Pretending not to know, Hangman: pre-registration document

Matan Mazor1, Ian Phillips2, & Chaz Firestone2

1 University College London

2 Johns Hopkins University

Author note

The authors declare no competing interests

Correspondence concerning this article should be addressed to Matan Mazor, 12 Queen Square, WC1N 3BG London. E-mail: [mtnmzor@gmail.com](mailto:mtnmzor@gmail.com)

Abstract

Generally, what we do depends on what we know. But sometimes we try to appear not to know something that we really do know. Such pretense behavior relies on counterfactual self-simulation — an understanding of how we would behave if our knowledge were different — and so provides an opportunity to investigate how well people can emulate a hypothetical knowledge state. Surprisingly, despite its immediate relevance to metacognition and theory of mind, little research has focused on quantifying pretense accuracy, relative to non-pretense behaviour. In previous experiments, we examined the ability to both produce and detect pretense behaviour using the game “Battleship” — normally played by searching for ships hidden behind cells in a grid. Here, we test the generalizability of our findings to a game where players reveal a hidden word by guessing letters (“Hangman”). In ‘pretend’ blocks, subjects will know what the hidden word is, but will try to play as if they lack this knowledge. Similar to our analysis of Battleship games, we will quantify subjects’ self-simulation ability by comparing their behaviour in pretend and non-pretend games. We will further test players’ ability to detect pretense, and examine the association between pretense ability and pretense-detection ability.

*Keywords:* Pretense, self-simulation, metacognition, theory of mind

*Word count:* 3634

Pretending not to know, Hangman: pre-registration document

# Motivation

The ability to intentionally deceive others relies on a capacity to reason about mental states (Frith & Frith, 2005). This is evident in a similar developmental trajectory for the acquisition of theory of mind and the ability to deceive and detect deception (Shultz & Cloghesy, 1981; Sodian, Taylor, Harris, & Perner, 1991; Wimmer & Perner, 1983), and a similar distribution of deception and theory of mind in the animal kingdom (e.g., Emery & Clayton, 2004; Hall & Brosnan, 2017). This link makes conceptual sense: to deceive others, one needs to understand that others can have different knowledge and beliefs than one’s own. Moreover, deception often involves pretense behaviour, which in turn relies on an ability to simulate and mimic one’s own behaviour under a counterfactual belief state. For example, in order to successfully deceive your friends into thinking that you were surprised by the birthday party they threw for you, it is not sufficient that you are able to reason about their mental states (“I know that they are planning a surprise party, but they don’t know that I know that.”) — you also need to convincingly simulate and mimic your hypothetical behaviour had you not known about the party (“Where would I look first? What would I say? How long would it take me to recover from the surprise?”). This reliance of pretense behaviour on self-simulation makes it an ideal opportunity to examine metacognitive knowledge about one’s own mental states, and the potential reliance of this knowledge on a self-simulation. By comparing non-pretend and pretend behaviour, we can ask which aspects of their cognitive processes subjects can simulate, and which aspects are not represented in their mental models of their own cognition.

In Exp. 1 (N=100, exploratory) and 2 (N=500, pre-registered), we examined pretense behaviour using an online version of the game “Battleship”. We quantified pretense quality by comparing patterns of cell selections in pretend and non-pretend games. We observed a striking ability to simulate ignorance: pretenders made rational cell selections given the limited knowledge state they simulated, and the timing of their selections qualitatively matched click-time patterns found in standard, non-pretend games. In a final ‘judge’ block, subjects were unable to identify pretend games in a two-alternative forced choice discrimination task, and the accuracy of their judgments was uncorrelated with the quality of their pretense.

In Exp. 3 (N=100, exploratory), we tested how general this pretense capacity is by comparing pretend and non-pretend games in an online version of the game “Hangman”. In Exp. 4, we will replicate and extend our results in a larger sample (N=500).

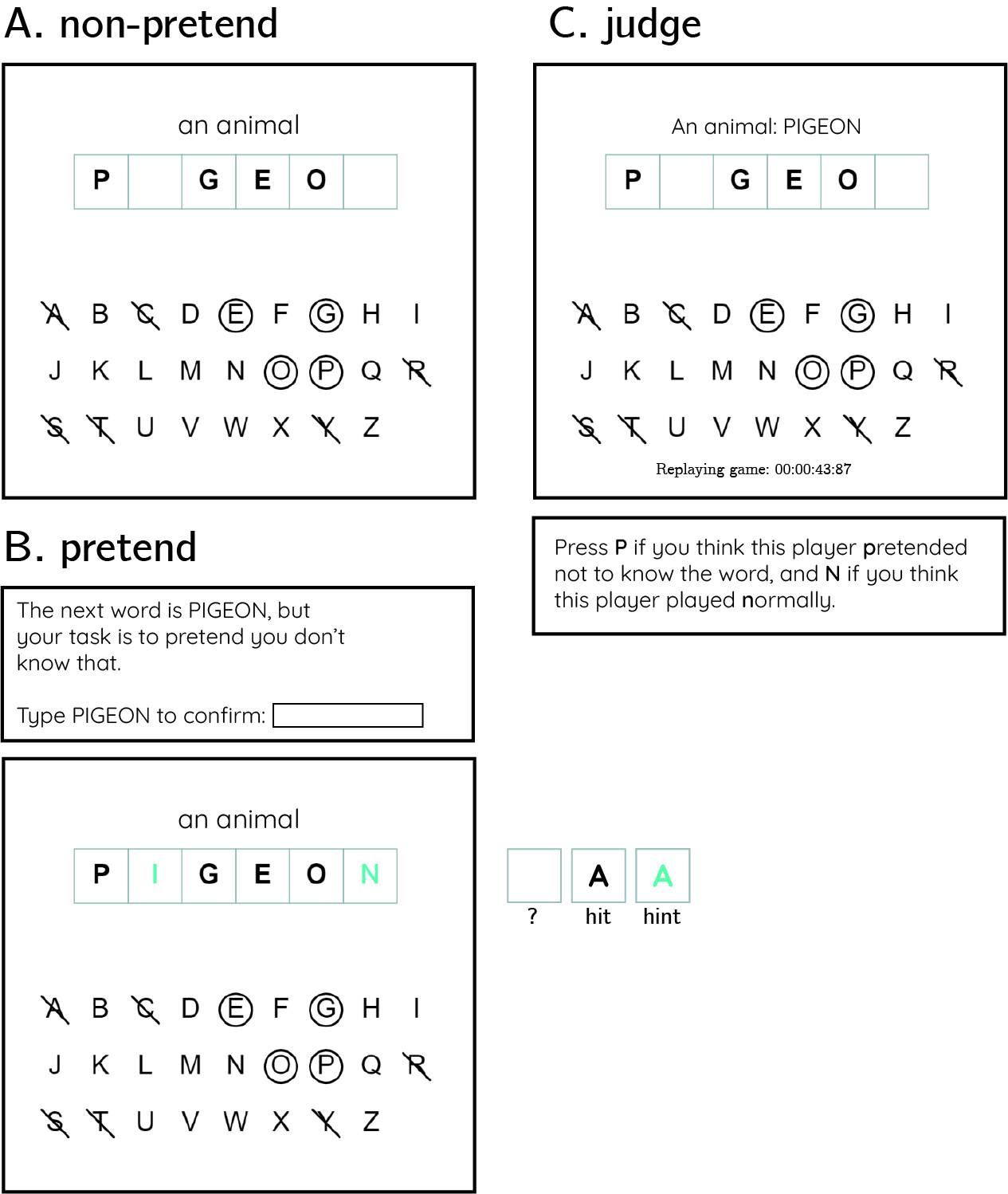
# Methods

We report how we determined our sample size, all data exclusions (if any), all manipulations, and all measures in the study.

## Participants

The research complies with all relevant ethical regulations and was approved by the Institutional Review Board of Johns Hopkins University. Participants will be recruited via Prolific and will give informed consent prior to their participation. They will be selected based on their acceptance rate (>95%), being native English speakers, and being located in the USA. We will collect data until we reach 500 included participants. The entire experiment will take 20 minutes to complete (median completion time in our pilot data: 19 minutes). Participants will be paid 3.15 USD for their participation, equivalent to an hourly wage of approximately 9.50 USD, in addition to a bonus payment.

## Procedure



*Figure* *1.*  Experimental Design: players will play 12 games of the game Hangman in two conditions presented in two blocks of 6 games. The order of blocks will be randomized between players. A: In non-pretend games, players will start the game not knowing the identity of the hidden word, and try to reveal the word with as few letter guesses as possible. B: In pretend games, players will be told the identity of the word, and their task will be to play the game as if they do not have this information. Before pretending, players will verify they had read the hidden word by typing it. C: Lastly, in 5 judge trials players will observe a replay of the game of a previous player, and try to tell whether the player was pretending or not.

The first instructions screen will inform participants that the experiment, based on the game Hangman, has three parts, and that the points they will accumulate will translate to a monetary bonus payment. They will then be presented with a leaderboard of previous players. Then, the rules of the game will be presented:

In the following game, your task is to reveal a hidden word or phrase by guessing letters. What makes the game difficult is that you can’t see the word; all you can see is a row of squares - a square for each letter. Use your mouse to make letter guesses. We will have five types of words: body parts, numbers, US states, fruit, and famous people. You will start each game with 15 points and lose one point for every guess of a letter that is not in the word.

We will then explain that “the words in this game are the kind of words that will be familiar to most English-speaking fifth-graders. We didn’t pick any strange or particularly difficult words”.

Once they respond correctly to a multi-choice comprehension question (“the goal of the game is to…”: “reveal the word with as few letter guesses as possible”), participants will play a practice round, revealing the word PIGEON (see Fig. 1A).

After the main instructions, comprehension question and practice round, participants will complete one pretend and one non-pretend block, each followed by one half-game (see below for details). The order of pretend and non-pretend blocks will be counterbalanced between participants. Each block will comprise five games played with five out of ten different words, and one half-game. The allocation of words to conditions will be randomized between participants, with the constraint that both pretend and non-pretend blocks will include exactly one word from each category. The order of words within a block will be randomized, except for the half-game, which will always be delivered at the end.

The ten words will include two number words (ELEVEN, NINETY SIX), two famous people (DALAI LAMA, TAYLOR SWIFT), two fruits (STRAWBERRY, LEMON), two body parts (TOOTH, HEAD), and two US states (MONTANA, IOWA).

### Non-pretend games.

In non-pretend games, participants will reveal a hidden word with as few letter guesses as possible. An online counter of the number of points will be displayed on the screen, deducting one point for every guess of a letter that is not in the target word. After each game, feedback will be given about the number of points obtained.

After completing the five games, participants will perform one half-game (see below for details).

### Pretend games.

Participants will be given the following instructions:

In the next part of the experiment, you’ll play 6 games where you reveal a hidden word by guessing letters.

However, this time your goal is different.

In this round, we’re going to tell you the word in advance, but **we want you to act like you don’t know this information**.

To see how good you are at this, we’re going to compare your games to the games of people who played normally, without knowing what the word was, and see how similar they are. We will measure which letters you click and the timing of your guesses; if your clicks look similar to people who played like normal (trying to reveal the word with as few guesses as possible, but without any hints), you’ll get bonus points. But if your games look different, you won’t get these bonus points. Your number of clicks in this part will not affect your bonus. Only your ability to play like you didn’t see the word in advance.

After one practice round, pretending not to know that the hidden word is PIGEON, and one comprehension question (“In this part of the experiment my goal is to…”: “play the game as if I don’t know what the word is so that I look like someone who had no hints”), participants will play five pretend games. Each game will be preceded by a short message informing subjects about the identity of the target word. To start pretending, players will be asked to type in the target word on their keyboard. After pretending, we will remind players that a game that looks similar to the games of participants who had no hints will be awarded 10 bonus points.

After completing the five games, participants will perform one half-game (see below for details).

### Half games.

In order to directly compare participants’ pretend and non-pretend games for identical belief states (true or pretended knowledge about the identity of the word), we will ask participants to also complete one pretend and one non-pretend game, given a partly finished game with some letters already guessed (they will be told that the computer made these guesses). The two half-game words will be one fruit: PAPAYA or BANANA, with guessed letters [A, E, I, O, M, T], and one body part: HAND, or HAIR with guessed letters (A, E, O, M, T, H, P). The assignment of category (fruit or body part) to condition (pretend and non-pretend), as well as the identity of the target word within each category (e.g., PAPAYA or BANANA), will be randomized between participants.

Instructions for the non-pretend half-game will be:

For the next game, the computer chose the first letters for you; you can take over from where it left off. Your challenge is to complete the game. Just like in the previous games, here also you will lose one point for each letter that you guess and is not in the word.

Instructions for the pretend half-game will be:

For the next game, the computer chose the first letters for you; you can take over from where it left off. Just like in the previous games, here also you will know what the word is, but your bonus points will depend on your ability to play as if you didn’t know the word.

### Judge trials.

In the final part of the experiment, participants will observe five games of previous players and determine who had hints and who didn’t. Instructions for this part are:

In this third and last part of the experiment, we ask you to be a judge for previous players, and see if you can tell which of the players were shown the word (but acted like they weren’t). We will show you 5 replays of the games of previous players. Your task is to decide whether they played normally or pretended. For each game that you get right, you will receive 10 points. Good luck!

Then, on each judge trial, one game of a previous player will be replayed in real time, with the target word presented above. For non-pretend games, only games from the group of participants that pretended in the second block (and played normally in the first block) will be chosen for presentation in this part. For both pretend and non-pretend games, only games shorter than 1.5 minutes will be presented. Judge participants will indicate their decision by pressing the P and N keys on their keyboard. After making a decision, participants will be informed whether they will receive the 10 points. Whenever a pretend game is classified as a non-pretend game, they will be informed that the pretender will receive these 10 points instead of them.

Lastly, participants will be asked the following debrief questions:

Did you have a strategy that you used for pretending you did not see the word? What was most difficult about pretending? How about telling between players who pretenders and who played for real - did you have a strategy for that?

And:

We would appreciate it if you could share any thoughts you had about the experiment, or anything we should take into account when analyzing your data.

### Randomization.

The order and timing of experimental events will be determined pseudo-randomly by the Mersenne Twister pseudorandom number generator, initialized in a way that ensures registration time-locking (Mazor, Mazor, & Mukamel, 2019).

## Data analysis

### Rejection criteria.

Participants will be excluded if all five of their pretend games were all hits. Unlike Battleship, here we decided not to exclude pretend games that were all hits. In Hangman, getting a perfect score is possible even when not pretending, as was the case in 27% of all non-pretend games in exploratory Exp. 3.

### Data preprocessing.

Data will not be preprocessed in any way.

### Hypotheses and analysis plan.

This study is designed to explore subjects’ capacity for self-simulation under a counterfactual knowledge state, and the limits of this capacity. In particular, we are interested in comparing non-pretend and pretend games and identifying similarities and differences. We will also assess subjects’ capacity to discriminate between pretend and non-pretend games, and how this capacity relates to their ability to pretend.

Unless otherwise specified, all hypotheses will be tested using a repeated measures t-test, with a significance level of 0.05.

All hypotheses will be tested separately for the group of participants that pretended in the first and in the second block. If results for the two groups diverge, analysis will be repeated on the first block of all participants (pretend games of pretend-first participants and non-pretend games of pretend-second participants), in a between-subjects manner, and results will be interpreted in light of this analysis.

*Hypothesis 1 (game duration)*: We will test the null hypothesis that pretend games take as long as non-pretend games. For each participant, the difference between median game duration in pretend and non-pretend games will be extracted. Subject-wise differences will then be carried over to a t-test at the group level.

*Hypothesis 2 (first click latency)*: We will test the null hypothesis that the first click in pretend games takes as long to execute as in non-pretend games. For each participant, a difference between median first click latency in pretend and non-pretend games will be extracted. Subject-wise differences will then be carried over to a t-test at the group level.

*Hypothesis 3 (number of misses)*: We will test the null hypothesis that pretend games are as long, in terms of total number of letter misses, as non-pretend games. For each participant, we will extract the difference between the median number of letter selections that do not appear in the target word in pretend and non-pretend games. Subject-wise differences will then be carried over to a t-test at the group level.

*Hypotheses 4-6 (click latency by click outcome)*: Letter selections will be classified based on whether they resulted in revealing a letter that appears in the target word (*hit*) or not (*miss*). We will then contrast median click latency between hits and misses separately for pretend and non-pretend games. We will test for a difference in selection latency as a function of outcome on pretend games, on non-pretend games, and for an interaction effect between click outcome (hit or miss) and condition (pretend or non-pretend) on click latency. The same analysis will then be repeated by classifying letter selections on the basis of whether the previous selection resulted in a hit or a miss (*Hypothesis 5*), and whether the next selection will result in a hit or a miss (*Hypothesis 6*).

#### Model based analysis.

The following analyses are designed to test for differences in game optimality between pretend and non-pretend games, and for a relationship between decision difficulty and click latency. To do so, we approximate optimal behaviour by approximating the posterior probability that a letter appears in the word, given available information. Critically, in modelling pretend games we do not treat hints as part of this available information, because an optimal player should ignore this information in choosing where to click next. Given this posterior, a rational player should choose letters that have a high posterior probability of appearing in the word.

To approximate the posterior probability of letters given a game state, we will follow the following procedure:

1. We will use the category information (e.g., ‘a fruit’), to obtain a probability-weighted list of category-compatible words (or names, in the case of famous people). The lists were obtained in the following way: for US states and number names, we used an exhaustive list (in the case of numbers, of numbers of 1-2 words), fruit names were taken from Wikipedia ([simple.wikipedia.org/wiki/List\_of\_fruits](https://simple.wikipedia.org/wiki/List_of_fruits)), famous people names from a [crowdsourced document](https://docs.google.com/spreadsheets/d/1mWez4KH0wueV_SSzdN16MMUG4qCa_nFFHToJcUvtJIA/edit#gid=4), and body-part words from existing prototypicality norms (Uyeda & Mandler, 1980). We extended Uyeda and Mandler’s list of body part words by adding all body part words commonly appearing in lists on the internet. In the case of famous people, names were given prior probabilities in proportion to the number of visits their Wikipedia entries received in 2021. The top 100 most popular entries were given a prior probability twice that of the next 100, three times that of the next 100, and so forth. All entries from 600 and on were given the same prior probability, seven times smaller than that of entries at the top 100 positions. In the cases of fruits and body parts, we used prototypicality norms from Uyeda and Mandler (1980) to assign higher prior probability to more prototypical words (mapped to lower numbers on the 1-7 scale used by Uyeda and Mandler). Words that were not included in the norm were given the maximum score (that is, the lowest perceived prototypicality), of 7. Body parts and fruits were included in both singular and plural forms. The plural forms were assigned a prototypicality score of 100, due to the fact that category names were presented in the singular form (‘a fruit’ and ‘a body part’). We then took the reciprocal for each score, and normalized it by the total sum to get a probability distribution over words . The full prior distributions for each category are included in this pre-registration.
2. The likelihood of a game state given a target word equals 0 when the word is inconsistent with the information available to the player (this includes the word length, the identity of letters that do not appear in the word based on previous guesses, and the identity and position of letters that do appear in the word based on previous guesses). When consistent, the likelihood is a non-zero quantity that is equal for all consistent words. Bayes rule will be used to extract the posterior over words given game state . The full non-zero distributions for the initial states of all games are included in this pre-registration.
3. The probability that an individual letter appears in the target word is the sum of posterior probabilities of words that contain this letter where returns 1 if appears in and 0 otherwise.
4. A posterior probability over letter selections will be obtained by dividing the probabilities of individual letters by their total sum .

To quantify this notion of optimality, before each cell selection we will compute the posterior probability that each of the unclicked letters appears in the target word, given the game state. Then, we will rank letters from highest to lowest according to their posterior probability and record the rank of the chosen cell.

*Hypothesis 7 (game optimality):* We will test the null hypothesis that pretend games are as optimal as non-pretend games. Mean click optimality will be contrasted between the pretend and non-pretend games of each player. Subject-wise differences will then be carried over to a t-test at the group level.

*Hypothesis 7B (letter frequency heuristic):* We will repeat the game optimality analysis, but this time using the ranked frequency of the selected letter in the English language (from high to low: E, T, A, O, I, N, S, H, R, D, L, C, U, M, W, F, G, Y, P, B, V, K, J, X, Q, Z, obtained from Wikipedia). We will test the null hypothesis that letters selected in pretend games are of similar frequency to those selected in non-pretend games. Mean ranked probability will be contrasted between the pretend and non-pretend games of each player. Subject-wise differences will then be carried over to a t-test at the group level.

*Hypothesis 8 (entropy-RT relation):* The entropy of the posterior map will be extracted before each cell selection using the following formula: . We will test for within-subject associations between click latency and the linear and quadratic expansions of posterior entropy in pretend and in non-pretend games by fitting a multiple regression model to the data of each subject. We will contrast subject-wise beta values for pretend and non-pretend games to test the null hypothesis that the association between posterior entropy and click latency is the same in both conditions.

#### Judge accuracy.

*Hypothesis 9 (judge accuracy):* Judge accuracy will be quantified as the number of correct guesses in the judge block, divided by 5. This quantity will be extracted separately for each subject and contrasted against chance (50%) at the group level.

*Hypothesis 9B (judge bias):* Judge bias will be quantified as the number of ‘pretend’ guesses in the judge block, divided by 5. This quantity will be extracted separately for each subject and contrasted against chance (50%) at the group level.

*Hypothesis 10 (judge-pretend relation):* Pretense quality will be quantified as the number of pretend games that were misclassified by other participants as non-pretend games, divided by the number of pretend games that were presented to other participants. Due to randomization, this denominator is different for different participants. We will test for a correlation between pretense quality and judge accuracy at the group level. To maximize statistical power, this hypothesis will be tested on the pooled data from both groups of participants.

#### Half games.

Our main focus in analyzing half-games will be the identity of the first letter selection. We will ask whether the distribution over letter selections is different in pretend and non-pretend games, and within pretend games, between the two words of each category.

*Hypothesis 12 (irrational letter selections):* Irrational letter selections are ones that are inconsistent with the current game state, such that . We will use a binomial test to test the null hypothesis that the group-level proportion of irrational letter selections is similar in pretend and non-pretend games.

*Hypothesis 13 (target word effect):* For this analysis, we will focus on half-games where the first letter selection was not irrational, and ask whether knowledge of the target word had an attraction or repulsion effect (or no effect at all) on letter selection. Within each word pair (HAIR0 and HAND, PAPAYA and BANANA), we will contrast the proportion of letter selections that are consistent with the first target word (I and R for HAIR, P and Y for PAPAYA) or the second target word (N and D for HAND, B and A for BANANA) in pretend against non-pretend games, and within pretend games, as a function of the target word. For example, we will compare the proportion of non-pretenders who clicked on B or N among those who clicked on B,N,P or Y when encountering \_A\_A\_A, and test whether this proportion is similar in pretend games for the same game state \_A\_A\_A, and within pretend-games, whether it is different as a function of the target word (PAPAYA or BANANA).

## Power calculations and sample size justification

With approximately 250 participants in each group (allocation to group will be determined pseudorandomly), we will have 95% statistical power to detect an effect size of 0.23 standard deviations in a repeated measures t-test (Hypotheses 1-9). In case of a conflict between the results of the two groups, a between-subjects comparison will be performed, using data from the first block only (pretend games from the pretend-first group and non-pretend games from the pretend-second group). For this between-subjects t-test, we will have 95% statistical power to detect an effect size of 0.32 standard deviations.

With 500 participants in both groups combined, we will have 95% statistical power to detect a Pearson correlation of 0.16 (Hypothesis 10).

# References

Emery, N. J., & Clayton, N. S. (2004). The mentality of crows: Convergent evolution of intelligence in corvids and apes. *Science*, *306*(5703), 1903–1907.

Frith, C., & Frith, U. (2005). Theory of mind. *Current Biology*, *15*(17), R644–R645.

Hall, K., & Brosnan, S. F. (2017). Cooperation and deception in primates. *Infant Behavior and Development*, *48*, 38–44.

Mazor, M., Mazor, N., & Mukamel, R. (2019). A novel tool for time-locking study plans to results. *European Journal of Neuroscience*, *49*(9), 1149–1156.

Shultz, T. R., & Cloghesy, K. (1981). Development of recursive awareness of intention. *Developmental Psychology*, *17*(4), 465.

Sodian, B., Taylor, C., Harris, P. L., & Perner, J. (1991). Early deception and the child’s theory of mind: False trails and genuine markers. *Child Development*, *62*(3), 468–483.

Uyeda, K. M., & Mandler, G. (1980). Prototypicality norms for 28 semantic categories. *Behavior Research Methods & Instrumentation*, *12*(6), 587–595.

Wimmer, H., & Perner, J. (1983). Beliefs about beliefs: Representation and constraining function of wrong beliefs in young children’s understanding of deception. *Cognition*, *13*(1), 103–128.