Metacognitive contributions to search termination: pre-registration document

Matan Mazor1 & Stephen M. Fleming1,2,3

1 Wellcome Centre for Human Neuroimaging, UCL

2 Max Planck UCL Centre for Computational Psychiatry and Ageing Research

3 Department of Experimental Psychology, UCL

Author note

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

Abstract

When searching for a target object among distractors, a robust finding in cognitive science is that search time varies as a function of the number of distractors and their similarity to the target object. This is true not only for the time taken to say a target is present, but also for the time taken to terminate a search and conclude that a target object is absent. However, while models of visual search have successfully characterized perceptual and cognitive contributions to target detection, the mechanisms underpinning when to terminate a search and conclude that a target is absent are poorly understood. In other words, when do people decide to give up and conclude that nothing is there? One hypothesis is that efficient search termination involves metacognitive knowledge such as the expected difficulty of the search, target salience, or fluctuations in one’s attentional state. By focusing on the first few trials in a visual search task, here we control participants’ ability to base their search termination on metacognitive knowledge before engaging with the task, and the effect of experience on the shaping of this knowledge. The results of this experiment will provide foundational information about the latent metacognitive variables contributing to search termination, and refine our understanding of how people judge the absence of stimuli.

Metacognitive contributions to search termination: pre-registration document

# Experiment 1

# Methods

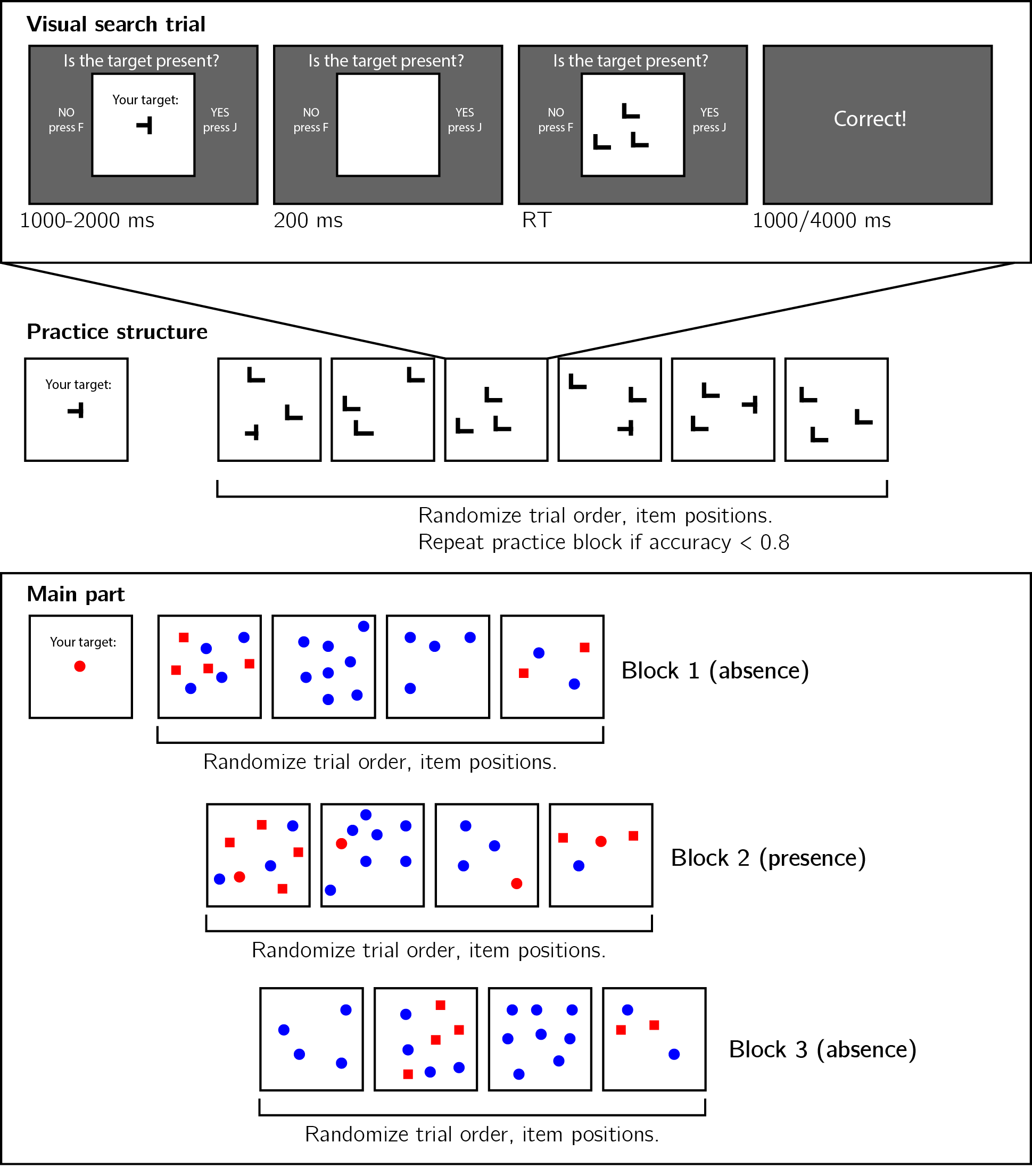
We report how we determined our sample size, all data exclusions (if any), all manipulations, and all measures in the study.

## Participants

The research complied with all relevant ethical regulations, and was approved by the Research Ethics Committee of University College London (study ID number 1260/003). 1187 Participants were recruited via Prolific, and gave their informed consent prior to their participation. They were selected based on their acceptance rate (>95%) and for being native English speakers. We collected data until we reached 320 included participants for each of our hypotheses (after applying our pre-registered exclusion criteria). The entire experiment took around 3 minutes to complete (median completion time in our pilot data: 3:06 minutes). Participants were paid £0.38 for their participation, equivalent to an hourly wage of £7.60.

## Procedure

Participants were first instructed about the visual search task. Specifically, that their task is to report, as accurately and quickly as possible, whether a target stimulus was present (press ‘J’) or absent (press ‘F’). Then, practice trials were delivered, in which the target stimulus was a rotated *T*, and distractors are rotated *L*s. The purpose of the practice trials was to familiarize participants with the structure of the task. For these practice trials the number of items was always 3. Practice trials were delivered in small blocks of 6 trials each, and the main part of the experiment started only once participants responded correctly on at least five trials in a block (see Figure 1).



*Figure* *1:*. Experimental design. Top panel: each visual search trial started with a screen indicating the target stimulus. The search display remained visible until a response is recorded. To motivate accurate responses, the feedback screen remained visible for one second following correct responses and for four seconds following errors. Middle panel: after reading the instructions, participants practiced the visual search task in blocks of 6 trials, until they had reached an accuracy level of 83% correct or higher (at most one error per block of 6 trials). Bottom panel: the main part of the experiment comprised 12 trials only, in which the target was a red circle. Unbeknown the subjects, only trials 5-8 (Block 2) were target-present trials, and the remaining trials were target-absent trials. Each 4-trial block followed a 2 by 2 design, with factors being set size (4 or 8) and distractor type (color or conjunction; blue circles only or blue circles and red squares, respectively).

In the main part of the experiment, participants searched for a red circle among blue circles or a mixed array of blue circles and red squares. Set sizes was 4 or 12, resulting in a 2-by-2 design (search type: color or colorshape, by set size: 4 or 12). Critically, and unknown to subjects, the first four trials were always target-absent trials (one of each set-size search-type combination), presented in randomized order. These trials were followed by the four corresponding target-present trials, presented in randomized order. The final four trials were again target-absent trials, presented in randomized order.

### Randomization.

The order and timing of experimental events was 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 were excluded for making more than one error in the main part of the experiment, or for having extremely fast or slow reaction times in one or more of the tasks (below 250 milliseconds or above 5 seconds in more than 25% of the trials).

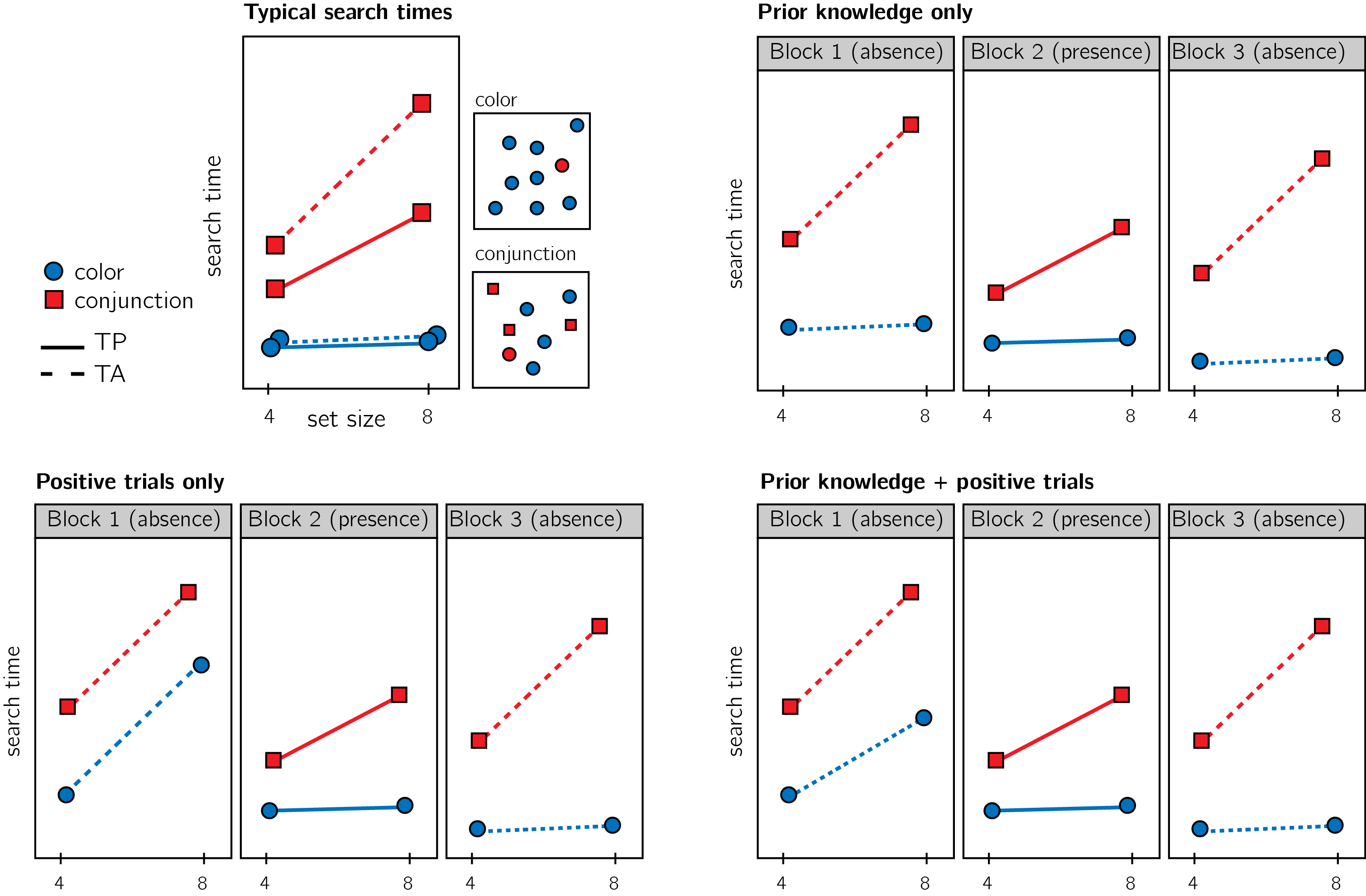
Error trials, and trials with response time below 250 milliseconds or above 1 second were excluded from the response-time analysis.

### Data preprocessing.

To control for within-block trial order effects, a separate linear regression model was fitted to the data of each block, predicting search time as a function of trial serial order (, with denoting the mean-centered serial position within a block). Search times were corrected by subtracting the product of the slope and the mean-centered serial position, in a block-wise manner.

### Hypotheses and analysis plan.

This study is designed to test several hypotheses about the contribution of metacognitive knowledge to search termination, the state of this knowledge prior to engaging with the task, and the effect of experience trials on this metacognitive knowledge. We outline some possible search time patterns and their interpretation in Fig. 2. In the next section we demonstrate our hypotheses and analyses on pilot data.



*Figure* *2:*. Top left: typical search time results in visual search experiments with many trials. Set size (x axis) affects search time in conjunction search, but much less so in color search. However, it is unclear whether this pattern is also true for the first trials in an experiment. Top right: one possible pattern is that the same qualitative pattern will be observed in our design, with an overall decrease in response time as a function of trial number. This will suggest that the metacognitive knowledge necessary to support efficient inference about absence was already in place before engaging with the task. Bottom left: an alternative pattern is that the same qualitative pattern will be observed for blocks 2 and 3, but not in block 1. This will suggest that for inference about absence to be efficient, participants had to experience some target-present trials. Bottom right: alternatively, some of the metacognitive knowledge is available prior to engaging with the task, and some is acquired by exposure to target-present trials. This will manifest as different slopes for conjunction and color searches in blocks 1 and a learning effect for color search between blocks 1 and 3.

Subject-wise search slopes were extracted for each combination of search type (color or conjunction) and block number by fitting a linear regression model to the reaction time data with one intercept and one set-size term.

Analysis comprised a positive control based on target-present trials, a test of the presence of a pop-out effect for color search in block 1, and a test for the change in slope for color search between blocks 1 and 3. All hypotheses were tested using a within-subject t-test, with a significance level of 0.05. Given the fact that we only have one trial per cell, one excluded trial is sufficient to make some hypotheses impossible to test on a given participant. For this reason, for each hypothesis separately, participants were included only if all necessary trials met our inclusion criteria. This means that some hypotheses were tested on different subsets of participants.

## Results

We used R (Version 3.6.0; R Core Team, 2019) and the R-packages *}base* [@}R-base], *}psycho* [@}R-psycho], *}reticulate* [@}R-reticulate], *base* (Version 3.6.0; @}R-base; R Core Team, 2019), *BayesFactor* (Version 0.9.12.4.2; Morey & Rouder, 2018), *brms* (Version 2.13.0; Bürkner, 2017, 2018), *broom* (Version 0.5.6; Robinson & Hayes, 2020), *coda* (Version 0.19.3; Plummer, Best, Cowles, & Vines, 2006), *cowplot* (Version 1.0.0; Wilke, 2019), *dplyr* (Version 1.0.0; Wickham et al., 2020), *forcats* (Version 0.5.0; Wickham, 2020), *ggplot2* (Version 3.3.1; Wickham, 2016), *lsr* (Version 0.5; Navarro, 2015), *Matrix* (Version 1.2.17; Bates & Maechler, 2019), *MESS* (Version 0.5.6; Ekstrøm, 2019), *papaja* (Version 0.1.0.9942; Aust & Barth, 2020, 2020, 2020), *purrr* (Version 0.3.4; Henry & Wickham, 2020), *pwr* (Version 1.3.0; Champely, 2020), *Rcpp* (Version 1.0.4.6; Eddelbuettel & François, 2011; Eddelbuettel & Balamuta, 2017), *readr* (Version 1.3.1; Wickham, Hester, & Francois, 2018), *stringr* (Version 1.4.0; Wickham, 2019), *tibble* (Version 3.0.1; Müller & Wickham, 2020), *tidyr* (Version 1.1.0; Wickham & Henry, 2020), and *tidyverse* (Version 1.3.0; Wickham, Averick, et al., 2019) for all our analyses.

Overall mean accuracy was 0.95 (standard deviation =0.06). The median reaction time was 623.98 ms (median absolute deviation = 127.37). In all further analyses, only correct trials with response time between 250 and 1000 ms are included.

*Hypothesis 1 (positive control)*: Search times in block 2 (target-present) followed the expected pattern, with a steep slope for conjunction search (, 95% CI , ) and a shallow slope for conjunction search (, 95% CI , ; see Fig. 3). The slope for color search was significantly lower than 10 ms/item and thus met our criterion for being considered ‘pop-out’ (, ). Furthermore, the difference between the slopes was significant (, ). This positive control served to validate our method.

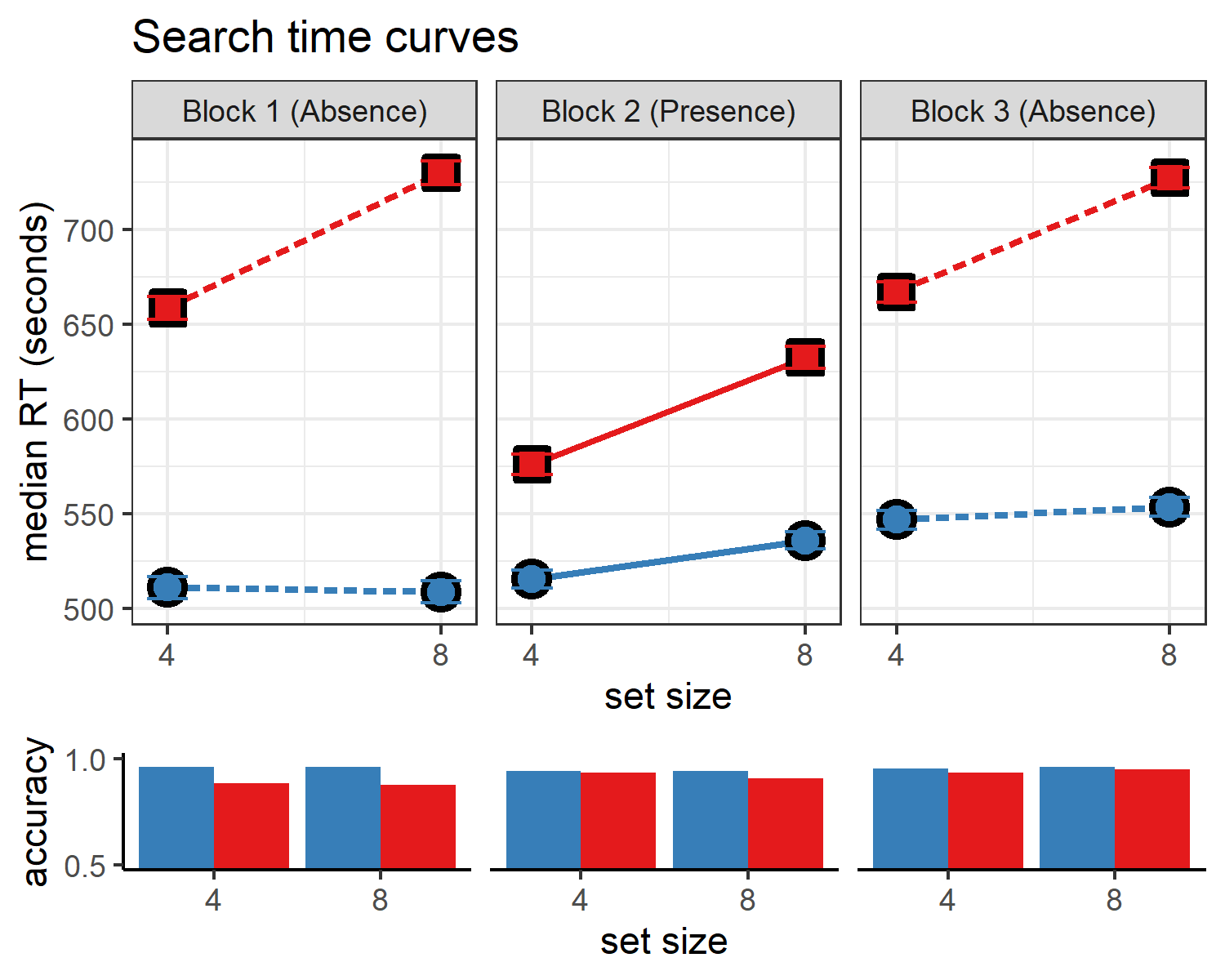
*Hypothesis 2*: Our central focus was on results from block 1 (target-absent). Here participants didn’t yet have experience with searching for the red circle. Similar to the second block, the slope for the conjunction search was steep (, 95% CI , ). A clear ‘pop-out’ effect for color search was also evident (, 95% CI , , , ). Furthermore, the average search slope for color search in this first block was significantly different from that of the conjunction search (, ), indicating that a color-absence pop-out does not depend on task experience.

Pre-registered hypotheses 3-5 were designed to test for a learning effect between blocks 1 and 3, before and after experience with observing a red target among blue distractors. Given the overwhelming pop-out effect for target-absent trials in block 1, no much room for learning was left. Indeed, results from these tests support a prior-knowledge only model (see Fig. 2).

*Hypothesis 3*: Like in the first block, in the third block color search complied with our criterion for ‘pop-out’ (, 95% CI , , , ), and was significantly different from the search for conjunction search (, ). This result is not surprising, given that a pop-out effect was already observed in block 1.

*Hypothesis 4*: To quantify the learning effect for color search, we directly contrasted the search slope for color search in blocks 1 and 3. We find no evidence for a learning effect (, ). Furthermore, a Bayesian t-test with a scaled Cauchy prior for effect sizes (r=0.707) provided strong evidence in favour of the absence of a learning effect ().

*Hypothesis 5*: In case of a learning effect for pop-out search, Hypothesis 5 was designed to test the specificity of this effect to color pop-out by computing an interaction of block number and search type. Given that no learning effect has been observed, this test makes little sense. For completeness, we report that the change in slope between blocks 1 and 3 was similar for color and conjunction search (, 95% CI , , , ).



*Figure* *3:*. Median search time by distractor set size for the two search tasks across the three blocks. Correct responses only. Error bars represent the standard error of the median.

## Additional analyses

In Experiment 1, we found a clear pop-out effect for color absence in the first trials of the experiment, before participants experienced color pop-out in target-present trials. As per our analysis, this reflects prior metacognitive knowledge about the expected efficiency of color search. In order to terminate the search immediately, participants must have known, implicitly or explicitly, that a red item would have popped out immediately. In the setting of this experiment, this knowledge could not be acquired in previous trials. However, an alternative account is that participants noticed the pop-out of the red distractors in the conjunction trials of block 1, and based their expectation for color pop out on those trials. This account can be directly tested by zeroing in on the subset of participants that performed the two color trials before the two conjunction trials in block 1. This subset of participants showed a clear pop-out effect (, 95% CI , , , ).

# Experiment 2

Experiment 1 provided unequivocal evidence for color-absence pop-out prior to experiencing color pop-out in the context of the task. We interpret this as indicating that task-naive adults had valid implicit or explicit metacognitive expectations about color pop-out. This metacognitive knowledge may be innate (acquired in the course of evolution, for example driven by the utility of color search for foraging), learned from previous visual experience (for example, first-person experience of attention being immediately drawn to distinct colors), or culturally acquired (for example, through language). Experiment 2 is designed to extend these findings to another stimulus feature that has been found to efficiently guide attention: shape. The time cost of additional distractors in shape search is under 10 ms (, 95% CI , and , 95% CI , for target-absent and target-present responses in our pilot data), rendering it a case of parallel, efficient search. It is possible however that unlike in the case of colour, the metacognitive knowledge that gives rise to the pop-out effect for shape-absence is acquired through experience with the task. In this experiment we also include an additional control for prior experience with visual search tasks.

# Procedure

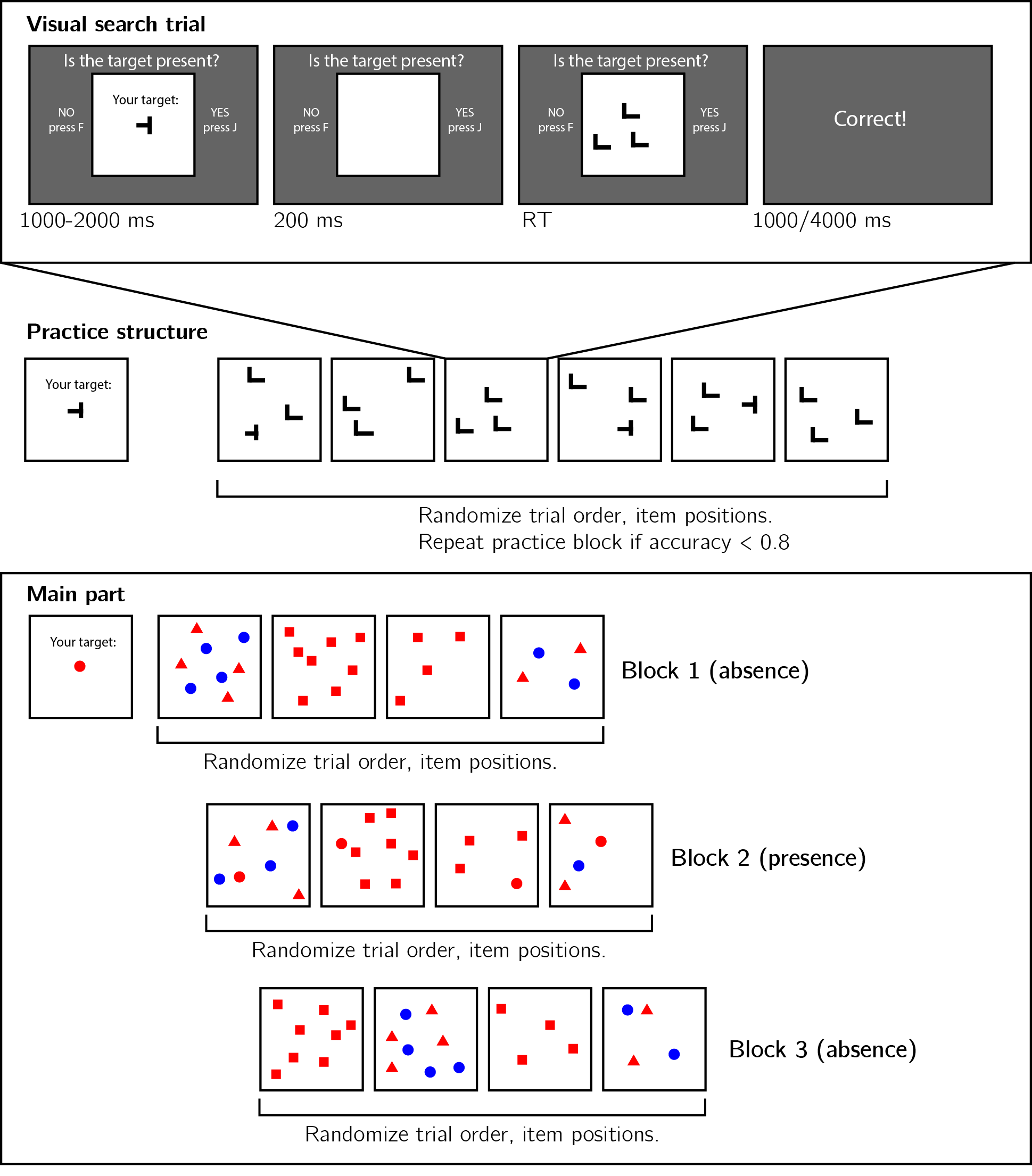
Experiment 2 will be similar to Experiment 1 with two changes. First, the distractors will be red triangles and blue circles for conjunction trials, and red squares for shape trials (see Fig. 4. Second, participants will be asked to complete a short debriefing questionnaire upon finishing the study. The first question (Q1) will be:

Some searches are easier than others. This means that participants find the red dot faster when hidden among some sets of distractors, compared to others. To the best of your ability, order the following distractor sets from easiest (fastest) to hardest (slowest).

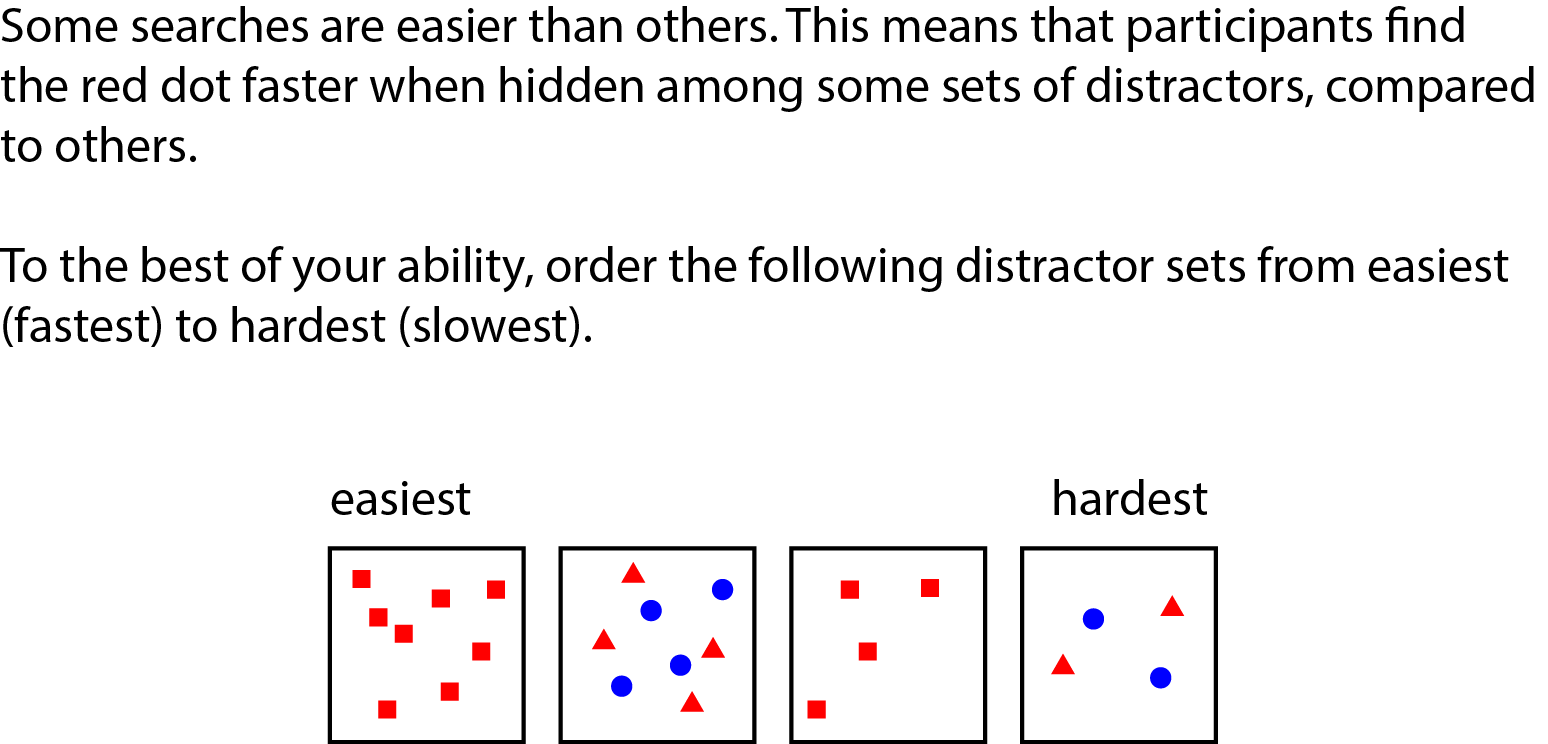
Participants will then be requested to drag the four distractor sets (4 red squares, 2 red squares and 2 blue circles, 8 red squares, 4 red squares and 4 blue circles) into position to reflect their perceived difficulty (see Fig. 5). The second question (Q2) will be:

Have you ever participated in a similar experiment before, where you were asked to search for a target item?

In case they answer ‘yes’, they will be further asked to report when they took part in the experiment, and provide as many details as they can remember about the experiment itself.



*Figure* *4:*. Experimental design. Top panel: each visual search trial started with a screen indicating the target stimulus. The search display remained visible until a response is recorded. To motivate accurate responses, the feedback screen remained visible for one second following correct responses and for four seconds following errors. Middle panel: after reading the instructions, participants practiced the visual search task in blocks of 6 trials, until they had reached an accuracy level of 83% correct or higher (at most one error per block of 6 trials). Bottom panel: the main part of the experiment comprised 12 trials only, in which the target was a red circle. Unbeknown the subjects, only trials 5-8 (Block 2) were target-present trials, and the remaining trials were target-absent trials. Each 4-trial block followed a 2 by 2 design, with factors being set size (4 or 8) and distractor type (color or conjunction; blue circles only or blue circles and red squares, respectively).



*Figure* *5:*. Debrief question Q1. Participatns will be asked to drag the four search arrays into position, to match their perceived difficulty.

## 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 et al., 2019).

# Data analysis

## Rejection criteria

Participants will be excluded for making more than one error in the main part of the experiment, or for having extremely fast or slow reaction times (below 250 milliseconds or above 5 seconds) in more than 25% of the trials.

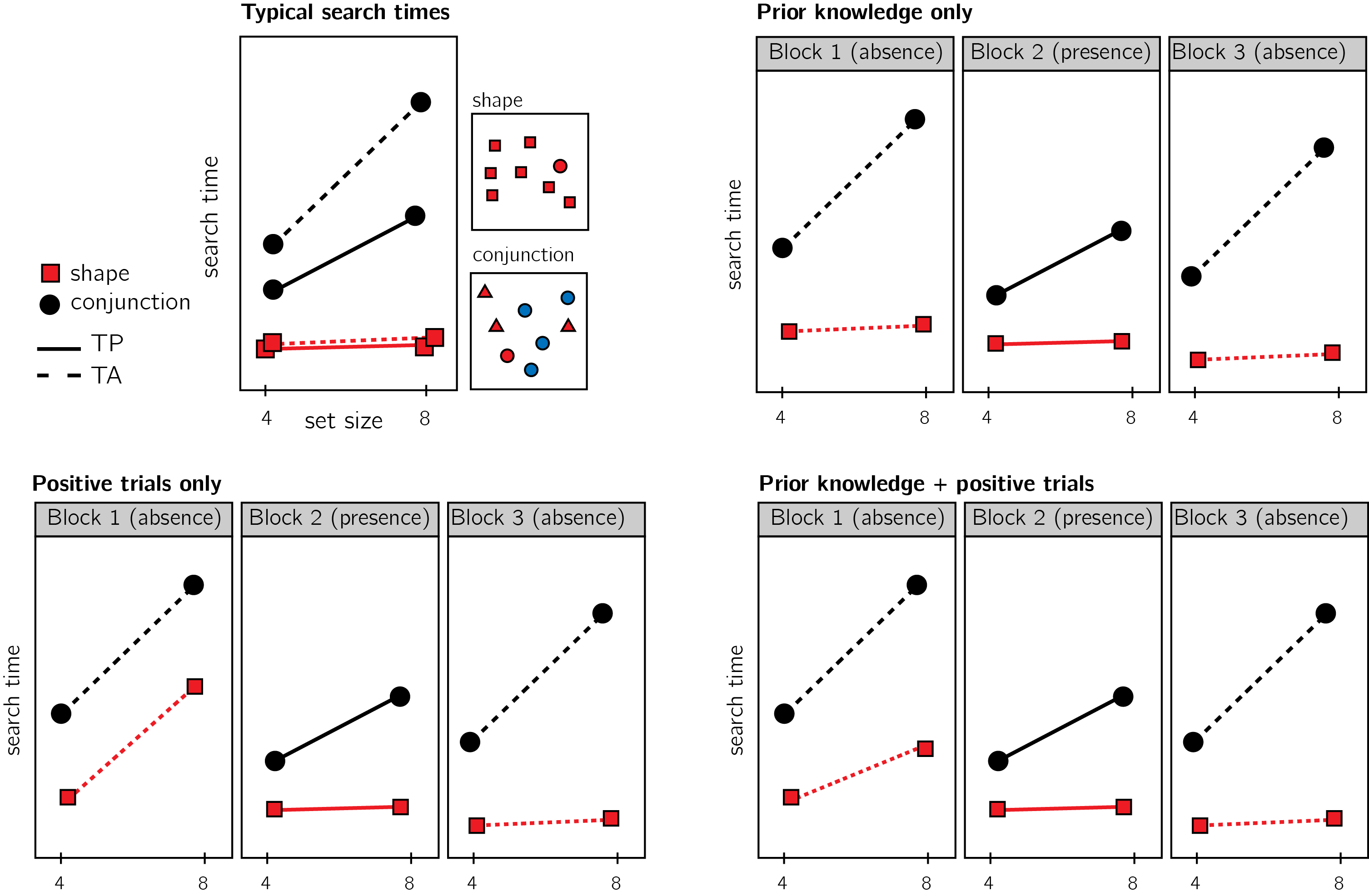
Error trials, and trials with response time below 250 milliseconds or above 1 second will be excluded from the response-time analysis.

## Data preprocessing

To control for within-block trial order effects, a separate linear regression model will be fit to the data of each block, predicting search time as a function of trial serial order (, with denoting the mean-centered serial position within a block). Search times will be corrected by subtracting the product of the slope and the mean-centered serial position, in a block-wise manner.

## Hypotheses and analysis plan

Experiment 2 is designed to test several hypotheses about the contribution of metacognitive knowledge to search termination in the case of color search, the state of this knowledge prior to engaging with the task, and the effect of experience trials on this metacognitive knowledge. We outline some possible search time patterns and their interpretation in Fig. 6.



*Figure* *6:*. Top left: typical search time results in visual search experiments with many trials. Set size (x axis) affects search time in conjunction search, but much less so in shape search. However, it is unclear whether this pattern is also true for the first trials in an experiment. Top right: one possible pattern is that the same qualitative pattern will be observed in our design, with an overall decrease in response time as a function of trial number. This will suggest that the metacognitive knowledge necessary to support efficient inference about absence was already in place before engaging with the task. Bottom left: an alternative pattern is that the same qualitative pattern will be observed for blocks 2 and 3, but not in block 1. This will suggest that for inference about absence to be efficient, participants had to experience some target-present trials. Bottom right: alternatively, some of the metacognitive knowledge is available prior to engaging with the task, and some is acquired by exposure to target-present trials. This will manifest as different slopes for conjunction and shape searches in blocks 1 and a learning effect for shape search between blocks 1 and 3.

Subject-wise search slopes will be extracted for each combination of search type (shape or conjunction) and block number by fitting a linear regression model to the reaction time data with one intercept and one set-size term.

Analysis will comprise a positive control based on target-present trials, a test of the presence of a pop-out effect for color search in block 1, and a test for the change in slope for color search between blocks 1 and 3. All hypotheses will be tested using a within-subject t-test, with a significance level of 0.05.

Given the fact that we only have one trial per cell, one excluded trial is sufficient to make some hypotheses impossible to test on a given participant. For this reason, for each hypothesis separately, participants will be included only if all necessary trials meet our inclusion criteria. This means that some hypotheses may be tested on different subsets of participants.

*Hypothesis 1 (Positive control)*: To validate our methods and the quality of our data, we will test for a difference between search slopes for shape and conjunction search in block 2 (target-present). Based on previous work we expect a steeper slope for conjunction than for shape search Treisman and Gelade (1980)]. This positive control will serve to confirm that these effects are detectable in a large sample, even with only one trial per cell.

*Hypothesis 2*: Pop-out for shape absence in block 1. Throughout our analysis, we will define pop-out as a search slope significantly lower than 10 ms/item. This cutoff was chosen based on empirical distributions of search slopes in feature search (Wolfe, 1998). We will test the null hypothesis that the search slope in the color search, block 1 (target-absent) equals to or is higher than 10ms/item, using a t-test. We will further test the null hypothesis that search slopes for shape and conjunction searches in blocks 1 are equal.

*Hypothesis 3*: Pop-out for shape absence in block 3. We will test the null hypothesis that the search slope in the shape search, block 3 (target-absent) equals to or is higher than 10ms/item, using a t-test. We will further test the null hypothesis that search slopes for shape and conjunction searches in blocks 3 are equal.

*Hypothesis 4*: Search slope for shape search changes between blocks 1 and 3. We will test the null hypothesis that the search slope in the shape search, block 1 (target-absent) equals to search slope in the shape search, block 1 (target-absent).

*Hypothesis 5*: The change in search slopes between blocks 1 and 3 is different for shape and for conjunction searches. To rule out a nonspecific change in search slope between blocks 1 and 3, We will compare the difference in search slopes for shape search between block 1 and 3 with the difference in search slopes for conjunction search for the same blocks.

In case we find a pop out effect for shape absence in block 1, we will make sure that it was not driven by experienced participants that have taken part in similar visual-search tasks in the past. this will be done by repeating our analysis on the subset of participants that reported not participating in a similar experiment before.

Responses to Q1 (perceived difficulty) will be used for post-hoc exploratory analysis, looking into explicit metacognitive insight into search-guiding metacognitive knowledge.

## Statistical power

Statistical power calculations were performed using the R-pwr package (Champely, 2020).

With a minimum of 320 participants for each hypothesis, we will have statistical power of 95% to detect effects of size 0.20

# References

Aust, F., & Barth, M. (2020). *papaja: Create APA manuscripts with R Markdown*. Retrieved from <https://github.com/crsh/papaja>

Bates, D., & Maechler, M. (2019). *Matrix: Sparse and dense matrix classes and methods*. Retrieved from <https://CRAN.R-project.org/package=Matrix>

Bürkner, P.-C. (2017). brms: An R package for Bayesian multilevel models using Stan. *Journal of Statistical Software*, *80*(1), 1–28. <https://doi.org/10.18637/jss.v080.i01>

Bürkner, P.-C. (2018). Advanced Bayesian multilevel modeling with the R package brms. *The R Journal*, *10*(1), 395–411. <https://doi.org/10.32614/RJ-2018-017>

Champely, S. (2020). *Pwr: Basic functions for power analysis*. Retrieved from <https://CRAN.R-project.org/package=pwr>

Eddelbuettel, D., & Balamuta, J. J. (2017). Extending extitR with extitC++: A Brief Introduction to extitRcpp. *PeerJ Preprints*, *5*, e3188v1. <https://doi.org/10.7287/peerj.preprints.3188v1>

Eddelbuettel, D., & François, R. (2011). Rcpp: Seamless R and C++ integration. *Journal of Statistical Software*, *40*(8), 1–18. <https://doi.org/10.18637/jss.v040.i08>

Ekstrøm, C. T. (2019). *MESS: Miscellaneous esoteric statistical scripts*. Retrieved from <https://CRAN.R-project.org/package=MESS>

Henry, L., & Wickham, H. (2020). *Purrr: Functional programming tools*. Retrieved from <https://CRAN.R-project.org/package=purrr>

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.

Morey, R. D., & Rouder, J. N. (2018). *BayesFactor: Computation of bayes factors for common designs*. Retrieved from <https://CRAN.R-project.org/package=BayesFactor>

Müller, K., & Wickham, H. (2020). *Tibble: Simple data frames*. Retrieved from <https://CRAN.R-project.org/package=tibble>

Navarro, D. (2015). *Learning statistics with r: A tutorial for psychology students and other beginners. (Version 0.5)*. Adelaide, Australia: University of Adelaide. Retrieved from <http://ua.edu.au/ccs/teaching/lsr>

Plummer, M., Best, N., Cowles, K., & Vines, K. (2006). CODA: Convergence diagnosis and output analysis for mcmc. *R News*, *6*(1), 7–11. Retrieved from <https://journal.r-project.org/archive/>

R Core Team. (2019). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing. Retrieved from <https://www.R-project.org/>

Robinson, D., & Hayes, A. (2020). *Broom: Convert statistical analysis objects into tidy tibbles*. Retrieved from <https://CRAN.R-project.org/package=broom>

Treisman, A. M., & Gelade, G. (1980). A feature-integration theory of attention. *Cognitive Psychology*, *12*(1), 97–136.

Wickham, H. (2016). *Ggplot2: Elegant graphics for data analysis*. Springer-Verlag New York. Retrieved from <https://ggplot2.tidyverse.org>

Wickham, H. (2019). *Stringr: Simple, consistent wrappers for common string operations*. Retrieved from <https://CRAN.R-project.org/package=stringr>

Wickham, H. (2020). *Forcats: Tools for working with categorical variables (factors)*. Retrieved from <https://CRAN.R-project.org/package=forcats>

Wickham, H., Averick, M., Bryan, J., Chang, W., McGowan, L. D., François, R., … Yutani, H. (2019). Welcome to the tidyverse. *Journal of Open Source Software*, *4*(43), 1686. <https://doi.org/10.21105/joss.01686>

Wickham, H., François, R., Henry, L., & Müller, K. (2020). *Dplyr: A grammar of data manipulation*. Retrieved from <https://CRAN.R-project.org/package=dplyr>

Wickham, H., & Henry, L. (2020). *Tidyr: Tidy messy data*. Retrieved from <https://CRAN.R-project.org/package=tidyr>

Wickham, H., Hester, J., & Francois, R. (2018). *Readr: Read rectangular text data*. Retrieved from <https://CRAN.R-project.org/package=readr>

Wilke, C. O. (2019). *Cowplot: Streamlined plot theme and plot annotations for ’ggplot2’*. Retrieved from <https://CRAN.R-project.org/package=cowplot>

Wolfe, J. M. (1998). What can 1 million trials tell us about visual search? *Psychological Science*, *9*(1), 33–39.