

## **Rethinking egocentric bias: a computer mouse tracking study of adult belief processing**

### **Supplementary information 2: Replicating main effects for spatial measures in an online study**

Richard J. O'Connor, Andrew P. Lucas & Kevin J. Riggs

University of Hull, Department of Psychology, Cottingham Road, Hull, HU6 7RX, UK

In Experiments 1 and 2 reported in the main text, we found no evidence of the specific interaction predicted by egocentric bias. Instead, we typically found evidence of main effects of scenario and question across errors, temporal and spatial measures. In Experiment 1, for errors and RT we found 2 main effects of question and scenario and no interaction. For the spatial measure PL, we also found 2 main effects and no interaction. For the spatial measures of MAD and AUC, however, we did find an interaction of question by scenario (although in the opposite direction predicted by egocentric bias) with the effect of question being significant for true belief scenarios only. In Experiment 2, we found both main effects and no interaction for all 3 spatial measures (thus PL showed consistency across both experiments). In Experiment 3, in which participants were instructed to attend to the belief of the agent, the effects of question on all spatial measures were removed or reversed, relative to the patterns seen in Experiment 1 and 2.

While some of these differences between experiments in the patterns of effects for MAD and AUC were predicted<sup>1</sup>, and we provide explanations for observed differences between experiments in the main text, an alternative possible explanation – raised by an anonymous reviewer – is that we observed differences between experiments because the effects of question and scenario for spatial measures are in fact not robust. On such an explanation, differences between experiments and our failure to find evidence of egocentric bias may be due to our paradigm and spatial measures – MAD and AUC in particular – being unable to consistently elicit main effects of question or scenario in the predicted direction (greater deviation answering belief compared to reality questions and greater deviation on false belief compared to true belief scenarios). If our paradigm

---

<sup>1</sup> Specifically the differences seen between Experiment 3 and the other two experiments: here we predicted that attending to beliefs should reduce the observed effects of question.

cannot robustly detect such main effects of question or scenario, then one might not expect to be able to detect an interaction between these factors, as necessary for detecting egocentric bias.

We provide arguments against this interpretation of our data in the main text. In addition, we report below data from an additional experiment that used the same video stimuli, design and methods as Experiment 1, with some minor changes to allow adaptation to online data collection. This online experiment was powered to detect main effects of question and scenario. If these effects are robust, one should expect in this additional study to see main effects of question and scenario, specifically larger MAD and AUC for reality compared to belief questions and larger MAD and AUC for false belief compared to true belief scenarios. This is indeed what we found, as reported below.

### Method

All data and analysis code are available at <https://researchbox.org/1024>. Research materials are available from the authors upon request.

#### *Participants*

Forty-one participants completed the study. After trial-level exclusions (see below), 27 participants contributed data to the final sample. The final sample of 27 participants therefore exceeded the minimum sample size of 18 that is recommended by Brysbaert (2019) for achieving 80% power to detect main effects with an effect size of  $d_z = .50$  in a  $2 \times 2$  repeated-measures design<sup>2</sup>; indeed, it reached the minimum sample size of 27 that is recommended by Brysbaert (2019) for achieving 80% power to detect an effect size of  $d_z = .40$  for such main effects.

Participants were asked to report their current gender identity, with the response options: “female”, “male”, “non-binary”, “prefer not to say”. Of the final sample, 11 identified as male and 16 identified as female. Participants’ mean age was 22.4 years (SD = 6.0 years). Twenty-four participants were ordinarily right handed

---

<sup>2</sup>As in Experiments 1 and 2, the use of  $d_z = .50$  is based on  $d_{av} = .40$ , the recommended estimate of the smallest effect size of interest by Brysbaert (2019), when  $r \geq .68$ . Note this experiment used the same video stimuli as in Experiment 1, where the effect of question on MAD for TB-scenarios was  $d_z = .59$ , and the effect of scenario on MAD for reality questions was  $d_z = .50$ . Thus using  $d_z = .50$  as an estimate of a minimum effect size of interest for detecting main effects was justifiable.

and 3 were ordinarily left handed, though all confirmed they normally used a computer mouse with their right hand. Participants were all undergraduate psychology students at the University of Hull and were compensated for their time with course credits. The study was approved by the University of Hull Faculty of Health Sciences ethics committee.

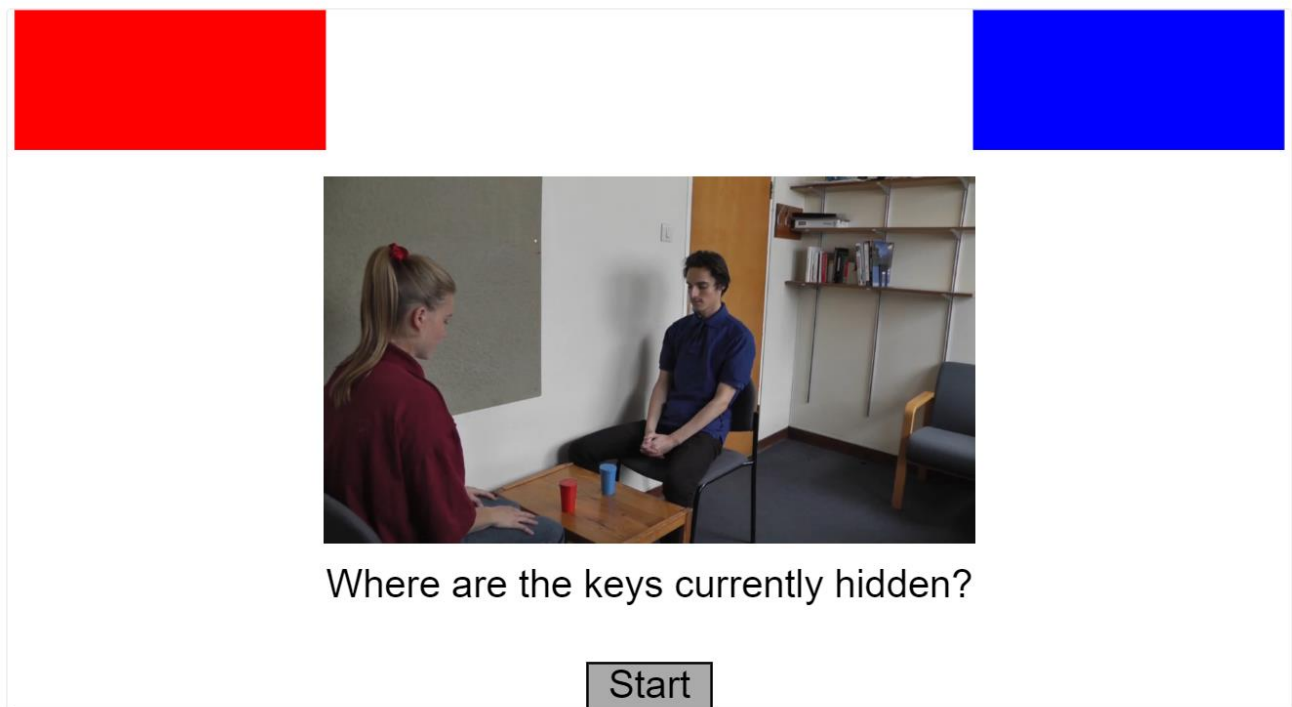
### *Stimuli and apparatus*

Participants completed the study online and therefore used their own computer devices. However, participants were instructed to use a physical computer mouse to complete the study, either with a laptop or a desktop computer. All participants confirmed that they did so at the end of the study. The study was presented to participants in their internet browser using lab.js (Henniger et al., 2021). Stimuli were presented to participants in full-screen browser mode with a constant width-to-height ratio, such that across variable display ratios stimuli were not distorted. Response boxes, presented in the upper left and right corners of the display, had a width: height ratio of 20:9 and were presented at 50% of the width of the participant's display. The start box, presented in the lower middle of the display, had a width: height ratio of 8:3 and was presented at 10% of the width of the participant's display. The study used the same 16 video stimuli as used in Experiment 1, presented in their original aspect ratio at 50% of the width of the participant's display. Question text was presented below the final frame of each video in size 32 font. Figure 1 provides the layout of these stimuli for a typical 16:9 display.

### *Design*

The experiment used the same 2 x 2 repeated-measures design as Experiments 1 – 3. However, the overall number of trials participants received was reduced from 128 to 80 trials. The decision to reduce overall trial numbers was made so as to reduce the overall length of the study – Experiments 1 – 3 typically took over an hour to complete – so as to make the study more suitable for online testing. Trials were delivered in a single block in which participants answered 20 belief questions (10 following false belief scenarios, 10 following true belief scenarios), 20 reality questions (10 following false belief scenarios, 10 following true belief scenarios) and 40 filler questions (20 following false belief scenarios, 20 following true belief scenarios). All 16 videos appeared

5 times, with all videos appearing at least once with each experimental question and twice with a filler question. Eight videos (four per scenario) appeared with an additional experimental question (four per question), and the other eight videos (four per scenario) appeared with an additional filler question. The correct answer was always 50% “red” and 50% “blue” for any given scenario and question combination, and thus unpredictable for participants. Videos and questions were randomly presented across the block.



**Figure 1.** Question screen (here using the reality question) as displayed to participants at the end of each trial. Layout of stimuli is for a typical 16:9 display (see details in the main text).

## *Procedure*

The procedure was the same as in Experiment 1, with the following changes. Due to restrictions on how browsers can interact with a user's cursor position, it was not possible to reset the mouse position to the bottom-centre of the screen when the question appeared at the end of each video. Therefore, participants were instructed that after clicking on the start box they needed to keep the mouse still until the question appeared at the end of each video. If a participant moved the mouse outside the start box while the video was playing, they received an early-movement warning message after they gave their response to the question. The message instructed them to keep their mouse still on future trials.

To ensure that participants fully understood the requirements of the task prior to beginning the main experiment, during the six practice trials participants received trial-by-trial feedback on their accuracy, whether they correctly kept the mouse still during the video, and whether they moved the mouse within the initiation time cut-off (i.e., within 1000ms of the question appearing, as in Experiment 1). If a participant in practice triggered either the early movement or initiation time warnings more than two times, or made more than two errors on the practice videos, they were required to complete the practice a second time.

In the main experiment, participants received warning messages after each trial if they moved the mouse outside the start box during the video or if they did not move the mouse within 1000ms of the question appearing after the video, but they did not receive trial-by-trial feedback on their accuracy. Participants were given the opportunity for breaks after every 20 trials. At these points they also received feedback on their overall accuracy on the task so far, in order to maintain engagement with the task. After completing all trials, participants finally completed an online debrief questionnaire asking them to confirm that they had used a physical computer mouse for the duration of the study and had not engaged in attempts to "cheat" during the study, specifically by making notes during the videos or using their hands to point at events on the screen (in Experiments 1 – 3 such strategies would not have been possible, due to the in-person presence of the experimenter). Participants were informed that their response to these questions would not affect their

allocation of course credit as compensation for completing the study. No participant reported engaging in those specific strategies.

#### *Data processing and exclusions*

Response accuracy, initiation time and response time was recorded on each trial by lab.js. Trial-by-trial mouse-tracking data were processed in the following steps, using the mousetrap package (Wulff et al., 2021) for R, version 4.2.2 (R Core Team, 2022; see also analysis script at <https://researchbox.org/1024>). These steps follow those used in Experiment 1, with some additional trial exclusion criteria required due to the online nature of the study.

First, in order to normalise mouse-path coordinates across varying screen sizes, raw mouse-path coordinates for each participant were recalculated as a proportion of the participant's screen size and mapped to a coordinate space with a minimum coordinate (bottom-left) of (-0.5, 0) and a maximum coordinate (top-right) of (0.5, 1). Then all responses to filler trials, trials with an incorrect response, trials where the participant moved the mouse outside of the start box and trials with an IT greater than 1000ms were excluded. Visual inspection of the remaining trajectories revealed a small number ( $n = 18$ ) of highly unusual trajectories in which the mouse moved down to the extreme bottom of the screen (minimum y-coordinate less than or equal to 0.03) prior to selecting a response: these trials were excluded. Mouse-path trajectories were then spatially normalised and remapped to the right-hand side of the screen (start points aligned to (0, 0) and end points to (0.5, 1)). Then trajectories were time normalised (each trajectory divided into 101 bins of equal temporal duration) on a trial by trial basis. MAD, AUC and PL were then calculated per trial. Any trial with a measurement of MAD, AUC, PL, IT or RT that was above or below 3 SD of the participant overall mean across experimental conditions was then excluded. Finally, participants ( $n = 14$ ) with fewer than 5 trials (i.e., 50%) remaining in any condition were excluded. For the final analysed sample, 849 trials remained (out of a maximum of 1080). On average, each participant contributed 31.4 trials, out of a maximum of 40 (78.6%).

## Results

Errors and temporal data can be seen in Figure 2. Mouse-path trajectories and spatial measures can be seen in Figure 3. These data were analysed in the same way as in Experiments 1 – 3. Temporal and spatial measures were analysed with 2 x 2 repeated-measures ANOVAs with the factors question (belief, reality) and scenario (false belief, true belief). Errors were analysed using mixed-effect logistic regression with factors of question and scenario and their interaction included as fixed effects.

### *Errors*

The final mixed-effect logistic regression model included random intercepts for participants and random slopes for the question effect: adding additional random effects resulted in singularity errors. In this model there was a significant main effect of question (coefficient = 0.54, SE = 0.21,  $z = 2.54$ ,  $p = 0.011$ ) and scenario (coefficient = 0.67, SE = 0.16,  $z = 4.21$ ,  $p < 0.001$ ), but no significant interaction (coefficient = 0.02, SE = 0.15,  $z = 0.10$ ,  $p = 0.92$ ). More errors were made on belief questions compared to reality questions, and more errors were made on false belief scenarios compared to true belief scenarios.

### *Spatial data*

For MAD, ANOVA showed significant main effects of both question ( $F(1, 26) = 6.47$ ,  $p = 0.017$ ,  $\eta_g^2 = 0.027$ ) and scenario ( $F(1, 26) = 12.2$ ,  $p < 0.002$ ,  $\eta_g^2 = 0.073$ ). Their interaction was not significant ( $F(1, 26) = 0.27$ ,  $p = 0.61$ ,  $\eta_g^2 = 0.001$ ). For AUC, ANOVA also showed significant main effects of both question ( $F(1, 26) = 4.53$ ,  $p = 0.043$ ,  $\eta_g^2 = 0.026$ ) and scenario ( $F(1, 26) = 8.59$ ,  $p = 0.007$ ,  $\eta_g^2 = 0.059$ ) and no significant interaction ( $F(1, 26) = 0.01$ ,  $p = 0.93$ ,  $\eta_g^2 < 0.001$ ). Finally, for PL, ANOVA showed a significant main effect of scenario ( $F(1, 26) = 11.5$ ,  $p = 0.002$ ,  $\eta_g^2 = 0.027$ ), but no significant main effect of question ( $F(1, 26) = 2.85$ ,  $p = 0.10$ ,  $\eta_g^2 = 0.009$ ) and no significant interaction ( $F(1, 26) = 0.18$ ,  $p = 0.68$ ,  $\eta_g^2 < 0.001$ ). For all spatial measures, Hartigans' dip statistic found no evidence of bimodality (for all measures within all conditions,  $p$ -values  $\geq 0.15$ ).

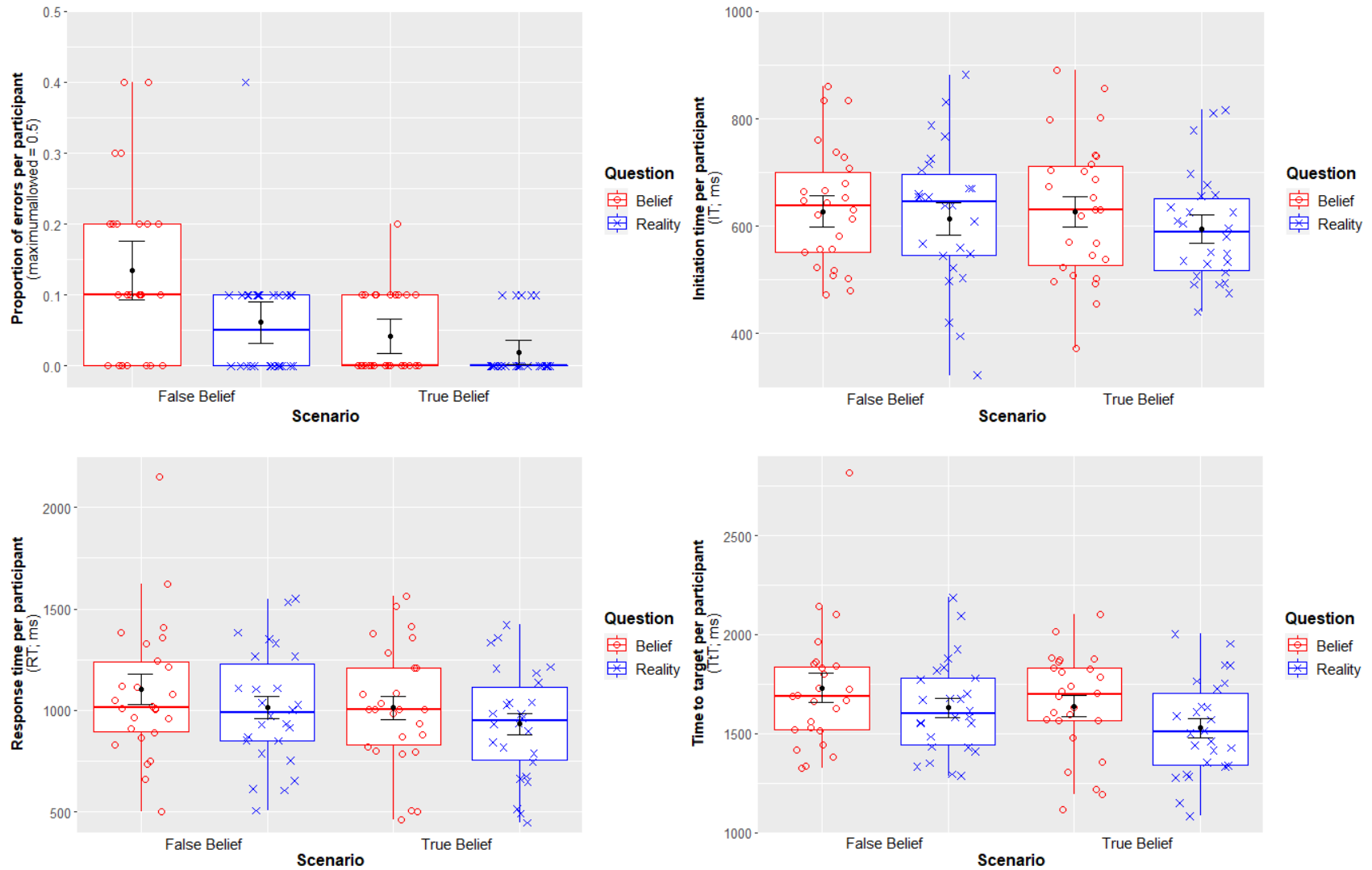
In all cases, the significant main effect of scenario reflected greater attraction towards the incorrect answer on false belief scenarios compared to true belief scenarios. For MAD and AUC, the significant main

effect of question reflected greater attraction towards the incorrect answer on belief compared to reality questions.

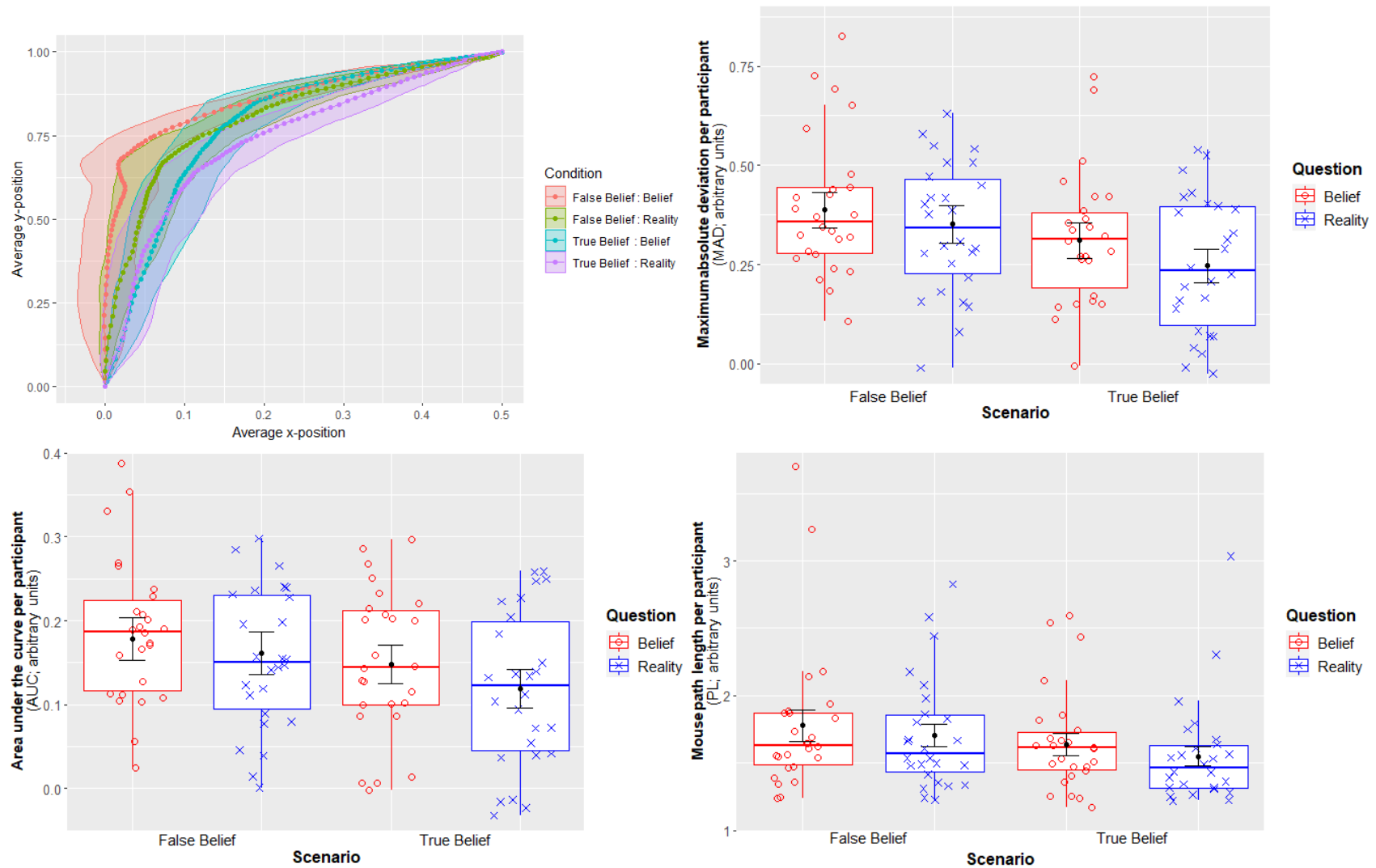
#### *Temporal data*

ANOVA for IT showed no significant main effect of question ( $F(1, 26) = 3.12, p = 0.089, \eta_g^2 = 0.012$ ) or scenario ( $F(1, 26) = 1.88, p = 0.18, \eta_g^2 = 0.005$ ). There was no significant interaction ( $F(1, 26) = 0.72, p = 0.40, \eta_g^2 = 0.002$ ).

For RT there was a significant main effect of question ( $F(1, 26) = 7.03, p = 0.013, \eta_g^2 = 0.016$ ) and scenario ( $F(1, 26) = 7.29, p = 0.012, \eta_g^2 = 0.015$ ), but no significant interaction ( $F(1, 26) = 0.14, p = 0.72, \eta_g^2 < 0.001$ ). Belief questions took longer to respond to than reality questions, and false belief scenarios took longer to respond to than true belief scenarios.



**Figure 2.** Error and temporal data. Note that TtT is provided here as a comparable measure to Reaction Time as typically reported in RT studies: TtT measures the time elapsed between the question appearing and participants clicking on their response. Boxes indicate 25th - 75th percentile (IQR), with the middle line indicating the median. The whiskers indicate  $\pm 1.5$  IQR. Each coloured ring / cross is an individual participant's mean. The black dot indicates the condition mean, with error bars representing 95% CI (i.e., 1.96 SE).



**Figure 3.** Mouse-path trajectories and spatial measures (MAD, AUC, PL). For trajectory plots, all trajectories are remapped such that the correct answer is on the right. Ribbons indicate  $\pm 1$  SE in the x-position around a particular time bin. Lines indicate the condition mean trajectory, collapsed across subjects. For the spatial measures, boxes indicate 25th - 75th percentile (IQR), with the middle line indicating the median. The whiskers indicate  $\pm 1.5$  IQR. Each coloured ring / cross is an individual participant's mean. The black dot indicates the condition mean, with error bars representing 95% CI (i.e., 1.96 SE).

## Discussion

Data from this additional online experiment were consistent with those in Experiments 1 and 2. In particular, significant main effects of both question and scenario were seen for the spatial measures of AUC and MAD, matching the direction of the effects observed in Experiments 1 and 2. Note that this additional experiment was not powered to detect an interaction between question and scenario. Rather, it was powered to detect the presence of main effects, in order to demonstrate that this paradigm and the measures used are capable of eliciting robust main effects of question and scenario in the predicted directions: greater deviation on belief compared to reality questions, and on false belief compared to true belief scenarios.

These findings therefore rule out the suggestion, of an anonymous reviewer, that differences in the question and scenario effects between experiments might simply reflect noise, and that our paradigm might not be able to robustly elicit effects of question or scenario. Across three experiments (Experiments 1, 2 and this online study) in which participants were under no instructions to attend to beliefs, we consistently observe main effects of question and scenario in the same direction across experiments. In Experiment 3, when under instruction to attend to beliefs, the effects of question were removed or reversed. The failure to observe any evidence of egocentric bias in Experiments 1 and 2, therefore, cannot be attributed to a lack of sensitivity of either the spatial measures used or our particular paradigm.

## References

- Henninger, F., Shevchenko, Y., Mertens, U. K., Kieslich, P. J., & Hilbig, B. E. (2021). lab.js: A free, open, online study builder. *Behavior Research Methods*, 1-18.
- Kieslich, P. J., & Henninger, F. (2017). Mousetrap: An integrated, open-source mouse-tracking package. *Behaviour Research Methods*, 49, 1652–1667.