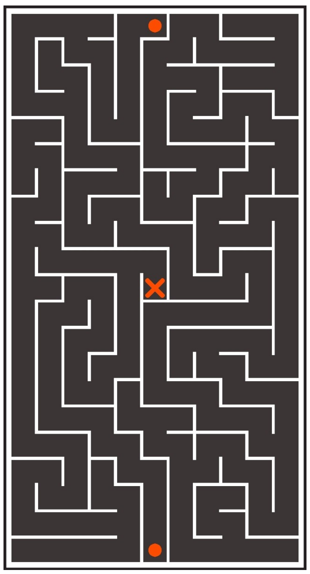
Maze Game Data Analysis

1. Introduction

1.1 Explanation of the Game

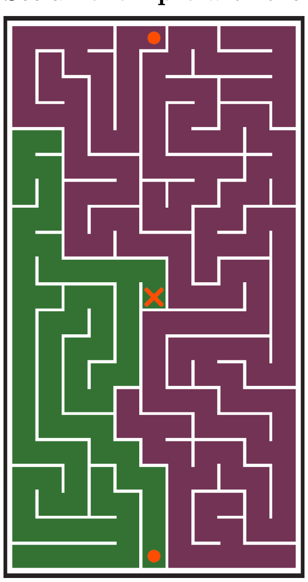
aMaze is a mobile puzzle game where players must navigate procedurally generated mazes under time pressure to earn the greatest score possible. Each maze in the game is unique, ensuring replayability, unpredictability and a challenge to any player in a short time game.

- *Maze:* When starting a session, the player enters a **block** of 10 **mazes**. In each maze, players are presented with a new randomly generated grid of 11 cells width and 21 cells height. At the central cell of the maze there is always an orange X that marks the treasure of the maze, as well as two orange dots that mark the two entrances centered at the top and bottom rows.



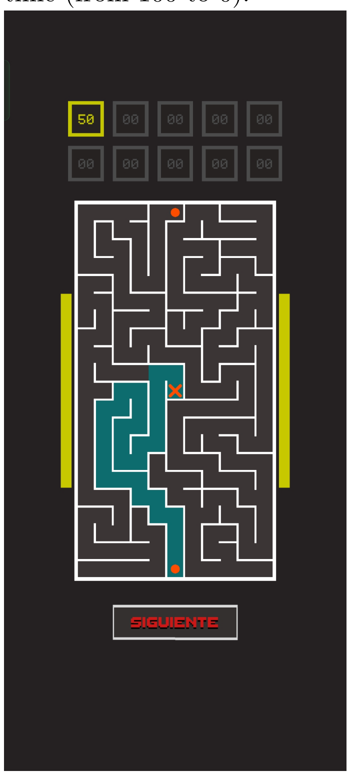
- *Algorithms:* Each maze is generated with two simultaneous backtracking algorithms starting at each entrance cell. This generates to fully separated paths, with only one connected with the central cell.

  See an **example** with the top algorithm painted in purple and the bottom algorithm painted in green:

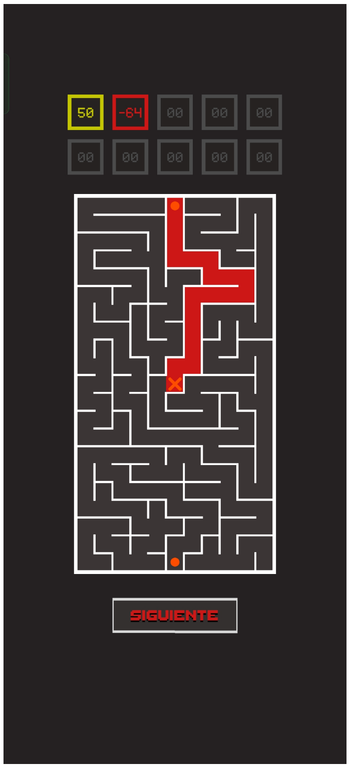


– *Task:* In each maze, the player must find out which entrance is connected to the central cell as fast as possible. The procedure of each maze is as follows:

1. The maze is revealed and a 10 seconds countdown starts. During each maze, the player receives continuous feedback. Two shrinking bars on the sides of the screen visually indicate the remaining time. At the top of the screen, small squares summarize the scores obtained in previous mazes within the block. The square corresponding to the current maze displays a live countdown of the exact points available, decreasing second by second. This dual feedback system allows players to monitor their progress both in the moment and across the block.
2. The player should tap the upper half of the maze if the central cell connects with the top entrance, or the lower half if it connects to the bottom entrance.
3. Whether the player gives a response or the 10 second timer is over, the maze ends and the score is registered at the corresponding score squares located at the top of the screen.
   The scores are determined as follows:
   * **Correct answer**: They gain an amount of points equal to the percentage of remaining time (from 100 to 0).

   * **Wrong answer**: Negative points equal to the percentage of remaining time (from -100 to 0).



- * **Time up:** Alternatively, if participants didn't give any response when the time is over, the score will become automatically -100.

4. Each score is stored at the top of the mazes, and after 10 mazes are finished the block ends and participants get the final score as the average score of all 10 mazes (with a minimum of 0).

This method of scoring encourages players to always try to give a response, but spend as much time as needed to have enough certainty that the side they select will be the correct one, balancing between the **probability** of reward (as the information gathered through time) and the **amount** of reward. The faster they respond, the greater will be the reward if they succeed, but also it will be more risky. This tension between speed and accuracy, added to the fast nature of the game (as each block usually takes between 30 seconds and 1.5 minutes to complete) makes the game perfect for short but engaging play sessions.

Finally, with this score, players may find a personal goal both individually by improving one's own scores across sessions and/or competitively, through leaderboards that rank players globally.

1.2 Aim of the Analysis

The aim of this report is to provide insight into how players interact with the game, identify any aspects that influence their success, and allow for predictions about how well they will perform and for how long will they play.

Therefore, we focus on player behavior and a bit of UX insight.

1.3 Research Questions / Hypotheses

Based on the aims above, the present exploratory report focuses on five broad questions:

- **RQ1 – Engagement and churn**
  - When and how often do players abandon the game after or during a block?

  Hypotheses

  - **H1a.** Incomplete blocks will be relatively rare and more frequent among casual players (those who play fewer blocks overall).

- **H1b.** Blocks that start with very negative scores (early wrong responses) will be more likely to be abandoned (higher block churn). Report_fixed_V2
- **H1c.** Both very low and very high block scores will be followed by shorter within-session pauses, whereas moderate scores will lead to slightly longer breaks.

- **RQ2 - Maze structure and difficulty**
  - Are the generated mazes "fair" in the sense that the correct exit side (top vs bottom) occurs with approximately equal frequency?
  - How do structural properties of the maze (e.g. length of the path, distribution of turns and crossings...) relate to response time and accuracy?

  Hypotheses

  - **H2a.** The correct exit side will be approximately balanced across mazes (aprox 50% top vs. bottom).
  - **H2b.** Longer paths will be associated with longer response times, for both correct and incorrect decisions.
  - **H2c.** Mazes with more complex geometry will tend to be harder, leading to lower accuracy and increased response times.

- **RQ3 – Behaviour and decision-making**
  - How accurate are players overall at identifying the correct exit, and do they show systematic response biases towards one side?
  - How is the speed–accuracy tradeoff expressed under the 10-second time pressure of the task?

  Hypotheses

  - **H3a.** Players will perform clearly above chance level overall (accuracy well above 50%).
  - **H3b.** Some individuals may show systematic bias towards one decision.
  - **H3c.** Very fast responses will be less accurate, with accuracy improving and stabilising for responses made at later times in the trial.

- **RQ4 – Learning across different time scales**
  - Do players show improvements in speed and/or accuracy across their overall experience with the game (long-term learning across days and sessions)?
  - Do they show "warm-up" or fatigue effects within sessions (short- to medium-term learning)?

  Hypotheses

  - **H4a.** As players accumulate experience (more trials played overall), response times will decrease while maintaining or slightly improving accuracy.
  - **H4b.** At the beginning of a session there may be be a short warm-up period with slightly slower and/or less accurate responses that stabilise after a few blocks.

---

2. Methodology and Data

2.1 Data Collection

The game creates 2 different files locally, which are updated at the end of each maze:

- **datapack.json** contains all the information of each cell, of each maze, of each block, along with participants' information. This is the data that we will analyze.

- **summary.json** contains only the information of each participants' block, without the specifics of the maze and cell data. This alternative file is used to more easily keep progression of each participants with only local files.

2.2 Main Variables

- Player-level variables:
  - **PlayerID:** Nickname that each participant chose personally.
  - **DeviceID:** Random 4-character identifier of each individual device.
- Maze-level variables:
  - **TrialID:** Identifier of the maze order within a block (from 1 to 10).
  - **TrialTimeStamp:** Time stamp of each individual maze.
  - **Score:** Each maze score goes from -100 to 100. It starts at 100 as the maze starts and decreases towards 0 across the 10 seconds players have to answer. If the response is correct the value is stored as positive and if the response is wrong the value becomes negative. If players do not respond before the 10 seconds, the score becomes automatically -100.
  - **RT:** The time in seconds players took to give a response. If a player does not respond, the RT is stored as NA.
  - **Correct:** Whether participants selected the correct side (TRUE) or the wrong side (FALSE).
  - **CorrectSide:** Which is the side of the correct path.
  - **PathLength:** How many steps is the path long.
- Cell-level variables:
  - **CellX:** Column coordinate of each cell.
  - **CellY:** Row coordinate of each cell.
  - **Algorithm:** Which algorithm "included" this cell, either Top or Bottom.
  - **NumInPath:** If the cell is part of the shortest path to the exit, it shows the number in steps that you might find it if following the path from the beginning. If it is not part of the optimal path, it shows NA.
- Block-level variables:
  - **BlockID:** Identifier of the number of blocks that the player performed (increasing from 1).
  - **BlockTimeStamp:** Time stamp of the start of the block of mazes.
  - **AvgScore:** Average of the scores of all completed mazes within a block.
  - **IsBlockComplete:** Whether the block reached the end (TRUE) or the player closed before finishing (FALSE).

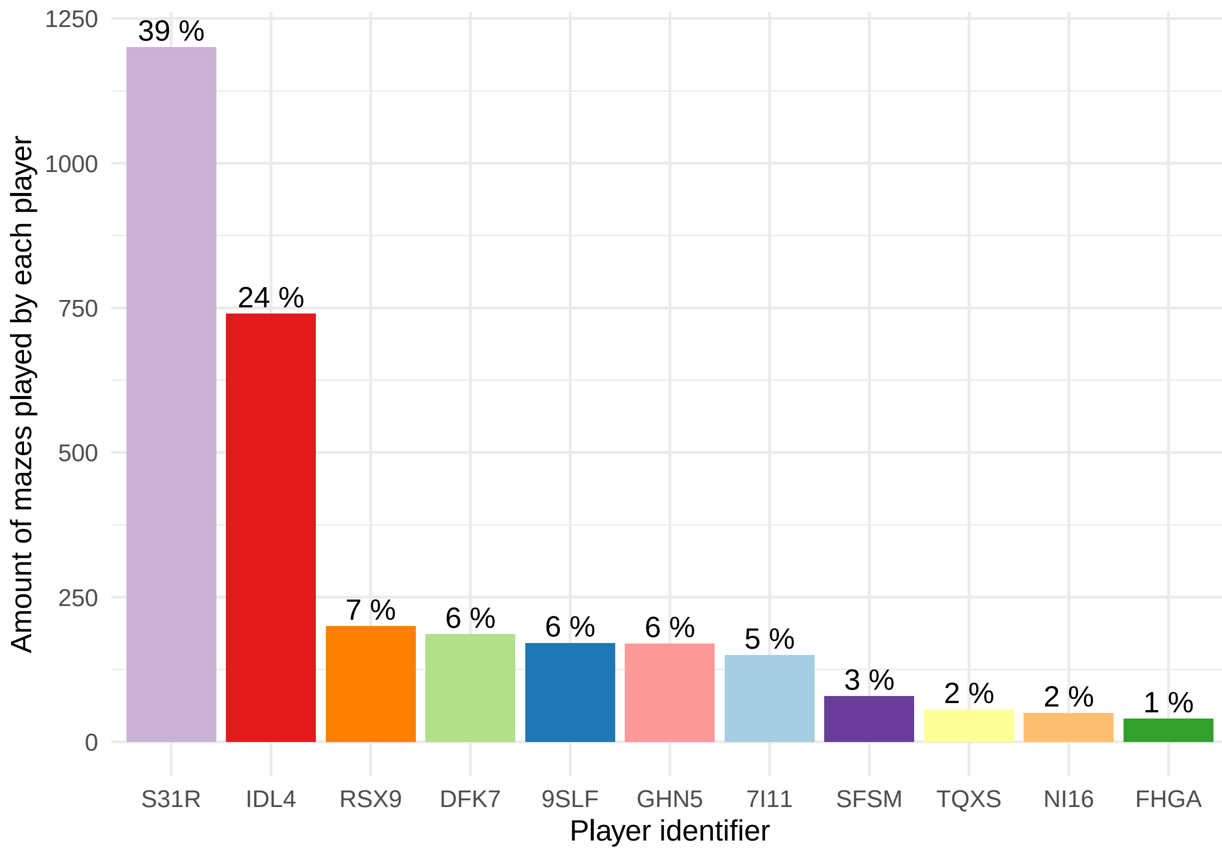
3. Analysis and Results

3.1 Summary Data

- **How many players played the game?**

We gathered data from a total of **3043 mazes** across **11 players** during an interval of **24** days.

- **How much did they play?**
  - We observe an uneven amount of mazes played between participants, with a few participants owning the majority of the games played.



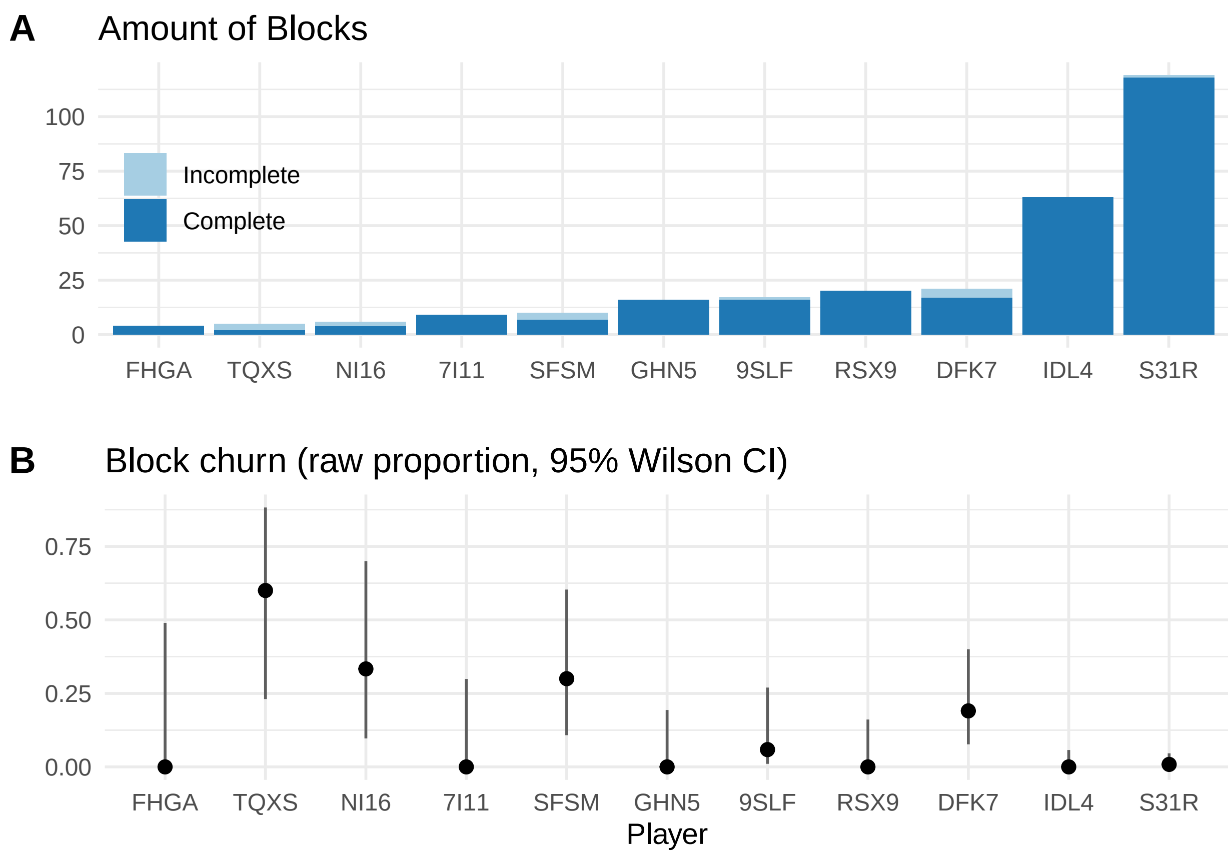
3.2 Engagement

A key aspect of the experience is not only *how much* people play, but *where/when* do they stop. We look at disengagement through two lenses. First, **block churn**: blocks of mazes that were started but not completed. Second, **session churn**: players who finish a block and then stop playing for a while (end of a session).

3.2.1 Incomplete blocks (Block churn) One example of this situation is when players are performing a block of mazes but decide to stop playing midway through the block.

We quantify *block churn* as the fraction of started blocks that ended incomplete. In our dataset, incomplete blocks are rare (*aprox 5%*) and unevenly distributed across players.

Panel A displays each player's *volume* of complete vs incomplete blocks. To correct for the lack of a greater sample, Panel B shows each player's *incomplete rate* by a shrunk estimate (that gently pulls very small samples towards the overall average, leaving well-measured players essentially unchanged) with 95% Wilson confidence intervals (which is more robust for small samples).



From this, we can conclude that more hardcore players (those that played more games overall) almost never leave a block unfinished, while higher incomplete rates are observed mostly on more casual players (with less mazes played overall) and come with wider uncertainty. However, although block churn exists in our sample, it doesn't seem to be the dominant disengagement pattern.

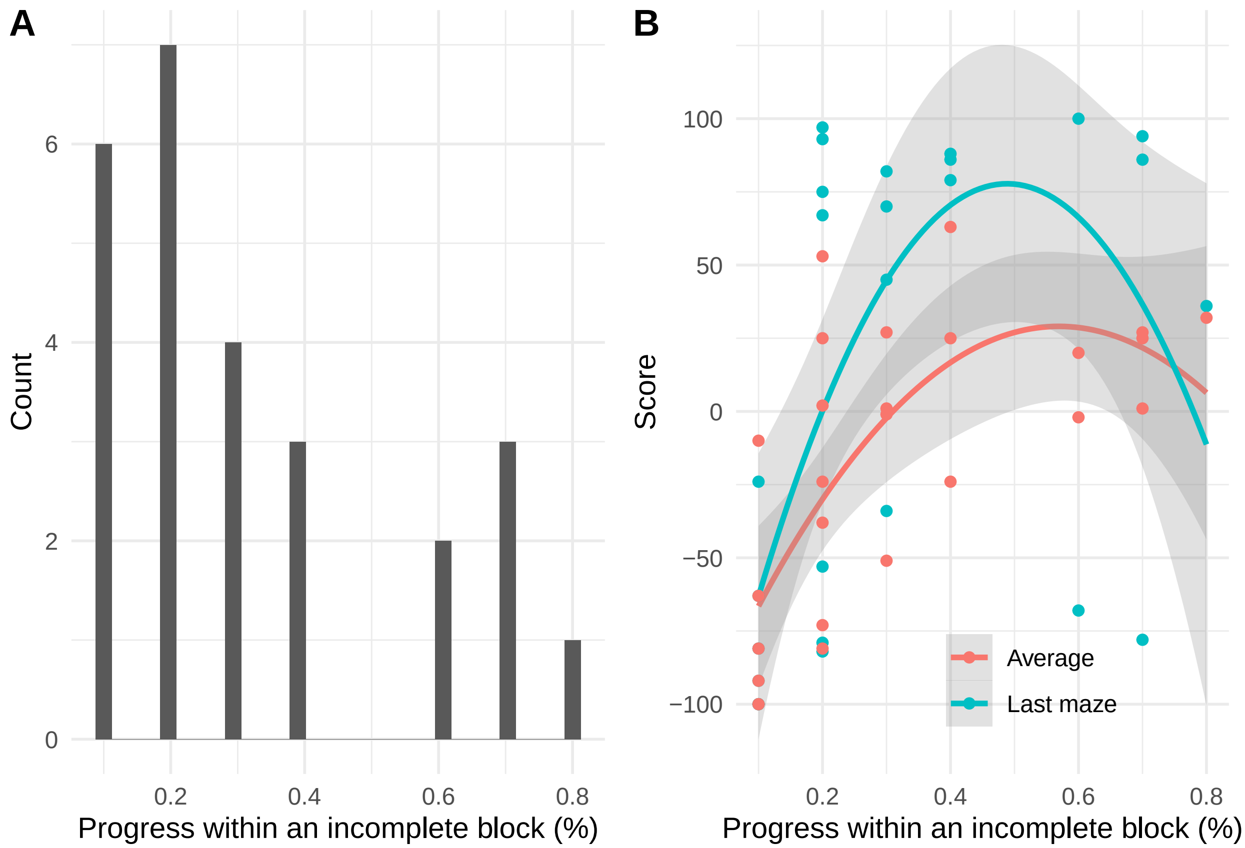
How far into the block do players quit?

This block churn could happen at different moments, for example a player might start with very low scores in the first couple of mazes and decide it not worth to carry on with the block, or they could try to get to the end and maybe predict a low final score and decide to quit from frustration.

Given the small number of incomplete blocks (N=*26*) we can only report descriptive evidence.

A bimodal distribution of churn can be identified (Panel A), one with few mazes (1 or 2) and a smaller one closer to the end of the block, at 7 trials. No blocks missing only one trial are found.

In terms of its relation to the score of the mazes played (Panel B), we find that churn in early mazes is always associated to negative score, which represents blocks that start with wrong responses. With churn later in the block, we find that although participants got a very good score in the last maze, the average score was still low, which implies that early mazes had to have a negative score (hence a wrong answer). This shows that the most important mazes to predict whether participants will churn are the first ones of the block.



3.2.2 Spacing between blocks (Sessions churn) Another conceptualization of disengagement could be the moment at which players stop playing successive games and leave the game.

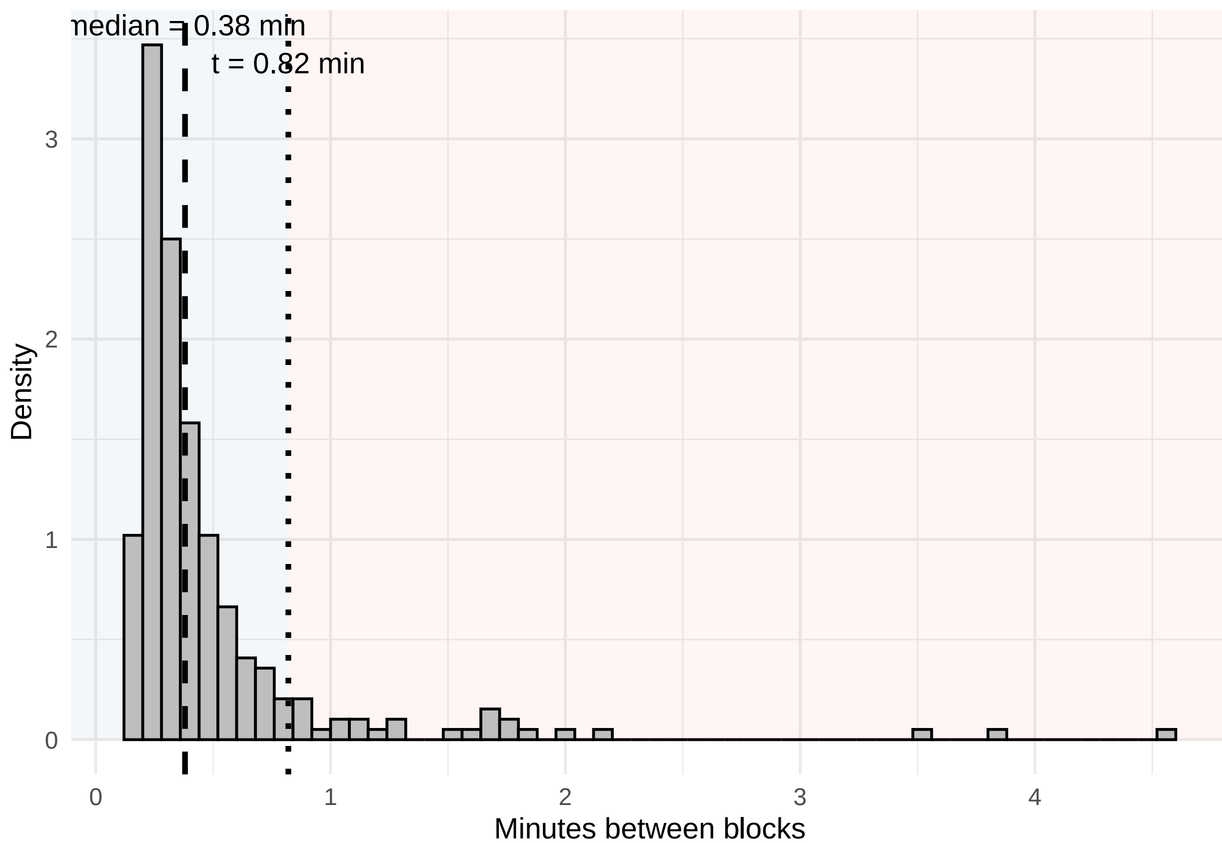
Here, the criteria is the gap between blocks. However, interpreting the time between blocks can be tricky. Players could be having a short rest between blocks of the same "gaming session", or they could completely separate their attention from the game, coming back at a later (although sometimes short) time to begin with a new session.

The key to define when a session has really ended relies on finding an adequate time threshold, after which we expect (at least in most cases) that the following block will be part of a new session. Once we identify this we can pinpoint the games where, although completing the block, led to players leave the game afterwards. This in turn can allow to study what factors can influence the disengagement to the game.

We compute inter-block gaps from the **end time** of a block (last maze start + RT) to the **start time** of the next block.

To differentiate between short within-sessions rests from session ends we use a robust, data-driven threshold (t) based on the Median and the Median Absolute Deviation. Gaps below the threshold keep the session as continued, gaps above the threshold are considered the division between two sessions. With this, we can pinpoint the moments at which players stopped playing.

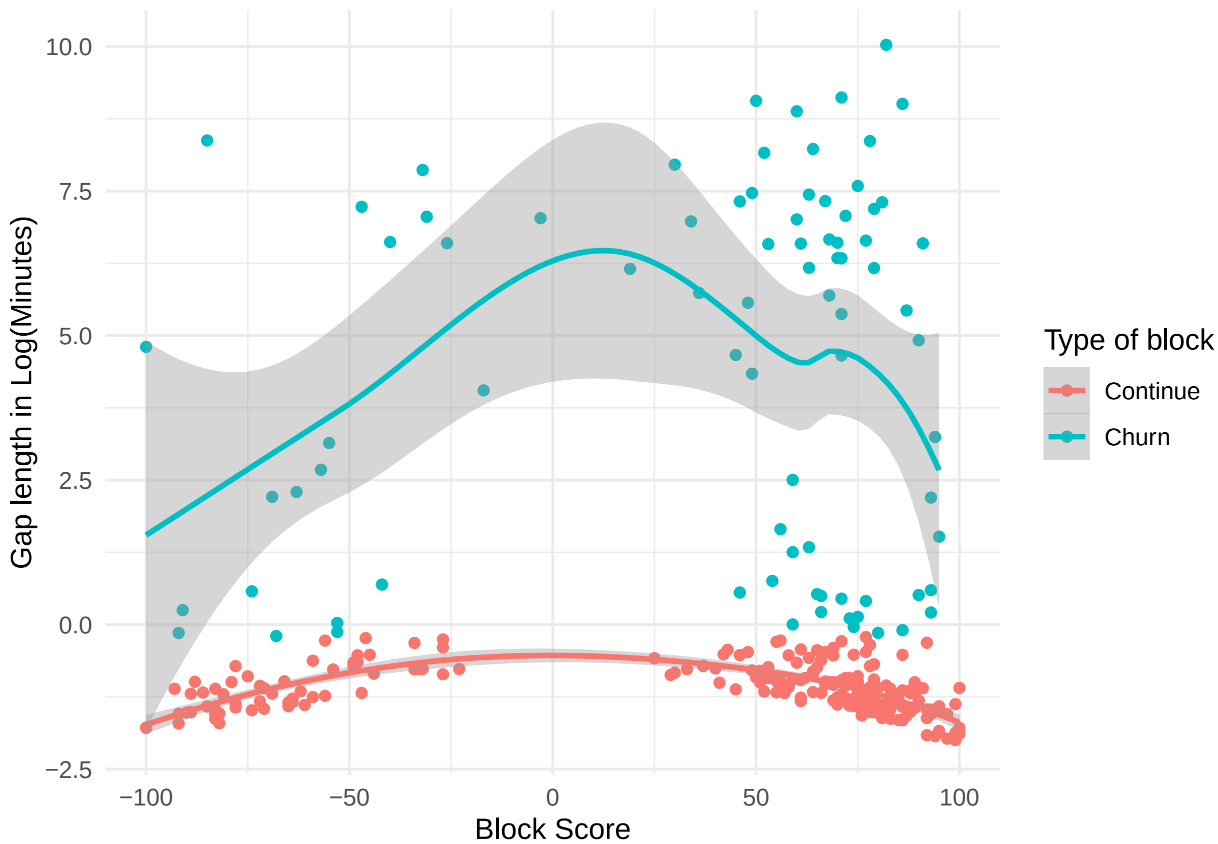
Here we show the distribution of the shortest gaps between blocks (longer gaps, such as hours or days are not even considered for representing the threshold). The vertical dashed line marks the **median** (0.38 min), and the dotted line marks the **threshold** (0.82 min).



Due to the huge scale difference in gaps between blocks, where most of them lie below 3 minutes but there are some observations to up to days, we applied a logarithmic transformation to the gaps length in minutes.

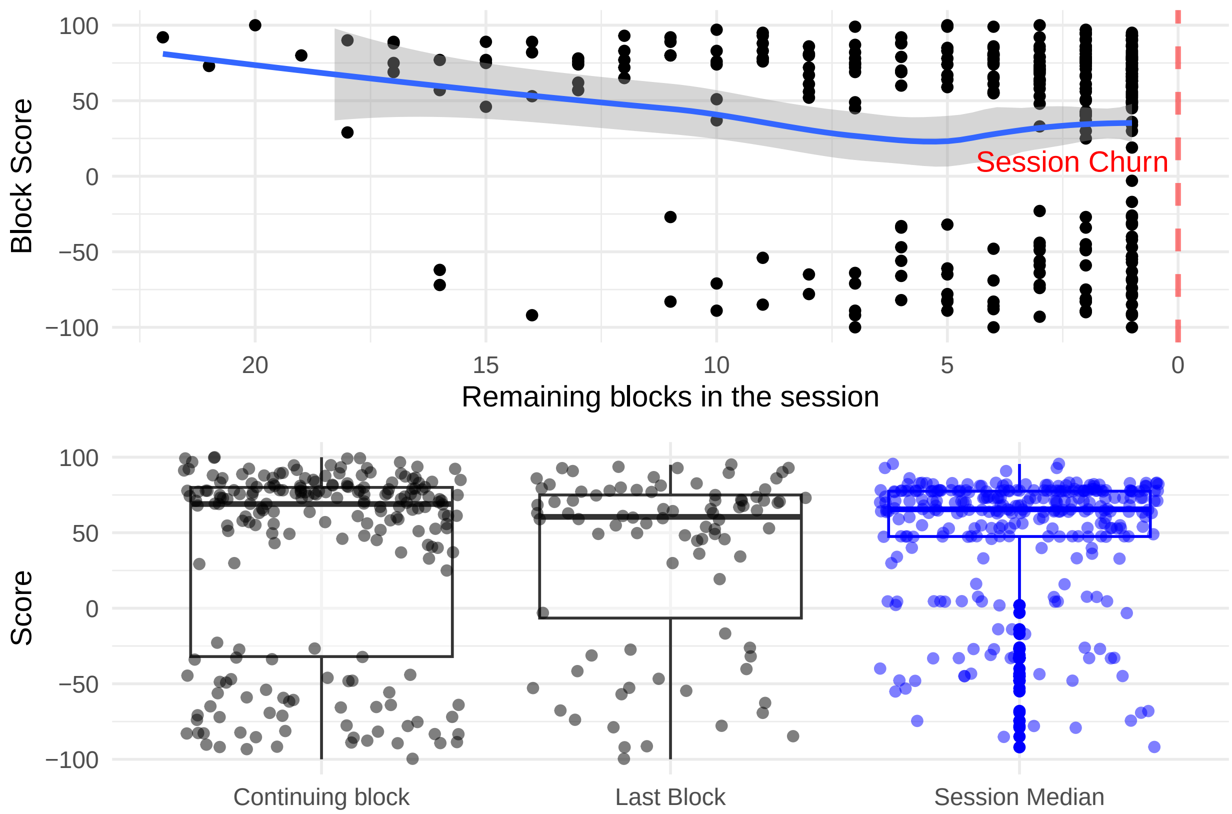
We find that within-session rests (gaps after blocks that do not end the session), show a weak inverted-U relationship with score. Pauses are shortest after a very low and very high scores, and slightly longer around mid-range scores. One possible explanation for this finding is that after a big success players are reinforced to keep playing as quickly as possible, while a very bad score drives them to improve that last game and surpass their own score. In both cases, players **engage** more with the game because of the **extreme scores**. In terms of session ends (by definition longer gaps), they follow a similar pattern, but with gaps more evenly distributed after higher scores.

Note that given the sample size and the difficulty at establishing a reliable threshold, these patterns are descriptive rather than inferential.

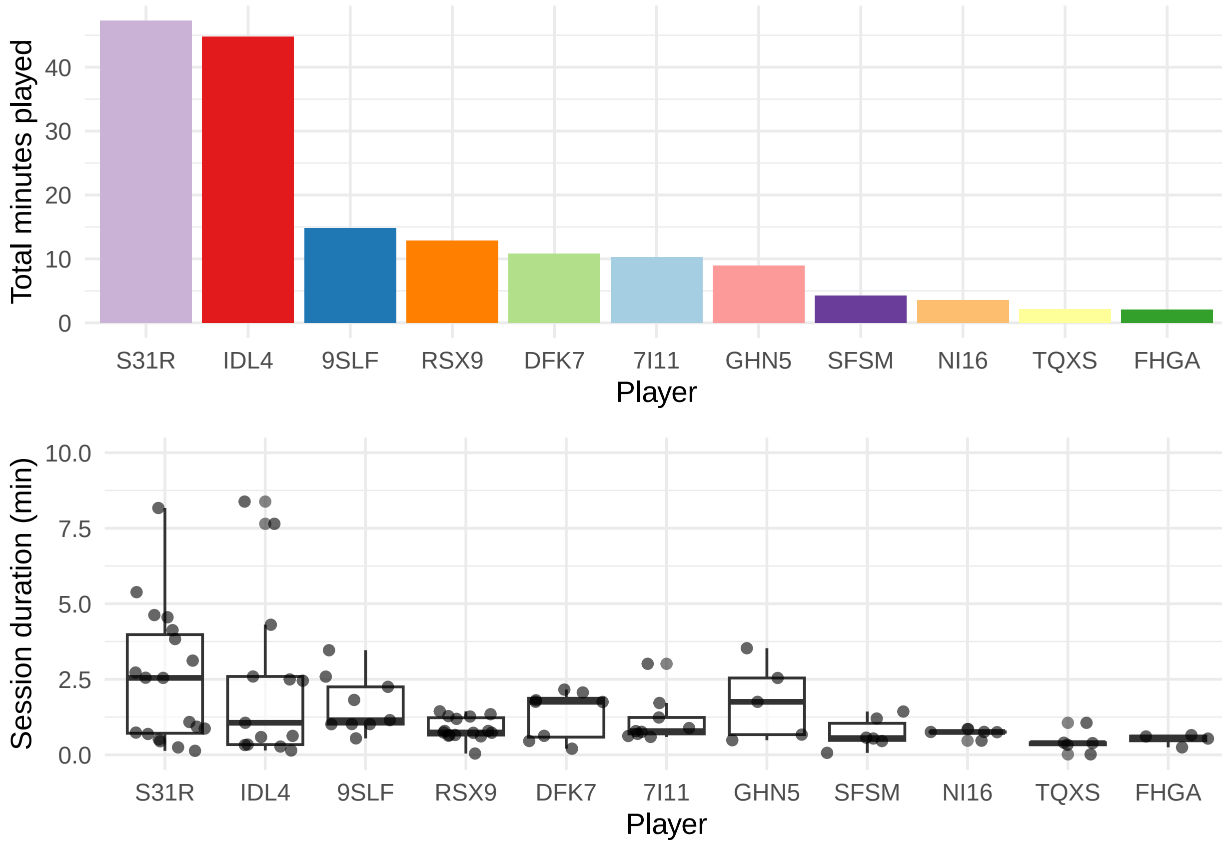


We also find that early blocks in a session tend to show better performance, that decays with later blocks presenting more varied scores, with increased prevalence of negative values. This shows that players might be getting tired after repeated blocks, and once their performance has worsened enough, they leave the game until a future session. Although our data is not sufficient, with enough amount of data this could be used to get a tailored prediction of when each player might abandon the game according to their evolution in performance.

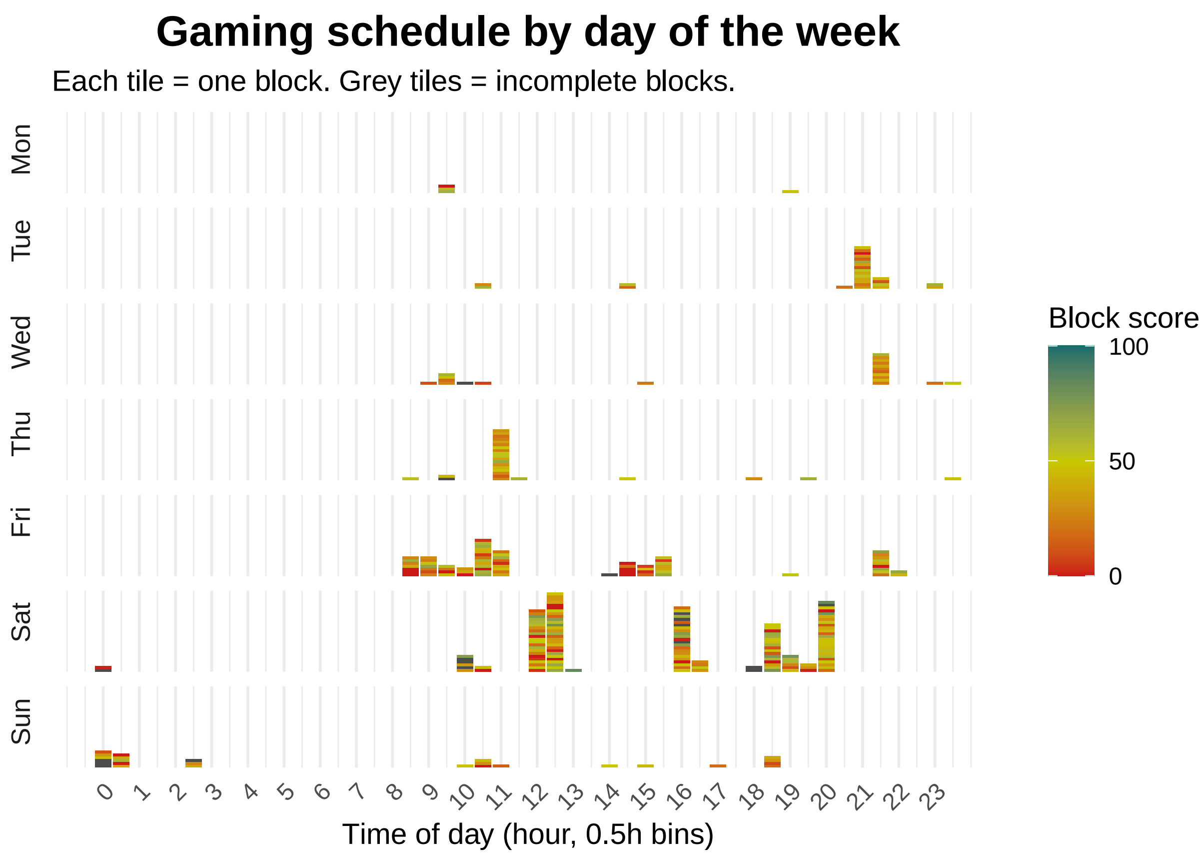
Also, although there are not significant differences between the last block and other blocks or even the session median, it is clear that the variability in scores is greater at these last blocks. Probably tiredness makes these games less representative of the real player level.



3.2.3 Time played Total time varies widely across players. Pairing minutes with sessions and average session length reveals distinct engagement patterns (few long sessions vs. many short ones).



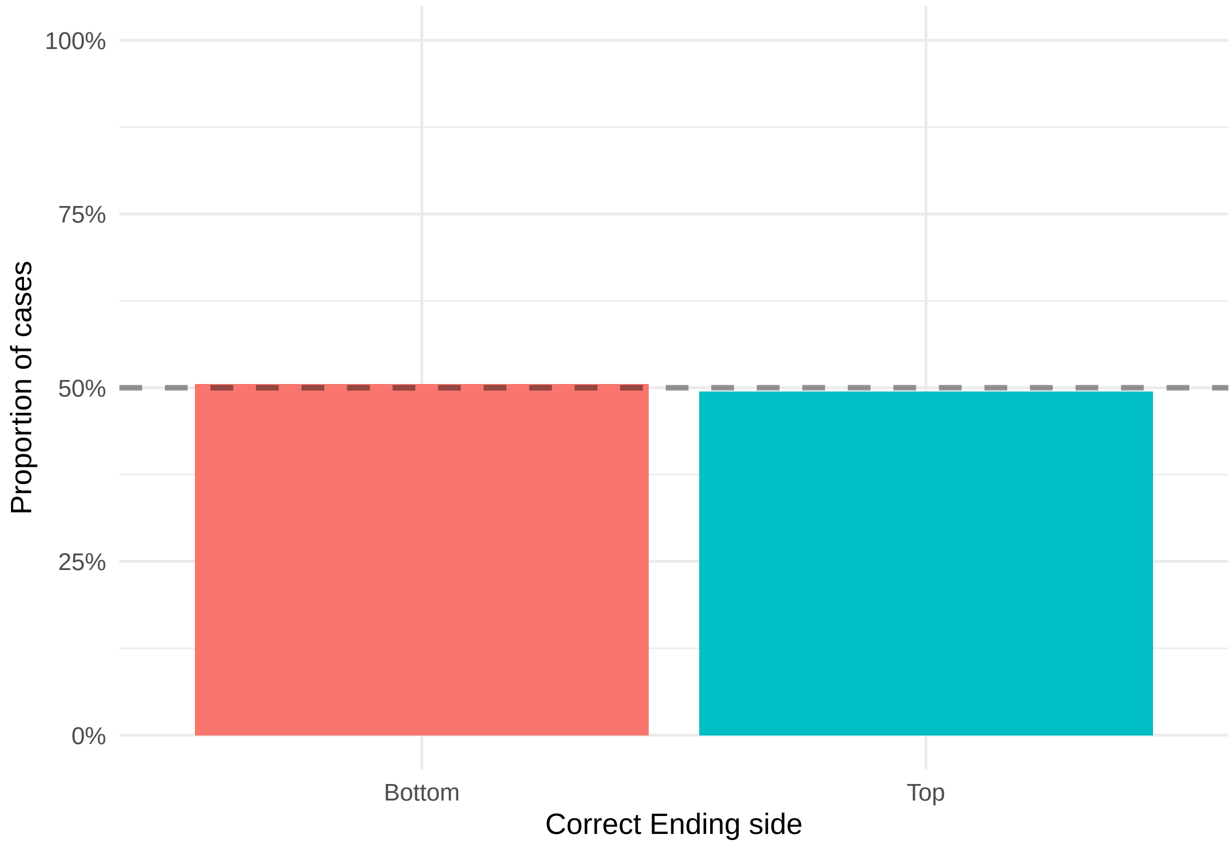
Total minutes played
40
30
20
10
0
S31R
IDL4
9SLF
RSX9
DFK7
7I11
GHN5
SFSM
NI16
TQXS
FHGA
Player
Session duration (min)
10.0
7.5
5.0
2.5
0.0
S31R
IDL4
9SLF
RSX9
DFK7
7I11
GHN5
SFSM
NI16
TQXS
FHGA
Player



Usually, the most frequent playing times are throughout the morning and in the evening. Saturday is a very packed day with lots of games taking place around noon, with a peak at 16, and also throughout the evening. Also, the night with more nocturnal players is the Saturday to Sunday night.

3.3 Maze Data

First, we check that mazes were "fair". For example, the end of the maze should be with equal probability at each side.

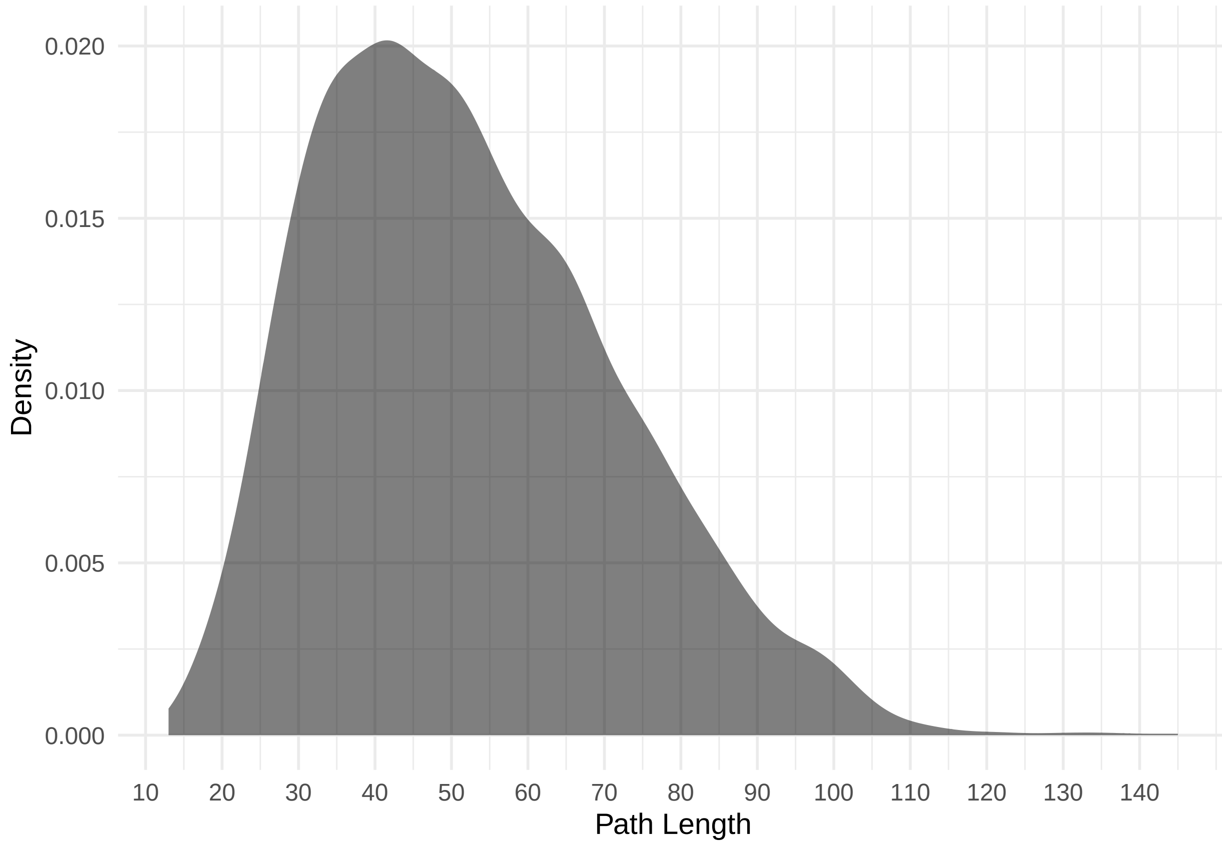


3.3.1 Path Length We can also generate metrics that could relate to maze difficulty.

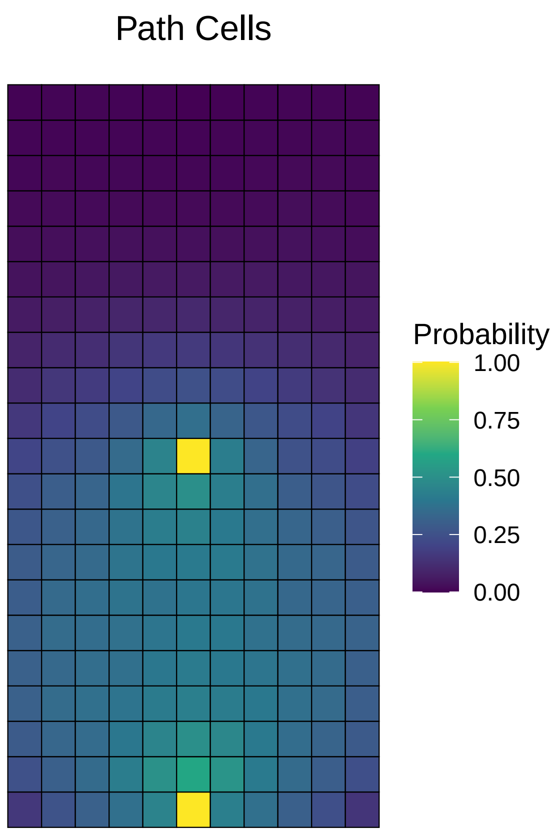
For example, if the path between the correct exit and the center of the maze is longer, it could explain longer response times and even more wrong responses. Especially if participants are using the strategy of following the path from the center towards the correct exit.

Considering the dimensions of the maze, the virtual minimum path length that could complete the maze would be 10 cells if those were casually generated all on a row towards the correct exit side. Regarding the maximum, we would not expect a path to fill all the maze (around 230 cells), but since it is a competition between two simultaneous algorithms, we would expect paths that take a lot to reach the exit to fill as maximum half the maze, which would be not much more than 100 cells.

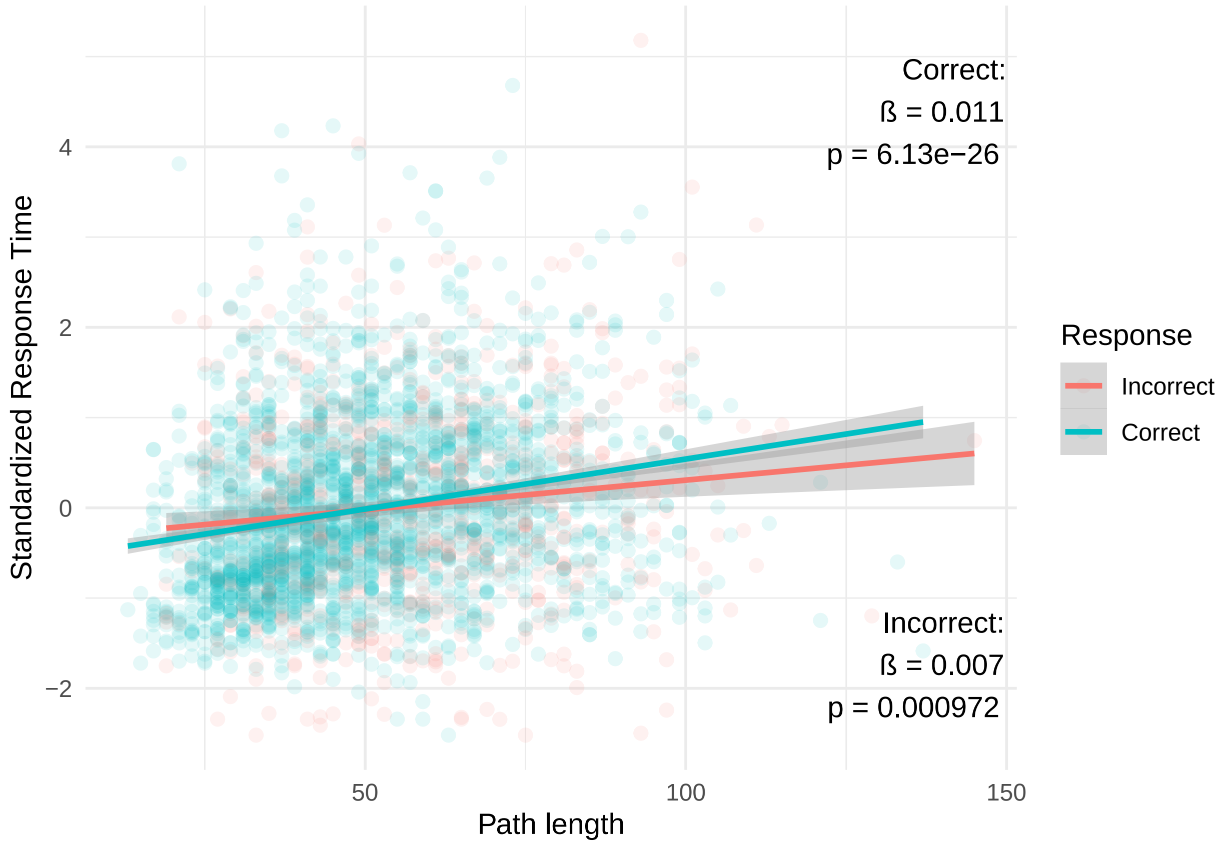
We find a right-skewed unimodal distribution with a peak around 40-45 cells.



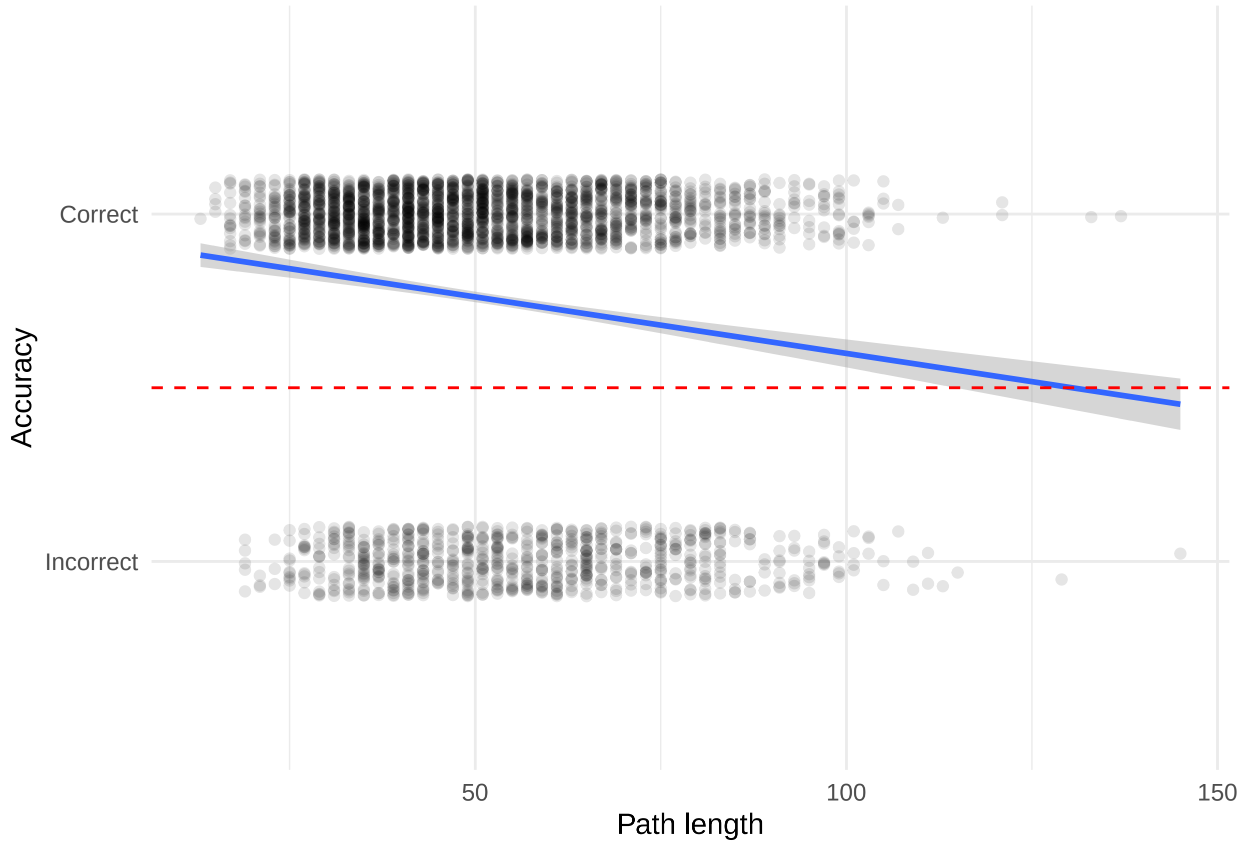
Here we show the most common cells that are part of a maze with the bottom exit. See that it spreads all the maze horizontally and can even go further than the center.



Under the hypothesis that longer paths could induce greater response times, we analyzed the relation between these two variables. We fitted a linear regression model predicting standardized RT from path length and correctness and found a significant positive relation both when responses were correct and incorrect. However, the magnitude of this effect on response times is small.



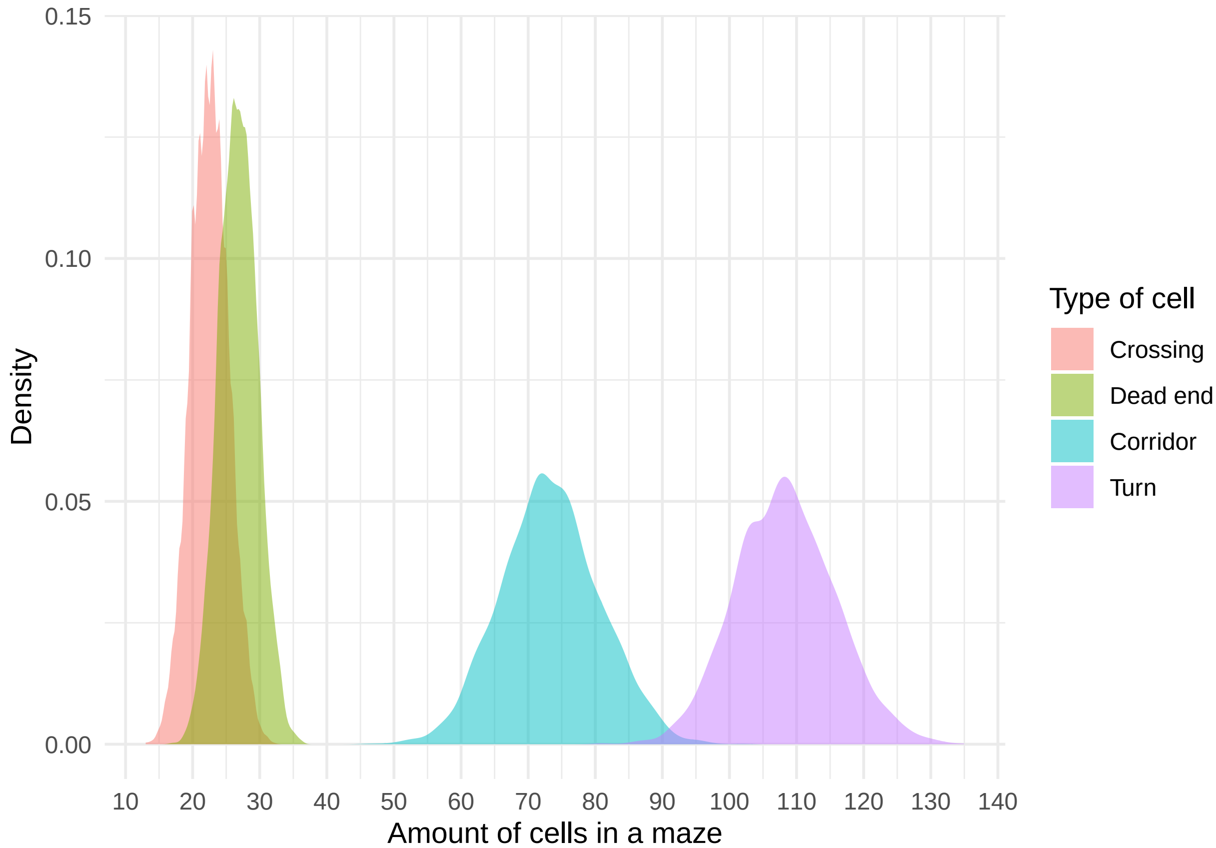
Accuracy on the other hand shows that is clearly affected by path length. We fitted a linear regression and found a significant ($p<.001$) negative relation (slope: -0.0033), meaning that longer paths decrease the probability of delivering a correct response. This proves that the length of the path is a relevant factor to consider difficulty in the game, and could be an element to balance if we want to keep players engaged.



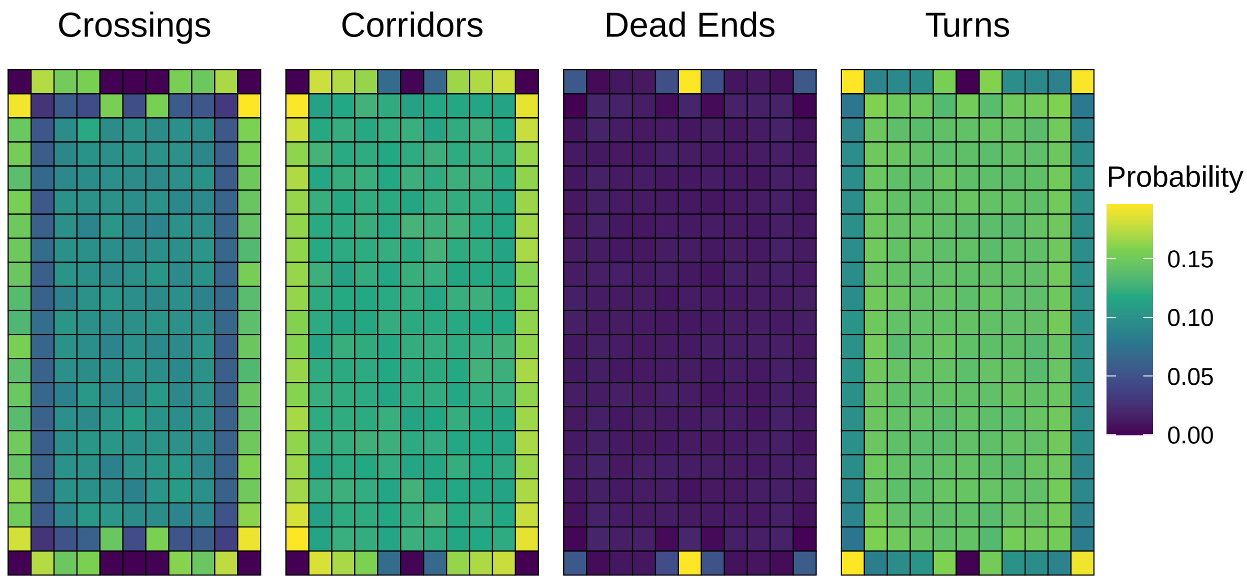
3.3.2 Dead ends and crossings Another element that can help or hinder players' navigation through the maze is the type of cells that are generated.

For example, dead ends might be a source of frustration, where players find out that they followed an erroneous path, making them backtrack their own steps. Crossings require a decision making step to carry on with any path. On the other hand, corridors and turns just need to be followed.

Taking into account that the maze is comprised of 231 cells, we find the most common type to be turns (almost half of the maze), followed by corridors (up to a third of the maze). This means that most part of the maze is usually comprised of easy-to-follow cells, with the amount of dead ends and corridors adding to not much more than 20% of the maze.

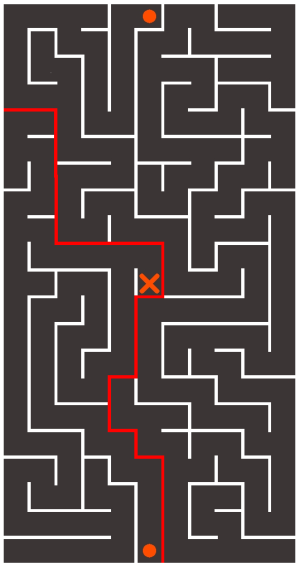


Also, we can show where each type of cell appears more frequently in a maze. Here you can see that crossings and corridors are pretty common around the maze, but especially at the borders. However, corners are in most cases turns and marginally dead ends. Also, turns are the most common type of cell in the center of the maze, and both starting points are always dead ends, so the first step is always towards the de center regarding of the side.



3.3.3 Horizon Up to this point, the elements we described should influence how players solve the game by following the path of the maze. However, there can be other strategies.

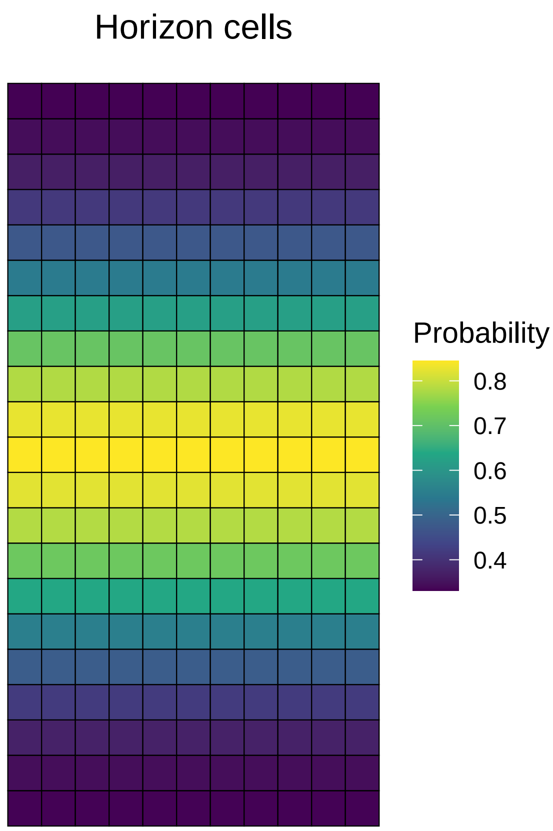
For example, due to the maze generation process, where two simultaneous backtracking algorithms are competing to gain terrain and include the central cell, a natural frontier between the areas "captured" by each algorithm is created. Imagine that this division is right behind the central cell, and is almost horizontal. In this case one could easily separate the valid area, solving the game without needing to follow every step of the path.



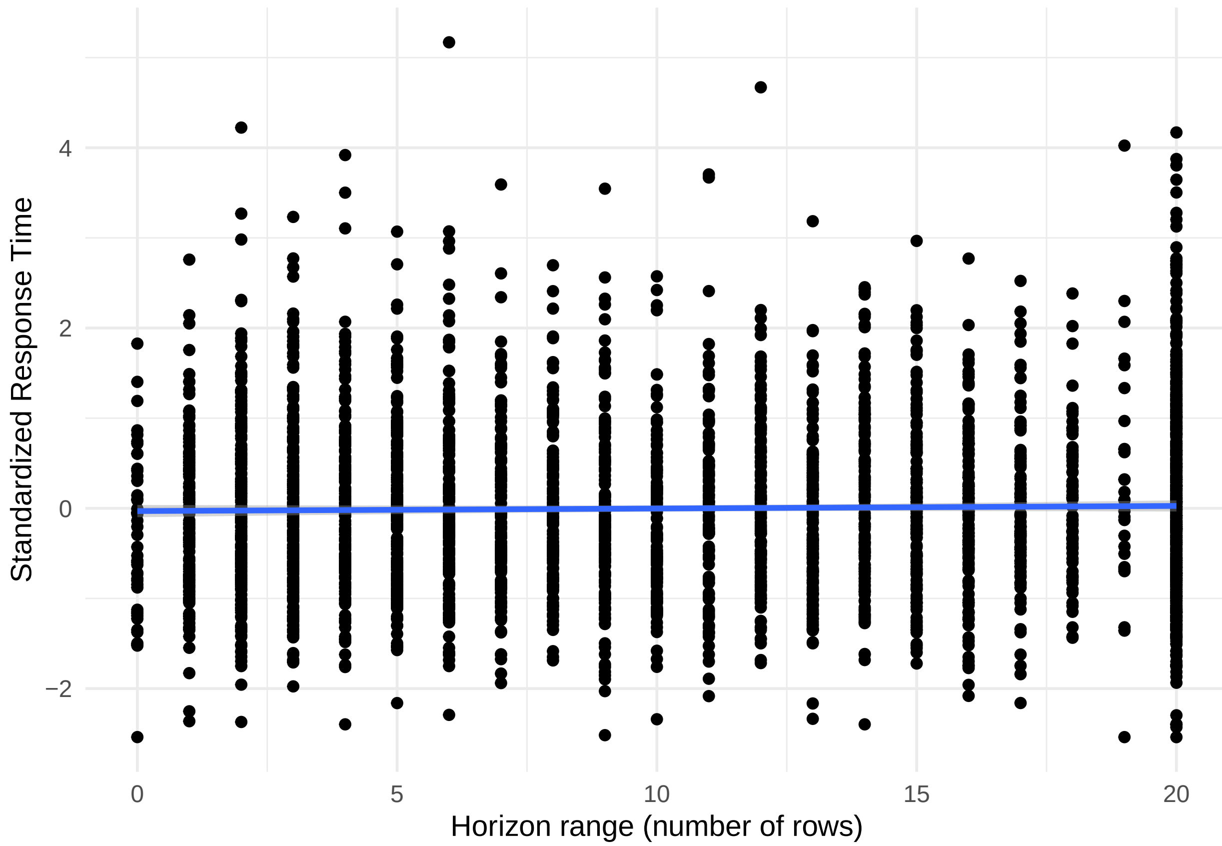
Fortunately, this is only a rare, but possible case. Note that there are many factors that could generate “easier” or “harder” mazes due to the random nature of its generation process.

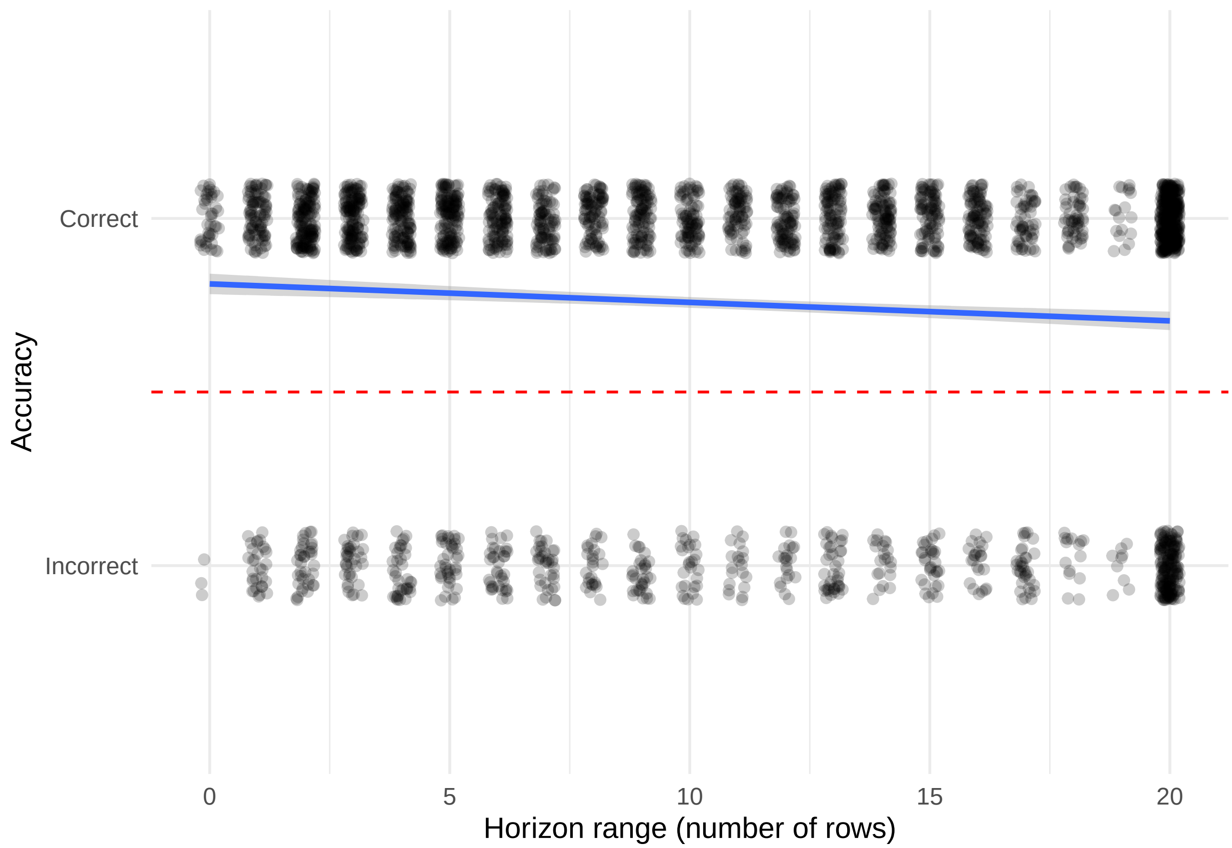
Just to be safe, we can check how centered this frontier usually is (how close to the central cell, therefore more salient) and how spread (vertically) it is around the maze (making it harder to discern).

Overall, the horizon is centered with an average range of 14 cells. However, this can vary on specific mazes, which could influence performance.



For example, we can see that standardized response times are affected by the range of the horizon, however, the accuracy is indeed reduced as the range increases (slope: -0.0053, $p<.001$ from the linear regression model). This means that players are less accurate when the path involves going back and forth in a wider range. Probably these acute turns are what makes them discard that path erroneously.

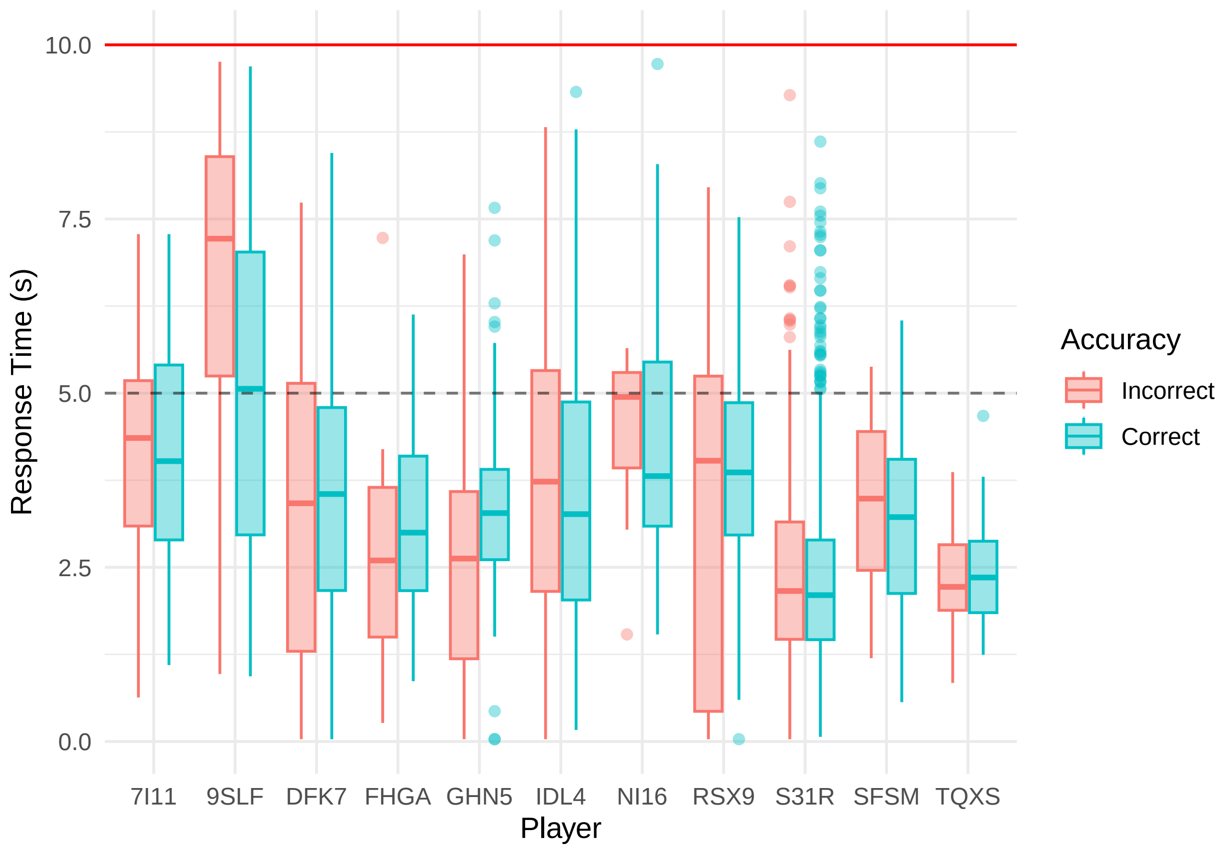




3.4 Behavior Data

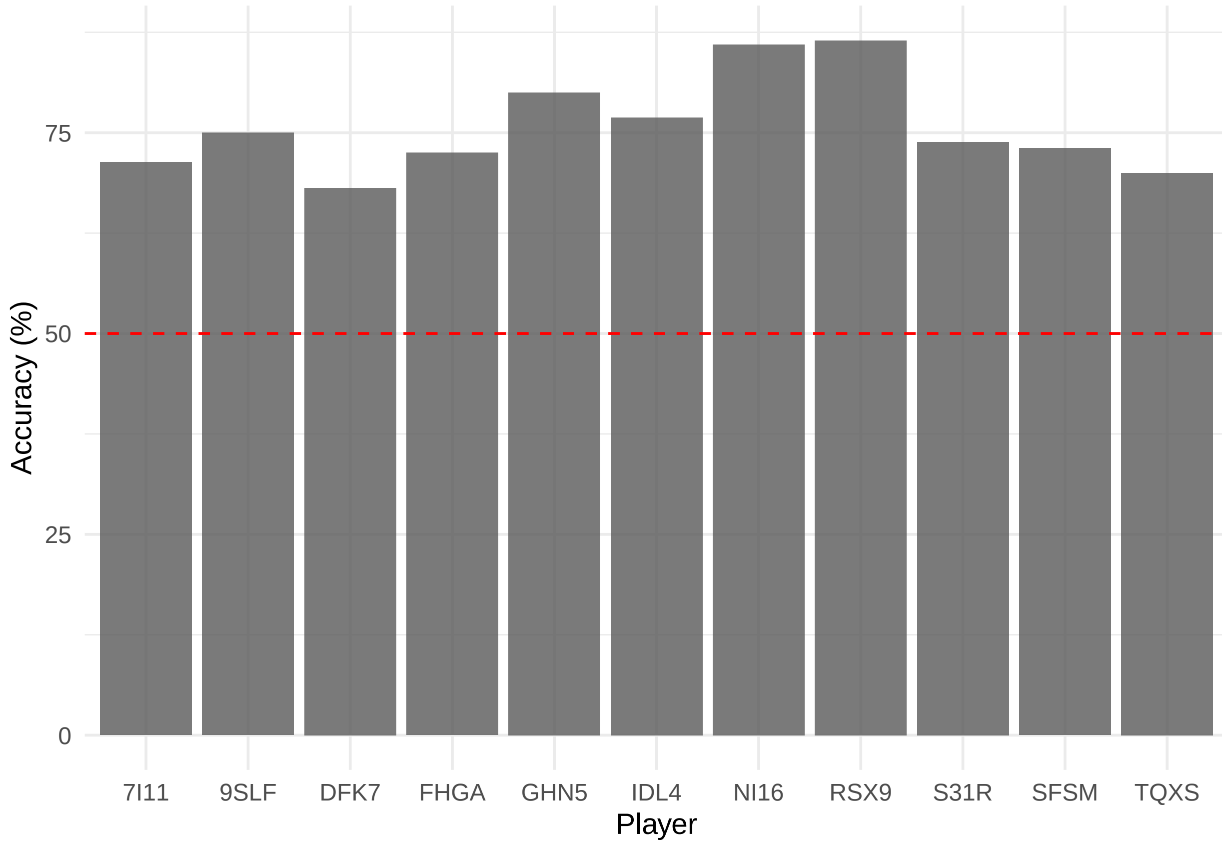
In terms of performance, we have two main metrics: Response Time and Accuracy.

3.4.1 Response Time First, we find that in most cases, players don't take more than half the available time (10 second maximum). Some are more consistent, with higher peaks of the distribution of response times, while others have a wider spread, showing more variability in their responses.



3.4.2 Accuracy First, we check how good player are at detecting which side does the path lead to.

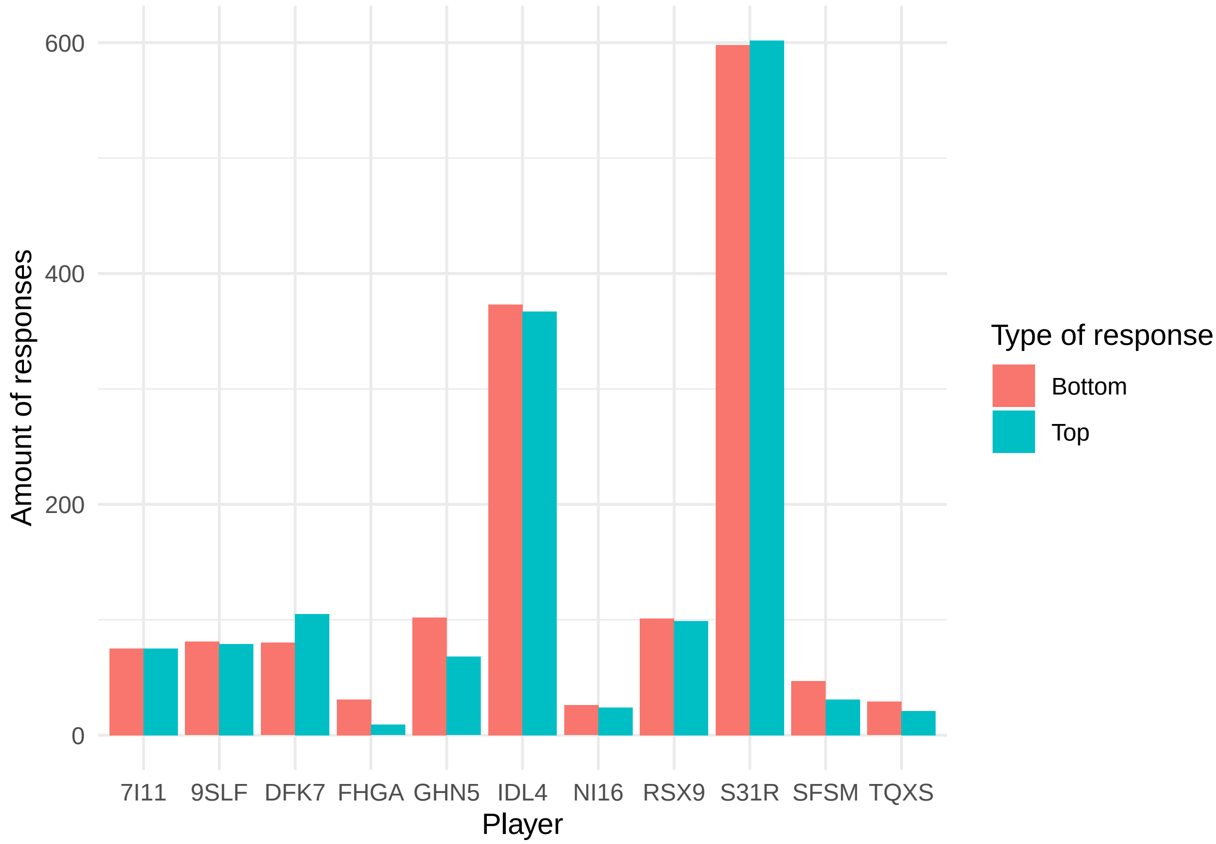
We can see that all players performed well above chance, between 68 and 86 % of times (average 75.7%), without much variability (SD 6.1)



Response Bias Also, a relevant phenomenon in decision paradigms of similar options is that players tend to choose systematically one over the other, regardless of any external information.

Since this is a factor very that is very specific to each individual, we have to look player by player whether the proportion of responses to each side is homogeneous.

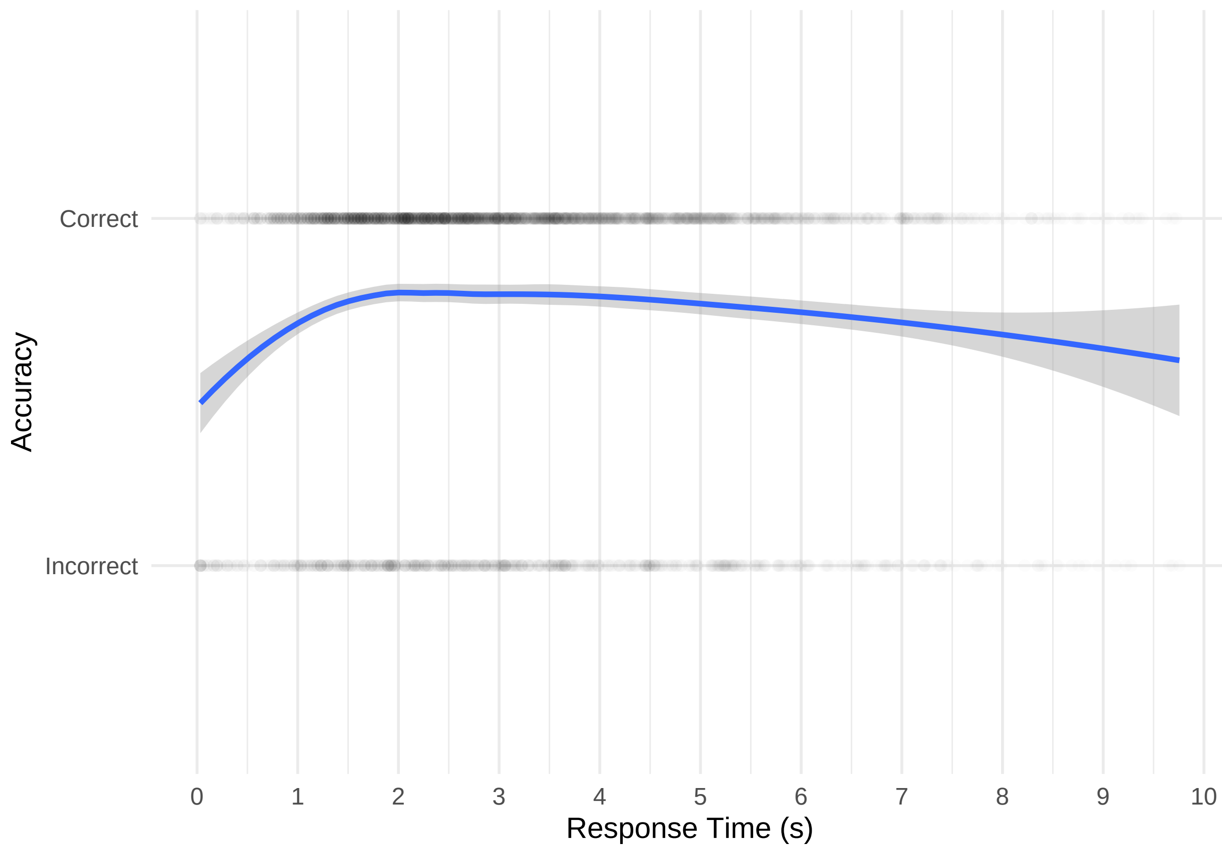
We can see that responses are very homogeneous, and those cases where a greater difference might be observed are also those who played less, so more data would be needed from them to determine if this is really a bias or just due to the random nature of the maze generation process.



Speed-Accuracy Tradeoff Another common phenomenon is that, the quicker the response, the fewer cognitive resources are allocated to the decision, probably leading to more erroneous responses.

What we find here is that quick responses between 0 and 1.5 seconds tend to decrease the probability of correct the shorter they become, while longer response times show a stable probability of correct response until almost the end of the trial.

This tells us that hasty decisions are more probable to produce a negative outcome, and also that 1.5 seconds seems to be enough to increase the probabilities of success without reducing the reward.



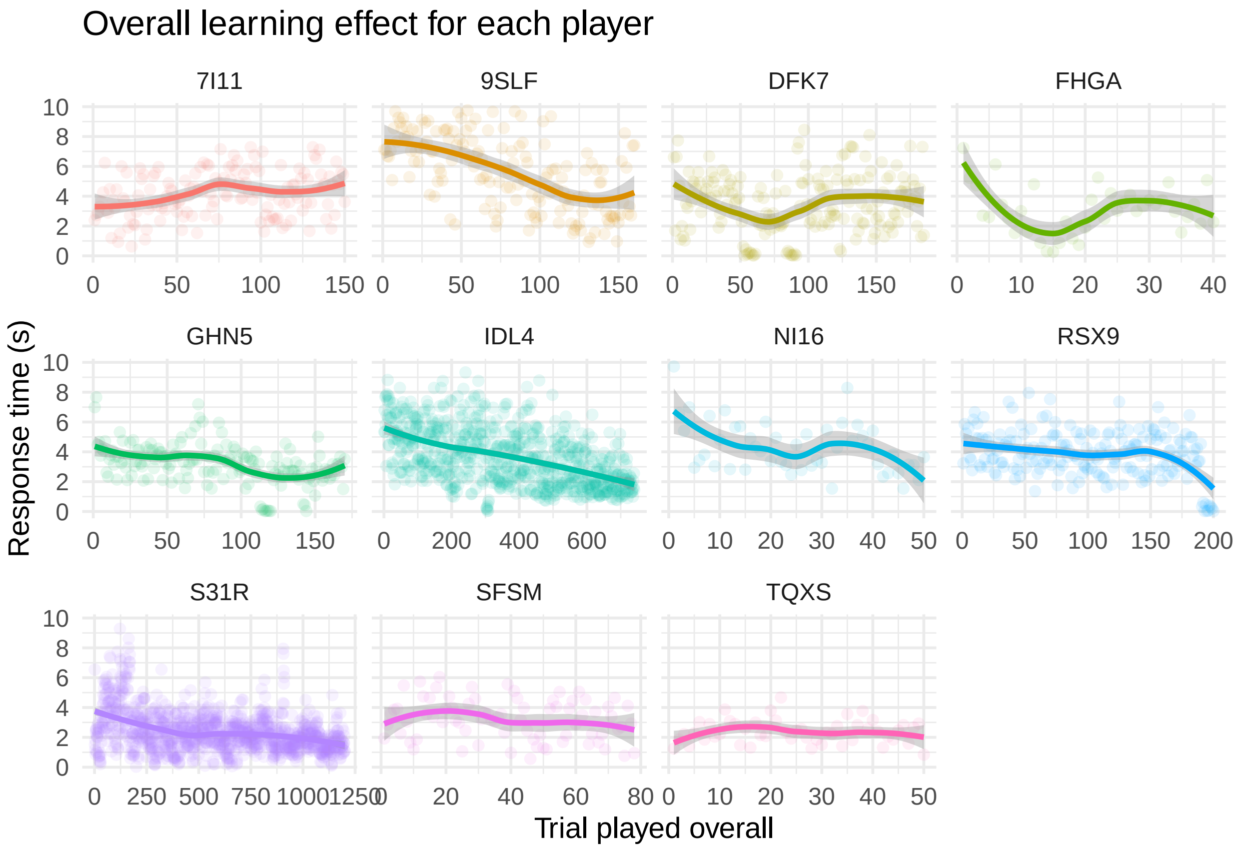
3.4.3 Learning Another factor that we could predict is the learning effect. This means that players, through repeated games, become better at solving the maze, either because their exploration speed increases, because they improve their decision process or because they just get rid of biases or bad strategies.

With this, we can explore learning effects on how their speed increases as well as whether they do more accurate responses as they play in different ranges; within one block, during the same session, or throughout all their gaming experience.

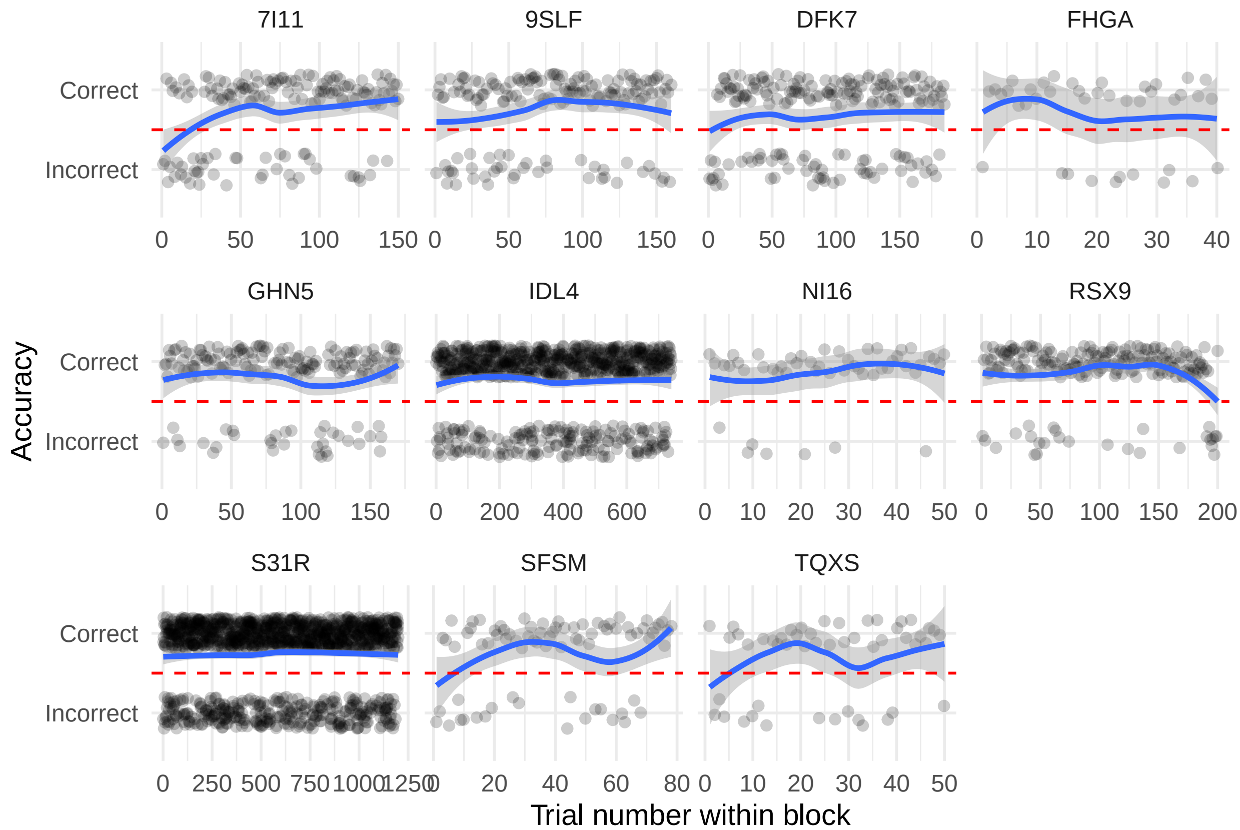
Overall Learning This type of learning relates to how players improve they performance as they play through even days. This represents long-term learning, and could be associated to finding better strategies and getting used to the layout of the mazes and the interface of the game.

Here we can observe in most cases a slight decrease in response time as players accumulate experience with the game. However, we must be cautious with these results, as there is a significant variability in the amount of games played between players.

Regarding accuracy, we see a tendency to increase in some player, but it is not the case with all players, and the effect is quite shallow. There may be a small learning, but not as clear or impactful as in response times.



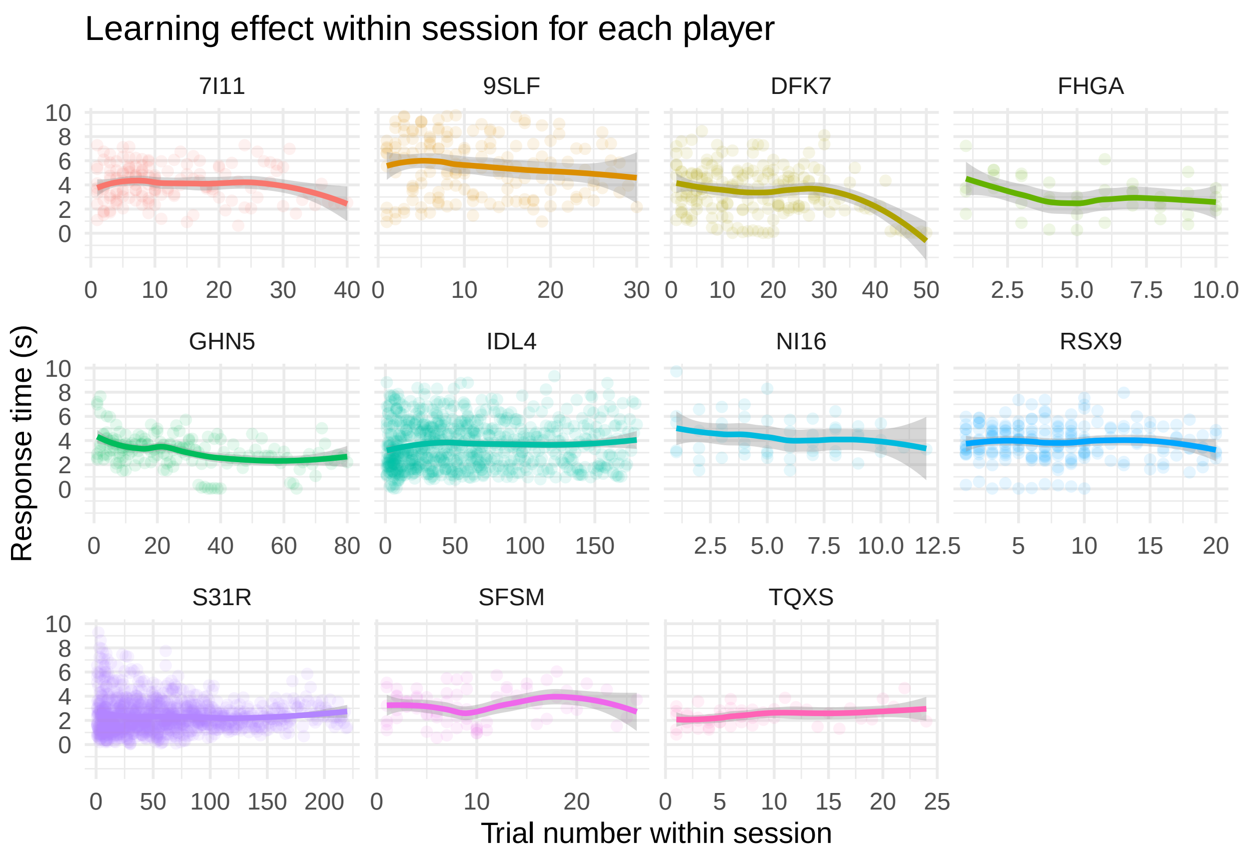
Overall learning effect for each player
7I11
9SLF
DFK7
FHGA
GHN5
IDL4
NI16
RSX9
S31R
SFSM
TQXS
Response time (s)
Trial played overall
10
8
6
4
2
0
0
50
100
150
0
50
100
150
0
50
100
150
0
10
20
30
40
0
50
100
150
0
200
400
600
0
10
20
30
40
50
0
50
100
150
200
0
250
500
750
1000
1250
0
20
40
60
80
0
10
20
30
40
50

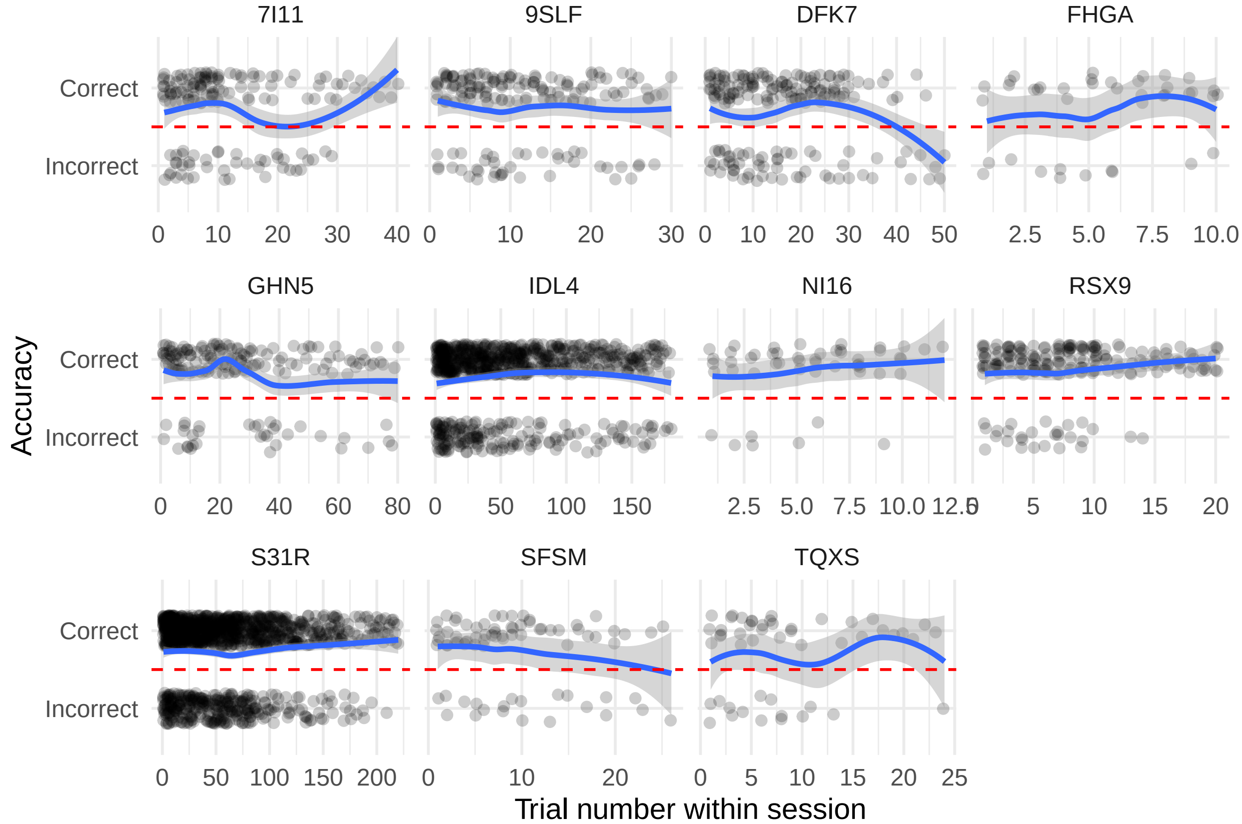


Session Learning This improvement is more associated to a "warming up" effect. Temporary improvements that represents how players evolve from starting to play to having accumulated some games on a row.

Similarly, a small decrease can be detected in the response time, although in most cases it is limited to the earliest trials of the session, probably showing that participants needs a few trials or blocks to get into the mindset to solve the mazes. However, this effect is less prominent that overall learning, and additional data from more players could help determine the consistency of this effect.

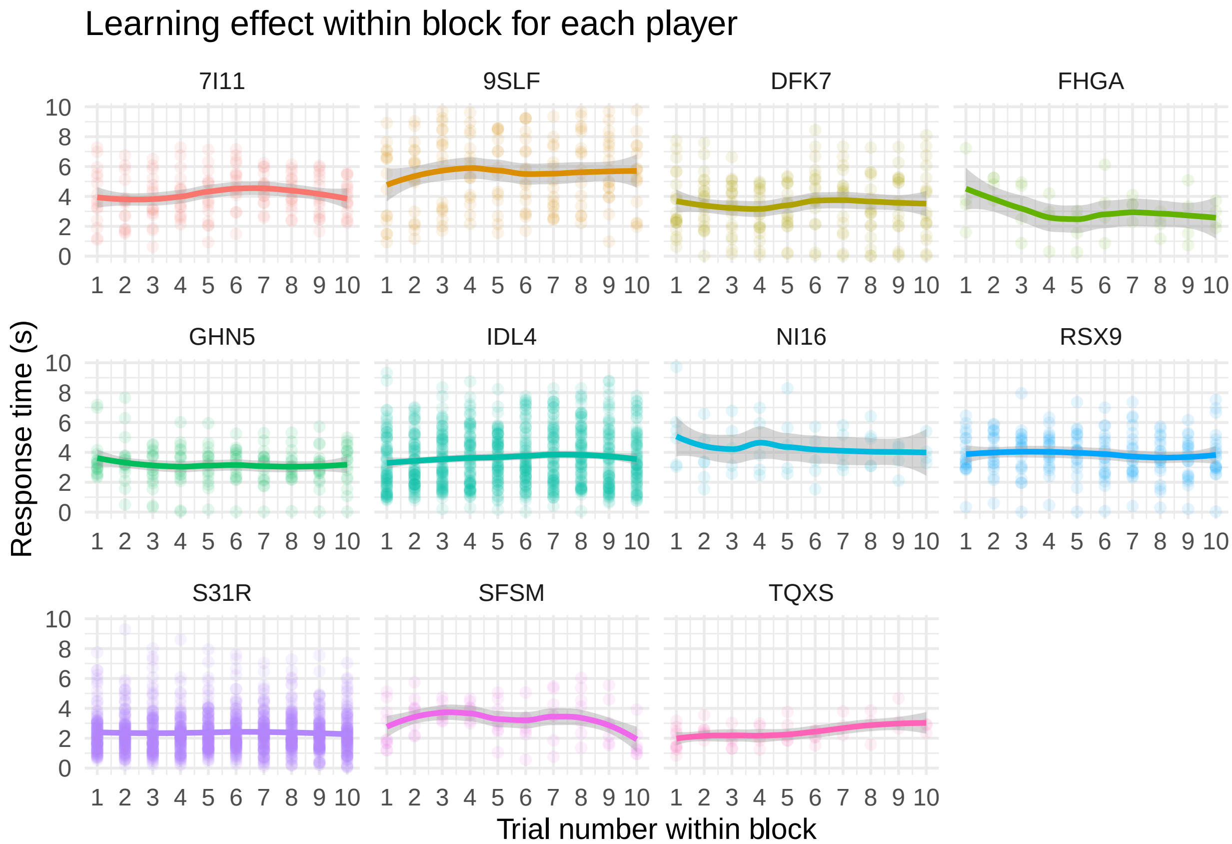
In terms of accuracy, we find different profiles between participants. Considering the lack of more data, the patterns are inconclusive.


Learning effect within session for each player
7I11
9SLF
DFK7
FHGA
GHN5
IDL4
NI16
RSX9
S31R
SFSM
TQXS
Response time (s)
Trial number within session

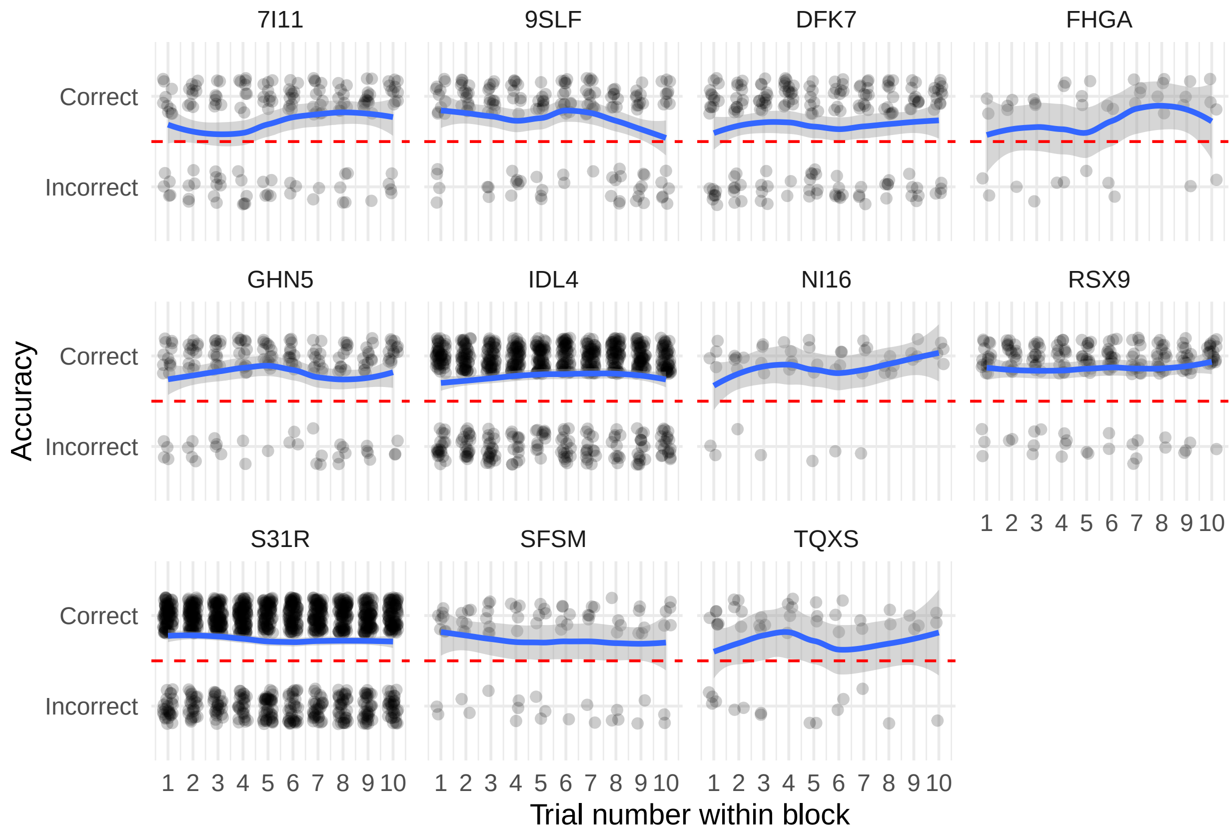


Block Learning Similarly to session learning, block learning is even more limited. Including the trials within the same block we evaluate how the pressure of the scores being accumulated towards the end of the block to be averaged into the final score takes effect in players' performance.

There does not seem to be any consistent pattern between the trial number and the response time or the accuracy, meaning that the time participants need to give an answer is not necessarily different through the course of a block nor are there specific trials where accuracy is improved.



Learning effect within block for each player
7I11
9SLF
DFK7
FHGA
GHN5
IDL4
NI16
RSX9
S31R
SFSM
TQXS
Response time (s)
Trial number within block



Future Analysis

- **Post-error behavior**
  - Effect of mistakes on subsequent RT and accuracy.
  - Serial dependencies.
- **More difficulty metrics**
  - Similarity between current and previous mazes.
  - Path width and center
  - Symmetry of the maze and/or the path
- **Identifying troublesome cells**
  - Detect areas that when being part of the path reduce accuracy of some players (as individual differencies).