

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

Supplementary information 1

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

University of Hull, Department of Psychology, Cottingham Road, Hull, HU16 4RS, UK

Contents

Outline.....	2
Experiment 1	4
Spatial data – AUC and PL	4
Temporal data – TtT	5
Sensitivity Analyses for all Bayes Factors calculated for Experiment 1.....	9
Exploratory analyses: performance on first trials and first half of the experiment.....	10
Experiment 2	12
Spatial data – AUC and PL	12
Temporal data – TtT	12
Sensitivity Analyses for all Bayes Factors calculated for Experiment 2.....	13
Exploratory analyses: performance on first trials and first half of the experiment.....	15
Experiment 3	17
Spatial data – AUC and PL	17
Temporal data – TtT	17
Comparison of Experiments 1 and 3.....	18
Errors	20
Spatial data – MAD, AUC and PL	20
Temporal data – IT, RT and TtT	21
Standardised time bin analyses of x and y coordinates	28
Bootstrapping simulations	28
Analysis of x and y coordinates in Experiments 1 – 3	30
References.....	37

Outline

We report below additional analyses not included in the main text. The following analyses were also preregistered, but were not included in the main text for ease of exposition. With few exceptions, the patterns observed in these additional data and analyses match those reported in the main text. None of these additional analyses provide any evidence for the presence of egocentric bias.

For Experiments 1 – 3, we report below analysis of:

- Area under curve (AUC): a spatial measure of the area bound by an idealised straight line response and the actual trial trajectory.
- Path length (PL): a spatial measure of the length of the trial trajectory
- Time to target (TtT): a temporal measure summing initiation time (IT) and response time (RT). Mouse tracking typically analyses IT and RT separately (Freeman & Ambady, 2010; Kieslich & Henninger, 2017). Summing these two temporal measures provides a measure of the overall time taken for participants to respond to the question, from the time when the question appeared to when participants clicked on a response box. This measure is therefore closer, conceptually, to response time as typically reported in RT studies.
- For all measures, the full analyses for the comparisons between Experiments 1 and 3.
- Standardised time bin analyses of the *x* and *y* coordinates of participants' response trajectories: this analysis compares between conditions the average *x* and *y* coordinates across percentile time bins in order to determine at what time point, if any, differences *x* and *y* positions emerge according to condition.

We also report the following additional analyses for Experiment 1 and Experiment 2 that were not pre-registered, but were suggested by anonymous reviewers or were conducted in response to their comments:

- For the above additional measures (AUC, PL, TtT), Bayesian analyses of the extent to which the data supported the null hypothesis over the specific interaction predicted by the presence of egocentric bias (i.e., that the difference between belief and reality questions would be greater in the false belief scenarios than the true belief scenarios). We also provide sensitivity analyses for all Bayes Factors calculated in the main text and in SI-1 to demonstrate that support for the null hypothesis is found across a range of prior widths.
- Analyses of the first trials from each condition.
- Analyses of trials from the first half of each experiment.
- Analyses including experiment half as an additional factor.

Experiment 1

Spatial data – AUC and PL

AUC data are shown in Figure 1. ANOVA of AUC found the same pattern as the MAD data reported in the main text. There were significant main effects of Question ($F(1, 75) = 6.73, p = 0.011, \eta_g^2 = 0.006$) and Scenario ($F(1, 75) = 19.5, p < 0.001, \eta_g^2 = 0.011$), and a significant interaction ($F(1, 75) = 5.46, p = 0.022, \eta_g^2 = 0.004$). The interaction was driven by a difference between belief and reality questions in the true belief scenarios ($t(75) = 5.52, p < 0.001, d = 0.633$), but not in the false belief scenarios ($t(75) = 0.411, p = 0.682, d = 0.047$). Furthermore, there was no difference between scenarios for responses to belief questions ($t(75) = 1.37, p = 0.174, d = 0.157$), however there was a difference for reality questions ($t(75) = 4.40, p < 0.001, d = 0.505$). Confirming further these results, Bayesian analysis of the mean differences found that the null hypothesis was 25.3 times more likely than the alternative hypothesis that the difference in AUC between Belief and Reality questions was *greater* in the false belief scenarios than in the true belief scenarios. Finally, Hartigans' dip statistic (HDS) found no evidence that the AUC data was bimodally distributed (all $ps \geq 0.484$). As in the analysis of MAD reported in the main text, analysis of AUC therefore found no evidence for egocentric bias in Experiment 1.

PL data are shown in Figure 2. ANOVA of PL found significant main effects of Question ($F(1, 75) = 10.3, p = 0.002, \eta_g^2 = 0.006$) and Scenario: $F(1, 75) = 5.67, p = 0.020, \eta_g^2 = 0.006$), but no significant interaction ($F(1, 75) = 2.21, p = 0.141, \eta_g^2 < 0.001$). Bayesian analysis of the mean differences found that the null hypothesis was 18.7 times more likely than the alternative hypothesis that the difference in PL between Belief and Reality questions was *greater* in the false belief scenarios than in the true belief scenarios. There was no evidence that the PL data were bimodally distributed (HDS: all $ps \geq$

0.432). While these effects on PL therefore differed somewhat from the MAD and AUC data (where significant Question by Scenario interactions were found), they still do not provide any evidence in support of the presence of egocentric bias.

Temporal data – TtT

TtT data are shown in Figure 3. ANOVA of TtT found significant main effects of both Question ($F(1, 75) = 52.2, p < 0.001, \eta_g^2 = 0.015$) and Scenario ($F(1, 75) = 76.7, p < 0.001, \eta_g^2 = 0.017$), but no interaction ($F(1, 75) = 0.402, p = 0.528, \eta_g^2 < 0.001$). Given that TtT is a composite measure of IT and RT, these effects reflect those observed for those measures in the main text (no effects for IT, but two main effects and no interaction for RT). Bayesian analysis of the mean differences found that the null hypothesis was 12.2 times more likely than the alternative hypothesis that the difference in TtT between Belief and Reality questions was *greater* in the false belief scenarios than in the true belief scenarios. The TtT data therefore also provided no evidence in support of egocentric bias.

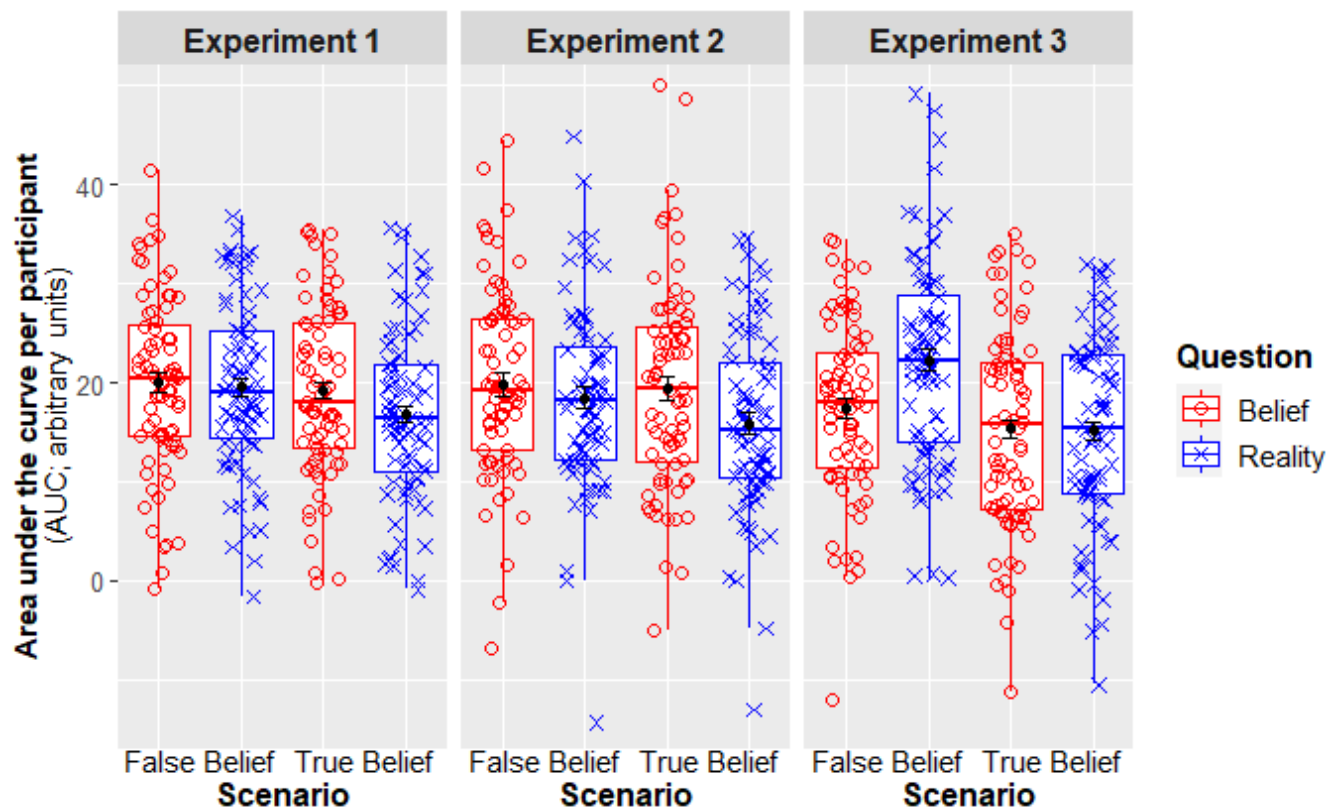


Figure 1. AUC data in Experiments 1 – 3. Boxes indicate 25th - 75th percentile (IQR), with the middle line indicating the median. The whiskers indicate ± 1.5 IQR. Each coloured ring / cross is an individual participant's mean. The black dot indicates the condition mean, with error bars representing 95% CI (i.e., 1.96 SE).

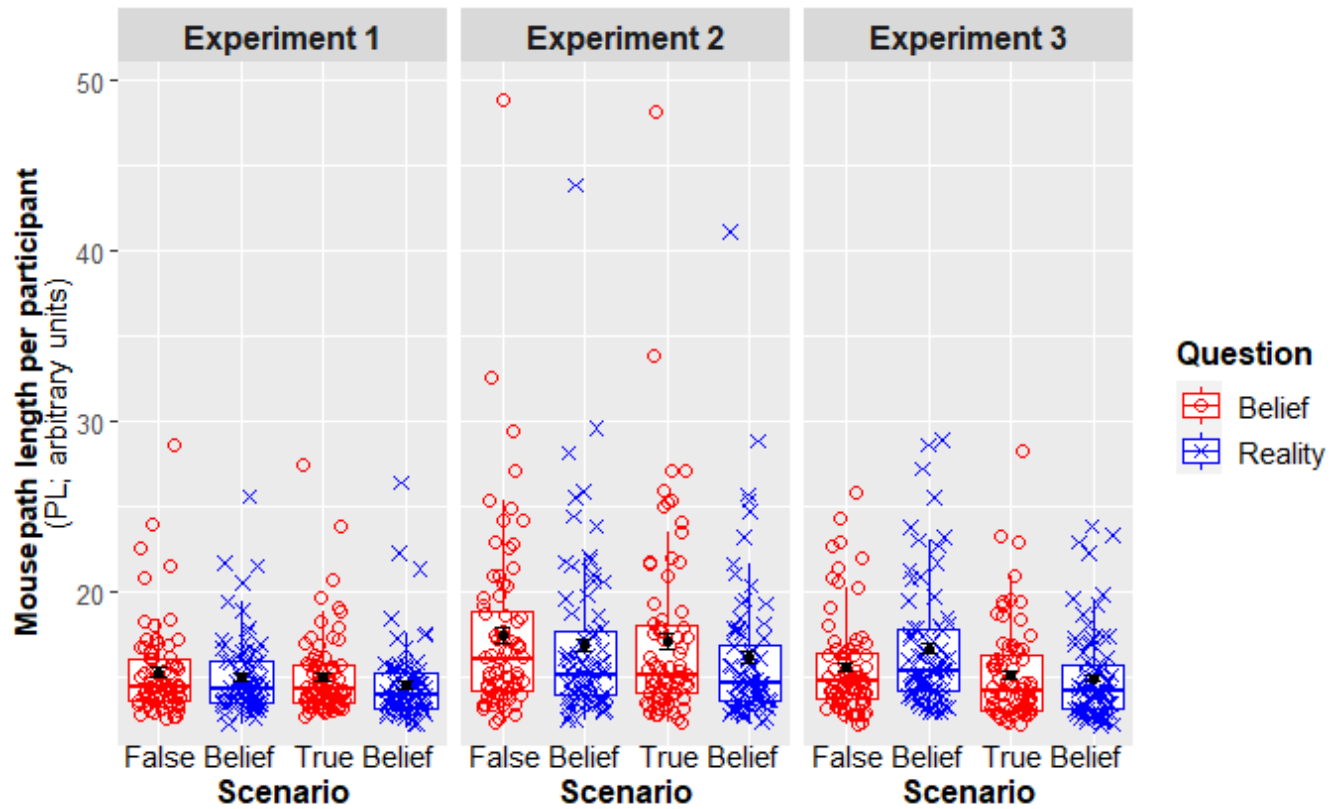


Figure 2. PL data in Experiments 1 – 3. Boxes indicate 25th - 75th percentile (IQR), with the middle line indicating the median. The whiskers indicate ± 1.5 IQR. Each coloured ring / cross is an individual participant's mean. The black dot indicates the condition mean, with error bars representing 95% CI (i.e., 1.96 SE).

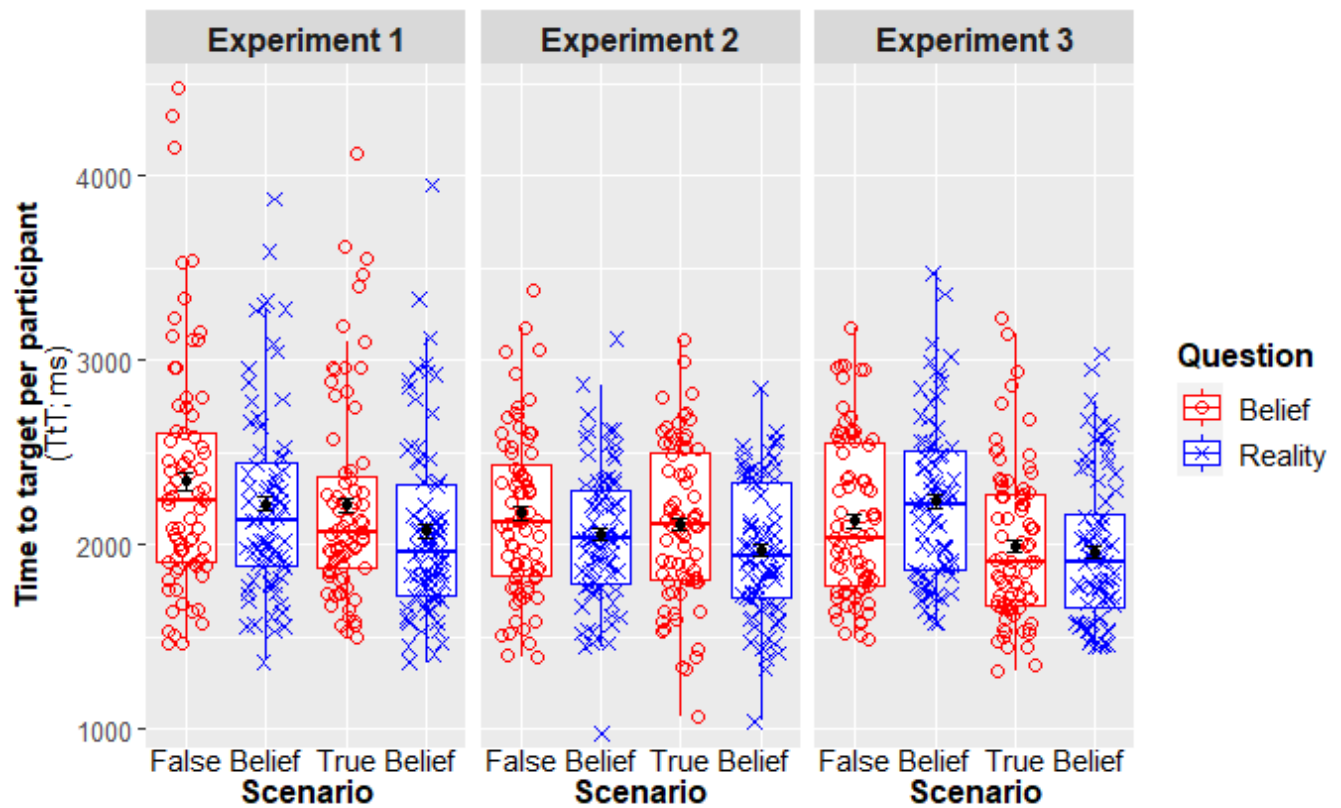


Figure 3. TtT data in Experiments 1 – 3. Boxes indicate 25th - 75th percentile (IQR), with the middle line indicating the median. The whiskers indicate ± 1.5 IQR. Each coloured ring / cross is an individual participant's mean. The black dot indicates the condition mean, with error bars representing 95% CI (i.e., 1.96 SE).

Sensitivity Analyses for all Bayes Factors calculated for Experiment 1

Figure 4 below reports sensitivity analyses for all Bayes Factors (BFs) calculated for Experiment 1. In all cases, the BFs reported in the main text and in SI-1 reflect support for the null hypothesis over the specific interaction predicted by egocentric bias: that the difference between belief and reality questions would be *greater* following false belief scenarios compared to true belief scenarios. BFs were calculated using Bayesian directional t-tests with a Cauchy prior of width 0.707 on mean differences for errors (after empirical logit transformation), MAD, AUC, PL, RT and TtT. The sensitivity analyses below demonstrate that for all measures, our data supported the null hypothesis across a range of reasonable prior widths.

Error (empirical logit transformed)

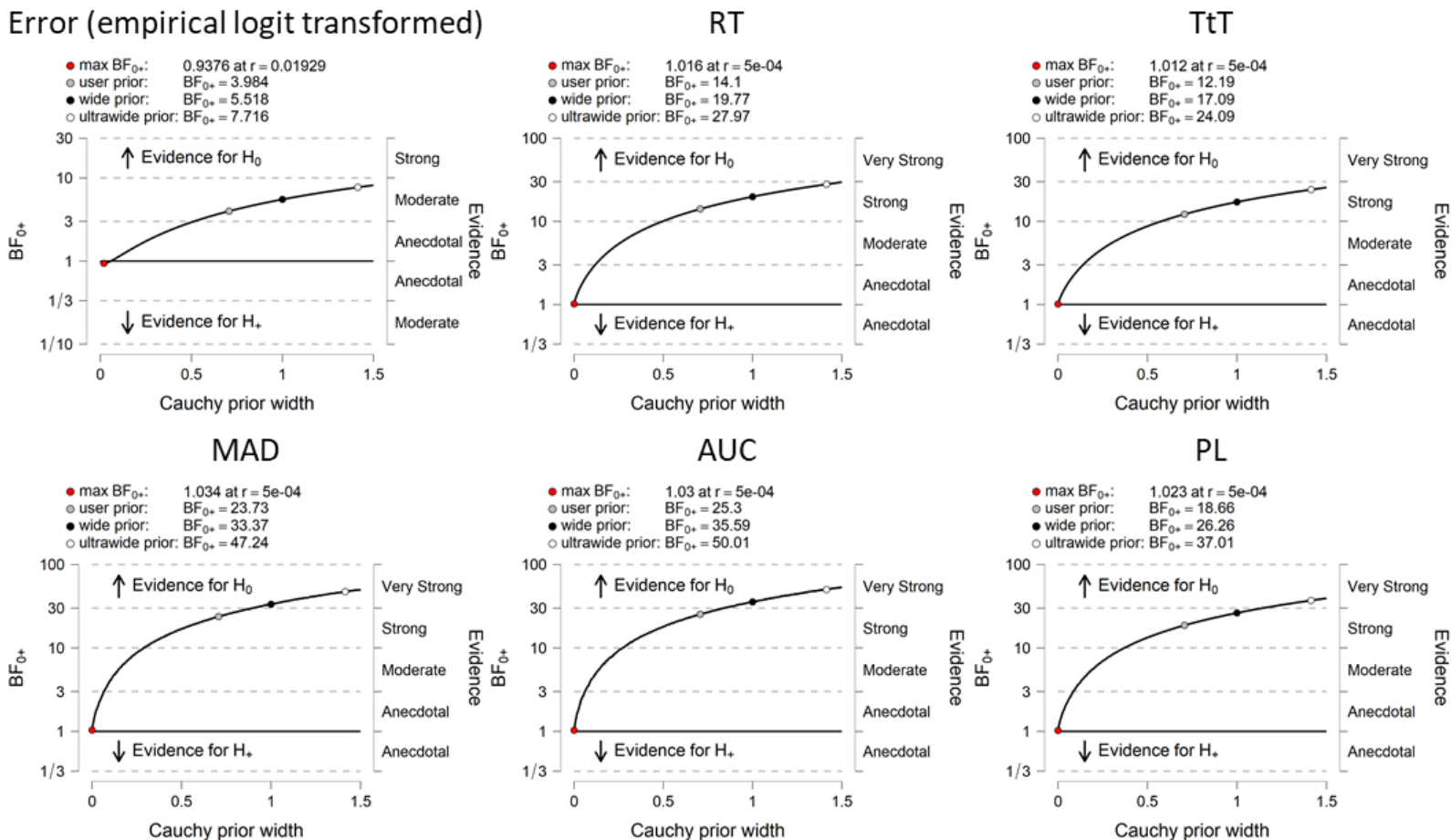


Figure 4. Bayesian sensitivity analyses for Bayes Factors (BF) calculated from Experiment 1 data. The “user prior” gives the BFs reported in our analyses, calculated with a width of 0.707.

Exploratory analyses: performance on first trials and first half of the experiment

Participant's performance on the first trial of each possible question by scenario combination was analysed to check whether there was any evidence for egocentric bias on these first trials. Participants' errors on these first trials (304 trials) were analysed using mixed-effect logistic regression. Models contained the fixed effects of question, scenario and their interaction and a random participant intercept only. Random participant slopes were not included given that each participant only provided 4 trials in total. This model found a significant main effect of scenario only (coefficient = 0.53, SE = 0.19, $z = 2.72$, $p = 0.007$), but no significant main effect of question (coefficient = 0.24, SE = 0.19, $z = 1.23$, $p = 0.22$) nor a significant interaction (coefficient = -0.15, SE = 0.19, $z = -0.75$, $p = 0.45$). Participants made more errors on false belief scenarios (False Belief : Belief 20% errors, False Belief : Reality 17% errors) compared to true belief scenarios (True Belief : Belief 11% errors, True Belief : Reality 5% errors) on the first trial.

For spatial (MAD, AUC, PL) and temporal measures (RT, TtT), first-trial performance was analysed only on correct trials that were not otherwise excluded (see main text for exclusion criteria). Participants with zero such trials or only one such trial were excluded ($n = 4$), leaving a total of 223 trials from 72 participants. Given that ANOVA analyses would require every participant to have four such trials (one trial per question by scenario combination), spatial and temporal measures were analysed using linear mixed models with the fixed effects of question, scenario and their interaction and a random participant intercept only. No fixed effects were found to be significant for any measure (for the interaction term, all $p > 0.59$; for all main effects, all $p > 0.08$).

Participant's performance on trials from only the first half of the experiment were also analysed, with errors analysed using mixed-effect logistic regression and MAD, AUC, PL, RT and TtT

analysed using 2 x 2 repeated-measures ANOVAs, as reported in the main text analyses of all trials. For every measure, significant main effects of scenario (all $p \leq 0.012$) and question (all $p \leq 0.008$) were observed, with all effects in the same direction as those reported in the main text and above for the analyses of all trials. There was no significant question by scenario interaction observed for any measure (all $p > 0.21$).

Finally, in order to assess whether participants changed their patterns of performance with regards to how question and scenario affected their responses, all analyses of error, MAD, AUC, PL, RT and TtT were reanalysed to include experiment half as a fixed factor with two levels (whether trials came from the first half or second half of the experiment), alongside all two-way and three-way interactions of half with question and scenario. If the effects of question or scenario were not stable across the time-course of the experiment, then one would predict significant interactions of half with these effects. No such interactions were observed on any measure (all $p > 0.12$).

In sum, the analyses of the first trials per condition therefore revealed no evidence for egocentric bias on any measure. However, with the exception of the analysis of errors, no significant main effects were seen either on the first trials, indicating that these analyses were most likely underpowered. Analysis of trials from the first half also revealed no evidence for egocentric bias. Those analyses did detect effects of question and scenario across all measures. With the exception of the question by scenario interaction for MAD and AUC (where analysis of all trials revealed a significant interaction in the opposite direction to that predicted by egocentric bias – this interaction was not significant when analysing trial from the first half of the experiment only), the effects observed in the first half matched those reported for all trials in the main text and above. Furthermore, there was no evidence that experiment half interacted with any of these effects. This therefore suggests that these effects were stable across the experiment.

Experiment 2

Spatial data – AUC and PL

The pattern of ANOVA effects for AUC matched those reported in the main text for MAD. Both main effects of Question ($F(1, 75) = 21.7, p < 0.001, \eta_g^2 = 0.018$) and Scenario ($F(1, 75) = 5.40, p = 0.023, \eta_g^2 = 0.005$) were significant, but there was no significant interaction ($F(1, 75) = 2.92, p = 0.092, \eta_g^2 = 0.003$). Bayesian analysis of the mean differences found that the null hypothesis was 20.4 times more likely than the alternative hypothesis that the difference in AUC between Belief and Reality questions was *greater* in the false belief scenarios than in the true belief scenarios. There was no evidence that the AUC data were bimodally distributed (HDS: all $ps \geq 0.767$). There was therefore no evidence from the AUC data in support of egocentric bias.

ANOVA of the PL data also showed the same pattern of significant main effects of Question ($F(1, 75) = 22.3, p < 0.001, \eta_g^2 = 0.008$) and Scenario ($F(1, 75) = 14.5, p < 0.001, \eta_g^2 = 0.002$), but no interaction ($F(1, 75) = 1.17, p = 0.283, \eta_g^2 < 0.001$). Bayesian analysis of the mean differences found that the null hypothesis was 15.5 times more likely than the alternative hypothesis that the difference in PL between Belief and Reality questions was *greater* in the false belief scenarios than in the true belief scenarios. There was no evidence that the PL data were bimodally distributed (HDS: all $ps \geq 0.993$). As with MAD and AUC, analysis of the PL data therefore provided no evidence in support of the presence of egocentric bias.

Temporal data – TtT

ANOVA of TtT found significant main effects of both Question ($F(1, 75) = 60.3, p < 0.001, \eta_g^2 = 0.025$) and Scenario ($F(1, 75) = 20.3, p < 0.001, \eta_g^2 = 0.006$), but no significant interaction ($F(1, 75) = 0.491, p = 0.486, \eta_g^2 < 0.001$). As a composite measure of IT and RT, this pattern of effects reflects

those observed in those two measures, as reported in the main text: for IT a significant main effect of Scenario, but no significant main effect of Question and no significant interaction; for RT a significant main effect of Question, but no significant main effect of Scenario and no significant interaction. Bayesian analysis of the mean differences found that the null hypothesis was 12.8 times more likely than the alternative hypothesis that the difference in TtT between Belief and Reality questions was *greater* in the false belief scenarios than in the true belief scenarios. TtT therefore provided no evidence in support of the presence of egocentric bias.

Sensitivity Analyses for all Bayes Factors calculated for Experiment 2

Figure 5 below reports sensitivity analyses for all Bayes Factors (BFs) calculated for Experiment 2. In all cases, the BFs reported in the main text and in SI-1 reflect support for the null hypothesis over the specific interaction predicted by egocentric bias: that the difference between belief and reality questions would be greater following false belief scenarios compared to true belief scenarios. BFs were calculated using Bayesian directional t-tests with a Cauchy prior of width 0.707 on mean differences for errors (after empirical logit transformation), MAD, AUC, PL, RT and TtT. The sensitivity analyses below demonstrate that for all measures, our data supported the null hypothesis across a range of reasonable prior widths.

Error (empirical logit transformed)

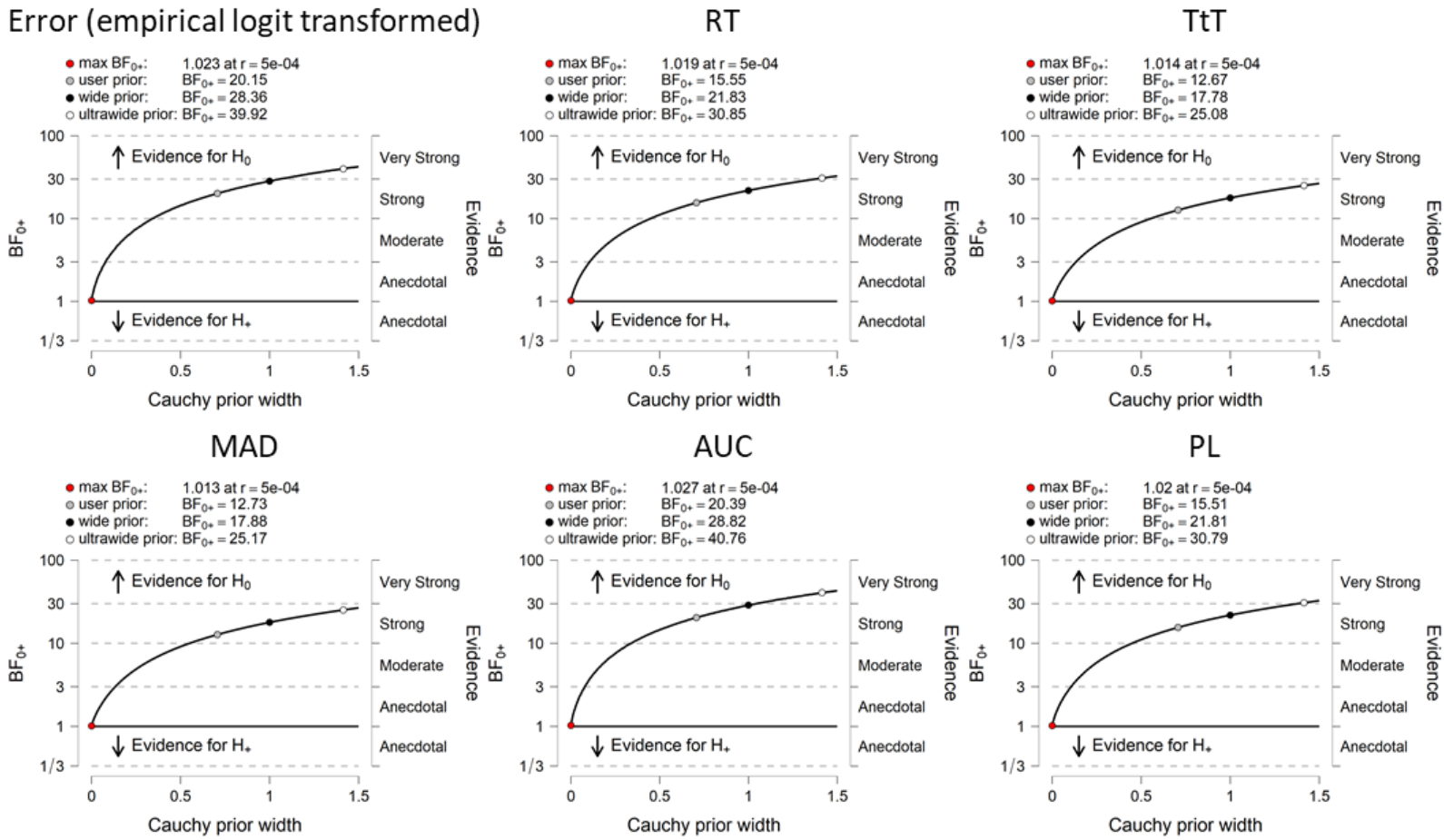


Figure 5. Bayesian sensitivity analyses for Bayes Factors (BF) calculated from Experiment 1 data. The “user prior” gives the BFs reported in our analyses, calculated with a width of 0.707.

Exploratory analyses: performance on first trials and first half of the experiment

As in Experiment 1, participant's performance on the first trial of each possible question by scenario combination was analysed. Participants' errors on these first trials (304 trials) were analysed using mixed-effect logistic regression as described above. The model found a significant main effect of scenario only (coefficient = 0.70, SE = 0.28, $z = 2.47$, $p = 0.013$), but no significant main effect of question (coefficient = 0.43, SE = 0.27, $z = 1.60$, $p = 0.11$) nor a significant interaction (coefficient = -0.51, SE = 0.27, $z = -1.88$, $p = 0.060$). Participants made more errors on false belief scenarios (False Belief : Belief 13% errors, False Belief : Reality 15% errors) compared to true belief scenarios (True Belief : Belief 11% errors, True Belief : Reality 3% errors) on the first trial.

For spatial (MAD, AUC, PL) and temporal measures (RT, TtT), first-trial performance was analysed only on correct trials that were not otherwise excluded (see main text for exclusion criteria). Participants with zero such trials or only one such trial were excluded ($n = 4$), leaving a total of 239 trials from 72 participants. Spatial and temporal measures were analysed using linear mixed models with the fixed effects of question, scenario and their interaction and a random participant intercept only. The only main fixed effect found to be significant for any measure was a main effect of question upon TtT (coefficient = 78, SE = 29, $t\text{-value} = 2.65$, $p = 0.009$), with belief questions responded to slower than reality questions on the first trial. For the interaction of question and scenario, no effects were significant for any measure (all $p > 0.11$).

Participant's performance on trials from only the first half of the experiment were also analysed, with errors analysed using mixed-effect logistic regression and MAD, AUC, PL, RT and TtT analysed using 2 x 2 repeated-measures ANOVAs. For errors, the final model found a significant main effect of scenario (coefficient = 0.50, SE = 0.24, $z = 2.08$, $p = 0.038$), and question (coefficient = 0.45, SE

= 0.18, $z = 2.43$, $p = 0.016$) but no significant interaction (coefficient = -0.38, SE = 0.24, $z = -1.60$, $p = 0.11$). Participants made more errors on belief questions (false belief scenario = 5%; true belief scenario = 4%) compared to reality questions (false belief scenario = 5% ; true belief scenario = 1%), and on false belief compared to true belief scenarios. For spatial measures and TtT, significant main effects of scenario (all $p \leq 0.018$) and question (all $p < 0.001$) were observed, with all effects in the same direction as those reported in the main text and above for the analyses of all trials, and no significant question by scenario interaction observed for any measure (all $p > 0.28$). For RT, a significant main effect of question was observed ($p < 0.001$), but no main effect of scenario ($p = 0.31$) nor an interaction ($p = 0.29$).

Finally, all analyses of error, MAD, AUC, PL, RT and TtT were reanalysed to include experiment half as a fixed factor with two levels (whether trials came from the first half or second half of the experiment), alongside all two-way and three-way interactions of half with question and scenario. No such interactions were observed on any measure (all $p > 0.096$).

In sum, the analyses of the first trials per condition therefore revealed no evidence for egocentric bias on any measure. However, with the exception of the analysis of errors and TtT, no significant main effects were seen on the first trials, indicating that these analyses were most likely underpowered. Analysis of trials from the first half also revealed no evidence for egocentric bias. Those analyses did detect effects of question and scenario across all measures. With the exception of the additional main effect of question observed for errors (analysis of all trials did not find this effect), the effects observed in the first half matched those reported for all trials in the main text and above for all measures. Furthermore, there was no evidence that experiment half interacted with any of these effects. This therefore suggests that these effects were stable across the experiment.

Experiment 3

Spatial data – AUC and PL

ANOVA of AUC found the same pattern of effects as those reported in the main text for MAD: there were significant main effects of both Question ($F(1, 75) = 11.3, p = 0.001, \eta_g^2 = 0.014$) and Scenario ($F(1, 75) = 36.8, p < 0.001, \eta_g^2 = 0.051$), qualified by a significant interaction ($F(1, 75) = 14.5, p < 0.001, \eta_g^2 = 0.015$). The interaction was driven by the same pattern of effects as in the MAD data: participants showed greater attraction to the incorrect response on reality questions compared to belief questions in false belief scenario ($t(75) = 3.73, p < 0.001, d = 0.428$), but not in true belief scenario ($t(75) = -0.341, p = 0.734, d = 0.039$). Finally, there was no evidence of bimodality in the AUC data (all HDS $ps \geq 0.929$).

ANOVA of PL data also found the same pattern of effects as those reported in the main text for MAD, with significant main effects of Question ($F(1, 75) = 8.28, p = 0.005, \eta_g^2 = 0.005$) and Scenario ($F(1, 75) = 41.2, p < 0.001, \eta_g^2 = 0.030$) qualified by a significant interaction ($F(1, 75) = 17.5, p < 0.001, \eta_g^2 = 0.010$). Again, the interaction was driven by significantly longer mouse paths for reality questions compared to belief questions in false belief scenario ($t(75) = 3.85, p < 0.001, d = 0.441$), but not in true belief scenarios ($t(75) = 1.53, p = 0.130, d = 0.175$). There was no evidence that the PL Data were bimodally distributed (all HDS $ps \geq 0.882$).

Temporal data – TtT

ANOVA of TtT found no significant main effect of Question ($F(1, 75) = 2.69, p = 0.105, \eta_g^2 = 0.001$), but did find a significant main effect of Scenario ($F(1, 75) = 122, p < 0.001, \eta_g^2 = 0.054$), qualified by an interaction ($F(1, 75) = 15.7, p < 0.001, \eta_g^2 = 0.005$). The interaction was driven by participants showing longer times to give their response to reality compared to belief questions in

false belief scenarios ($t(75) = 3.37$, $p = 0.001$, $d = 0.387$), but not in true belief scenarios ($t(75) = 1.51$, $p = 0.136$, $d = 0.173$). TtT therefore showed a composite pattern of the IT and RT effects reported in the main text: no significant main effects of Question for either temporal measure, a significant main effect of Scenario for both measures, and a significant interaction for RT but not IT. As reported in the main text, the interaction for RT resulted from significant differences between reality and belief questions in false belief scenarios but not in true belief scenarios.

Comparison of Experiments 1 and 3

Across the measures reported above and in the main text, participants in Experiment 3 showed a different pattern of performance to those in Experiment 1. In order to assess whether these differences across experiments were themselves statistically significant, for spatial and temporal measures we conducted mixed $2 \times 2 \times 2$ ANOVAs including Experiment (Experiment 1, Experiment 3) as an additional between-subjects factor alongside the within-subject effects of Question (Belief, Reality) and Scenario (false belief scenario, true belief scenario), as reported in Table 2. Where significant three-way interactions are reported, these are decomposed into two further 2×2 mixed ANOVAs, investigating the effects of Experiment $\times$ Question separately for each scenario, as reported in Table 3. For errors, we carried out mixed-model logistic regression with Experiment, Question and Scenario, and all interaction terms, as fixed effects. Participant intercepts and slopes for Question, Scenario and their interaction were included as random effects. Regression coefficients for the fixed effects from that model are reported in Table 1.

Note that for all analyses we are primarily concerned with the interaction of Experiment with Question: significant three-way or two-way interactions involving the factors of Experiment and Question would indicate that participants did indeed respond differently in Experiment 3 when

instructed to attend to the female actor's beliefs, compared to Experiment 1 where they were under no such instruction. As argued in the main text, this would suggest that participants in Experiment 1 were not spontaneously prioritising encoding the female actor's beliefs, and that the effects of Question observed in Experiment 1 cannot be solely attributed to incidental differences in the question probes used for belief and reality. Such interactions were indeed observed across all spatial measures, RT and TtT, as explained below.

Table 1. Fixed effects for error data mixed-model logistic regression.

Fixed Effect	<i>Coefficient Estimate</i>	<i>SE</i>	<i>Z value</i>	<i>p value</i>	Significance
Experiment	-0.25	0.11	-2.23	0.026	*
Question	0.30	0.19	1.59	0.11	—
Experiment × Question	-0.11	0.094	-1.21	0.23	-
Scenario	1.23	0.19	6.54	< 0.001	***
Experiment × Scenario	-0.13	0.088	-1.47	0.14	-
Question × Scenario	-0.22	0.19	-1.18	0.24	-
Experiment × Question × Scenario	-0.03	0.095	-0.36	0.72	—

Note. Significance codes: — = non-significant, * = significant at the $p \leq 0.050$ level, ** = significant at the $p \leq 0.010$ level, *** = significant at the $p \leq 0.001$ level.

Errors

The only significant fixed effects were those of Scenario, with more errors on false belief compared to true belief scenarios, and of Experiment, with more errors in Experiment 1 than Experiment 3¹.

Spatial data – MAD, AUC and PL

There was a significant three-way interaction for MAD. The follow-up ANOVAs further found significant Experiment x Question interactions for both false belief scenarios and true belief scenarios. In Experiment 1 in false belief scenarios participants showed no significant difference in attraction towards the incorrect answer between belief and reality questions, but in Experiment 3 participants showed more attraction on reality compared to belief questions. In Experiment 1 in true belief scenarios participants showed greater attraction on belief compared to reality questions, but in Experiment 3 no significant difference in attraction between questions was observed.

The ANOVA of AUC did not find a significant three-way interaction. However, there was a significant two-way interaction of Experiment x Question. As can be seen in Figure 1 above, and as reported in the analyses of Experiments 1 and 3 above, this interaction was the result of participants in Experiment 1 showing greater attraction towards the incorrect answer on belief compared to reality

¹ This pattern differed from the results of comparisons made by the pre-registered ANOVA for errors, which found a significant three-way interaction between Experiment, Question and Scenario. Follow-up 2x2 ANOVAs for false belief scenarios found a significant main effect of Experiment, qualified by an Experiment x Question interaction: in Experiment 1 participants made more errors on belief questions compared to reality questions, but the reverse pattern was shown in Experiment 3, with more errors on reality compared to belief questions. In true belief scenarios, however, the follow-up ANOVA revealed only a main effect of question, with belief questions showing more errors than reality questions, and no interaction.

questions (albeit only significantly so in true belief scenarios), but on Experiment 3 participants showed greater attraction towards the incorrect answer on reality compared to belief questions (albeit only significantly so in false belief scenarios).

The ANOVA of PL did find a significant Experiment x Question x Scenario interaction. Decomposing that three-way interaction, we observed interactions of Experiment and Question in both the false belief scenarios and true belief scenarios. As can be seen in Figure 2 above, and as reported in the analyses of Experiments 1 and 3 above, these interactions can be explained by participants in Experiment 1 producing responses with a greater PL for belief compared to reality questions, whereas in Experiment 3 participants produced responses with a greater PL for reality compared to belief questions in false belief scenarios, and no difference between questions in true belief scenarios.

Temporal data – IT, RT and TtT

For IT, the 2 x 2 x 2 ANOVA only found a main effect of scenario, with participants taking longer to initiate their first movement on false belief compared to true belief scenarios. No other effects were significant.

For RT, a significant three-way interaction was observed. The follow-up ANOVAs found significant Experiment x Question interactions for both false belief scenarios and true belief scenarios. In Experiment 1 in both false belief and true belief scenarios participants took longer to respond to belief compared to reality questions. However, in Experiment 3 participants took longer to respond to reality compared to belief questions in false belief scenarios, and showed no significant difference in RT between question types in true belief scenarios.

For TtT a significant three-way interaction was also observed. Decomposing that three-way interaction, we observed interactions of Experiment and Question in both the true belief scenarios and the false belief scenarios. As can be seen in Figure 3 above, and as reported in the analyses of Experiments 1 and 3 above, these interactions can be explained by participants in Experiment 1 taking longer to answer belief compared to reality questions, whereas in Experiment 3 participants answered belief questions faster than reality questions in false belief scenarios, and showed no significant difference in TtT between questions in true belief scenarios.

Table 2. Summary of $2 \times 2 \times 2$ ANOVA for MAD, AUC, PL (spatial data) and IT, RT, TtT (temporal data)

Measure	Effect	<i>F</i>	<i>p</i>	η_g^2	Significance
MAD	Experiment	0.099	0.753	< 0.001	–
	Question	0.195	0.659	< 0.001	–
	Experiment $\times$ Question	16.2	< 0.001	0.008	***
	Scenario	60.8	< 0.001	0.029	***
	Experiment $\times$ Scenario	11.4	< 0.001	0.006	***
	Question $\times$ Scenario	21.7	< 0.001	0.009	***
	Experiment $\times$ Question $\times$ Scenario	4.69	0.032	0.002	*
AUC	Experiment	1.09	0.298	0.006	–

Measure	Effect	<i>F</i>	<i>p</i>	η_g^2	Significance
AUC	Question	1.09	0.297	< 0.001	—
	Experiment × Question	18.0	< 0.001	0.010	***
	Scenario	55.3	< 0.001	0.029	***
	Experiment × Scenario	10.6	0.001	0.006	***
	Question × Scenario	19.9	< 0.001	0.009	***
	Experiment × Question × Scenario	3.26	0.073	0.001	—
PL	Experiment	2.06	0.154	0.012	—
	Question	0.109	0.742	< 0.001	—
	Experiment × Question	17.9	< 0.001	0.005	***
	Scenario	40.3	< 0.001	0.017	***
	Experiment × Scenario	9.79	0.002	0.004	**
	Question × Scenario	18.4	< 0.001	0.005	***
	Experiment × Question × Scenario	6.82	0.010	0.002	**
IT	Experiment	0.063	0.802	< 0.001	—
	Question	1.11	0.294	< 0.001	—

Measure	Effect	<i>F</i>	<i>p</i>	η_g^2	Significance
IT	Experiment × Question	1.44	0.232	< 0.001	–
	Scenario	10.7	0.001	0.002	**
	Experiment × Scenario	0.738	0.392	< 0.001	–
	Question × Scenario	2.06	0.153	< 0.001	–
	Experiment × Question × Scenario	0.270	0.604	< 0.001	–
RT	Experiment	2.27	0.134	0.014	–
	Question	14.0	< 0.001	0.002	***
	Experiment × Question	37.3	< 0.001	0.005	***
	Scenario	154	< 0.001	0.020	***
	Experiment × Scenario	3.88	0.051	< 0.001	–
	Question × Scenario	14.1	< 0.001	0.002	***
	Experiment × Question × Scenario	6.12	0.014	< 0.001	*
TtT	Experiment	3.10	0.080	0.019	–
	Question	15.6	< 0.001	0.003	***
	Experiment × Question	36.4	< 0.001	0.007	***

Measure	Effect	<i>F</i>	<i>p</i>	η_g^2	Significance
TtT	Scenario	198	< 0.001	0.030	***
	Experiment × Scenario	5.65	0.019	< 0.001	*
	Question × Scenario	9.98	0.002	0.001	**
	Experiment × Question × Scenario	4.97	0.027	< 0.001	*

Note. Significance codes: – = non-significant, * = significant at the $p \leq 0.050$ level, ** = significant at the $p \leq 0.010$ level, *** = significant at the $p \leq 0.001$ level. All *df* (1, 150).

Table 3. Summary of 2×2 ANOVA, decomposing the significant three-way interactions in *Table 1* for MAD, PL, RT and TtT

Scenario	Measure	Effect	<i>F</i>	<i>p</i>	η_g^2	Significance
False Belief	MAD	Experiment	0.516	0.474	0.003	—
		Question	7.05	0.009	0.011	**
		Experiment $\times$ Question	11.7	< 0.001	0.017	***
	PL	Experiment	4.47	0.036	0.025	*
		Question	6.27	0.013	0.005	*
		Experiment $\times$ Question	14.7	< 0.001	0.012	***
	RT	Experiment	1.35	0.247	0.009	—
		Question	0.069	0.793	< 0.001	—
		Experiment $\times$ Question	28.4	< 0.001	0.009	***
True Belief	MAD	Experiment	1.74	0.189	0.011	—
		Question	0.591	0.443	< 0.001	—
		Experiment $\times$ Question	27.8	< 0.001	0.011	***
	PL	Experiment	1.57	0.212	0.010	—
		Question	25.0	< 0.001	0.008	***
		Experiment $\times$ Question				
	RT	Experiment				
		Question				
		Experiment $\times$ Question				
	TtT	Experiment				
		Question				
		Experiment $\times$ Question				

Scenario	Measure	Effect	<i>F</i>	<i>p</i>	η_g^2	Significance
True Belief	MAD	Experiment × Question	7.55	0.007	0.002	**
		PL	0.348	0.556	0.002	—
		Question	18.9	< 0.001	0.005	***
		Experiment × Question	4.36	0.039	0.001	*
	RT	Experiment	3.42	0.066	0.022	—
		Question	43.6	< 0.001	0.007	***
		Experiment × Question	13.4	< 0.001	0.002	***
		TtT	4.80	0.030	0.030	*
	TtT	Question	41.3	< 0.001	0.009	***
		Experiment × Question	17.2	< 0.001	0.004	***

Note. Significance codes: — = non-significant, * = significant at the $p \leq 0.050$ level, ** = significant at the $p \leq 0.010$ level, *** = significant at the $p \leq 0.001$ level. All *df* (1, 150).

Standardised time bin analyses of x and y coordinates

The spatial measures reported in the main text (MAD) and in the supplementary analyses above (AUC and PL) each provide a measurement indexing overall attraction towards the incorrect answer on any given trial. These measures therefore summarise spatial attraction across all time-points within a given trial. While these measures are widely used in mouse tracking, a possible concern is that by only giving one data-point per trial, they may be insensitive to differences between conditions that may emerge only in a specific time-window in participants' responses. We therefore provide below an analysis of the spatial coordinates (x and y) of participants' responses at every percentile time-bin (for example of this approach, see: Bartolotti & Marian, 2012; Dale et al., 2007; Farmer, Anderson, et al., 2007; Farmer, Cargill, et al., 2007; Incera & McLennan, 2016; Marghetis et al., 2014; Spivey et al., 2005; Yamamoto et al., 2016).

For each participant, average x and y coordinates for each of the 101 standardised time bins were computed for each experimental condition. Then at each of the 101 time bins, a 2 x 2 ANOVA was performed with the factors Question and Scenario for both x and y coordinates. These analyses therefore test at which time bins, if any, main effects or an interaction of Question and Scenario emerge for Experiments 1 – 3. The presence of significant interaction effects at any time bin were followed-up with subsequent t-tests. These analyses therefore allowed us to confirm for Experiments 1 and 2 that there was no evidence of egocentric bias at any given time bin.

Bootstrapping simulations

Given that this analysis produces 101 ANOVAs for both the x and y coordinates for each experiment, we followed a widely used (e.g., Bartolotti & Marian, 2012; Dale et al., 2007; Farmer, Anderson, et al., 2007; Farmer, Cargill, et al., 2007; Incera & McLennan, 2016; Marghetis et al., 2014;

Spivey et al., 2005; Yamamoto et al., 2016) bootstrapping simulation method to avoid the problem of multiple comparisons (and increasing Type I error rate). This method provides an estimate of how many consecutive significant time bins one would expect by “chance” (5%) alone from a given dataset.

Recall that for each participant, our experimental dataset had four trajectories (one per experimental condition), with each trajectory divided into 101 standardised time bins. The bootstrapping simulation was performed as follows. First, for each experimental condition, at each time bin, four parameters were derived: the mean x and y position, and the associated standard deviations. These parameters were used to draw normal distributions, which were randomly sampled from 76 times, using the `rnorm` function of base R, to create a simulated dataset of an equivalent size to our experimental dataset (76 participants, each with four trajectories, each with 101 time bins). This was repeated 10000 times, to create 10000 simulated datasets.

These simulated datasets were then analysed as follows. At each time bin, a within-subjects 2×2 (Question $\times$ Scenario) ANOVA was performed to determine if there was a significant effect for the two main effects and the interaction on position at each time bin. The maximum run length of consecutive significant time bins for each effect was recorded.

The length of these runs of significant time bins represented the length of the run that occurred by chance alone in simulated data with parameters identical to our experimental dataset. For comparison to our experimental dataset, we looked for the longest run size that occurred in less than 5% of our simulated datasets (i.e., in less than 500 out of 10000 simulated datasets). If we then recorded in our actual data a run length for a given effect larger than this threshold, we could be confident that a run of that size was statistically significant.

Analysis of x and y coordinates in Experiments 1 – 3

Below we report for each Experiment where runs of significant effects occurred, where these were above the threshold from our boot strapping simulations, for both x and y coordinates. Figures 6 and 7 below provide plots of mean x and y coordinates per time bin for each condition for each Experiment. Differences in x coordinate positions between conditions reflects differences in the degree of attraction towards the incorrect answer, pulling participants' movement away from the correct answer (see discussion in Spivey et al., 2005; Bartolotti & Marian, 2012). Differences in y coordinate positions between conditions, however, is believed to reflect differences in the overall degree of attraction for *both* correct and incorrect answers combined, as both answers are pulling movement in the same upwards direction (see discussion in Spivey et al., 2005; Bartolotti & Marian, 2012). The presence of egocentric bias would predict greater attraction from the incorrect answer specifically when answering belief questions in false belief scenarios, resulting in, for a given time bin, x coordinate positions closer to the incorrect answer and potentially y coordinate positions closer to the top of the screen (reflecting the increased overall attraction from both answers combined). We therefore focus in the analyses of Experiments 1 and 2 below on runs of significant effects for the Question by Scenario interaction term and the interpretation of any such effects where present.

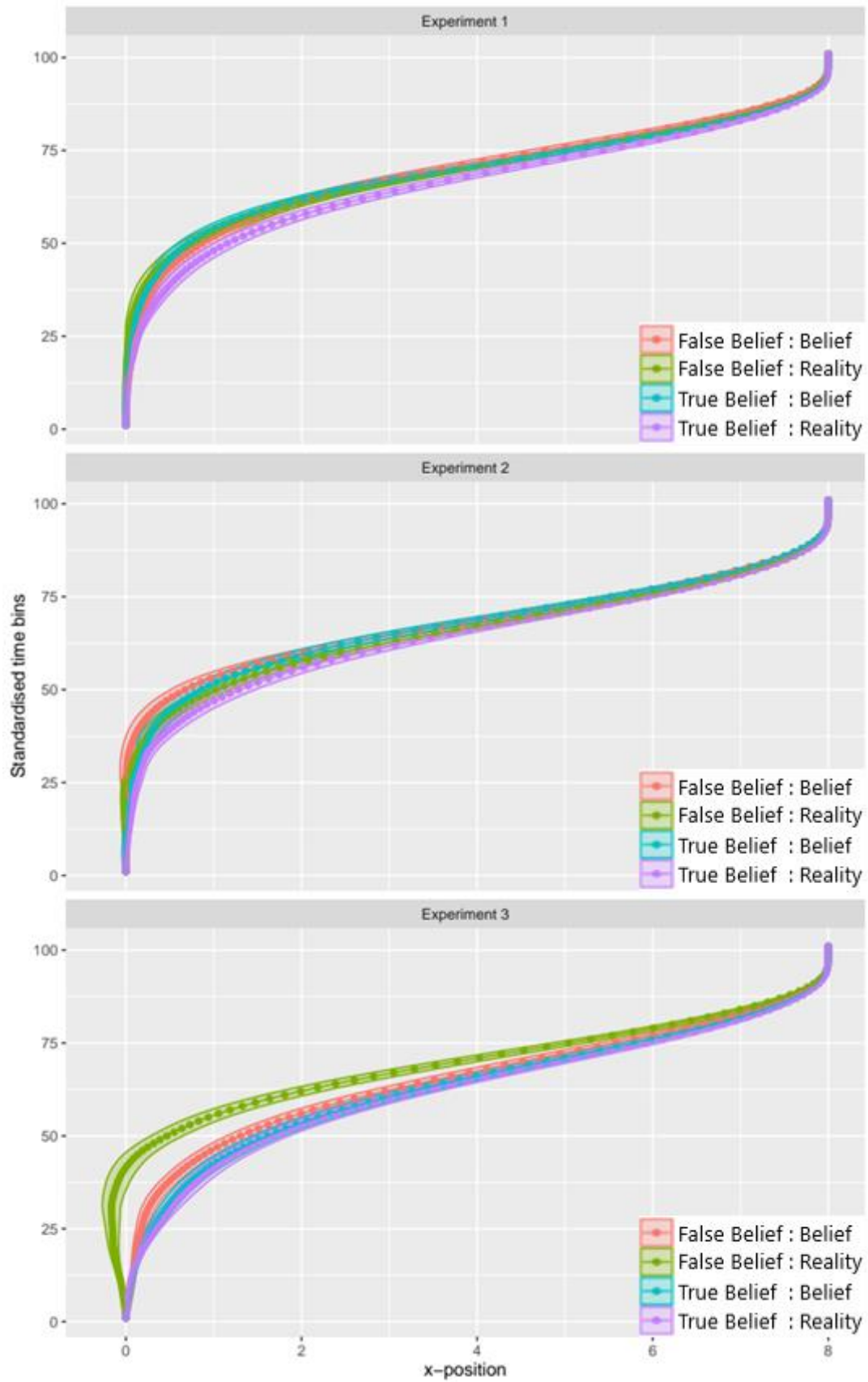


Figure 6. Experiments 1 – 3 x-position against standardised time bins.

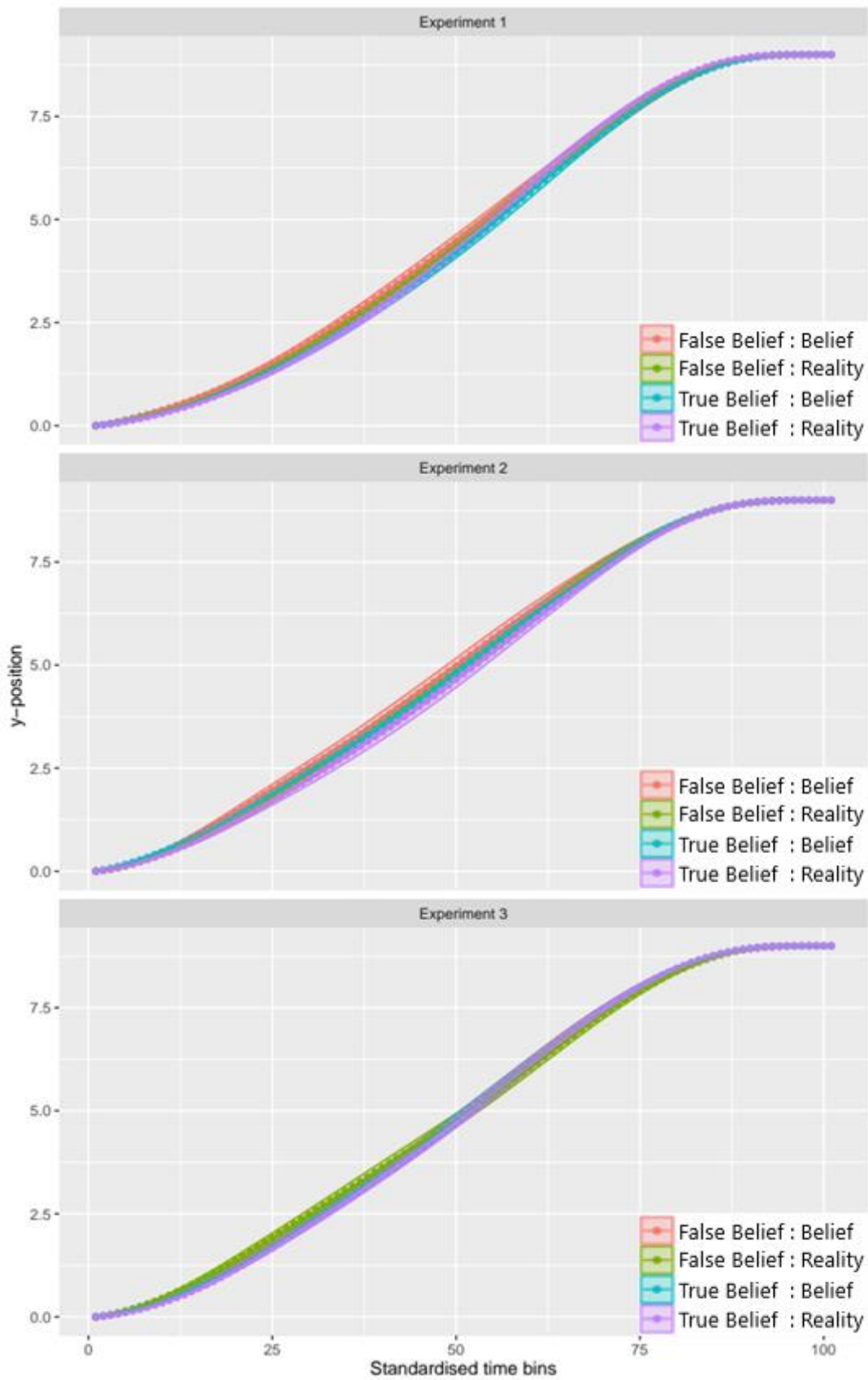


Figure 7. Experiment 1 – 3 standardised time bins against y-position

Experiment 1

The interaction term of the x position ANOVA was significant between time bins 19 and 71 inclusive (53 consecutive bins). This compares with a bootstrapping-derived threshold of eight consecutive time bins for the interaction term, and was thus statistically significant. Predominantly overlapping with this interaction there was also a significant main effect of Question from time bin 53 to time bin 96 (inclusive) and of Scenario from time bin 36 to 43 (inclusive) and from time bin 57 to 95 (inclusive).

To analyse the interaction, we performed *t*-tests comparing x coordinates for the belief and reality questions separately for the false belief and true belief scenarios at those time bins which showed a significant interaction effect. There were no significant differences across these time bins between question types in the false belief scenarios. In the true belief scenarios, there was a difference in x position from time bin 22 until time bin 71, with x positions in the belief condition closer to the incorrect answer compared to those for the reality condition. This confirmed that the interaction we observed was *not* driven by an egocentric bias, and matched the effects reported for the spatial measures reported above and in the main text.

ANOVA with the y position data showed no interaction at any time bin, thus providing no evidence for egocentric bias. The bins showed a main effect of Question from bin 5 to bin 32 and from bin 68 to bin 91 inclusive (significant against a threshold of 5). A main effect of Scenario was also observed from bins 3 to 57 inclusive (significant against a threshold of 5 also).

Experiment 2

ANOVA with the x-position data showed no significant interaction, again confirming that no evidence for egocentric bias was seen at any time point in participants' responses. There was a main

effect of Question between bins 39 and 95 inclusive (significant against a threshold of 12). The Scenario main effect was significant from bins 34–51 (significant against a threshold of 6).

There was an interaction in the y -position ANOVA from bins 2-5 (4 consecutive time bins). While small in length, this compares with a bootstrapping-derived threshold of three consecutive time bins for the interaction terms, and was thus statistically significant. To interpret the significant interaction, follow-up t -tests found that during this run of 4 bins there were no differences between belief and reality questions for the false belief scenarios. However, for the true belief scenarios, there were significant differences between belief and reality questions across this run of bins, with y positions for belief questions closer to the top of the screen than those for reality questions. This confirmed that the interaction we observed in these four time bins was *not* driven by an egocentric bias. Finally, outside of that short run of a significant interaction term, there were main effects of Question between bins 2 and 63 (significant against a threshold of 4), and main effects of Scenario between bins 26 and 69 (significant against a threshold of 4).

Experiment 3

While the data from Experiment 3 in-and-of themselves do not provide a direct test of the presence of egocentric bias, for completeness we report the results of the time bin analysis here demonstrating over which time bins significant effects were found. ANOVAs of the x -position showed an interaction from bins 3 – 90 inclusive. This compares with a bootstrapping-derived threshold of 18 consecutive time bins for the interaction term, and was thus statistically significant. Predominantly overlapping with that interaction, main effects of Question were also observed for bins 2-83 and main effects of Scenario for bins 9-93.

To interpret the interaction, *t*-tests on the *x* position data were performed comparing belief to reality questions separately for the false belief and true belief scenarios. For the false belief scenarios, significant differences were observed from bin 3–89, with *x* positions on reality questions closer to the incorrect answer. For the true belief scenarios, however, significant differences were only observed in a small run of 5 consecutive bins towards the end of the trial (77–81). Here, we observed that belief was closer to the incorrect answer than reality was. This pattern therefore matches that observed in the spatial measures reported above and in the main text, with greater attraction towards the incorrect answer observed for reality questions compared to belief questions in false belief scenarios, but the absence of such differences throughout the majority time-course of participants' responses in true belief scenarios.

In the *y*-position data, there was an interaction from time bins 2–38 inclusive. This compares with a bootstrapping-derived threshold of 6 consecutive time bins for the interaction term, and was thus statistically significant. No Question main effect was observed at any time bin, but a main effect of Scenario was observed in time bins 3–39 (overlapping with the interaction term) and in a short series of 8 consecutive time bins (86–93) towards the end of the trial (significant against a threshold of 4).

To interpret the interaction term, *t*-tests on the *y* position data were performed comparing belief to reality questions separately for the false belief and true belief scenarios. Within the range of bins specified by the interaction, in false belief scenarios significant differences were observed between belief and reality questions for time bins 2–37, with *y* positions for reality questions closer to the top of the screen than those for belief questions. In the true belief scenarios, however, significant differences were only observed in a short run of consecutive time bins in the earliest part of the response (time bins 3–15), where *y* positions for responses to belief questions were closer to the top

of the screen than those for reality questions. This pattern broadly matches that reported for the spatial measures above and in the main text, with differences between reality and belief responses observed in false belief scenarios, but such differences being typically absent during the majority of participants' responses in true belief scenarios.

References

- Bartolotti, J., & Marian, V. (2012). Language learning and control in monolinguals and bilinguals. *Cognitive Science*, 36(6), 1129–1147. <https://doi.org/10.1111/j.1551-6709.2012.01243.x>
- Dale, R., Kehoe, C., & Spivey, M. J. (2007). Graded motor responses in the time course of categorizing atypical exemplars. *Memory and Cognition*, 35(1), 15–28. <https://doi.org/10.3758/BF03195938>
- Farmer, T. A., Anderson, S. E., & Spivey, M. J. (2007). Gradiency and visual context in syntactic garden-paths. *Journal of Memory and Language*, 57(4), 570–595. <https://doi.org/10.1016/j.jml.2007.04.003>
- Farmer, T. A., Cargill, S., Hindy, N., Dale, R., & Spivey, M. (2007). Tracking the continuity of language comprehension: Computer mouse trajectories suggest parallel syntactic processing. *Cognitive Science*, 31(5), 889–909. <https://doi.org/10.1080/03640210701530797>
- Freeman, J. B., & Ambady, N. (2010). MouseTracker: Software for studying real-time mental processing using a computer mouse-tracking method. *Behaviour Research Methods*, 42, 226–241. <https://doi.org/10.3758/brm.42.1.226>
- Incera, S., & McLennan, C. T. (2016). The time course of within and between-language interference in bilinguals. *International Journal of Bilingualism*, 22(1), 88–99. <https://doi.org/10.1177/1367006916644688>
- Kieslich, P. J., & Henninger, F. (2017). Mousetrap: An integrated, open-source mouse-tracking package. *Behaviour Research Methods*, 49, 1652–1667. <https://doi.org/10.3758/s13428-017-0900-z>
- Marghetis, T., Núñez, R., & Bergen, B. K. (2014). Doing arithmetic by hand: Hand movements during exact arithmetic reveal systematic, dynamic spatial processing. *Quarterly Journal of Experimental Psychology*, 67(8), 1579–1596. <https://doi.org/10.1080/17470218.2014.897359>
- Spivey, M. J., Grosjean, M., & Knoblich, G. (2005). Continuous attraction toward phonological competitors. *Proceedings of the National Academy of Sciences*, 102(29), 10393–10398. <https://doi.org/10.1073/pnas.0503903102>

Yamamoto, N., Incera, S., & McLennan, C. T. (2016). A Reverse Stroop Task with Mouse Tracking.

Frontiers in Psychology, 7. <https://www.frontiersin.org/articles/10.3389/fpsyg.2016.00670>