following game metrics are event-based:

- **Location triggers**
  Several areas in the game environment have triggers that log when the player character enters or leaves the area. Triggers are implemented as invisible spheres around a given point, with the sphere size corresponding to the size of the area of interest. All regions of interest have such location triggers, as do the starting and destination areas (i.e., the ninja master that awaits participants at the end of the primary path).
- **Out of bounds**
  When the player character jumps into the chasm, their position is reset with accompanying effects indicating that they left the playable environment. In addition to the time this event occurred, the player position is also logged.
- **Player stuck reset**
  As with many games that simulate real-life physics, participants may find themselves stuck in the virtual geometry. The game keeps track of whether participants can move their character and resets their position to a nearby flat terrain after a few seconds if the player character gets stuck. Such occasions are logged with a timestamp and the player character's location at that time. It should be noted that this metric is primarily used to identify potential outliers in the recorded participant data.
- **Movement modifier inputs**
  Participants can use either their keyboards or mouse to make the player character jump or sprint. These events are logged when the respective input is entered.

By capturing these game metrics, the participant behavior in the game can be largely reconstructed for any given point in the experiment. This is particularly helpful when testing the experiment procedure to determine if there are problems in the game code or in how data is logged. For the actual data evaluation, however, most game metrics are used to derive measures that illustrate the play session as a whole. The following measures are derived from game metrics:

- **Durations**
  By using the timestamps of location triggers, stay durations that provide information about the participant behavior can be derived. Stay durations are logged to track how much time was spent, in sum, at a given region of interest. The duration of the entire game session is also tracked, as well as time spent before and after reaching the master.

- **Aggregate counts**
  The overall amount of "out of bounds" and "player stuck reset" events are summed up and tracked for each participant.

- **Session statistics**
  Some game metrics that are tracked at a regular interval are also expressed as session statistics to allow for comparison with measures from other participants. In each case, statistics are created for the entire session, as well as for the timespan before and after players have reached the master. The jumping and running of the player character are tracked in percentage over the session length. Since jumping is a relatively brief event, the percentages are generally low but enable direct comparison between participants with varying play durations. The following statistics are derived for position delta and path distance: mean value, standard deviation, median, and median absolute deviation. For rotation deltas, the same statistics are calculated but without median and median absolute deviation, as camera rotation often happens in short bursts, with the median resulting in zero in many cases.

7.1.4.2 In-Game Player Feedback

During the experiment, a feedback screen pops up at predetermined times to ask participants about their emotional state and to rate their curiosity on a unitless sliding scale. The pop-up screen is broken up into two sub-screens, with each screen focusing on a single question (figure 7.2). The separation into two individual sub-screens is meant to reduce bias when answering the second question. The first sub-screen presents participants with a question requiring them to enter text to progress. The question states:

> "In a word, what is your current emotional state?"

Once participants have entered some text, they can click a button to continue to the second sub-screen, where they are asked to rate their curiosity on a scale from "not curious at all" to "very curious". A slider can be placed and dragged on that scale. The slider position results in a measure of `0` for "not curious at all" and `1` for "very curious". The sliding scale does not provide any numerical feedback to participants but internally logs their input at two decimal places (e.g., `0.75` when the slider is three-quarters towards the "very curious" end of the scale). The slider does not snap to predefined increments, a decision that prevents participants from comparing their current measure to previous feedback. The sliding scale lets participants deliberately capture curiosity and acts as a momentary snapshot of their emotional state. Before participants click on the sliding scale, the slider handle is not visible and thus does not indicate a default neutral point. This increases the likelihood that ratings around the midpoint are set deliberately by participants and not the result of unreflected convenience.

Figure 7.2: The two sub-screens of the in-game player feedback interface.

Once participants leave the starting area in *Shinobi Valley*, an internal timer starts and determines when to present the in-game feedback screen to them. The timer only progresses when no user interface is shown on the screen (such as the feedback or help screen). The timer interval is set to the sequence: `1`, `3`, `5`, `8`, `12`, `17`, `23`, `30`, `38`, `47`, and `57`. This means that the first feedback screen is triggered after one minute, then two minutes later (minute `3`), then again two minutes later (minute `5`), and finally, with a linear increase of 1 additional minute between subsequent intervals. Participants are not expected to play for such a long time, but if they do, their feedback will be requested for up to one hour of playing time. Naturally, the amount of collected feedback is thus dependent on how long participants play. In the pilot study of *Shinobi Valley* the average play time was around 12 minutes. This is also why there are two intervals of 2 minutes between triggering the feedback screen rather than just one, as it provides more feedback at a time when most participants are likely to see it.

It should be noted that the player character needs to be on the ground for the feedback screen to trigger. If the character is "airborne", the feedback screen is delayed until the character is on the ground again. This decreases the likelihood of the feedback screen appearing in the middle of a jump. The delay is at most 1-2 seconds, such as when players jump from a mountain which causes the longest time spent "ungrounded". The reason for implementing such a delay is that taking the control away from participants during a game is already a distracting event that can impact their affective state negatively, at least momentarily. The impact would likely be more substantial if the screen were triggered in the middle of an action.

7.1.4.3 General Player Data

After playing *Shinobi Valley*, participants take an online post-play survey that involves questions about themselves, their experience with video games in general, and their experience with *Shinobi Valley*. The survey is similar to the one used in the pilot study, with some minor adjustments.

Demographic questions include age and gender identity, both of which are not mandatory. For gender identity, participants can provide their preferred identity as free text.

Participants are asked to self-identify as one of four types of video game player: "Novice", "Casual", "Core", or "Expert". These terms are not further explained and are chosen to reflect common terminology among game-playing audiences. As an example, the notion of a "core player" is likely not a familiar term to participants that are not very familiar with video game culture. This makes it more likely for them to self-classify as either "casual" or "novice" players, terms that are less domain specific.

Another question asks participants to estimate how often they have played video games in the last year. A note informs participants that they should include games played on mobile devices but exclude non-digital games such as board games or physical card games.

Possible options are:

- Never
- Less than 1 hour combined over the entire year
- 1 hour over the entire year on average
- 1 hour per month on average
- 1 hour per week on average
- 1 hour per day on average
- More than 1 hour per day on average

Participants are further asked about what game they are reminded of after playing *Shinobi Valley*, as well as what their favorite video game is. Both questions do not need to be answered but provide complementary information regarding a participant's perspective toward video games and the experiment. The next question inquires whether participants chose to leave the primary path and enter free text about why they did or did not. This is followed by a question about whether any game elements stood out to them, both negatively or positively. A subsequent comment field allows participants to comment freely on the game or the experiment.

The questions mentioned above are shown directly to participants and require their input. However, the online questionnaire also captures additional information about each participant. Based on the IP address, geolocation is logged that can be translated into a city and country belonging to that location. A URL parameter that is part of the experiment link is used for logging how participants were sampled and allows for later comparison between participants that were invited via crowdsourcing platforms and those that were reached via snowball sampling. Finally, the time it took participants to complete the questionnaire is measured at three points to estimate whether responses have been taken in a reflected manner. This allows for filtering out participants that responded much faster than the majority as outliers.

7.1.4.4 GUESS Questionnaire

The Game User Experience Satisfaction Scale (GUESS) is a validated scale for various factors contributing to video game experience. It is already used as part of the pilot study and described in more detail in Chapter 6.

In contrast to the pilot study, the GUESS questionnaire in this part of the study does not feature the sub-scales *Usability / Playability*, *Audio Aesthetics*, and *Visual Aesthetics*. The sub-scales of *Narratives* and *Social Connectivity* were already excluded before and are also not part of this experiment. The rationale for involving fewer sub-scales is that the game is no longer being assessed regarding its ability to simulate an entertainment game. By involving fewer questionnaire items, the overall duration of the experiment could be reduced.

Overall, participants rate a total of 26 statements, providing results in four sub-scales: *Play Engrossment*, *Enjoyment*, *Creative Freedom*, and *Personal Gratification*.

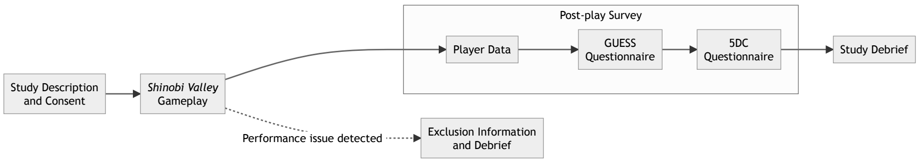
7.1.4.5 5DC Questionnaire

The *Shinobi Valley* experiment is designed to elicit the desire to explore through localized design interventions. Whether or not participants develop such a desire is, in part, influenced by their disposition to become curious. To assess this disposition, the post-play survey includes the Five-Dimensional Curiosity Scale (5DC), which describes an individual's general tendency to become curious (Kashdan et al. 2018). The scale and the questionnaire are further described in Chapter 4.

Within the post-play survey, items of the 5DC are presented to participants in sets of five items per page. The order in which the items are presented is randomized for each participant.

7.2 Procedure

In the study, *Shinobi Valley* is used as an online game and can be played in most modern browsers that support WebGL 2. All experiment steps are online and presented to participants through an experiment website that guides them along the way. The experiment was approved by the *Ethics Review Committee of the Faculty of Science* at *Leiden University*.



The first step is a description of the experiment and a request for consent for logging player data. Participants are not explicitly told that curiosity or exploration is the main focus of the research. Instead, they are informed that they will play a video game for research and how long the study will take ("around 15-20 minutes").

Before continuing to the game, the experiment website checks whether the browser window is large enough for the game to be played in full resolution. If the window is smaller than the game's dimensions (1920 by 1080 pixels), the participant cannot proceed and instead sees a message to play the game on a larger display or resize their window. If the resolution is sufficient, the participant can proceed to the game.

Within the game phase, participants are informed that *Shinobi Valley* should be played with headphones. Participants then proceed with playing the game, starting with the tutorial. The individual game phases are further described in Chapter 6. Participants are randomly sorted into a condition group and assigned a starting position in the game.

During play, the game periodically checks the frame rate at which it runs. When it registers that game has been running at less than 15 frames per second over a more extended period, the game stops and informs the participant that a performance issue has been detected. This check aims to ensure that participants experience the game at a minimum viable frame rate, given that subpar game performance can impact measured player behavior. Although participants are informed as part of the study description that a performant computer is required, it can be erratic to rely on participants to judge whether their computer performs sufficiently well.

While participants play the game, their actions are logged and submitted to a central experiment server. Periodically, an in-game survey appears as an in-game window that cannot be dismissed until it has been filled out.

Participants finish the game by finding the ninja master and waiting for the end of their meditation session. Upon completing the game, the experiment website automatically directs them to the post-play survey.

7.3 Results

The statistical tests conducted in this study use a Bayesian approach (O'Hagan 2008) and are calculated using *JASP* (JASP Team 2020; Marsman and Wagenmakers 2017). The reported *Bayes Factor* (BF_{10}) indicates the probability of the presence of an effect versus the absence (Schönbrodt and Wagenmakers 2018). A BF value over 1 indicates that the tested hypothesis is more likely than the null hypothesis. A value of 1 means that there is an equal chance of the hypothesis being different from the null hypothesis as there is of them being similar. A value below 1 indicates that the null hypothesis is more likely. Unlike classical hypothesis testing, a Bayesian test can indicate the likeliness of the null hypothesis rather than only rejecting it (O'Hagan 2008).

This thesis uses BF synonymously with BF_{10}. However, indexes are provided when not testing against the null hypothesis. Following standard practices (Jeffreys 1961), the study considers $BF>3$ as *moderate evidence for* a hypothesized effect (i.e., at least three times higher likelihood of a hypothesized effect versus no effect). A result of $BF<0.33$ indicates moderate evidence *against* the hypothesized effect (effectively, at least three times higher likelihood against a hypothesized effect). These values roughly correspond to a significance level of $p<0.05$ being interpreted as a statistically significant measure.

In this study, a *two-sided Bayesian T-test* is used to determine whether observations are significantly different. A *two-sided Bayesian Pearson correlation* is used to assess significant relationships between measures. Given the absence of well-informed (and sourced) prior beliefs, the default values for uninformed priors provided by *JASP* are used (Cauchy prior with of 0.707). Data for statistical tests is prepared using the *Pandas* package in Python (Reback et al. 2021).

A report of statistical tests and the underlying data can be found in the Open Science Framework (OSF) repository of this study[1], including settings that were used to calcu-

[1] OSF repository: https://doi.org/10.17605/OSF.IO/MVR37

late the results. The following sections describe noteworthy results derived from the experiment data. First, the participant sample and general observations are described. This is followed by results relevant to the experiment factors and the recorded nuisance variables. Finally, results relating to the performance of the four Pattern Instantiation Region (PIR) sets (extreme points, visual obstructions, out-of-place elements, and spatial connections) and their interrelation with experiment factors are presented.

7.3.1 Descriptive Statistics and General Observations

Overall, 389 participants took part in the experiment, out of which 266 completed the game and the post-game survey. Incomplete measures, and participants accounting for the fastest 2.5% of survey completions compared to the median, were excluded. Of the remaining $n=254$ participants, 48% identified as female ($n=122$), 50.8% as male ($n=129$), and 1.2% ($n=3$) identified as non-binary. The mean age was 31.8 ($SD=10.8$, $range=[18, 69]$). 35% of participants were recruited via *MTurk* ($n=89$), 31.1% through snowball sampling and social media ($n=79$), 28.7% from *Prolific* ($n=73$), and 5.1% from *SurveyCircle* ($n=13$). Female participants recruited via *Prolific* balanced out otherwise male-dominant demographics (without *Prolific*, 69.6% of participants were male).

To recall, the study uses a between-subjects 2x2x2x2 factorial design, with play direction and gender recorded as nuisance variables. Each participant contributes data to each factor and is randomly assigned one of the two conditions (with 50% probability for all experiment factors and play direction but 70% probability for playing with patterns versus 30% without).

The participant breakdowns for the individual experiment factors are:

- Played with patterns 72.4% ($n=184$; vs. $n=70$ without patterns)
- Played with goal statement 51.2% ($n=130$; vs. $n=124$ without)
- Played in alien environment 48.8% ($n=124$; vs. $n=130$ in nature environment)
- Played with assured financial compensation 63.8% ($n=162$; vs. $n=93$ without).
- Played in A→B direction 50.8% ($n=129$; vs. $n=125$ in B→A direction)

The participant breakdown closely matches the randomization percentages set as part of the experiment design. Participant procurement for the non-compensation group fell slightly short of the even-split target.

Table 7.1: 'Glasgow ratings' based on translated emotion words, split out over the three emotion dimensions arousal, dominance, and valence. For each, individual means and standard deviations are listed, which in turn provide means and standard deviations across all participants. Data points are split out to ratings 'before' and 'while' waiting.

	Arousal				Dominance				Valence			
	Indiv. Means		Indiv. SDs		Indiv. Means		Indiv. SDs		Indiv. Means		Indiv. SDs	
	M_M	SD_M	M_{SD}	SD_{SD}	M_M	SD_M	M_{SD}	SD_{SD}	M_M	SD_M	M_{SD}	SD_{SD}
Before Waiting	5.18	1.48	0.23	0.51	5.62	1.10	0.20	0.44	6.30	1.80	0.33	0.72
While Waiting	4.41	1.31	0.71	0.68	4.90	1.02	0.50	0.56	4.70	1.88	0.75	0.90

The average frequency of playing games when converted to was ***Mn=5.5*** (***SD=1.5***), corresponding to an in-between of "1 hour per week on average" and "1 hour per day on average" . The average gamer type self-identified as between "casual" and "core" (***Mn=2.6***, ***SD=0.9***).

The average play time was ***Mn=10.3*** minutes (***SD=3.6***, ***range=[6.6, 30.4]***). The mean play time before waiting was ***Mn=3.7*** minutes (***SD=3.7***, ***range=[0.6, 27.7]***), vs. ***Mn=6.5*** while waiting (***SD=1.4***, ***range=[2.7, 12.4]***).

Results of overall GUESS ratings (assessed on a scale of 1 for 'worst' to 7 for 'best') were:

- Creative Freedom: ***Mn=4.2*** (***SD=1.2***) — compare to ***Mn=5.15*** in pilot
- Play Engrossment: ***Mn=3.9*** (***SD=1.3***) — vs. ***Mn=4.78*** in pilot
- Enjoyment: ***Mn=3.7*** (***SD=1.6***) — vs. ***Mn=5.18*** in pilot
- Personal Gratification: ***Mn=4.3*** (***SD=1.3***) — vs. ***Mn=5.05*** in pilot

All GUESS ratings are close to the scale's midpoint but generally lower than the game's ratings during the pilot study.

The average in-game curiosity rating was ***Mn=0.6*** (***SD=0.2***, ***range=[0.01, 1.0]***); slightly above the scale mid-point of ***0.5***. Ratings steadily decreased over the play session, with ***Mn=0.71*** (***SD=0.2***) for the first rating, going down to ***Mn=0.49*** (***SD=0.3***; at ***n=34*** due to different play lengths) for the fifth rating moment.

Players used ***n=104*** unique emotion words to rate their emotional state when providing curiosity ratings. Before waiting, the most frequent responses were "curious" (***12.6%***), "calm" (**8%**), and "happy" (**8%**). While waiting, the most frequent responses were "bored" (**24.4%**), "annoyed" (**7.5%**), and "curious" (**6.4%**). Translation of the words to emotion dimensions via the Glasgow Norms resulted in means and SDs for each dimension (i.e., arousal, dominance, and valence) for each participant. SDs indicate the "emotional range" of words provided by the participant throughout the play session. From the individual participant results, means and SDs of each emotion dimension can be calculated across participants as well, resulting in a "mean of individual means", "SD of individual means", "mean of individual SDs", and "SD of individual SDs".

The resulting "In-ame *Glasgow* Ratings" are listed in table 7.1. A *paired-sample Bayesian T-Test* of *Glasgow* ratings before and after waiting shows *decisive evidence* for measures differing between the two experiment phases (**BF>1k**; i.e., **1000** times more likely than no difference). Individual means are higher before waiting for all emotion dimensions, while individual SDs are lower before waiting.

Most measures of player behavior differ notably before waiting, compared to while waiting; such as play duration (shorter before waiting; **BF>1k**), movement speed (**BF>1k**), and camera motion (**BF>100**), and spatial entropy (**BF>1k**).

Results of qualitative data, gathered in the form of coding participant comments, are shown in table 7.2.

7.3.2 Fixed Factor Results

In order to examine the impact of fixed factors on player behavior and emotional experience, a *Bayesian ANOVA test* was carried out for several dependent variables. These include game metrics (i.e., distance traveled from the path, distance traveled from the destination, play duration, position, rotation, and instances of going out of bounds, i.e., jumping into the chasm) and the *Glasgow* emotion ratings. Such a test results in a list of models (comprised of different combinations of the fixed factors) that have a likelihood of explaining differences in a specific measure (expressed as BF_M). For each fixed factor (and possible combination of fixed factors), a likelihood is calculated that they are part of a model that explains the difference (expressed as BF_{incl}). Finally, post hoc T-Tests show the likelihood of a fixed factor contributing to differences in a particular measure (expressed as BF_{10}).

Table 7.2: Notable 'tagged' comments with total counts, as well as split by 'with' or 'without' pattern and goal condition. Rows in the upper section are coded from reasons given for leaving the path, rows in the middle section are coded from elements that stood out to participants, and rows in the bottom section are valence of comments left for impression of the game as a whole. The table only contains comments that were given by at least 10 participants in total, and that are relevant to the posed hypotheses. Rows are sorted by overall count within each section.

			Pattern split		Goal split	
Tag	Meaning	counts	with	w/o	with	w/o
explore	explore in general	142	105	37	67	75
wait	to pass time while waiting	80	51	29	61	19
landmark	explore a landmark (unspecified target)	45	40	5	25	20
boredom	to alleviate boredom	32	21	11	17	15
mountain	to go to specific PIR: mountains	26	25	1	15	11
boundaries	to test limits of environment	26	17	9	13	13
expect	to find an expected game element that is not implemented	23	15	8	11	12
rocks	to go to OOP PIRs	16	16	–	7	9
scenery	to look at aesthetic elements in the environment	10	7	3	6	4
fog	to go to specific PIR: ground fog	10	10	5	3	7
chasm	to go to explore chasm (possibly incl. cliff cave PIR)	10	6	4	7	3
landmark	game area: any landmark	68	63	5	30	38
noInteraction	lack of interactive elements in the game	60	43	17	35	25
scenery	aesthetic elements of the game environment	58	41	17	29	29
wait	negative experience of having to wait	32	19	13	16	16
mountain	game area: mountains	32	30	2	17	15
cave	game area: caves	27	27	–	13	14
relaxing	overall atmosphere is experienced as calming	20	15	5	13	7
fog	game area: fog	19	19	–	9	10
rocks	game area: stone stacks	13	13	–	3	10
noReward	lack of validation for actions by the player	12	11	1	5	7
noGoal	lack of purpose or goal	11	6	5	3	8
val-pos	Valence of comment: positive	57	44	13	31	26
val-neg	Valence of comment: negative	36	22	14	12	24
val-neutral	Valence of comment: neutral	30	23	7	13	17
val-mix	Valence of comment: mixed positive and negative	18	12	6	10	8

Table 7.3: Fixed factor ANOVA results: Best models of Bayesian ANOVA for dependent variables, if model Bayes factor $BF_M > 3$. Results are split out for 'before waiting' and 'while waiting'. Individual fixed factors in the model are sorted in descending order by probability of inclusion in the model (BF_{incl}). Post Hoc T-Test results (BF_{10}) are included as superscripts if at least $BF_{10} > 3$, reported in steps: >3, >10, >100, >1k (1000). Interaction effects (e.g. [Goal × Pat]) do not have an associated BF_{10}. Fixed factor abbreviations are: Pattern (Pat), Compensation (Comp), and Environment (Env).

	Before Waiting		While Waiting	
Dependent Variable	Best Model (in order of incl. probability)	BF_M	Best Model (in order of incl. probability)	BF_M
Spatial Entropy	$Goal^{1k}$ + Pat^{10} + [Goal × Pat]	25	$Comp^{10}$ + $Goal^{10}$ + [Comp × Goal]	33
Path Dist. (M)	$Goal^{1k}$ + Pat^{100} + [Goal × Pat] + $Comp^{10}$	38	Pat^{1k} + [Goal × Comp] + $Comp^{3}$ + Goal	48
Path Dist. (SD)	$Goal^{1k}$ + Pat^{10} + $Comp^{10}$ + [Goal × Pat]	22	[Comp × Goal] + $Comp^{10}$ + Goal	31
Destination Dist. (M)	$Goal^{10}$ + Pat^{3} + [Goal × Pat] + Env	12	$Goal^{100}$ + [Goal × Comp] + $Comp^{3}$	34
Destination Dist. (SD)	*Null*	50	$Goal^{10}$+[Goal × Pat]+$Comp^{3}$+[Goal × Comp]+Pat	12
Duration	$Goal^{1k}$ + Pat^{3} + [Goal × Pat]	38	*Null*	26
Position Delta (M)	Goal	14	$Comp^{100}$ + Goal + [Comp × Goal]	17
Position Delta (SD)	$Goal^{1k}$	13	$Comp^{1k}$ + Pat	32
Rotation Delta (M)	$Goal^{10}$	25	*Null*	18
Rotation Delta (SD)	$Goal^{3}$	39	*Null*	21
Out of Bounds	$Goal^{10}$ + Pat + [Goal × Pat]	19	$Comp^{100}$	33
Glasgow Arousal (M)	*Null*	48	Pat^{3}	25
Glasgow Arousal (SD)	$Goal^{100}$ + Pat	16	$Goal^{3}$	39
Glasgow Dom. (M)	*Null*	24	*Null*	13
Glasgow Dom. (SD)	$Goal^{3}$ + Pat	12	Pat^{10} + $Goal^{10}$	18
Glasgow Valence (M)	*Null*	22	Pat^{3} + $Comp^{3}$	20
Glasgow Valence (SD)	$Goal^{10}$ + Pat + Comp	7	$Goal^{10}$ + Pat^{10}	33

The *Bayesian ANOVA* tests were run across the fixed factors "patterns", "goal", "environment", and "compensation". Two nuisance factors ("gender" and "play direction") were added to the null model and were thus included in all tested models. The exact settings are included in the OSF repository[2].

Table 7.3 shows an overview of the tests, the best model to explain differences in each variable, and the BF_{10} value of each fixed factor where $\boldsymbol{BF_{10} > 3}$.

7.3.3 Differences Between Design Patterns

To compare differences between the design of patterns, measures relating to individual *Pattern Instantiation Regions* (PIRs) are grouped into sets: *Extreme Points* (EXP), *Visual Obstructions* (OBS), *Out-of-place Elements* (OOP), and *Spatial Connections* (SPC).

[2] OSF repository: https://doi.org/10.17605/OSF.IO/MVR37

Each PIR set consists of three instances in the game environment. Results for differences between patterns are based on ***n=184*** (corresponding to 70% of the sample), as participants without patterns do not contribute any relevant data. PIR set measures are based on player activity in a predefined radius around individual PIRs of 8 game engine units (roughly equivalent to 8 meters). For each PIR, three measures are calculated:

- Spatial entropy (dispersion of player movement)
- Visit count (unique entries into a PIR lasting at least 1 second)
- Stay duration

Since only movements within a confined radius are considered for each PIR, spatial entropy only indicates player movement within the PIR radius.

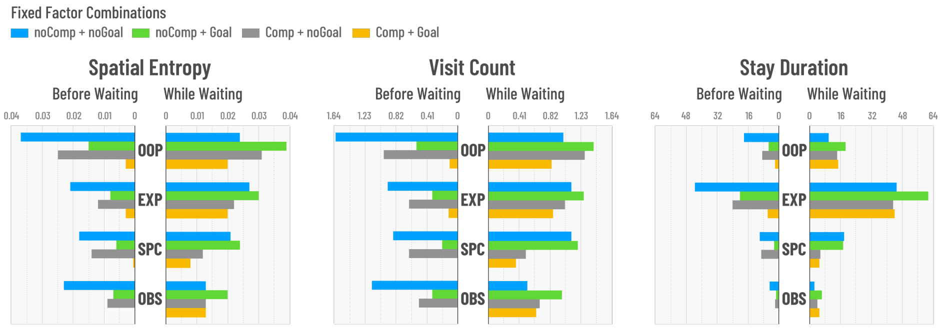


Figure 7.3: Graphs showing mean measures of spatial entropy (left), visit count (middle), and stay duration (right) of the four PIR sets: *Out-of-place Elements* (OOP), *Extreme Points* (EXP), *Spatial Connections* (SPC), and *Visual Obstructions* (OBS). For each of the three graphs, bars extend left to illustrate measures "before waiting" and right for "while waiting". Color coding of the bars indicates the combination of compensation and goal condition for a measure.

Bayesian repeated measures ANOVA tests across PIR sets show decisive evidence (***BF>1k***) for differences in PIR sets beyond the impact of subject factors (goal statement, environment aesthetic, or compensation). Figure 7.3 shows visual graphs of PIR set measures for different fixed factor combinations, as well as for before and while waiting. ANOVA tests show strong evidence that the "environment aesthetic" factor has no effect on these measures and is thus not included in the figure.

Table 7.4: PIR Sets repeated measures ANOVA results: Best models of Bayesian ANOVA for dependent variables, if model Bayes factor $BF_M > 3$. The top half of the table lists results for 'before waiting', the lower half for 'while waiting'. Individual fixed factors in the model are sorted in descending order by probability of inclusion in the model (BF_{incl}, see supplementary material for values). Post Hoc T-Test results (BF_{10}) are included as superscripts if at least $BF_{10} > 3$, reported in steps: $>$3, $>$10, $>$100, $>$1k (1000). Interaction effects (e.g. [Goal $\times$ Pat]) do not have an associated BF_{10}. Post Hoc comparison for individual PIR sets are listed in the last column and sorted by means. Fixed factor abbreviations are: Pattern (Pat), Compensation (Comp), and Environment (Env).

Dependent Variable	Best Model (in order of inclusion probability)	BF_M	PIRs in order of means
Before Waiting			
PIR Spatial Entropy	PIRs + Goal^{1k} + Comp^{1k} + [PIRs $\times$ Goal]	40	OOP $>$ EXP = OBS = SPC
Stay Duration	PIRs + Goal^{1k} + [PIRs $\times$ Goal] + Comp^{10} + [PIRs $\times$ Comp]	115	EXP $>$ OOP $\approx$ SPC $>$ OBS
Visit Counts	Goal^{1k} + PIRs + Comp^{1k}	27	OOP $>$ EXP = OBS = SPC
While Waiting			
PIR Spatial Entropy	PIRs + Comp^{1k}	41	OOP $\approx$ EXP $>$ OBS = SPC
Stay Duration	PIRs	18	EXP $>$ OOP $\approx$ SPC $>$ OBS
Visit Counts	PIRs + Comp^{100} + [PIRs $\times$ Env] + [PIRs $\times$ Comp] + Env	27	OOP = EXP > OBS = SPC

Table 7.4 shows the ANOVA results, with the last column indicating the order of means of individual PIR sets. In some cases, PIR sets are statistically equal (e.g., PIR spatial entropy before waiting: OOP is higher, but EXP, OBS, and SPC can be considered equal), even if differences seem to exist according to the means (as shown in figure 7.3). In such cases, between-subject factors such as goal and compensation are a likely cause, according to the best model. For the aforementioned example, goal, compensation, as well as an interaction effect between PIR sets and goal statement impact the measure. For an interpretation of differences between PIR sets regardless of other factors, "PIRs Post Hoc comparisons" in table 7.4 display the most probable results.

It should be noted that patterns do not perform uniformly. Therefore, the results of PIR sets should be understood as including some performance bias by individual pattern implementations. Table 7.5 lists stay durations and visit counts for individual PIRs of the four PIR sets. *Ground Fog* stands out as having been visited three times more often

Table 7.5: Visit counts and overall stay duration for individual PIRs, sorted by PIR sets and in order of mean visit count across participants. Visits are individual instances of at least > 1 sec time spent in a PIR. Stays are listed in percent of a player's session length.

Design pattern	visited by (n)	Mn_{visit}	SD_{visit}	$Mn_{\%stay}$	$SD_{\%stay}$
(OOP) Stone Stack A	42.9% (79)	1.5	0.8	1.74	1.99
(OOP) Stone Stack B	44.6% (82)	1.4	0.7	1.45	2.03
(OOP) Stone Spiral	45.7% (84)	1.3	0.6	1.76	1.80
(EXP) Mountain B	48.4% (89)	1.3	0.6	7.08	4.10
(EXP) Mountain C	44.0% (81)	1.2	0.5	4.94	4.71
(EXP) Mountain A	31.0% (57)	1.1	0.3	3.81	3.06
(SPC) Cliff Cave	28.8% (53)	1.6	0.8	2.83	2.01
(SPC) Hill Path	29.3% (54)	1.2	0.5	1.14	0.91
(SPC) Mountain Cave	24.5% (45)	1.1	0.3	1.29	1.57
(OBS) Ground Fog	49.5% (91)	1.4	0.6	0.73	0.60
(OBS) Forest B	18.5% (34)	1.2	0.5	0.79	0.90
(OBS) Forest A	19.6% (36)	1.0	0.2	0.65	0.61

than other patterns in OBS. *Mountain A* stands out in EXP for fewer visits and shorter stay durations.

7.3.4 Notable Correlations

Bayesian Pearson correlations were calculated to provide context for measures for which a statistical correlation was not assured (e.g., path distance and spatial entropy are logically correlated). PIR spatial entropy decisively correlates with PIR visits and stay durations (**BF>1k**) and is thus used to assess correlations between PIR sets and other metrics.

7.3.4.1 Before waiting: PIR sets

All PIR sets correlate positively with *Glasgow* SDs of all emotion dimensions (**BF>1k**, highest correlation for OOP) but not with *Glasgow* means. They positively correlate with participants going out-of-bounds (i.e., jumping into the chasm) for OOP, SPC, and OBS (all **BF>1k**, highest correlation for OOP), but not EXP. Camera movements positively correlate with all PIR sets entropies (**BF>1k** for all but SPC, which had **BF>30**); here, the highest correlation was with EXP. PIR sets do not correlate with game frequency or player type.

7.3.4.2 While waiting: PIR sets

Positive correlation with game frequency and player type for EXP (***BF>30***), OBS (***BF>30***), and SPC (***BF>3***); but not OOP. *Glasgow* measures do not correlate, except for EXP and arousal SD (***BF>3***). Positive correlation with out-of-bounds events for OBS (***BF>3***) and EXP (***BF>1k***). Camera movements positively correlate for EXP and SPC (both ***BF>1k***).

7.3.4.3 Non-correlations

Some correlation results are notable for their *lack* of correlation with other measures. 5DC measures showed evidence for a lack of correlation with game metrics, in-game curiosity ratings, or Glasgow ratings. The only exception is the *Thrill Seeking* dimension, which correlated with variation in camera rotation before waiting (***BF>10***). PIR set measures did not correlate with age, GUESS ratings, or in-game curiosity ratings.

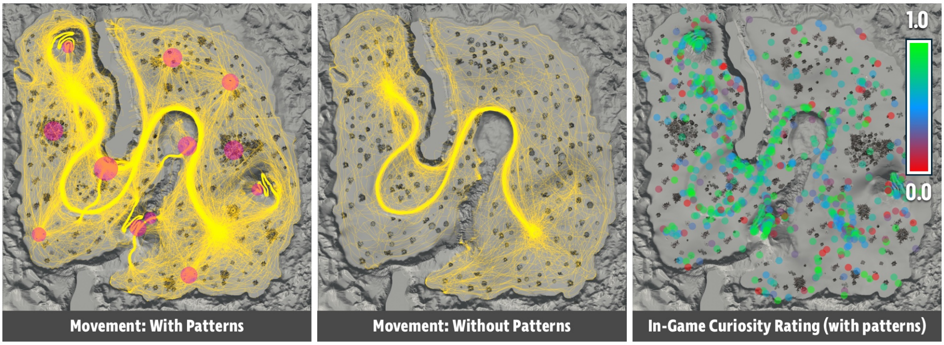


Figure 7.4: Player movement paths for the "with patterns" (left, includes pattern locations) and "without patterns" (middle) experiment factors; and distribution of in-game curiosity ratings (right).

7.3.5 Nuisance Variables

Play direction was recorded as a nuisance variable but evaluated through a *Bayesian Student T-Test* to assess its impact on measures. Most measures did not differ by play direction. Of note are path distance mean (***BF>100***) and SD (***BF>1k***), both especially "while waiting" (***BF≈1k***), and differences in the OOP "while waiting" measures visit count (***BF>1k***), stay duration (***BF>3***), and PIR spatial entropy (***BF>1k***). Here, the proximity of *Stone Stack A* to Zone A likely provided a more potent attractor than an equivalent OOP pattern for players heading toward Zone B.

Participant gender was reduced to female and male participants to assess the impact on measures through a T-Test. Measures impacted by gender differences are likely affected by a difference in “gamer type” (***BF>1k***) and “play frequency” (***BF>3***), with female players having lower measures in both due to differences in sampling distributions. GUESS measures Engrossment, Enjoyment, and Personal Gratification differ (all ***BF>30***, all higher for female participants). In-game curiosity ratings are increased for female players (***BF>100***), as are Glasgow ratings for arousal (***BF>3***) and valence (***BF>3***) while waiting. Female players had shorter (***BF>30***) and fewer (***BF>10***) visits to SPC PIRs while waiting, as well as fewer visits (***BF>3***) to EXP PIRs while waiting. Finally, female players moved slower (***BF>30***) and had fewer camera movements (***BF>100***).

7.4 Discussion

The primary goal of this study is to examine the effect of level design patterns for spatial exploration on player behavior and experience. Across multiple measures, results show that the presence of patterns indeed influenced how players interacted with the environment, and that patterns had an emotional impact. However, how exactly players were influenced depended on other factors. While results suggest that environment aesthetic has little to no impact, having a goal and being compensated evidently affects exploration. In some circumstances, the relative impact indeed exceeds that of pattern presence.

It is crucial to discuss player behavior over two phases of the game: before waiting for the master, and while waiting for the master to stop meditating. It can be hypothesized that a player’s motivation for exploration and emotional experience shifted at this point, although how exactly depends on the condition group. As such, the following sections discuss the differences between these two game phases.

Two nuisance variables were assessed in the study: play direction and gender. Play direction impacted player behavior, as players generally seemed to explore more while waiting when they ended in Zone A. Although the map was designed to be reversible, Zone A has an *Out-of-Place* element nearby, whereas Zone B features a *Visual Obstruction* element. Based on the popularity of patterns from this set, *Stone Stack A* likely provided a stronger motivation for exploration, whereas Zone B had fewer interesting PIRs in close range. Aside from a preference for visiting *Stone Stack A*, however, the overall impact appears limited. In addition to play direction, the demographic of players also had some impact. Female players generally had a higher emotional investment in the game and scored their game experience higher (GUESS). Neither of these impacted the results in a way that made them specific to players based on gender.

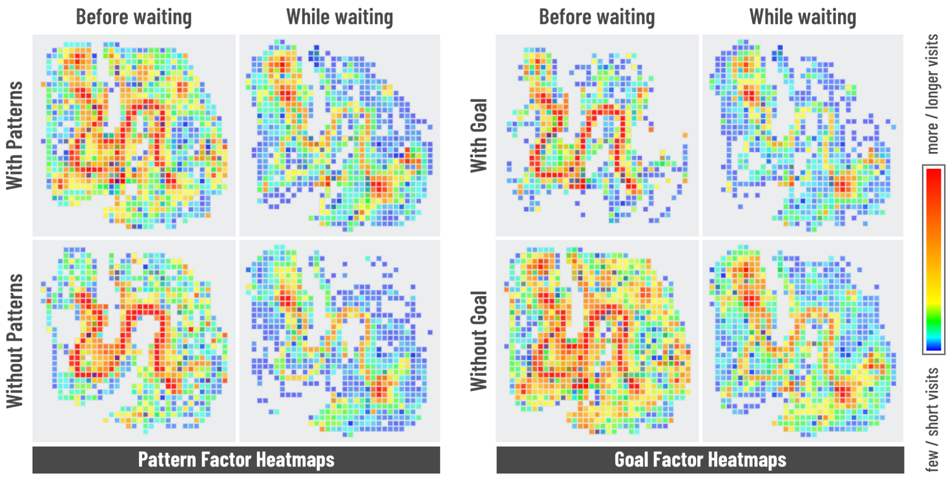


Figure 7.5: Heat maps of player presence split out by the experimental factors of goal statement and pattern presence.

7.4.1 Impact of Patterns

In general, level design patterns caused participants to venture further away from the path and further away from their destination (i.e., the master), resulting in movement across the environment that was overall more dispersed. These differences in exploratory behavior are visible in the visualizations of player presence (see figure 7.4 and figure 7.5) and confirmed by the statistical analysis. Although these results come with certain caveats (discussed in the following sections), patterns affected exploratory behavior. As such, *H1a is supported by the results of the study*. The impact of individual patterns is discussed in more detail later in this section.

In addition to affecting behavior, it was hypothesized that the patterns would also positively affect participants' emotional experiences. This was measured in two ways: through the in-game capture of emotion words and the post-game measures. However, the GUESS showed little to no differences between conditions. This is not necessarily a fault of the GUESS but possibly a side effect of the study design. Due to the decision to make participants wait for five minutes in an attempt to gather data under different circumstances, players likely grew bored. This is supported by *Glasgow* ratings of the in-game reported emotion words, which saw a decrease in valence, dominance, and arousal while waiting compared to before waiting. This suggests that the last five minutes could have colored participants' overall experience, eliminating any differences the GUESS might have uncovered. However, there is a difference in comments that participants made: those who played with patterns were more likely to comment positively about the game overall.

The results also show an impact of patterns on the in-game *Glasgow* ratings derived from emotion words. This effect, however, is primarily visible in the spread (i.e., standard deviation) of the ratings. Before waiting, *Glasgow* means are not affected. On the other hand, the impacts on spread were primarily due to the goal condition and only to a lesser extent due to the presence of patterns. While waiting, patterns had a small impact on *Glasgow* means, which happened in combination with receiving assured compensation. Patterns also impacted the spread of dominance and valence in combination with the goal condition.

Overall, the results suggest that patterns impacted the range of emotions expressed through the in-game ratings, with the emotional range increasing in the presence of patterns. Before waiting, a lack of patterns did not necessarily result in a more subdued emotional experience. Participants that cared to explore did so, while those that did not were driven by finding out what they were supposed to do, rather than being negatively impacted by a lack of patterns. However, the presence of patterns likely led some participants to explore or at least made an emotional impact.

While waiting, the presence of patterns is responsible for higher arousal and valence. This happens in combination with compensation, suggesting that participants that rushed to the stated goal (more likely driven by extrinsic, financial motivation) found elements that interested them while waiting as compared to those playing without patterns. It is unclear, however, whether the impact was due to exploratory curiosity or the desire to alleviate boredom. Participant comments suggest a mix of motivations, e.g., "left path to explore" (55.9%), because of "having to wait" (31.1%), to "explore a landmark" (17.7%), or "due to boredom" (12.6%). The spread of emotional values was increased due to patterns while waiting, suggesting more highs and lows in their play experience.

Having patterns elicited more comments from participants, suggesting that participants with patterns felt more strongly about their experience and the effort they put into the study. Although they more commonly commented on leaving the path to explore, they were also more likely to mention the lack of interaction in the game. It can be hypothesized that the presence of patterns created expectations. The relative novelty of encountering PIRs was likely not perceived as a reward in itself. This is why the emotional impact was more tied to fluctuations rather than an overall increase. Players probably enjoyed the moments when they were engaging in exploration but may have experienced disappointment when their efforts went unrecognized by the game system (either through a reward or by encountering actual interactive content).

Based on these findings, there is sufficient *statistical support for H1b*, albeit with some caveats. The presence of PIRs alone is not sufficient for increasing emotional investment. *Instead, they afford a possibility for exploration that, when realized, increases emotional investment.*

7.4.2 Impact of Goal Statement

Whether or not players were given a goal had a substantial impact on the effects of patterns. Before waiting for the master, among participants without a goal, those with patterns ventured further from the path than those without. They were also further away from the destination point (the master), i.e., they moved around more as they closed the distance to the destination. Entropy measures confirm this observation.

These differences were severely reduced, however, when a goal was introduced. When given a goal, the presence of patterns had a considerably lower impact on any of the measures. Although it was hypothesized that having a goal would reduce exploration, the magnitude of the difference is surprising. The goal formulation is not very specific and relatively subtle; it only informs the player that they are in search of their master and that he awaits them. Additionally, signs along the path point in the master's direction. The path itself, however, is present in all conditions. It can be argued that the path always hints at a goal, as it is reasonable to assume that a path leads to somewhere of interest. Thus it is not unreasonable to suspect they would follow the path as one of their first actions.

Nevertheless, participants not provided with a goal likely spent the first part of the game, at least in part, in search of what to do. Overall, in the presence of patterns, participants without a goal were still more likely to leave the path behind and explore the environment's boundaries than those having no patterns present. A possibility is that PIRs are of interest to participants irrespective of whether or not a goal is provided, but that the difference is masked by participants without a goal figuring out what to do.

While waiting, the data indicates a different behavior. Generally, participants with patterns exhibited more exploratory behavior than those without. Particularly interesting, however, is that participants with a goal explored more than those without. This is perhaps because participants with a goal explored less before finding the master. They were focused on accomplishing their goal when the game started. Once the game provided them with a new goal (i.e., waiting for the master to finish meditating), they felt free to explore. At this point, participants with patterns explored more than those without. Participants that had already explored before because they did not have a goal, did not feel the need to explore as much once they had to wait. Based on these findings, *H2 is considered to be supported by the results.*

When considering emotional impact, players without a goal had higher fluctuations in their emotional experience. However, this is only true before waiting for the master. While waiting, the opposite is true: player affect fluctuated more when an explicit goal was given. Being uncertain about the game's goal likely creates more potential for emotional investment, as players take it upon themselves to find out what the game is about. As the other data suggests as well, participants with a goal were likely more focused on achieving it. Once they are asked to wait, they are presented with a new situation.

Meanwhile, players with no goal were "given" one while waiting. While fluctuations in emotional investment do not necessarily indicate that players enjoy the experience more overall, the concept of designing for interest curves or experiential fluctuations in a game (e.g., Schell 2008) is understood as a strategy to increase overall interest. The results show that having a goal had little impact on the player affect means.

Despite the findings, it is not necessarily the case that a goal reduces exploration in all situations. In part, the results may be due to the nature of the experiment. While players may prioritize a goal or "quest" when playing an entertainment game, it is also not unlikely for them to abandon it in favor of freely exploring an (open) environment when presented with one — likely, this depends on the player and their play style. There is a possibility that, due to participants knowing they were participating in a study, they expected to receive instructions on what they were supposed to accomplish. Those who received a goal prioritized it, thinking it would be necessary to complete it to finish the study successfully. Those who did not receive a goal could have been motivated by finding out what they should do.

As such, it is possible that participants' motivation to explore was not strictly one of curiosity but impacted by seeking a purpose in the context of the study. However, if this were their only motivation, one would expect players to remain on the path, as it gave an implicit indication of where such a purpose might be found. Since this is not the case, it is plausible that patterns still had an impact and that players were not solely motivated by their desire to complete the experiment.

7.4.3 Impact of Compensation

As stated before, it is possible that the addition of an extrinsic reward, i.e., monetary compensation, could influence intrinsically motivated exploration. In addition to players looking for what they were supposed to do to finish the study, the addition of an assured monetary reward could have motivated players to finish the study as quickly as possible. If this were the case, such participants would be expected to put in minimal effort, spending as little time as possible and exploring only to a minimal extent.

Before waiting, whether or not people were compensated had limited effect. Some differences can be seen in how far participants ventured from the path, although patterns and goal statement had a more substantial influence. While waiting, however, compensation was more likely to influence measured behavior. It was the most likely measure for several dependent variables, including spatial entropy, distance from the path, and distance from the destination. Participants were also less likely to see what would happen if they jumped into the chasm. Overall, this shows that participants who were assured compensation were less likely to leave their destination once they had to wait, suggesting they wanted to get to the end as quickly as possible.

Regarding emotional experience, compensation had little influence. However, it was the most significant contributor to differences in valence means while waiting. Interestingly, valence was higher on average for participants who were assured compensation (note that this is the only measure that was significantly higher among people who were being compensated). This suggests that participants who were assured compensation were more content knowing they would progress the experiment within a particular time. Likely, those who were not assured compensation (especially those without a goal who had already spent more time exploring) were more annoyed at “being made” to wait. The presence of patterns, in turn, mitigated this somewhat.

7.4.4 Analysis of Patterns

Besides examining the overall impacts of patterns, the study aims to investigate the impacts of the individual patterns themselves. To this end, only the data from participants with patterns in their game environment was analyzed in detail. This was also the rationale for randomizing with a 70% chance of having patterns.

Before waiting, participants without a goal or compensation interacted with patterns the most; they tended to visit PIRs and stayed for a while. On the other hand, participants who had both a goal and were being compensated barely visited any PIRs and did not stay long at those they visited. This suggests that they were trying to get to the master as quickly as possible. Participants without a goal and with compensation visited more PIRs and stayed longer than those with a goal and without compensation. It appears that participants without a goal searched for one, while the compensation drove them forward. Those who were not being compensated took more time to explore the PIRs, even if they had a goal statement.

While waiting, participants are more active in exploring design patterns. Overall, participants visited more PIRs and stayed longer while waiting. Since they knew they had time to spend at this point, they took more time to look around. Participants with a goal who were not compensated visited the most PIRs and stayed the longest. This is in line with previously discussed findings that those with a goal spent less time getting to the master and then used the waiting time to explore. Similarly, exploration of patterns remained lowest among participants with a goal who were compensated — although they were also likely to get to the master early, they were less willing to venture farther away again to explore patterns in the environment.

7.4.4.1 Out of Place (OOP)

PIRs of this pattern attracted the most visitors and with relatively short visits. Spatial entropy, however, was generally high. This suggests that the pattern causes local exploration, where an object of interest is examined from multiple angles. The reason could be to examine the visual qualities of the object itself or to find out what function it could serve. In many games, OOP elements guide players to rewards (e.g., collectibles and upgrades for the player character), interaction opportunities, or game narrative progress. In such cases, finding out what can be done at these elements is a game mechanic in itself (e.g., requiring players to perform specific actions to make progress). Continued engagement in these regions could indicate the desire to gather more clues about these elements. Furthermore, all patterns except OOP were more attractive to players with higher game experience. This could suggest that PIRs of this pattern have universal appeal, regardless of gaming experience, while others attract only more experienced players.

7.4.4.2 Extreme Points (EXP)

EXP PIRs were visited by many players and caused them to stay longer than other patterns. One reason for this could be that players use the higher vantage points to get a lay of the land. It allows players to visually explore the environment, i.e., gain an understanding of it without having to travel there. This is perhaps especially the case for players without a goal, who stayed the longest at EXP PIRs, and before waiting. While waiting, stay durations are relatively similar regardless of fixed factors. It could be that reaching these patterns is challenging, and, as a result, succeeding in that challenge can feel rewarding in itself, causing players to take a moment to appreciate the result of their effort. The EXP pattern correlating with the spread of arousal ratings suggests that interaction with the pattern has moments of varying excitement. Reaching a high place also allows players to see their surroundings from a new and interesting perspective. As such, participants may stay there for aesthetic reasons (i.e., to enjoy the view), as is indicated by a correlation with increased camera rotation. Even if the game does not provide any specific interactivity, surveying the environment from a vantage point can be experienced as an engaging activity. Out of all patterns, EXP PIRs were most often mentioned by participants as reasons to leave the path or as noteworthy features.

7.4.4.3 Spatial Connections (SPC)

PIRs in this pattern were not visited as often as EXP or OOP patterns, and participants stayed for short periods. However, they stayed longer at SPC PIRs than at OBS PIRs. The *Cliff Cave* had the most prolonged stay duration. Compared to the others, it offered a unique vantage point of the environment (i.e., from within the chasm). It is possible that providing an exciting view contributes to the appreciation of a PIR, as indicated by participants looking around more — neither of the other SPC PIRs provided a "better" view than the mountains. It is possible that understanding how spaces connect can offer an intrinsic reward similar to that of exploring EXPs. Out of all PIRs, caves were commented on fairly frequently (nearly as much as mountains). This suggests that they stood out to participants and left an impression. However, this did not translate into as many visits. It could be that fewer participants noticed the SPC PIRs.

Additionally, because their entrances are more hidden (while mountains were readily apparent), the effort to figure them out was perhaps too demanding for some. Games tend to implement this pattern not as a way to guide players but to present them with a challenge that often involves a reward. The fact that players looked around more when at these PIRs may also suggest they were looking for something. Since there was nothing to find (a fact commented on often by participants), it is possible players did not feel the need to bother with figuring out these PIRs once the lack of reward had become apparent.

7.4.4.4 Visual Obstructions (OBS)

OBS PIRs were visited the least of all patterns (with one exception), and players stayed the shortest. Considering the results of the spatial entropy measure, it would also appear that players primarily ran through these PIRs, possibly without even registering them. *Ground Fog* stands out as an exception, as it was visited by more players than any other individual PIR. It was also relatively often commented on by participants after the game. It is conceivable that the two dense forests were not understood as places for potential exploration but rather as natural boundaries. Of course, it is also possible that the potential was understood but not sufficiently appealing. A challenge in successfully implementing this pattern stems from the fact that visual obstructions must still appear surmountable. Games often enclose the interactive space with environmental obstructions that communicate to players that they cannot be overcome (which is also the case for the game in this study). OBS patterns can easily be misread as areas that cannot be explored. As a result, games tend to implement this pattern more often for secrets (i.e., additional content designed to be encountered by a small selection of players) and not to guide player progress. This would explain why this pattern was more likely to be explored by players with higher game experience. Based on having played other games, they were possibly better equipped to recognize the PIRs as potentially interesting. Alternatively, it could also be that less experienced players avoided the possibility of danger as their vision was obscured.

7.4.5 Impact of Trait Curiosity

As stated in *H3*, there was a possibility that a general disposition for curiosity (i.e., trait curiosity) would impact exploratory behavior. The study results show that curiosity dimensions did not correlate with exploratory behavior measures (except for camera rotation with *Thrill Seeking*) or emotional experience. Whether or not a player experienced curiosity for exploration in the game did not seem to be impacted by their general disposition. It is possible that the threshold for engaging in exploratory behavior in a game like this is relatively low or that measures of trait curiosity in the physical world do not correspond to game environments. Based on these results, *H3 must be rejected.*

7.4.6 Measures of Exploration

Measures in this study involved validated psychometric instruments (i.e., GUESS, 5DC), game metrics, and an exploratory in-game measure of curiosity that has yet to prove its viability in further game user research studies. Especially the use of in-the-moment measures of emotional states through a combination of a curiosity scale and interpreting the affect of emotion words has, to the author's knowledge, not been described to measure exploration in video games before. The study results suggest that in-game curiosity ratings correlated with how players assess their game experience. However, these ratings also showed a wide variance across players, suggesting that unexplored factors contribute to being in a state of curiosity. One weakness of the measure is that it was gathered at fixed points in time instead of taking measures at specific player actions or locations in the game.

The interpretation of emotion words provided a more substantial basis for evaluating affective states that accompany exploration than the curiosity scale. It can be hypothesized that curiosity in games may be a short state that is more difficult to self-assess than the use of emotion words. Ultimately, measuring curiosity remains a challenging task — it is either determined by indirect measures or measured by interventions that can impact the measure by their mere presence. Nevertheless, the results of this study provide evidence that a combination of behavioral and affective measures can present insights into curiosity for exploration.

7.5 Limitations

The results of any study should be considered within the limitations of its design. As stated previously, games are complex systems with many interacting elements. With the design of *Shinobi Valley*, one of the goals was to strike a balance between controlling for confounding factors while still giving players the experience of playing an entertainment game. However, this meant that the game lacked many typical characteristics of games featuring spatial exploration. Level design patterns are not usually used in isolation. Instead, they guide players toward specific objects, objectives, or interaction possibilities. As such, they are likely to raise expectations in players that their exploratory behavior will somehow be acknowledged or rewarded. This is shown in the study's qualitative data, which suggests that many participants expected to find something as the result of their exploration. While exploratory behavior probably was not impacted (due to the short duration of the game), emotional investment likely was because the game did not provide the satisfaction that entertainment games do. Since this study intended to investigate curiosity-driven exploration and not specifically to entertain players, this is considered acceptable. However, future studies should carefully consider implementing rewards (if it benefits their purposes), especially when player entertainment is the focus of the study.

Concerning this, there is also the matter of distinguishing between assured and potential compensation and considering potential compensation by way of a random draw as being different. Although there was a significant difference between the two forms of compensation, the possibility cannot be excluded that a *potential* extrinsic reward carried an effect of its own.

Another limitation that should be noted is that the patterns tested in this study were formulated based on the design of 3D, open-world games. As such, the results cannot be generalized to all types of games. However, the patterns can likely be implemented in different virtual environments (e.g., smaller, "closed" levels) and other game forms. It is, for example, possible to have hard-to-reach places or out-of-place objects in a 2D game. Their implementation will require careful thought, however, and whether or not they are experienced similarly as they were in this study remains to be investigated.

Even a game as simple as *Shinobi Valley* adds layers of complexity to empirical assessment. The decision to include additional variables, rather than only focusing on the presence or absence of level design patterns, complicated the study considerably. Testing the game with only one independent variable would have been easier but would also miss essential findings related to interaction effects with, for example, a stated goal. As discussed previously, level design patterns do not occur in a vacuum. Therefore, it was important to include at least some aspects integral to many games rather than draw conclusions from a very narrowly designed experiment. However, this makes the gathered results more complex to interpret, a challenge inherent to disentangling player experience. Future studies should be designed with an awareness of this challenge.

Another decision that could have had a negative impact was to have more participants play with patterns present to increase the sample size for that condition. Naturally, this skewed the sample sizes of the condition groups. It is generally recommended that the groups be of similar size when conducting ANOVA tests between conditions. Instead, some condition groups had relatively few participants. For example, the smallest combined condition group (no pattern, goal, nature, no compensation) had 5 participants, while the largest group (pattern, goal, alien, compensation) had 32 (in general: $Mn=16$, $SD=8.9$). Because of this range, any interaction effects detected in the data lend themselves to further study with more participants and should be considered in future work.

Another consideration in the study design was the decision to make participants wait after encountering the master. Although this decision provided interesting findings relating to how the stated goal and level design patterns interacted, it cannot be said that exploration while waiting happened solely due to curiosity and was not, at least in part, motivated by boredom. As stated previously, curiosity is challenging to capture. Although the study uncovered interesting results, it cannot fully interpret the motivations of players. Likely, they are multi-dimensional. More thorough in-game measures (e.g., observation, questions, think-aloud protocol) might assist with unpacking the player experience further. However, a balance must always be struck between an experiment's invasiveness and a measure's thoroughness.

Finally, the GUESS questionnaire was only carried out after participants completed the game. Players filled out the survey after possibly spending the final minutes of the game annoyed or bored. As a result, the GUESS was likely influenced by these final minutes and not particularly useful in assessing differences between conditions. Although the in-game measures offset this somewhat, these do not assess game experience as thoroughly as the GUESS. While making players wait provided additional opportunities for collecting data, the use of any post-game questionnaire should be considered carefully if a similar design is used in future studies. Even if the study design is different, there is an inherent challenge in relying on post-game measures when capturing temporary states, such as curiosity.

7.6 Conclusion

This study uncovered empirical evidence for level design patterns eliciting curiosity-driven exploration in players. The impact was affected by an explicit goal statement and whether assured compensation was provided. In the absence of such design patterns, players engaged in less exploration and formed fewer expectations about being rewarded for doing so. Involving a goal statement strongly impacted players' likelihood of engaging in exploration. Participants were most engaged in curiosity-driven exploration when patterns in the environment provided opportunities, and the game's goal was left sufficiently ambiguous to pay attention to the larger environment. Players' comments showed that game exploration is understood as a mechanic in itself. It can be concluded that the dramatic principle of *Chekhov's Gun* in literature (Rayfield 1999) also applies to elements that invoke exploration: if something promises to be an interesting area for exploration, it should provide acknowledgment to players when they do so. Failing to do so results in a negative emotional response. This is already a common practice in game development, and the exclusion of rewards in this study was primarily motivated by reducing confounding experiment variables.

Based on the experiment results, whether or not players explore was not impacted by their general disposition for curiosity. This could mean that the threshold for developing curiosity was not very high in the game experiment or that general disposition is not a strong predictor for a curious state in a video game. The study provided evidence that exploration motivated by boredom differs from curiosity-driven exploration. Additionally, curiosity-driven exploration can have different motivations, such as looking for rewards, interaction possibilities, or testing the environment's boundaries. Design patterns can only increase the likelihood of curiosity, not enforce it. By controlling for environmental aesthetics, the findings of this study should apply not just to this specific implementation but to other game environments as well. Finally, whether or not participants were compensated also affected exploratory behavior and should be considered, especially in study designs looking into player behavior and gameplay experience.

With this study, there is now empirical evidence for the efficacy of a design practice that is already common in video games. Based on the analysis of in- and post-game measures, the study provides a conceptual framework for understanding the impact of individual patterns and mapping their efficacy in light of related factors, such as having a stated goal or behavior during waiting time. Through an experimental design that incorporates and evaluates various elements common to games, the study illustrates many complexities that result from the interaction between such elements. Although this results in a more complex research narrative, it is an account that can inform future empirical studies of player experience and shows the need to explore variables whose influence may otherwise be ignored. This does not mean that every game research should incorporate a variable on whether or not players have an explicit goal.

Instead, designers and researchers ought to consider what effects their decisions might have on players, especially when studying player experience and complex emotional constructs.

Future work may expand on the lexicon of design patterns that invoke curiosity for exploration. Given the vast design space of creating game worlds, more data from different implementations of the discussed patterns is needed to support or scrutinize this study's findings. The promise of engaging in this work is a better theoretical understanding of how to design for curiosity-driven exploration intentionally. In time, such work can also support efforts for better procedural creation of video game content or even real-world implementations of explorable architecture, such as the design of playgrounds or amusement parks.