Games User Experience - An experimental playtest-study with two external studios

Fabian Fröding
University of Skövde
Skövde, Sweden
fabian.froding@gmail.com

***Abstract*—This report describes and reflects a collaboration between a group of students and two game studios. The collaboration consisted of helping the studios test their games.**

The study resulted in a successful collaboration with one of the studios, while the collaboration with the other studios ended due to communication problems. The methodology created by the students introduced useful methods that can be re-used for future projects.

Concepts and methods from existing research in playtesting were also described and compared to the methodology used in this study.

Keywords—video games, user experience, playtesting.

I. Introduction

This project consisted of conducting playtest sessions to help two studios, namely YCJY and Warcry Interactive (referred to as *Warcry* for the reminder of this report), to test their games.

YCJY asked us to test a game under NDA (Non-disclosure agreement). Because of the NDA, our group served as internal testers for YCJY and provided them with subjective feedback focused on the UI (User-interface) of their game, and were therefore not allowed to recruit external test-participants.

Warcry asked us to test their FPS-game (first person shooter) "Only Lead Can Stop Them" (referred to as *"Lead"* for the remainder of this report). Specifically, Warcry wanted to evaluate these specific areas of their game: *weapons* and *enemies,* and later on also *UI* and *controls*.

II. Study Design & Execution

After our initial meetings with the studios, we had an internal meeting where we planned how to lay out the remaining time of the study period. Our main form of communication happened over Discord, and we used Trello to manage tasks and set internal deadlines. We agreed on a start and end-date for recruiting participants and conducting the experiments. After this period was over, we were not allowed to conduct any more experiments and focused on combining and analyzing the data.

To evaluate the areas that Warcry wanted to test, we designed an observational-based experiment consisting of three phases, *pre-play session interview, play session* and *post-play session interview.*

The pre-play session interview included questions about the test-participant to gather demographic data such as age and occupation. In the play session we let the participant play the game with an aim of a 20-minute think-aloud session. The play sessions were recorded for the purpose of extracting data after the experiments by re-watching the recording. The post-play session included questions relating to the enemies and weapons of the game to gain some qualitative data. The reason for having these specific questions after the play session was to avoid having the participants focusing just on those aspects of the game while playing. We mitigated the risk of inconsistencies in the experiments by relying on a collective document with instructions on how to conduct the experiment. Finally, the data gathered from the experiments were imported into a shared data document in which we also produced data visualization graphs to present to Warcry.

Furthermore, our group decided to draw heat maps based on player death locations. On our request, Warcry provided us with top-down screenshots of the levels in the game. We would then look at the recording to mark down locations where the test-participants died. These heatmaps would then be combined by a chosen group member to draw a combined heatmap for each build and level.

Halfway through the study, a suggestion was made to categorize the participants into two player categories: casual and experienced. Warcry liked the idea, so we decided to go through with it. We relied on the post-play session interview questions related to the FPS-genre to categorize the participants into the appropriate category (see Appendix for details).

Because of the NDA that our group had with YCJY, we did not design a study for their game. Instead, we tested their game in weekly iterations in which we evaluated existing UI and looked at how they could draw inspiration from UI in other games. We wrote our thoughts and feedback on an online document shared with the developers. Since we acted as testers for all of the iterations, we stated to the developers that our feedback would become biased in the sense that we would eventually become "expert"-testers for their game, in which the developers responded that that is what they wanted.

The experiments for Warcry were conducted individually by our group members to maximize the number of participants recruited. If we were to do the experiments as a group, the scheduling for these experiments would be less flexible, thus requiring more time and planning. In total we had 15 participants. Some tests were done in the tester's home, some in our own homes and some online.

Personally, I managed to execute my experiments without any interruptions or problems. In addition, as project manager I received no reports from my fellow group members about any difficulties during the testing phase.

As a side-effect of watching the recordings, our group observed several in-game bugs. Even though Warcry did not explicitly ask for reporting bugs, most of our group members noted down these bugs and reported them during the meetings with Warcry in an iterative manner, which was appreciated by the developers.

Because of the case with YCJY, most of the remainder of this report will focus on describing the results from the experiment conducted for Warcry.

III. Data and Results

By watching the recordings from the play sessions, we counted certain events related to enemies and weapons:

- Number of player-kills by each enemy type.
- Using time (in seconds) for each weapon in the game.
- Put markers for each death location of the player on a map of the current level.

Originally, we had planned to also track the idle time of the test-participants. However, it became too difficult to distinguish exactly which behavior counted as idle and not in combination with tracking the time between idle and not idle as players switched back and forth in the matter of seconds, so this idea was discarded.

The using time for each weapon was not measured in exact seconds, since the test-participants regularly switched weapons back and forth for just a few seconds. Instead, we measured the weapon using time when the participants were using a weapon for at least 10 seconds or more. This duration was then rounded to the closest 10th-number. This procedure might not have been exactly the same for each group member.

The process of re-watching the recordings to extract the required data was a time-consuming and tedious process. Early during the study, one group member created a script that would run in parallel with the game and automatically count the events. The group member sent the script to the developers who then integrated the script to their source code. However, due to the changes in the different builds we received from the developers, the script was rendered useless for most of the study. This obstacle did not greatly impair our work process, since it was mostly an alternative to save time.

Most of the data between the casual and experienced players looked almost the same. One notable exception was that experienced players were more prone to die from self-kills, such as using weapons that could harm the player if fired too close to nearby objects. This could be because the experienced players were less careful (see Figure 5 and 6 in the Appendix). Another notable difference was that experienced players used the "Rifle"-weapon much more than casual players (see Figure 8 and 9 in the Appendix). Furthermore, the enemy type "Soldier" killed the players far more often than the rest of the enemy types as can be seen in Figure 4 in the Appendix. The reasons for these results might have been several. The placing and availability of weapons and enemies as well as the proportion of different enemies might have factored into the results. Ultimately, the developers were free to analyze the results and visualizations as they wanted, since they had more internal knowledge of the intention behind these areas of the game.

IV. Discussion

In my personal opinion, the most important part of the project was to gather and present the quantitative data to Warcry. This opinion was not shared with all group members; some insisted that the qualitative feedback was more important. This was not a problem since we all agreed to present both types of data. In the end, Warcry showed interest in all areas of the final presentation.

Our group particularly excelled at planning and executing the experiments for Warcry. However, we could have put more effort into our communication with YCJY, maybe then the communication would not have been cut. However, we all agreed that it was better to focus on one studio and do the job well, rather than focus on both and do the jobs half-heartedly. This decision was also based on the fact that we did not hear

anything from YCJY even after they stated that they would contact us when their next build was ready.

If the study would have been extended, it would mainly be based on the feedback from our clients. Since Warcry was satisfied with our efforts, we can assume that the existing methodology worked well. Therefore we could have re-used the existing study design to recruit even more participants. If the study would have been redone I would have liked to attempt to recruit more testers to make the visualization of the quantitative data more robust and interesting. As project manager, I would also have chosen to use the platform "Jira", instead of Trello, for the purpose of learning a new technology.

The study design is applicable to both smaller and larger studios, since it focuses on analyzing specific aspects of a game such as enemies or weapons. However, larger studios would most likely have better recruitment possibilities and would have larger datasets with visualizations that would be more representative.

In general, the study did not require any particular resources that hindered the progress.

V. Conclusion

The final presentation for Warcry yielded positive feedback from the client, indicating that the study design and execution was successful. Since our group managed to execute the study without any major setbacks, and since the client was satisfied with the presented results, I would rely on this study design in future projects as well.

The collaboration with YCJY consisted of some qualitative but biased feedback from us as testers. Since the communication ended, we did not receive any feedback of how satisfied they were with the testing. However, during some of the earlier meetings they commented that our feedback on the game and it's UI was useful, which indicated some level of satisfaction from the developers.

VI. Lessons Learned

The main takeaway from this project would be the importance of communication with the client. Both the client and consultants are responsible for making the collaboration successful. In our case, the communication with Warcry was very fulfilling, while the communication with YCJY was less successful. As mentioned before, it was mostly because of the obstacle of the NDA, and not because of either communication problems by us or YCJY. The NDA took away the aspect of external testing which lowered the motivation both in our own group but also the client themselves.

Personally, I found the visualization of the quantitative data to be the most interesting part of the study. For future projects, I would like to have expanded this area in some way, such as visualizing more areas that the client wanted to test, or make more complex visualizations of specific areas demanded by the client(s). For example, the heat maps could have had marker types for each enemy type, which would have indicated what enemy type killed the player at the marked location, or even the weapon that the player was using at the time of death, which might have indicated the usefulness of some weapons.

The suggestion by our supervisor halfway through the study to categorize the participants into two categories and produce separate data visualizations for these groups was unexpected. However it was more of a positive "request" since I personally thought it was very interesting, rather than being stressful.

VII. Connection to GUX

Mirza *et al.* [1] discusses challenges that indie studios face when testing their games. One of the current challenges that companies face in playtesting is to gather, analyze and visualize data in an effective way. This applies to our study since we used an almost-identical procedure. Mirza also conducted an experiment with interview questions that only consisted of open-ended questions [1]. For our own interview, we did use yes/no questions for our pre-play interview, but that was only to gain raw data of the participant's demographics. For the post-play interview we also mainly relied on open-ended questions to reduce the risk of short answers.

Personally, I am a big advocate of not using consumers/your target audience as testers and am against the concept of beta-testing. Exposing consumers to early builds of the game spoils and ruins the experience when the official game is released. Creating a game is ultimately about delivering a product. When creating a movie, you do not expose the anticipating audience to parts of the movie before it is completely finished, doing so would spoil and ruin the final product. Why should it be any different with games? Instead, I think the better approach is to use an internal, dedicated testing-team. This approach is also mentioned by Mirza [1], and that it is a common approach for large studios and that it can be difficult for smaller indie-studios due to resource-limitations. In our case, we could describe ourselves as external testing-consultants. In a real-business scenario, the testing-consultants would probably have charged the clients

for the work. This is a common strategy used by large corporations to reduce the risk of biased feedback from exclusively using internal testing teams [1].

In another study [2], Mirza presents a method to visualize a combination of quantitative and qualitative data that is comparable to heatmaps. The visualization requires measurements of player arousal (player engagement), which would not have been feasible for the scope of this study, but is nonetheless worth discussing for future projects. The method tracks player movement and draws a yellow-red spectrum line that reflects the player's arousal at that location, where yellow represents low arousal and red high arousal. This method would have been interesting to use for our heatmaps in Warcry's FPS game. In particular, it could indicate several aspects of the game such as level design, the impact of enemy types in player's arousal (by checking if certain enemies are clustered at locations with high arousal) and if certain weapons were used during periods with high arousal.

Judeth Oden Choi *et al.* [3] conducted a case study that identified certain missteps during playtesting workshops. Specifically, the lack of purposefulness of the playtesting was identified as the recurring theme of the missteps, which included not setting specific player experience goals, not choosing appropriate playtesting methods and not using the collected data to improve the next iteration of playtests or improving the game design. For the next workshops, the researchers emphasized focus on setting player experience goals, formulating and testing hypotheses for those goals, and using the collected data in an effective manner. There are several aspects from Choi's findings that could have been useful for our case in the playtest for Warcry. Especially, asking our clients what the intended experience of various game elements/aspects was would have provided us with specific things to look for and ask our test-participants about, instead of formulating the interview-questions based on our own intuition. Then formulating the intended experiences from our clients as hypotheses and then testing these hypotheses might have been an interesting method to try. In addition, presenting our method to collect and present the data to Warcry *before* starting the playtesting phase would have been ideal, rather than doing it halfway-through that phase. Luckily, Warcry was satisfied when they heard our proposed methodology regardless.

Based on these findings from previous research related to playtesting, our group would have benefited from looking up existing playtesting methods and ways to visualize data prior to designing the playtest experiment, rather than designing our own from scratch.

VIII. Author's Individual Contribution

Individually, I contributed to the project in the following ways:

- Acted as project manager
    - Managed tasks and deadlines on Trello
    - Lead most of the conversations during Discord meetings with the clients.
    - Scheduled meetings with both our clients and our internal meetings.
    - Initiated the preparation and suggested the structure for presentation slides.
- Planned and suggested the structure and execution of the experiments.
- Conducted 3 of the 15 experiments.
- Participated in all presentations and all except one meeting with our supervisor.
- Helped out one member who had trouble counting the data from the recordings and inserting this data into the shared data sheet.

Overall, I enjoyed taking on the role as project manager and am satisfied with my effort that I put into the project.

References

[1] P. Mirza-Babaei, N. Moosajee, B. Drenikow, "Playtesting for Indie Studios", Ontario Tech University, October 2016.

[2] P. Mirza-Babaei, G. Wallner, G. McAllister, L. E. Nacke, "Unified visualization of quantitative and qualitative playtesting data", Extended Abstracts on Human Factors in Computing Systems, April 2014.

[3] J. Oden Choi, J. Forlizzi, M. G. Christel, R. Moeller, M. Bates, J. Hammer, "Playtesting with a Purpose", 2016 Annual Symposium, October 2016.

Pre-play session interview questions
How old are you?
What is your current occupation?
Do you are have you studied [or worked with] games?
How often do you play video games?
What (kind of) games do you play?
Do you play FPS-games? How much?
Have you played the first Doom-game or Wolfenstein 3D?

Table 1: Pre-play session interview questions.

Post-play session interview questions
What do you think of the game?
What do you think about the difficulty?
What was the hardest part of the game?
What do you think about the enemies?
What do you think about the weapons?
What do you think about the UI and the controls?
Anything more you want to add?

Table 2: Post-play session interview questions.

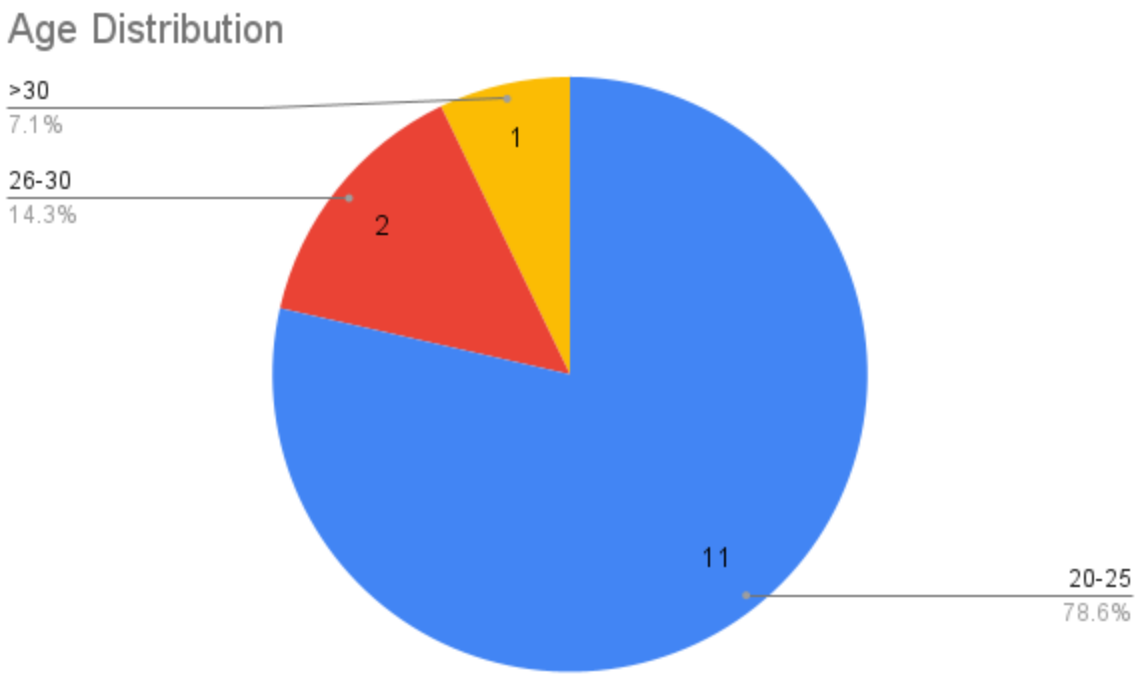

Figure 1: Age distribution of the test-participants.

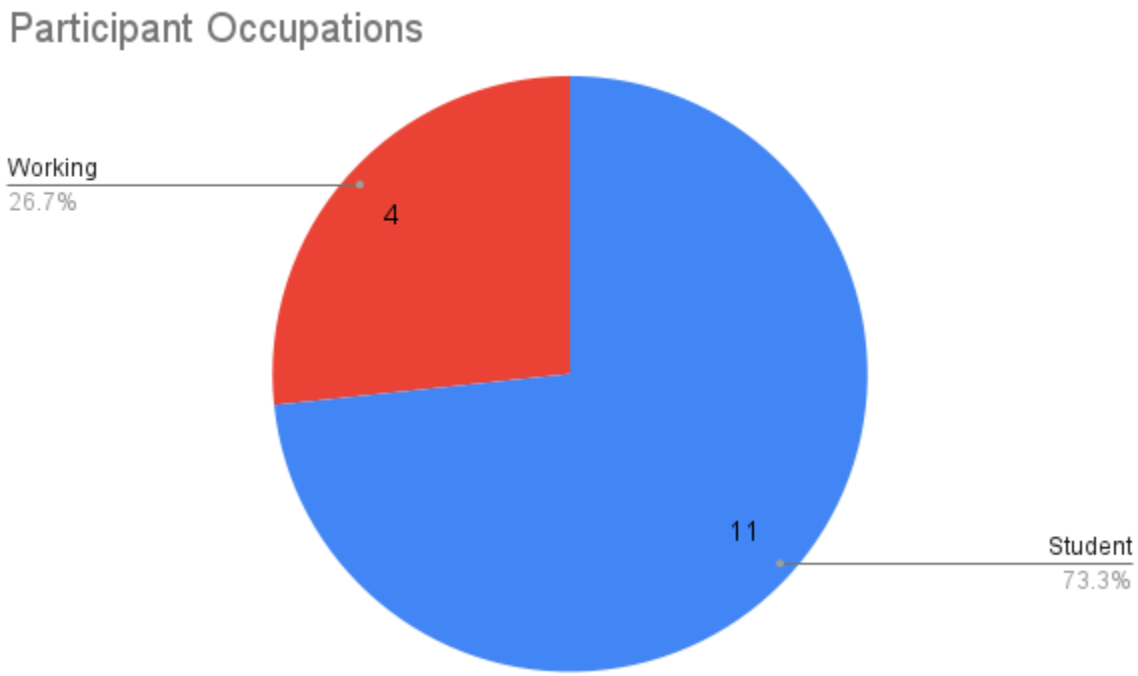

Figure 2: Participants that worked or studied.

Figure 3: Participants that had prior experience working or studying with games.

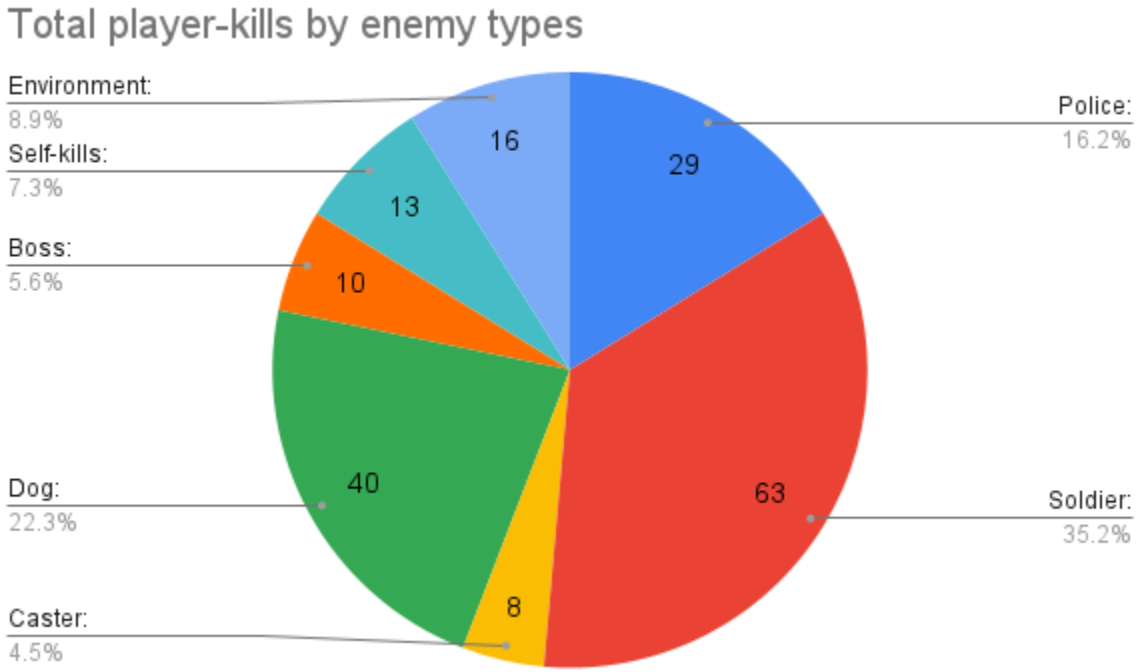

Figure 4: Total player-kills by enemy types

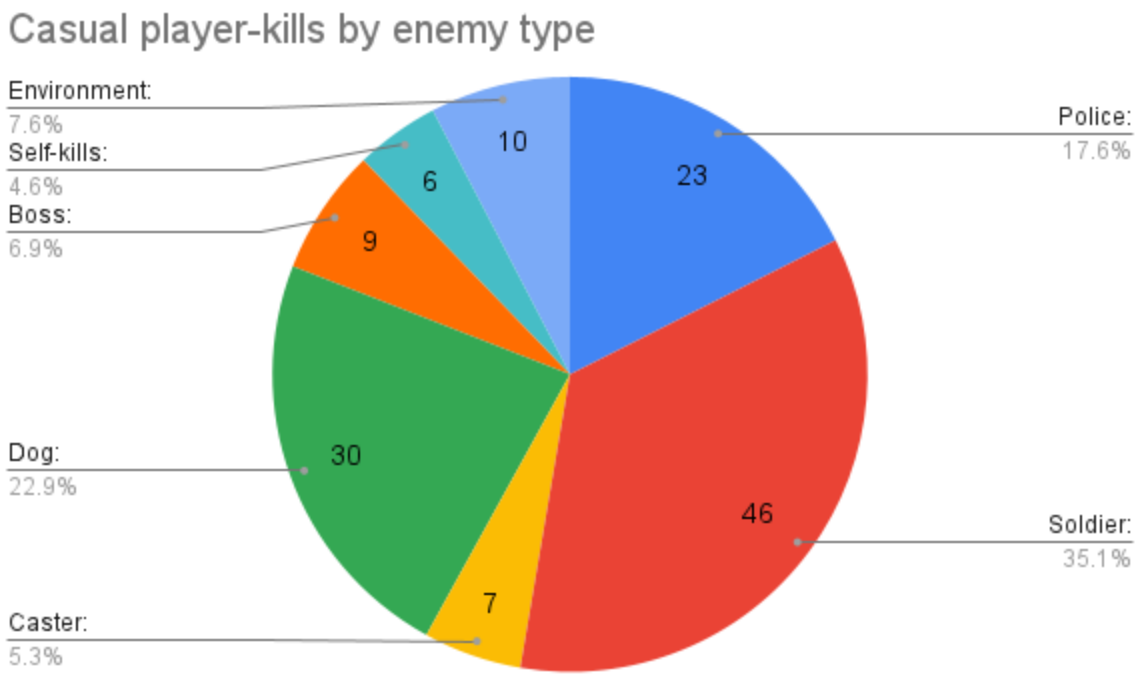

Figure 5: Casual player-kills by enemy types.

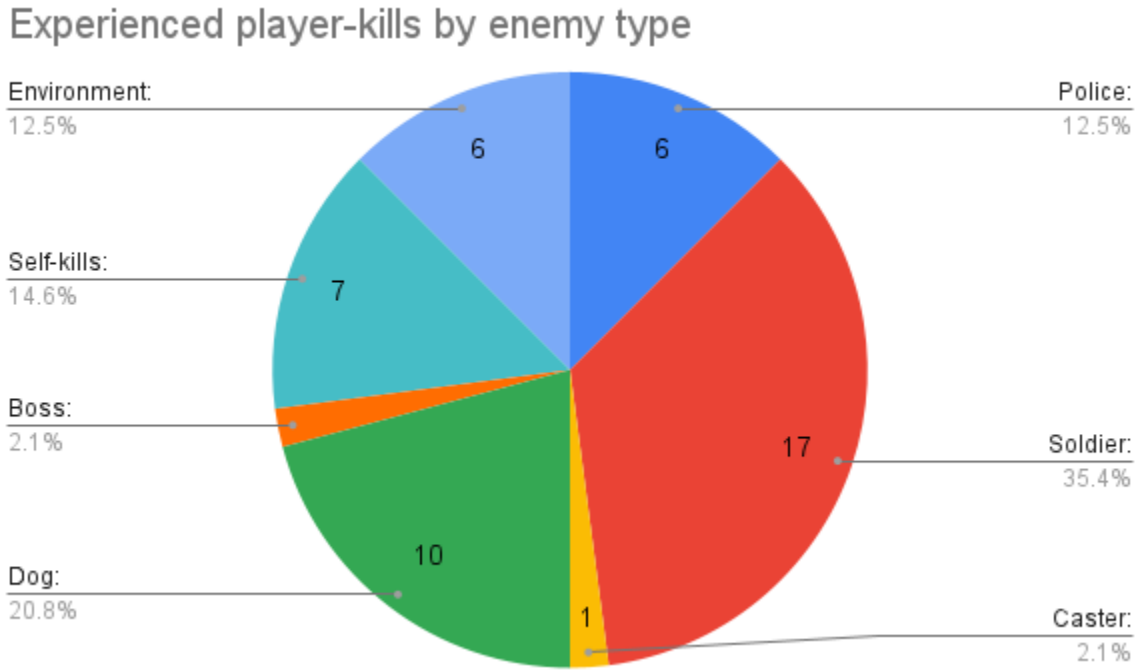

Figure 6: Experienced player-kills by enemy types.

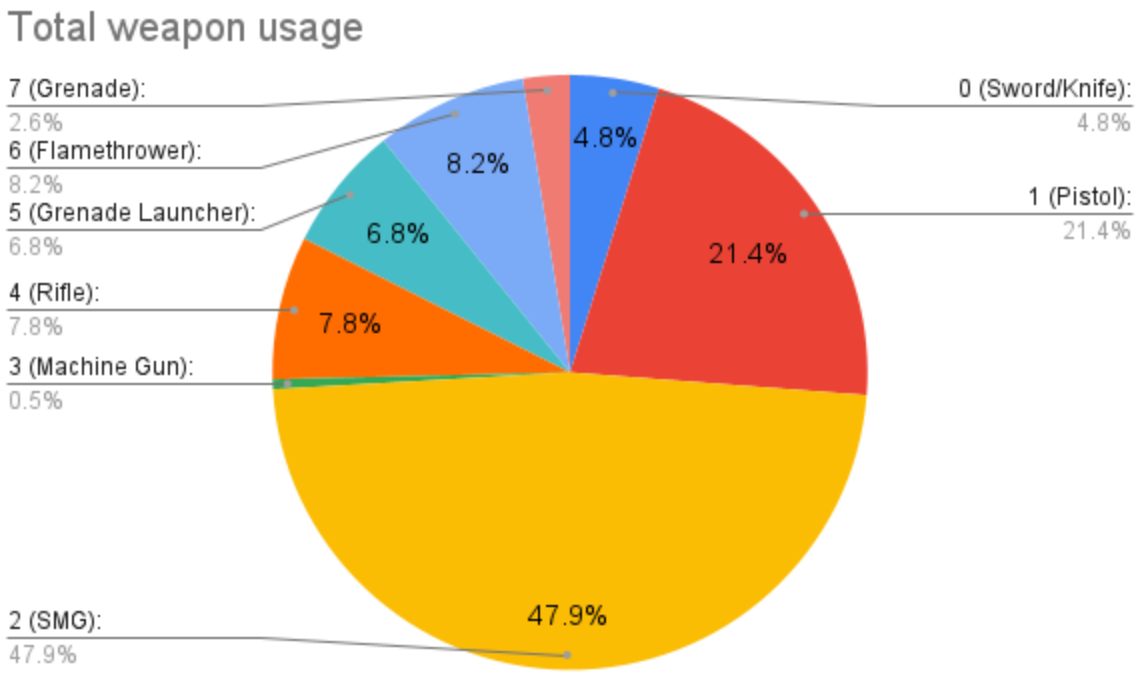

Figure 7: Total weapon using time.

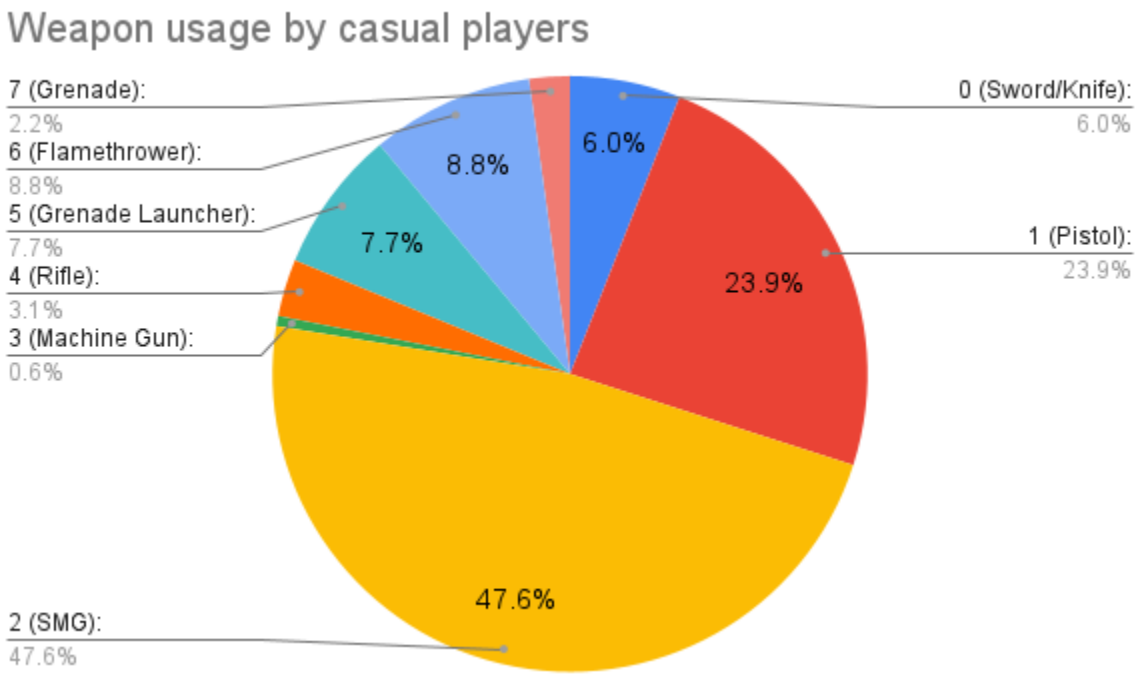

Figure 8: Weapon usage by casual players.

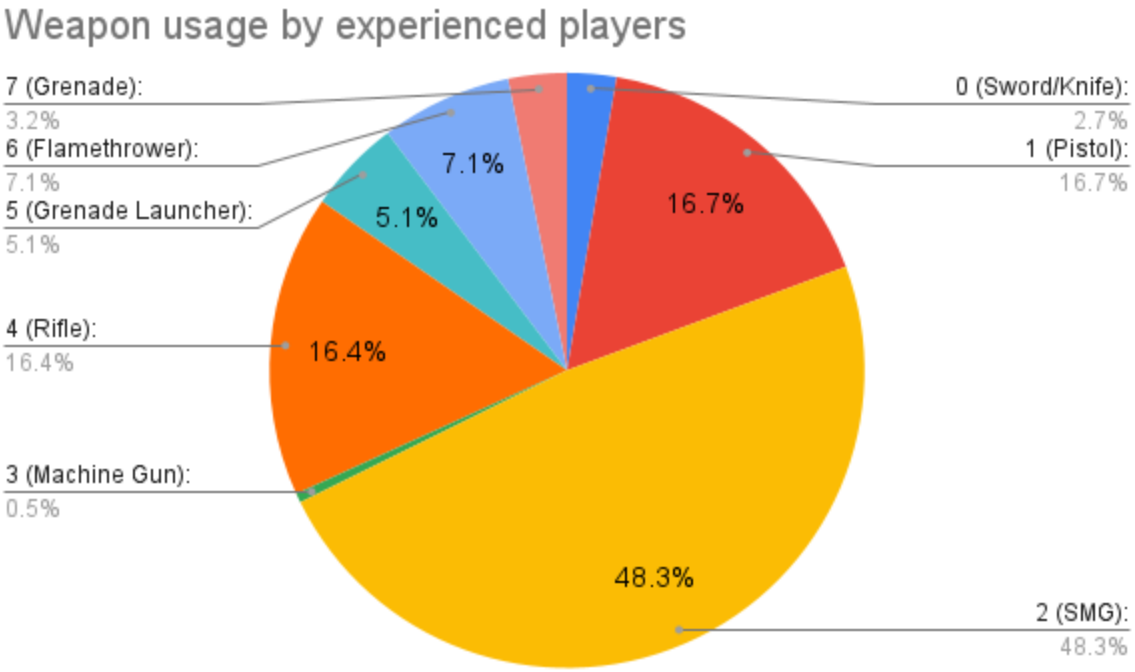

Figure 9: Weapon usage by experienced players.

Figure 10: Heatmap of build 1, level 1.

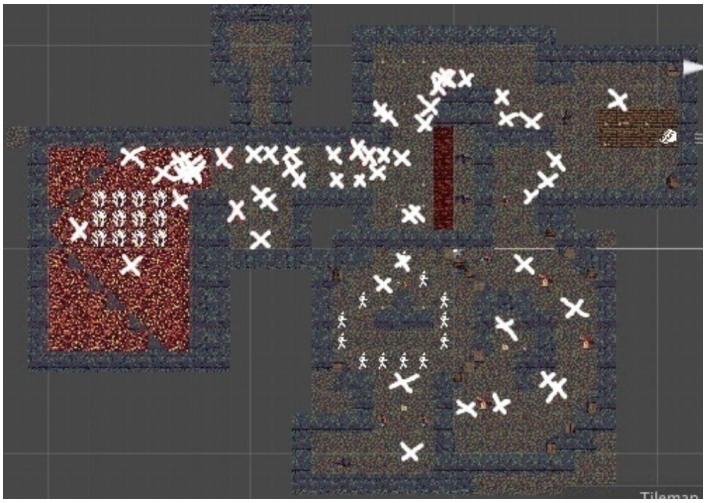

Figure 11: Heatmap of build 1, level 2.

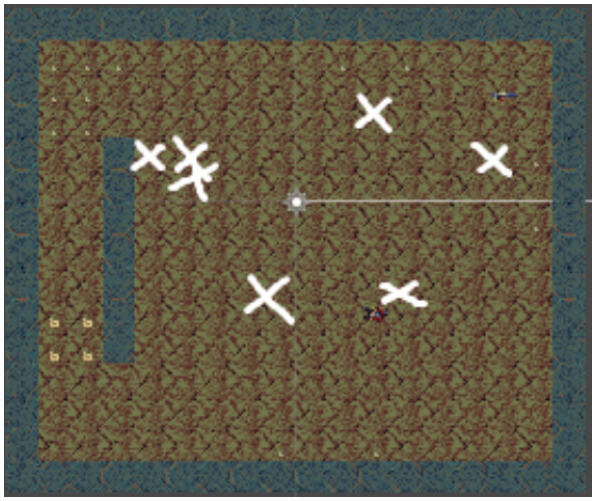

Figure 12: Heatmap of build 1, level 3.

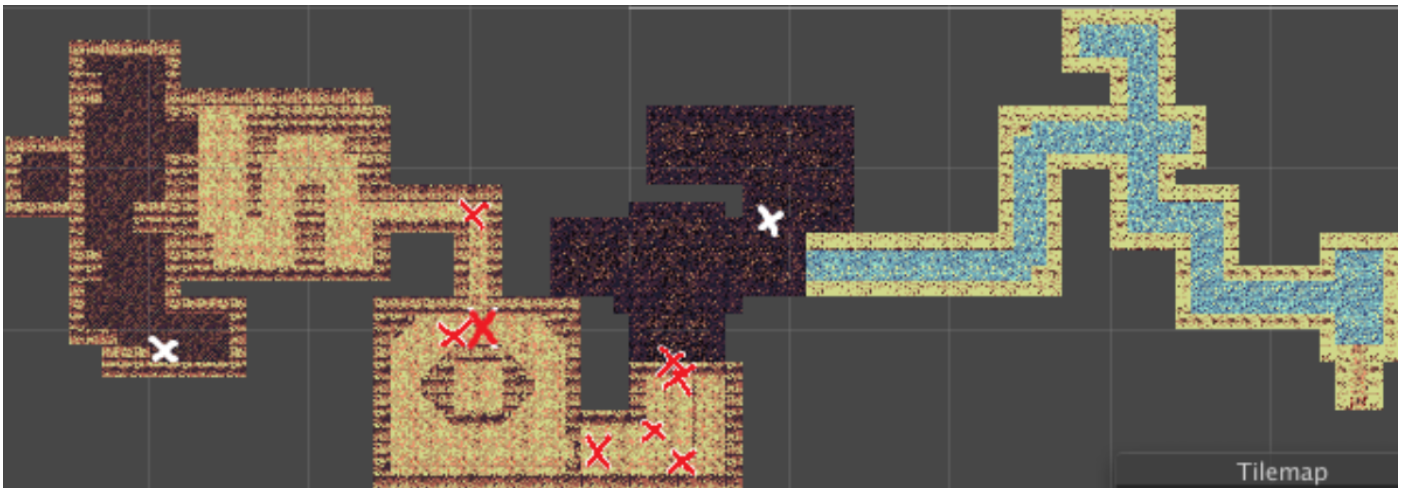

Figure 13: Heatmap of build 2, level 1.

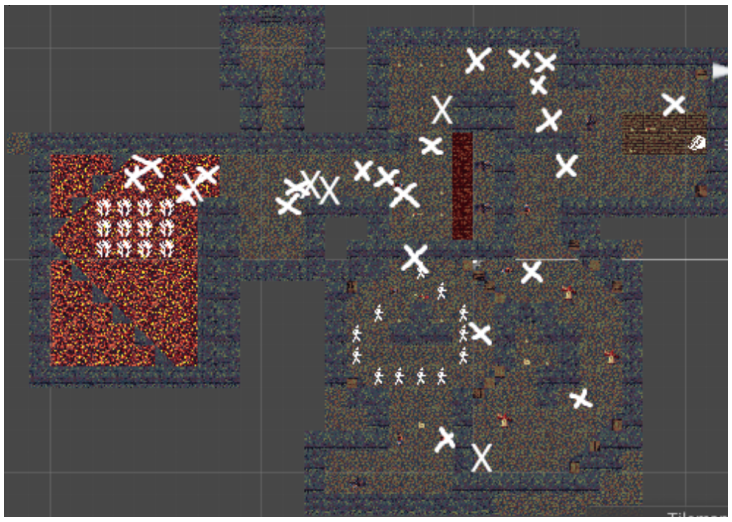

Figure 14: Heatmap of build 2, level 2.

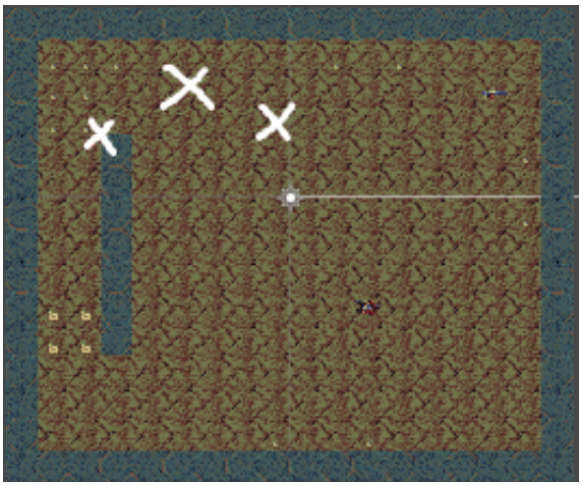

Figure 15: Heatmap of build 2, level 3.