*1. Methods*

*1.1 Participants*

After data cleaning (see below, 1.4), the total number of participants was *N* = 48 (Mean age = 19.4, SD = 1.976, 37 female, 11 male, 0 other). Due to external circumstances (COVID-19) the groups were slightly unbalanced, with 20 participants in the baseline group and 28 participants in the mind-tied group. All provided informed consent and had normal to corrected vision, and no history of any neurological pathology. The experimental procedure received ethical approval from the Victoria University of Wellington School of Psychology Human Ethics Committee.

*1.2. Stimuli and Apparatus*

The eye-tracking setup was analogous to previous experiments. We also used the same video stimuli as in previous experiments. However, in contrast to previous experiments the videos were cut off right before the hand appeared out of the occluder. This was done by ending the video loop 8 frames prior to the “hand appearing” event in each video.

Unlike previous experiments, this experiment featured a question prompt after each trial upon participants were asked to respond vocally. This was the case for both experimental conditions. Answers were recorded using a (insert mic here), as well as noted down manually by the experimenter. The voice records were stored and analysed off-line after the main experiment sequence was completed using Google voice recognition software.

In the mind-tied group, this experiment also included a secondary task. The point of this task was to exert cognitive load upon the participants during the measurement period of the experimental trials. The task consisted of an auditory discrimination task (see 1.3). Tones consisted of two 30 ms segments of a simple sine wave sampled at either 800 or 1200 ms. Tones were offset with 266 ms. The tones were selected at random, in counter-balanced fashion (i.e., in 50% of cases the tones would match, in the other 50% of cases they would be different).

*1.3 Procedure*

Up until first calibration of the Eyetracker, the experimental procedure was the same as in previous experiments. However, after the first calibration participants were presented with the instructions for the main body of the experiment, which also introduced the explicit prediction task (see below). In the mind-tied condition, another instruction was added in regards to the tone detection task. That is, participants were instructed to memorise if they thought the tones they heard were the same tone (i.e., the same frequency) or different from each other. We then tested if the participants understood the instructions by having them complete a random test-trial. This test trial would be repeated (each time with a new random video) until the participants showed sufficient understanding of the tasks.

A total of 60 trials were displayed in the main section of the experiment. These were divided over 5 blocks of 12 trials each. Per block, we displayed 4 false belief trials with 2 whole hand grip and 2 precision grip trials, 4 true belief trials with 2 whole hand grip and 2 precision grip videos, and 4 no-grip trials, of which 2 were false belief and 2 were true belief (although this is only nominal in this condition as there is no motor information to signal intention of the agent). Between blocks the participant had an opportunity to disengage from the eye-tracker setup and take a break. The participant could choose when to resume the experiment. The eye-tracker was recalibrated before a new block of experimental trials was commenced.

Each trial featured the same videos as in previous experiments, with the notable exception that the video display loops were cut off 8 frames before the hand appeared out of the occluder. As such, the participants never at any point saw the hand actually reaching for a cup. During phases 1 and 2, there was no difference with previous experiments. However, in the mind-tied condition the tones were presented after the fixation arrow disappeared. Specifically, the first sound was displayed 3 frames after the occluder dropped, while the second tone was displayed 11 frames after this trigger event. After the video was over, a prompt appeared with the explicit prediction question. Participants had 3 seconds to respond with a vocal “left” or “right”, at normal conversational level. In the mind-tied group, this prompt was followed up with a prompt that asked if the tones they heard were the same or different. Participants responded to this question by a keypress on the keyboard: “S” for “same” and “D” for “different”.

After the final block, we had a short debriefing session. We had the same two binary response questions (yes/no) as in previous experiments. Specifically, “When the cups are raised, is the agent able to see the cups?" and "When the cups are lowered, is the agent able to see the cups?".

We also included a third open ended question: “What made you decide in picking left or right on the prediction task?”. If the participant indicated that they based their answer on hand preshaping, we would ask the follow up question: “did you take into account that the agent couldn’t see the cups being swapped?”. Answers were coded by the experimenter based on similarities in participants’ responses. We were of course most interested in responses that indicated that the participants used either the hand preshaping, or the hand preshaping in combination with the agent’s belief state (i.e., if they were aware that the agent could not see the cups being swapped).

*1.4 Data-analysis*

To a large extent, the methodology we used was analogous to that in previous experiments, with a few exceptions due to differences in design. All experimental data (eye-tracking data and behavioural data) were imported into R for analysis. We withheld participants from analysis that had a wrong answer on the essential debrief questions or had an incomplete number of data points (2 belief conditions *x* 3 grip conditions, 6 data points). On a trial level, we removed all trials that had 1) a framerate lower than 59.6 FPS, 2) were not completed, 3) had timing issues due to eye-tracker and video streams being mismatched, or 4) had no fixations during our Time of Interest (TOI) in either one of the Areas of Interest (AOI).

AOIs were constructed in analogy with previous experiments. TOI was defined as the interval between the disappearance of the fixation arrow and the end of the video.

For each trial, we calculated a Differential Looking Time Score to represent overall gaze bias. First of all, we calculated the sum of the durations of all fixations within each of the AOIs (defined as the ‘target’ AOI and the ‘distractor’ AOI), per trial. DLTS was then defined, per trial, as: the total looking time for the target AOI, minus total looking time for the distractor AOI, divided by the total length of the TOI.

*Where:*

*t = total looking time per AOI during TOI*

*l = total duration of TOI*

This results in a score between -1 and 1, where -1 equals the full length of TOI spent fixating on the distractor AOI, and 1 equals the full length of the TOI spent looking at the target. 0 would hence indicate no overall bias towards either AOI during the TOI. These DLTS scores per trial were then averaged per belief *x* grip combinations, for a total of 6 data points per participant (2 belief conditions *x* 3 grip conditions).

For accuracy on the question at the end of the video (*“on what side of the screen will the hand appear?”*), we coded answers that matched the outcome of the video according to the agent’s beliefs as 1, and the alternative as 0. As such, if a participant watching a whole-hand-grip false-belief video answers this question with the belief state of the agent in mind (i.e., reasoning akin to *“since the large cup has been swapped without the agent knowing it, and the agent wants to grab the large cup, the hand should appear on the side of the small cup”*), theanswer would be marked as 1. If the participant had reasoned in line with motor matching (*“the large cup is on the right, so the agent will grab the large cup”*), the answer would be marked as 0. We averaged answers per belief *x* grip combinations, for a total of 6 data points per participant (2 belief conditions *x* 3 grip conditions). Stated otherwise, per belief *x* grip condition (for 6 data points per participant), we calculated the proportion of correct answers.

In order to test our hypotheses on the effects of our experimental variables on both measures, we modelled our data using Linear Mixed-effects Models (LMM) using the *lme4* package in R. Group was a between-subjects fixed categorical factor, while belief and grip were within-subjects fixed categorical factors. Participants, as experimental units, were treated as a random factor and thus included in the model. We recognise that using regression models for proportional data (the variable ‘accuracy’) is based on binary data; hence we also checked the analyses under the assumption of a Gamma distribution in a General Linear Mixed Effects Regression Model, slightly offsetting 0 values. This did not lead to different conclusions as compared with applying an LMM. As such, it seemed justified to treat accuracy as a continuous response variable.

For follow-up tests on the belief *x* grip interaction factor, we used a Holm-corrected linear hypotheses testing function from the *multcomp* package in R. In order to compare different levels of grip to a baseline level, we essentially included a 4th level for this variable, which was equal to the baseline for that measurement (0 for DLTS, .5 for accuracy). We then fitted a LMM with only grip as fixed factor, with the baseline level as reference level (intercept) to estimate differences in outcome measurements between these levels and the reference level. Estimates for the degrees of freedom of these t-tests were calculated using the Satterthwaite method.

In the correlational analyses we used the Spearman’s rank-order correlation coefficient as neither DLTS nor accuracy on the explicit task were normally distributed. As such, the correlation coefficient reflects monotonic association rather then linear association between these outcome variables. We therefore use both linear model fits and LOESS (non-parametric local regression) fit in our visual representations of this correlation.

All reported p-values on any other follow-up tests comparing different levels of experimental variables (see *2.*6) were corrected for multiple comparisons using Holm’s method.

*2. Results*

*2.1 Differential Looking Time Score*

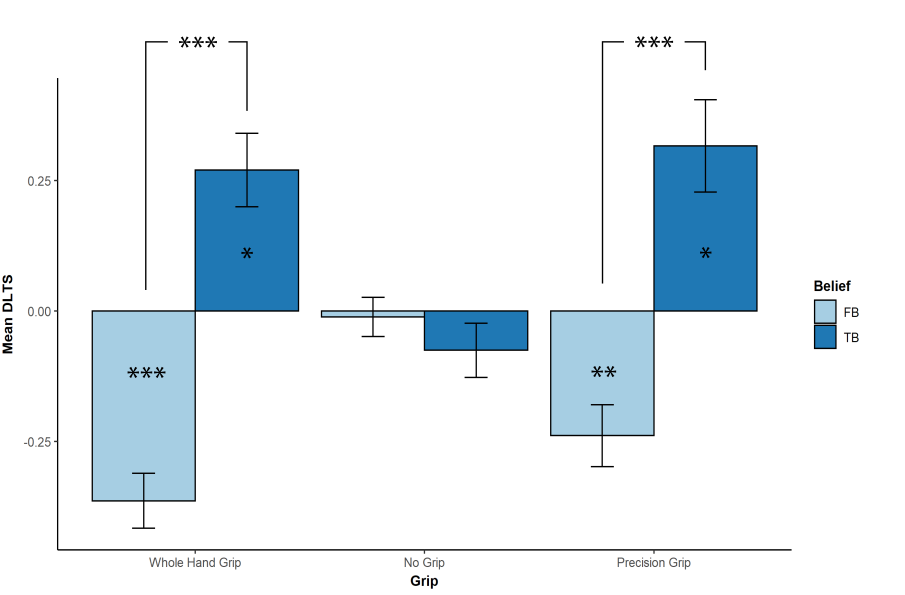
First of all, we tested the full model including the factors group (baseline or mind-tied), belief (true belief or false belief), grip preshaping (whole hand grip, precision grip or no grip), plus the interaction effect of belief and grip. The factor ‘participant’ was included as random factor. An ANOVA of this model revealed a significant main effect revealed a significant main effect of belief, F(2,235) = 50.136, p<.001, and a significant interaction effect between the factors belief and grip, F(2,235) = 15.57, p<.001. The main effect of group was not significant, F(1,46) = 0.762, p = .387, nor was the main effect of grip, F(2,235) = 1.899, p = .152.

We further also modelled the data in a similar fashion per group as independent sample. In this model, we included belief, grip, and the belief *x* grip interaction effect. An overview of the results of the analyses follows below.

*2.1.1 Baseline group*

Modelling the data of the baseline group separately, we found a significant main effect of belief, F(1,114) = 21.676, p<.001, a significant main effect of grip, F(2,114) = 4.948, p = 0.009, as well as a significant interaction effect between belief and grip, F(2,114) = 7.889, p <.001.

We found that within the whole hand grip condition, there was a significant difference between the DLTS scores in the true belief condition and those in the false belief condition, *t* = 4.676, p<.001. Similarly, we found a significant difference in the precision grip condition, with DLTS scores being significantly higher in the true belief condition as compared to the false belief condition, *t* = 4.098, p<.001. No such difference was found in the no grip condition, *z* = -.473, p = .637.

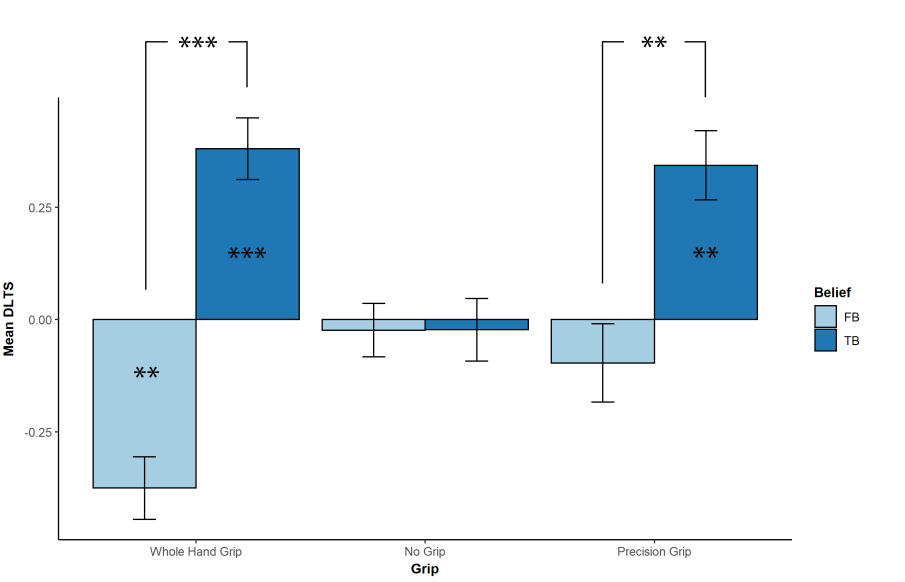


*Figure X. DLTS scores in the baseline group, per belief and grip condition. Error bars represent standard error of the mean. Stars represent the significance of the mean differences. Stars on a bar represent the significance of the mean of the belief/grip combination as compared to the baseline of 0 (no bias towards either target). Those above represent the difference between means per belief condition, i.e. comparing the true and false belief conditions per grip condition.*

We also compared DLTS scores for every combination of belief and grip (i.e., TB-WHG, TB-NG, TB-PG, FB-WHG, FB-NG, FB-PG) to the baseline of 0, which indicates no bias towards either target or distractor. For the true belief condition, DLTS scores were significantly higher than baseline (i.e., gaze was biased towards the target as compared to the distractor) for the whole hand grip condition, *t* = 2.0474, p = .045, as well as for the precision grip condition, *t* = 2.400, p = .019, but not for the no grip condition, *t* = -.574, p = .568. When looking at the DLTS scores for the false belief condition, we found that scores are significantly lower than baseline (i.e., gaze was biased towards the distractor as compared to the target) for the whole hand grip condition, *t* = -4.351, p<.001, and the precision grip condition, *t* = -2.859, p = .006, but not for the no-grip condition, *t* = -.138, p = .891.

*2.1.2 Mind-tied group*

When modelling the data of the mind-tied group independently, we found similar results as in the baseline group. Both the main effect of belief , F(1,135) = 33.13, p<.001 and grip, F(2,135) = 5.761, p = .004, were significant, as was the belief *x* grip interaction effect, F(2,135) = 8.335, p<.001.



*Figure X. DLTS scores in the mind-tied group, per belief and grip condition. Error bars represent standard error of the mean. Stars represent the significance of the mean differences.*

We also ran a follow up analysis of this model, comparing the scores in the true belief condition to those in the false belief condition per grip conditions. We found a significant difference in the whole hand grip condition (i.e., TB-WHG vs FB-WHG), *z* = 5.738, p<.001, and the precision grip condition (TB-PG vs FB-PG), *z* = 3.343, p<.01, but not for the no-grip condition, *z =* .008, p = .999.

We also compared every belief-grip combination with the baseline of zero. This revealed that mean DLTS scores were significantly higher in the true belief condition for both the whole hand grip (TB-WHG), *t* = 3.487, p<.001] and the precision grip (TB-PG), *t* = 3.149, p<.01. conditions, but not the no-grip condition (TB-NG), *t* = -.207, p = .837. When looking at the false belief cases, we found that mean DLTS scores were significantly lower than baseline in the whole hand grip condition (FB-WHG), *t* = -3.256, p<.01, but mean DLTS scores in the precision grip (FB-PG), *t* = -.838, p = .404, and the no-grip (FB-NG), *t* = -.205, p .838 conditions did not differ from baseline.

*2.2 Accuracy of explicit responses*

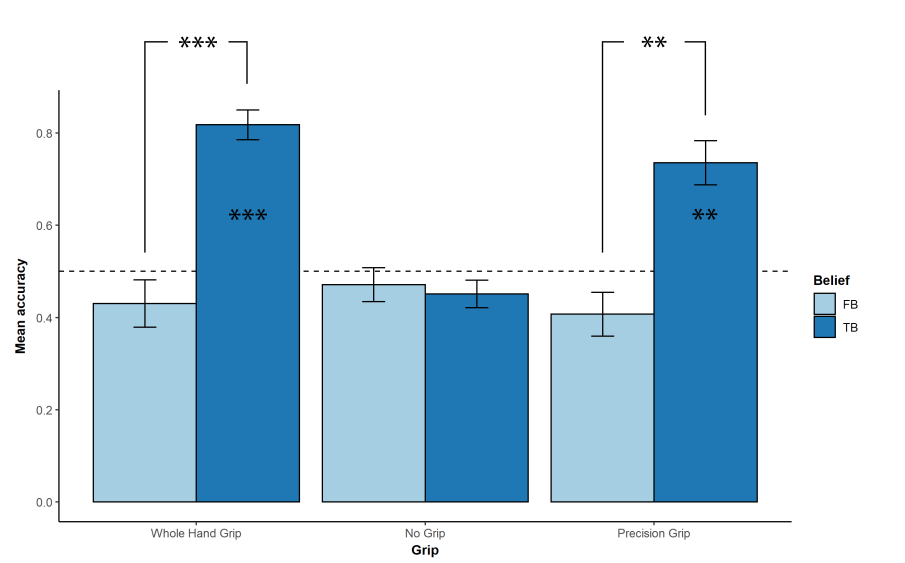
We followed a similar methodology here as with Differential Looking Time Scores (DLTS). First, we tested a model including the factors: group, belief, grip, plus the interaction effect between belief and grip. Again, ‘participant’ was included as a random factor. We found no significant effect of group, F(1,46) = .041, p = .84. Similar to the DLTS results, we found a significant affect of belief, F(1,235) = 45.215, p<.001, and a significant interaction effect of belief *x* grip, F(2,235) = 17.964, p<.001. Contrary to the DLTS results, we did find a significant main effect of grip here, F(2,230) = 3.46, p = .021.

Again, we modelled the data per group as independent samples as there was no main effect of group. We will take a closer at these below.

*2.2.1 Baseline group*

When modelling the data of the baseline group separately, we found a significant main effect of belief, F(1,114) = 17.859, p<.001, a significant main effect of grip, F(2,114) = 8.802, p<.001, as well a significant interaction effect between belief and grip, F(2,114) = 5.76, p<.01.

Again, we used a series of follow-up tests to take a closer look at the main effect of belief, per grip condition. We found that the accuracy of explicit responses was significantly higher in the true belief condition as compared to the false belief condition in the whole hand grip condition, *z* = 4.145, p<.001, as well as in the precision grip condition, *z* = 3.512, p<.01. However, no such difference was observed in the no-grip condition, *z* = -.211, p = .99.

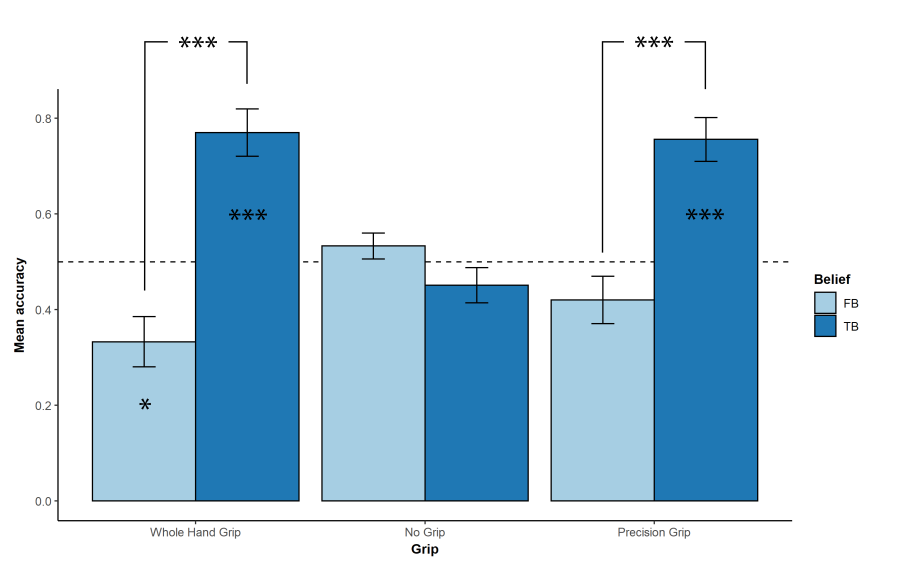


*Figure X. Accuracy of explicit responses scores in the baseline group, per belief and grip condition. Error bars represent standard error of the mean. Stars represent the significance of the mean differences. Stars on a bar represent the significance of the mean of the belief/grip combination as compared to chance level (.50). Those above represent the difference between means per belief condition, i.e. comparing the true and false belief conditions per grip condition.*

We also tested per belief-grip condition if participants performed higher or lower than chance level (.5). This analysis revealed that while in the true belief condition, participants scored significantly higher than chance level in both the whole hand grip (TB-WHG), *t* = 4.574, p<.001, and the precision grip condition (TB-PG), *t* = 3.387, p<.01, but not in the no-grip (TB-NG) condition, *t* = -.702, p = .485. However, in the false belief condition, participant performance did not differ from chance level in either the whole hand grip (FB-WHG), *t* = -.899, p = .371, the precision grip condition (FB-PG), *t* = -1.202, p = .234, or the no-grip condition (FB-NG), *t* = -.375, p = .71.

*2.2.2 Mind-tied group*

In the group that performed a secondary task, we found a significant main effect of belief, F(1,135) = 30.8, p<.001, a significant main effect of grip, F(2,135) = 10.454, p<.001, as well as significant interaction effect between belief and grip, F(2,135) = 12.188, p<.001.



*Figure X. Mean accuracy of explicit responses in the mind-tied group, per belief and grip condition. Error bars represent standard error of the mean. Stars represent the significance of the mean differences.*

Again, we looked at the main effect of belief by means of a series of follow-up tests, looking at the difference between the mean accuracy in the true and the false conditions per grip condition. We found that accuracy was significantly higher in the true belief condition as compared to the false belief condition in both the whole hand grip condition, *z* = 5.529, p<.001, and the precision grip condition, *z* = 4.237, p<.001, while no such difference was found in the no-grip condition, *z* = -1.036, p = .656.

We also looked at how participants performed compared to chance level. This follow-up analysis shows that in the true belief condition, accuracy is significantly higher than chance level in both the whole hand grip condition (TB-WHG), t = 3.992, p<.001, and the precision grip condition (TB-PG), *t* = 3.778, p<.001, but not the no-grip condition, *t* = -.721, p = .473. In the false belief condition, we found that the accuracy of participant responses was significantly lower than chance level in the whole hand grip condition (FB-WHG), *t* = -2.508, p = .014, while performance did not differ from chance level in both precision grip condition (FB-PG), t = -1.192, p = .237, and the no-grip condition (FB-NG), t = .497, p = .62.

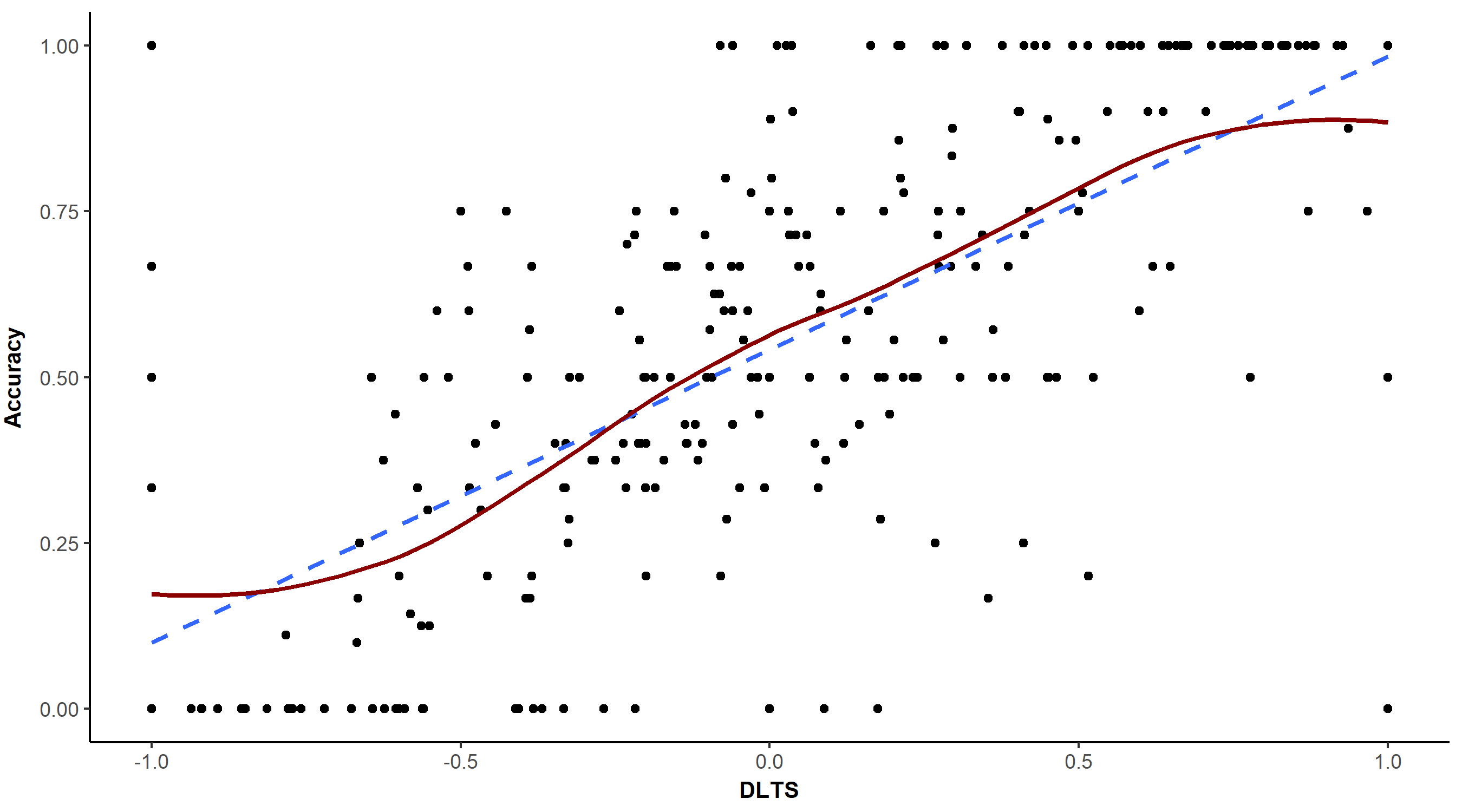
Note that this section might see revision for statistical reasons

*2.3 Correlation analysis between DLTS and accuracy of explicit responses*

We were also interested in comparing looking time bias (as measured by DLTS) and the accuracy of explicit responses in regards to the intention of the agent to reach for either the target cup or the distractor cup. In order to do so, we looked at the correlation patters between DLTS and response accuracy.

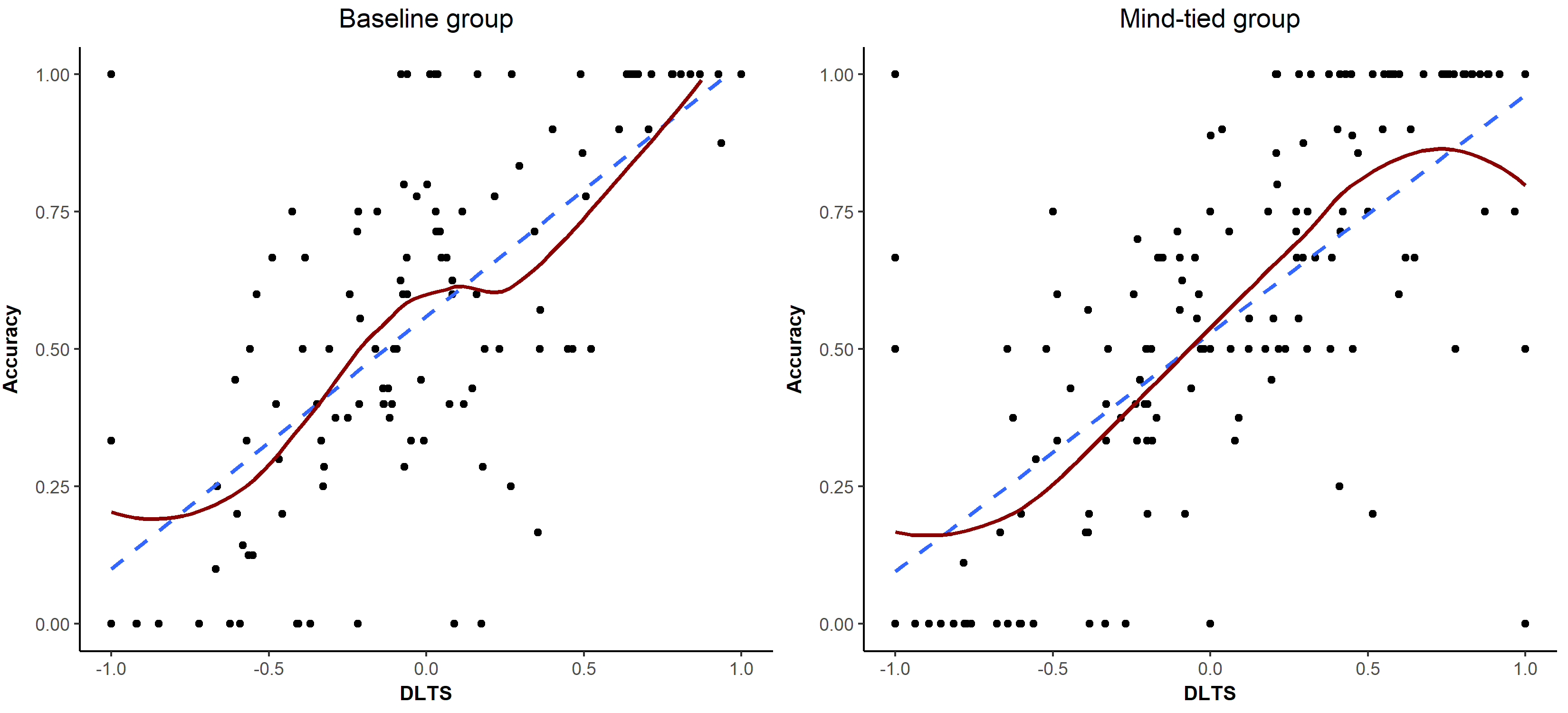
*2.3.1 Overall correlation*

First we looked at the correlation between DLTS and accuracy of explicit responses across all group, belief and grip conditions. Since at this level neither DLTS nor accuracy was normally distributed (as per Shapiro-Wilks normality tests) as well as bounded, we performed a Spearman’s rank correlation analysis. The results show a significant positive relation between DLTS and accuracy, *rs*= .703, p<.001, 95% CI = [.619;.779]. Note that the confidence interval is based on a 1000 replications bootstrap resampling method.



*Figure X. Correlation between DLTS and accuracy of explicit responses. The dashed blue line represents a fitted linear model, while the red line represents a LOESS (non-parametric local regression) fit. Based on this graph, we can assume linearity of the relationship.*

Next, we looked at the correlation between DLTS and accuracy per group using a similar methodology. This correlation was significant in the baseline condition, *rs*= .67, p<.001, 95% CI = [.525;.788], as well as in the mind-tied group, *rs*= .719, p<.001, 95% CI = [.588;.812].



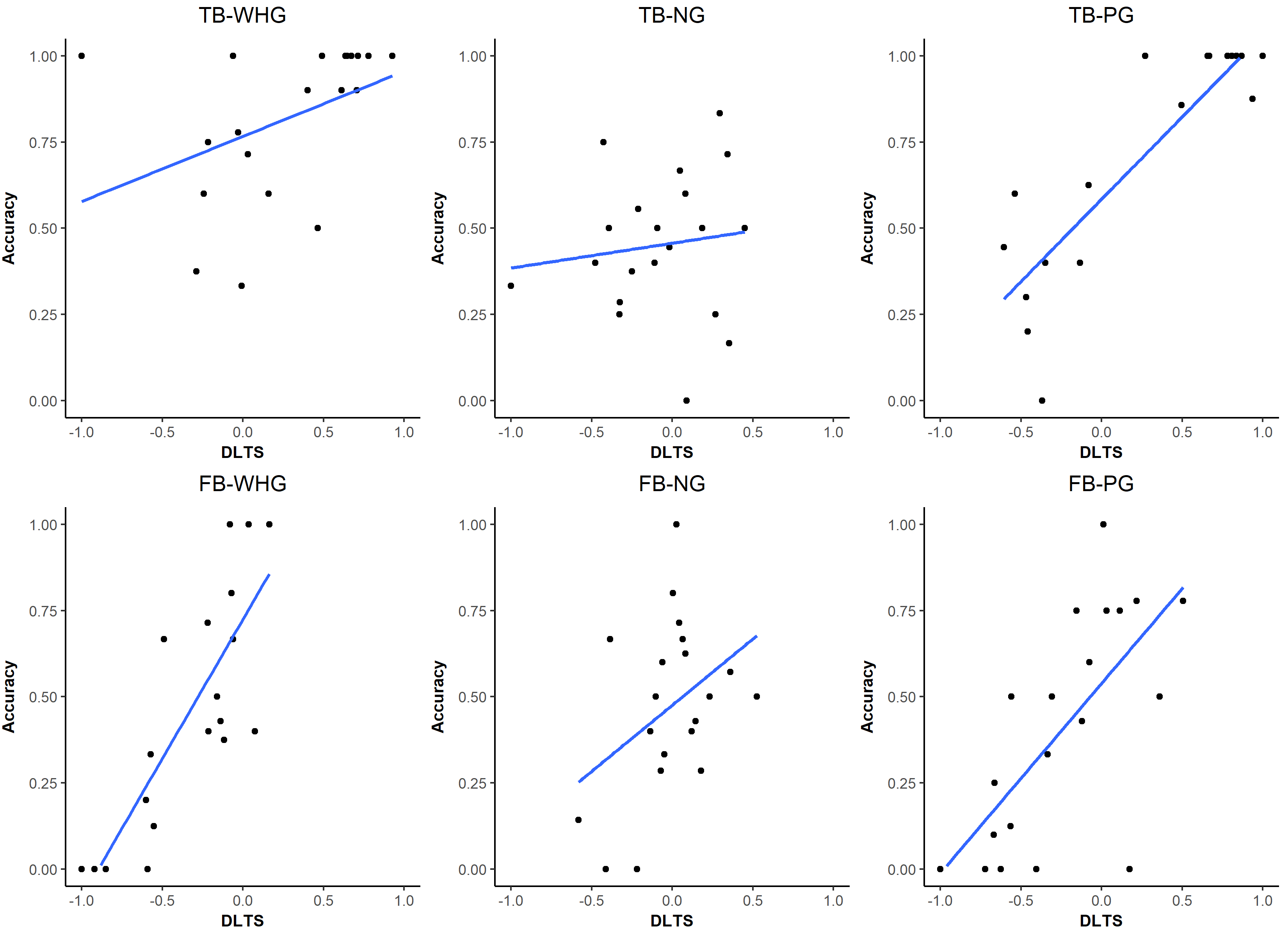
*Figure X. Correlation between DLTS and accuracy of explicit responses, per group. The dashed blue line represents a fitted linear model, while the red line represents a LOESS (non-parametric local regression) fit.*

We were also interested in exploring the correlation patterns between DLTS and accuracy of explicit responses in function of the different levels of group, belief and grip. We will look at the results of these analyses below.

*2.3.2 Baseline group*

We investigated the correlation patterns for every belief-grip combination in the baseline group on its own. The results of this analysis can be found in the table below, along with a visual representation of the relationship between DLTS and accuracy. Note that due to the relatively small sample size (N = 20) and large variability due to outliers in some conditions, the 95% confidence intervals are very large, signalling that we should be cautious with interpreting these correlation estimates. Please refer to Figure X for a more detailed look at the impact of the correlation patterns.

|  |  |  |  |
| --- | --- | --- | --- |
| *Condition* | *rs* | *95% CI* | *p* |
| TB: Whole Hand Grip | .532 | .105;.846 | .016 \* |
| TB: No Grip | .113 | -.351;.598 | .637 |
| TB: Precision Grip | .802 | .562;.911 | <.001 \*\*\* |
| FB: Whole Hand Grip | .837 | .536;.939 | <.001 \*\*\* |
| FB: No Grip | .311 | -.21;.677 | .182 |
| FB: Precision Grip | .693 | .306;.916 | <.001 \*\*\* |

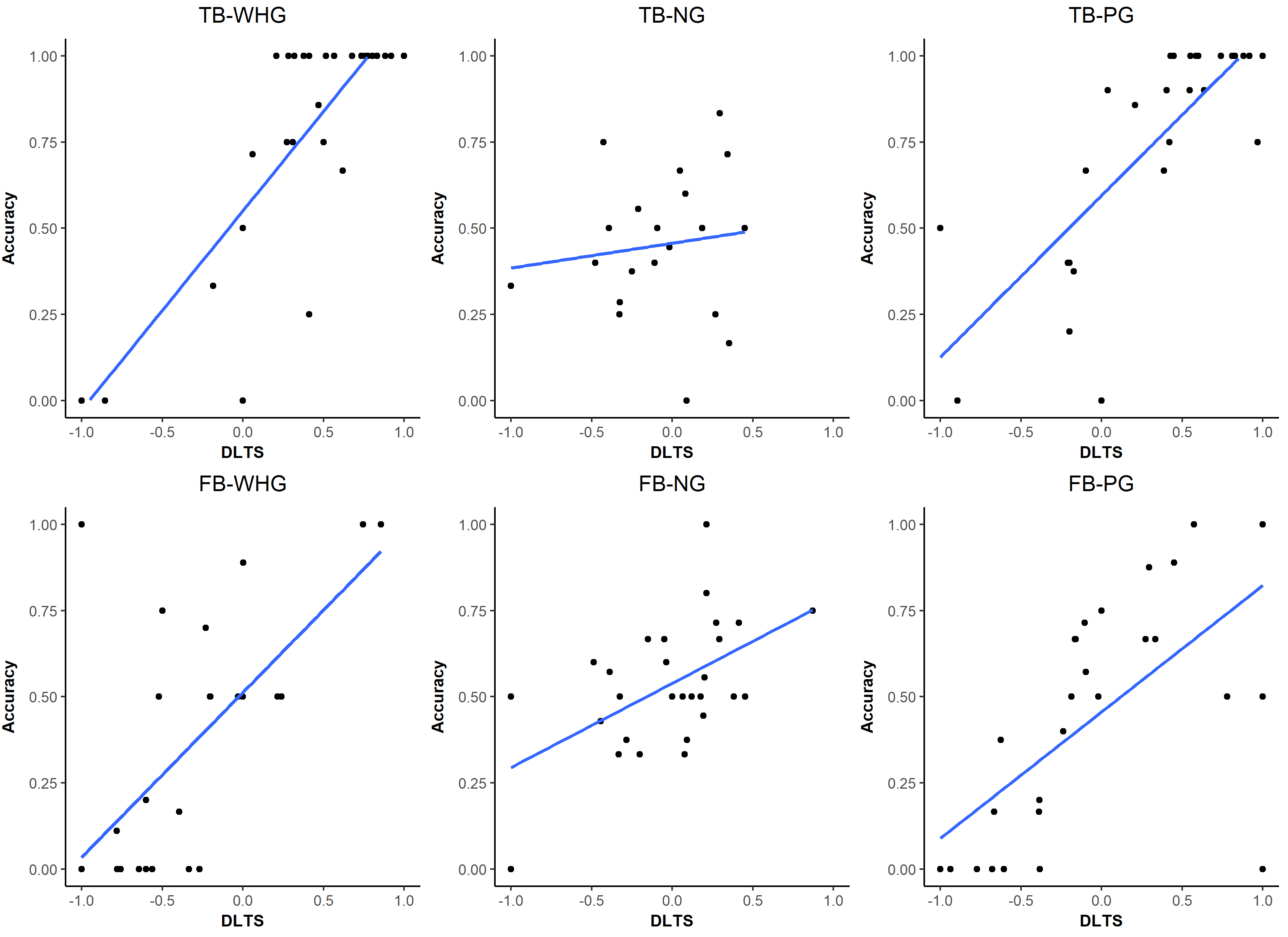


*Figure X. DLTS and accuracy of explicit responses plotted against each other, per belief and grip combination, for the baseline group only. Blue lines represent the linear model fit for each plot.*

*2.3.3 Mind-tied group*

Similar to the baseline group, we also did a further analysis of the correlation patterns in the mind-tied group. Details of this analysis can be found in the table and graph below. Note that here too confidence intervals are rather large in some conditions.

|  |  |  |  |
| --- | --- | --- | --- |
| *Condition* | *rs* | *95% CI* | *p* |
| TB: Whole Hand Grip | .701 | .442;.875 | <.001 \*\*\* |
| TB: No Grip | .272 | -.181;.653 | .162 |
| TB: Precision Grip | .809 | .579;.909 | <.001 \*\*\* |
| FB: Whole Hand Grip | .622 | .239;.868 | <.001 \*\*\* |
| FB: No Grip | .484 | .114;.729 | .01 \*\* |
| FB: Precision Grip | .71 | .388;.917 | <.001 \*\*\* |



*Figure X. DLTS and accuracy of explicit responses plotted against each other, per belief and grip combination, for the baseline group only. Blue lines represent the linear model fit for each plot.*

Note: the problem with the correlations is that some data points are derived of only a single trial, which can then become serious outliers, although this is less of an issue with Spearman’s correlation method.

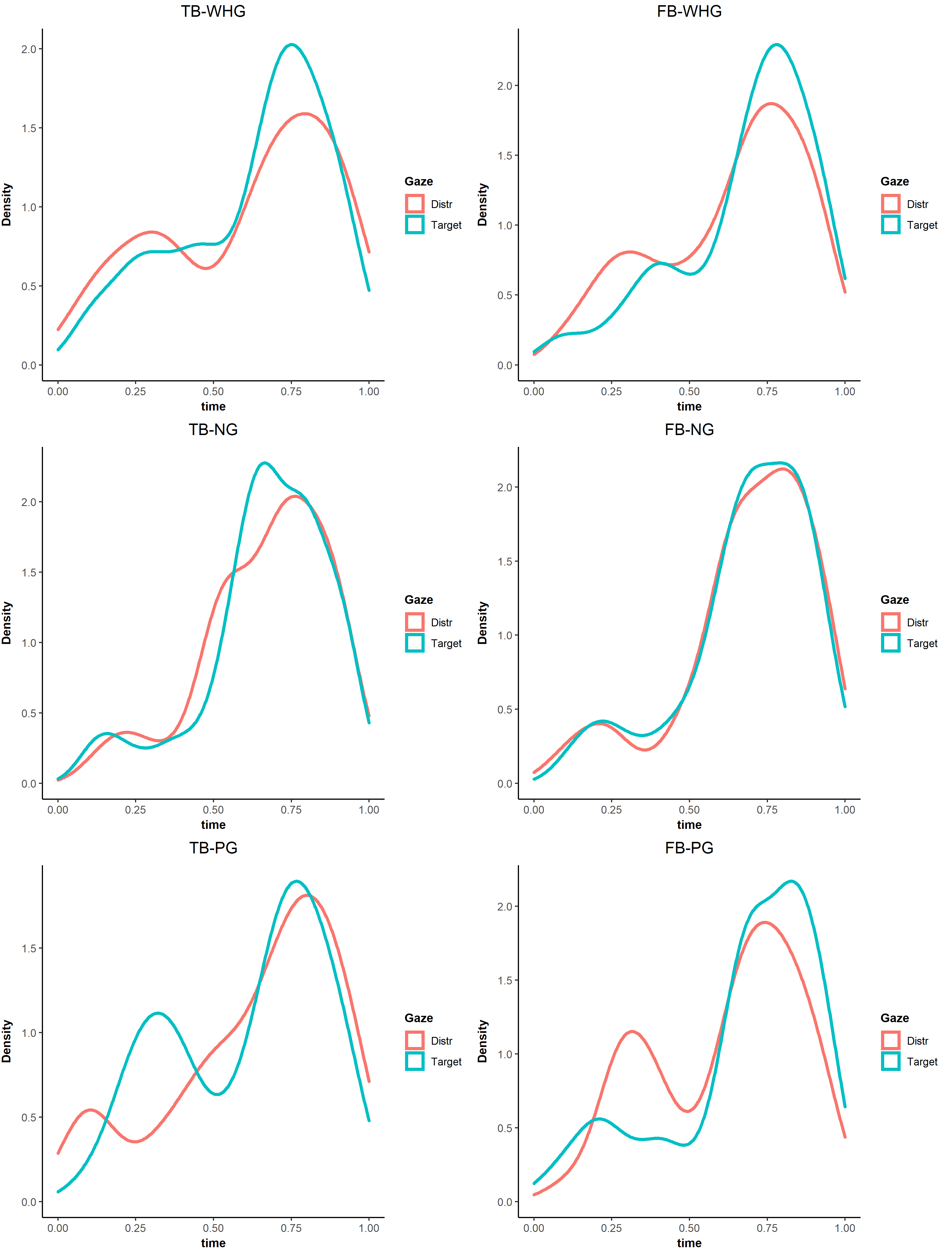
*2.4 Temporal visualisations of gaze behaviour*

Thus far our analysis of the data has focused on singular fixed measures of looking behaviour throughout the Time of Interest (TOI) of the trial. While Differential Looking Time Scores (DLTS) give us a good overall indication of the general bias of looking behaviour, this measure does not provide us with an indication of how gaze shifts during the TOI. To mitigate this shortcoming, we explored gaze behaviour in function of time. We did this by visualising fixation density, i.e. plotting fixations during the measurement period in function of their onset relative to start of the TOI. However, since considering only fixation onset reduces the amount of information to merely the number of fixations within a specific timeframe, we weighted every fixation with its duration. Hence, the density plots below consider the onset of every fixation, weighted by the duration of that fixation.

Another issue is that since videos were recorded live, their length and event timeline varied slightly. As a result, the length of the TOI varied over videos (mean duration = 4.232 seconds, standard deviation = .503). In order to mitigate this issue, we standardised fixation onsets by dividing their onset relative to the start of the TOI by the total length of the TOI. This results in a timeline between 0 and 1, where every fixation onset is represented in time as its relative position within the total length of that TOI.

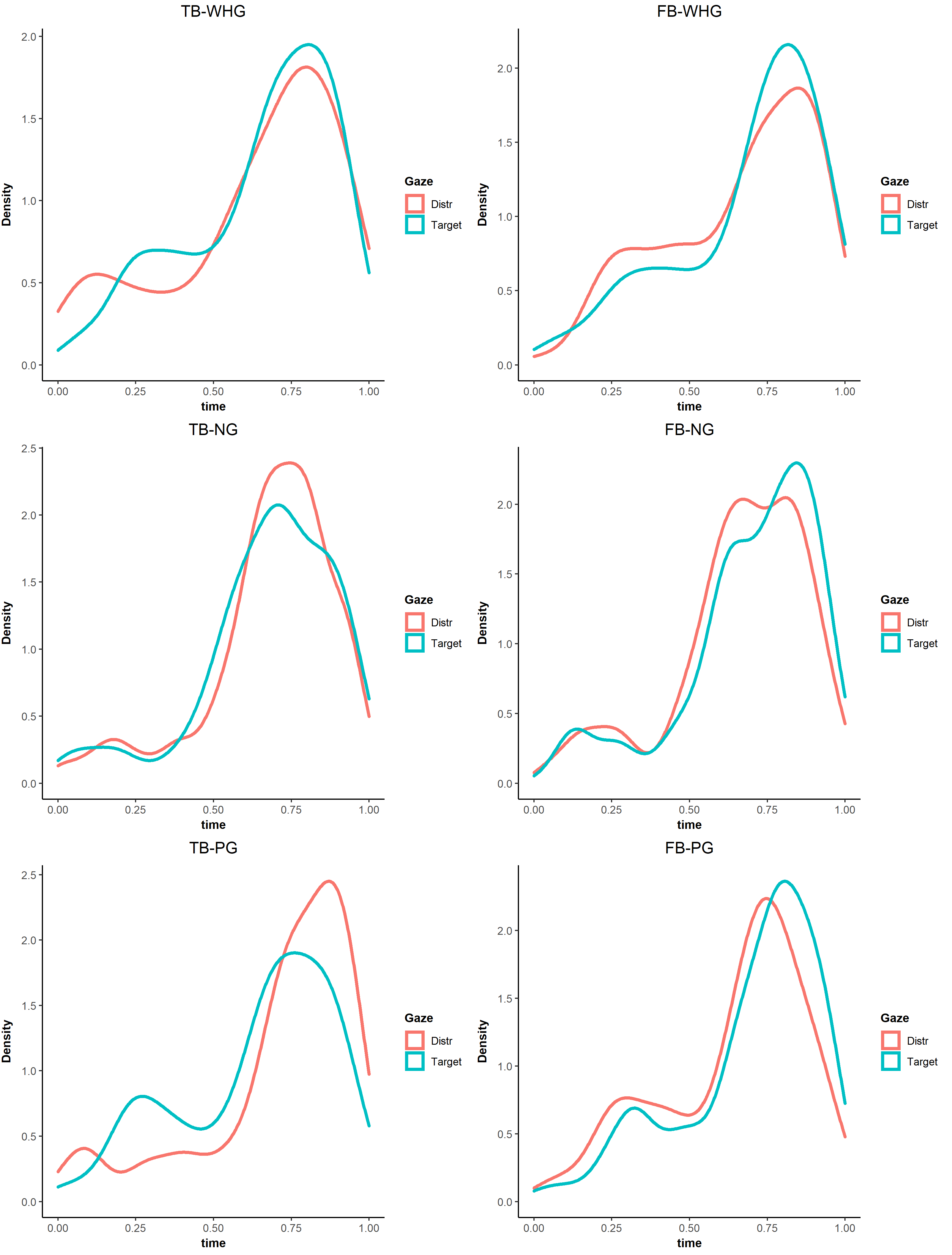
Note that we only considered fixations that fell within either the target AOI, i.e. the AOI that matches with the intended target of movement given the belief state of the agent, or the distractor AOI.

*2.4.1 Baseline group*



*Figure X. Density of fixations (weighted by their duration) over time in the baseline group, per AOI, in function of belief and grip condition combinations.*

*2.4.2 Mind-tied group*



*Figure X. Density of fixations (weighted by their duration) over time in the mind-tied group, per AOI, in function of belief and grip condition combinations.*

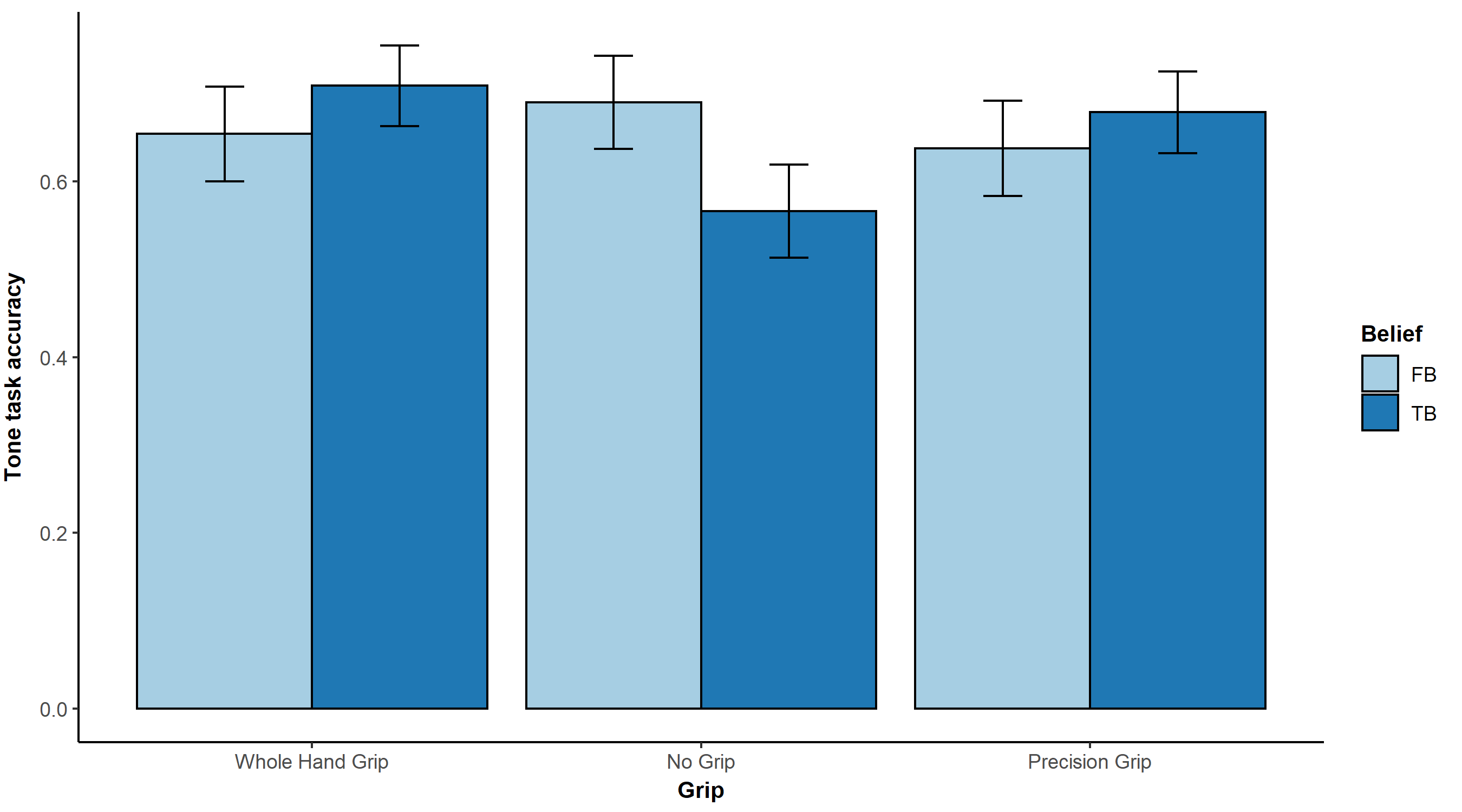
*2.5 Accuracy of secondary tone task*

We also had a look at performance on the secondary tone detection task in the mind-tied group. We found a global accuracy of 63.224%, and a mean accuracy over participants of 65.29%, standard deviation = 17.31%. Mean accuracy is significantly higher than chance level, as per a Wilxocon signed rank test, V = 343.5, p<.001, 95% CI = [.579;.727].

We tested if performance on the tone detection task was correlated to performance on the explicit mentalising task. There was, however, no significant correlation between both variables, *rs* = .077*,* p = .319, 95% CI = [-.060;.227].

We also tested if performance on the tone detection task was correlated to Differential Looking Time Scores (DLTS). However, there was no linear or monotonic relationship between these variables either, *rs* = .096*,* p = .217, 95% CI = [-.062;.243].

We also tested a linear mixed effect model, but found that neither belief, F(1,135) = .072, p = .789, grip, F(2,135) = .782, p = .459, nor their interaction, F(2,135) = 2.669, p = .073, had a significant effect on tone task accuracy.



*Figure X. Accuracy on the tone detection task in the mind-tied group in function of belief and grip conditions. Error bars represent the standard error of the mean.*

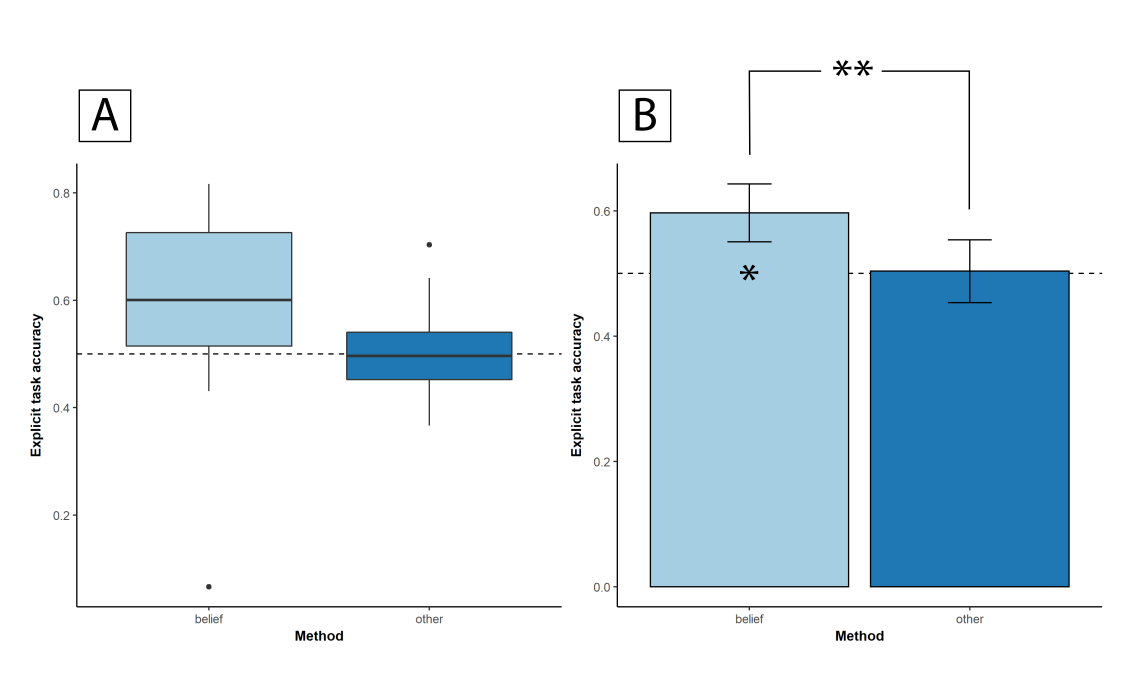
*2.6 Method of prediction*

During the debriefing session, we asked participants what motivated their decision on the explicit response task. Of the total of 48 participants, N = 20 indicated that they kept the beliefs of the agent into account, N = 16 decided purely based on the shape of the hand, N = 4 said to pick at random, N= 3 reported that they estimated based on perceived motions of the hand, N = 1 purely went for the large cup, and N = 4 had answers that were non-conclusive. For the sake of simplicity, we distinguished between mentalising (belief-based) methods and non-mentalising (other) methods only. An overview of the distribution of participant’s methods can be found in the table below.

|  |  |  |
| --- | --- | --- |
| *Method* | *Baseline group* | *Mind-tied group* |
| Belief | 11 | 9 |
| Other | 9 | 15 |
| None | 0 | 4 |

*2.6.1 Method of prediction and accuracy of explicit responses*

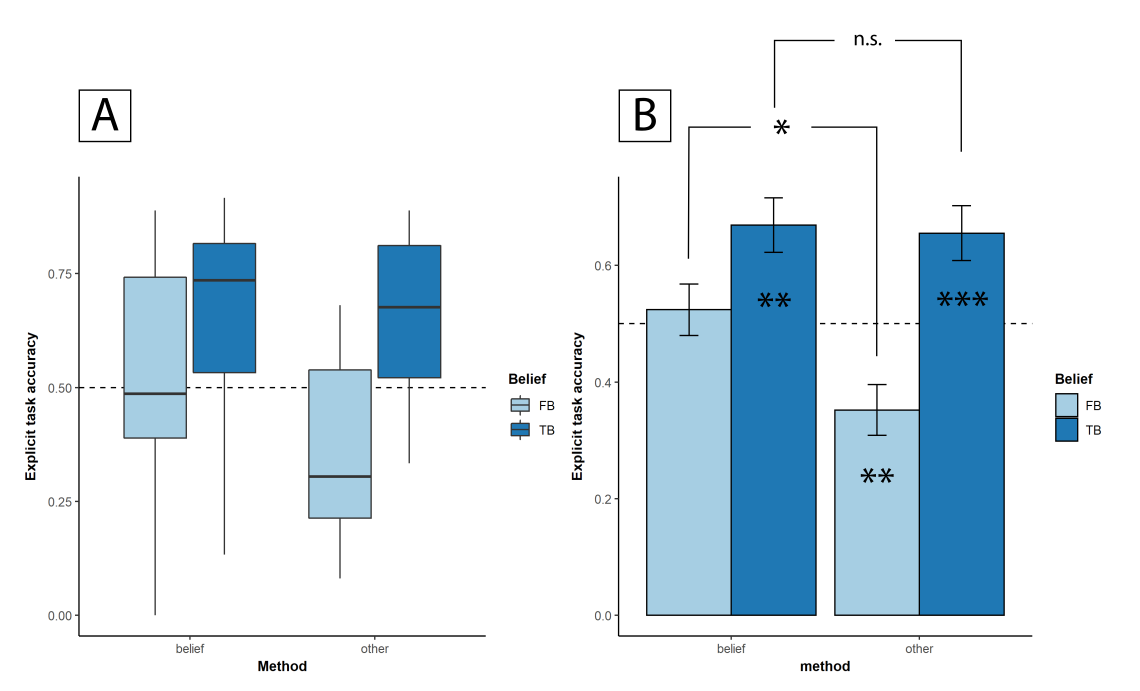
Of course, one would expect that participants that reported using a belief-based strategy to answer to the question “on what side of the screen will the hand appear?”, to have a better performance on this task as compared to participants who implemented a different strategy as a ‘correct’ answer was defined based on the belief state of the agent. Therefore, we tested if performance in the mentalising participants was significantly greater than chance level (50%). This was indeed the case (as per Wilcoxon Signed Rank Test), V = 174, p = .011, 95% CI = [.547;Inf]. We also tested whether or not accuracy on the explicit responses was significantly different from chance level in the non-mentalising participants (two-sided), which was not the case, *t*(23) = .234, p = .817, 95% CI = [.471;.535]. We also found that performance was significantly higher in the mentalising participants as compared to the non-mentalising participants, W = 346, p = .009, 95% CI = [.042;Inf]. All p-values reported above were Holm-corrected for multiple comparisons.



*Figure X. A) Boxplot of accuracy on the explicit response task, per method of responding per participant. B) Bar plot representing mean accuracy on the explicit response task, per method of responding. Error bars represent the standard error of the mean. Stars on the bar represent significance level of the mean difference as compared to chance level (50%). Note that chance level is represented by the dashed horizontal line.*

We investigated further by adding the factor ‘belief’. We expected that participants should show similar response patterns in the true belief condition regardless of response strategy, as mentalising strategies and motor-based strategies should arrive at the same conclusion. However, we do expect that participants that reported using mentalising strategies should have a better performance than non-mentalising participants in the false belief condition, as here mentalising-based and motor-based strategies should lead to different predictions.

First of all, we compared explicit response accuracy to chance level (50%). All p-values were Holm-corrected for multiple comparisons testing. In the true belief condition, we found that both the mentalising participants, V = 167, p = .008, 95% CI = [.606;.792], and the non-mentalising participants, *t*(23) = 4.423, p<.001, 95% CI = [.583;.728], scored significantly higher than chance level. However, in the false belief condition, we found that while performance of mentalising participants did not differ from chance level, *t*(19) = .459, p = .652, 95% CI = [.414;.636], participants that did not mentalise performed significantly lower than chance level, *t*(23) = -4.032, p = .002, 95% CI = [.276;.428].

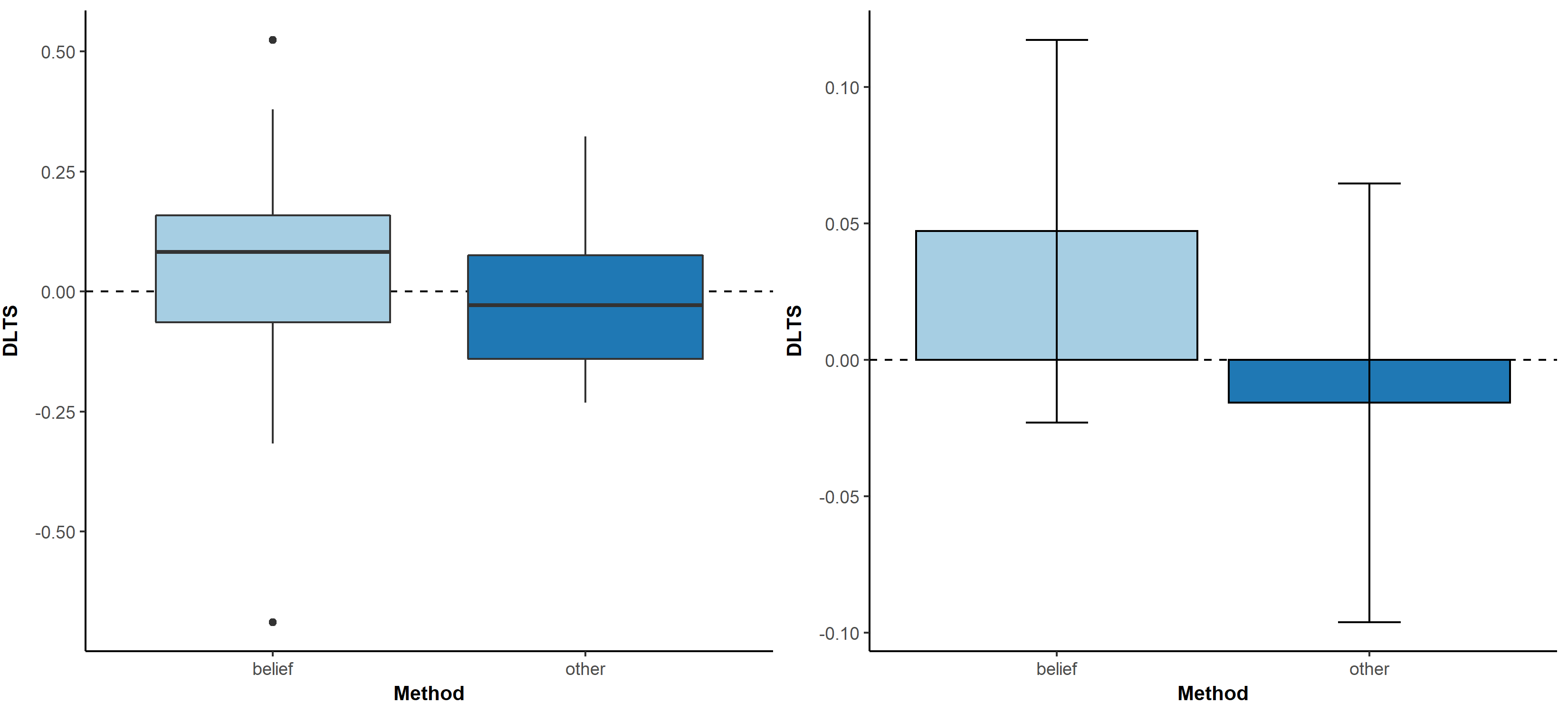


*Figure X. Figure X. A) Boxplot of accuracy on the explicit response task, per method of responding per participant and belief condition. B) Bar plot representing mean accuracy on the explicit response task, per method of responding and belief condition. Error bars represent the standard error of the mean. Stars on the bar represent significance level of the mean difference as compared to chance level (50%). Note that chance level is represented by the dashed horizontal line.*

In line with our hypotheses, we found that performance did not differ between mentalising and non-mentalising participants in the true belief condition, W = 252, p = .786, 95% CI = [-.069;.131]. However, in the false belief condition, mentalising participants performed significantly better than non-mentalising participants, *t* = 2.691, p = .011, 95% CI = [.064;Inf].

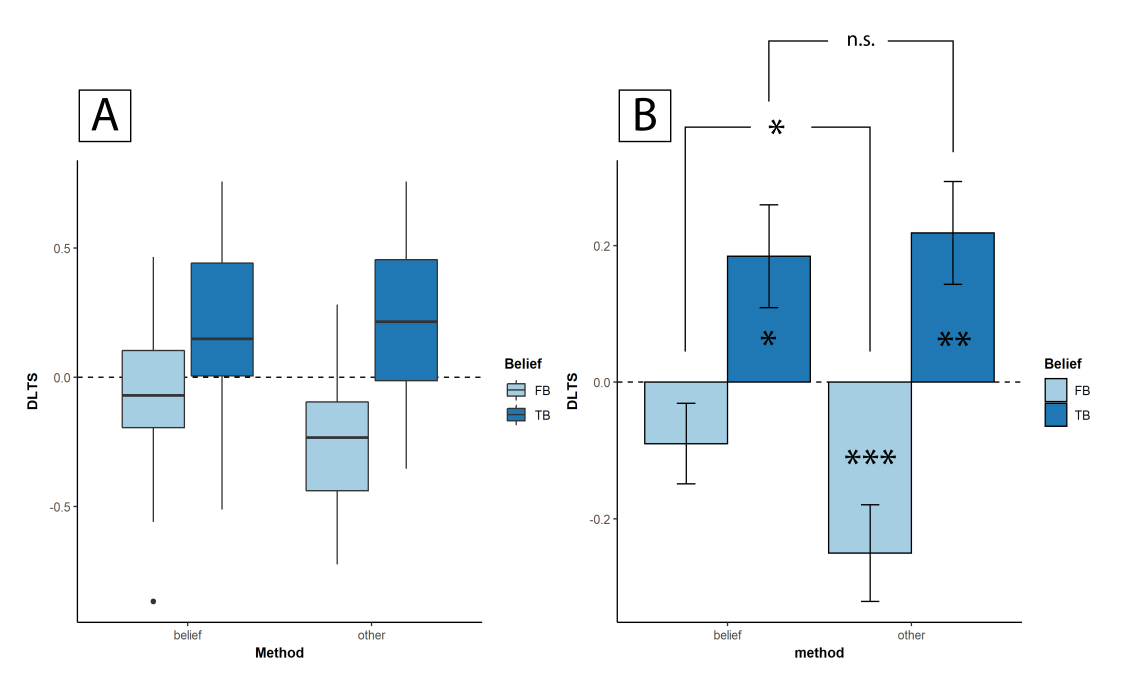
*2.6.2 Method of prediction and Differential Looking Time Scores (DLTS)*

We were also interested in the relationship between method of responding to the explicit response task and looking behaviour, as measured by DLTS. Again, we hypothesised that DLTS would be significantly higher than baseline level (0, no bias towards either target or distractor) in the mentalising participants. This was, however, not the case, *t*(19) = .815, p = .518, 95% CI = [-.053;Inf]. We also tested if non-mentalising participants were overall biased towards either target or distractor. This was not the case either, *t*(23) = -.511, p = .614, 95% CI = [-.079;.048]. Nor was there any difference in mean DLTS scores between both groups, *t* = .959, p = .518, 95% CI = [-.048;Inf].



*Figure X. A) Boxplot of DLTS, per method of responding per participant. B) Bar plot representing mean DLTS, per method of responding and belief condition. Error bars represent the standard error of the mean. Note that baseline level (0, no bias) is represented by the dashed horizontal line.*

Again, we investigated added the factor ‘belief’ into the analysis. We hypothesised that in the true belief condition, both mentalising and non-mentalising participants should have DLTS scores biased towards the target (i.e., larger than 0), and that scores in this condition should not differ over group. However, in the false belief condition, mentalising participants should have a DLTS score greater than 0, while non-mentalising participants should have a DLTS score lower than 0. Moreover, mentalising participants should have higher DLTS scores than non-mentalising participants.



*Figure X. Figure X. A) Boxplot of DLTS, per method of responding per participant and belief condition. B) Bar plot representing mean DLTS, per method of responding and belief condition. Error bars represent the standard error of the mean. Stars on the bar represent significance level of the mean difference as compared to baseline level (0, no bias). Note that baseline level (0, no bias) is represented by the dashed horizontal line.*

We found that in the true belief condition, both mentalising participants, *t*(19) = 2.377, p = .028, 95% CI = [.050;Inf], and non- mentalising participants, V = 253, p = .003, 95% CI = [.109;Inf], had a DLTS significantly higher than 0, indicating a significant bias towards the target. In the false belief condition, however, mentalising participants’ DLTS scores did not differ from baseline, *t*(19) = -1.391, p = .18, 95% CI = [-.225;.045], whereas non-mentalising participants showed DLTS scores that were significantly lower then baseline 0, *t*(23) = -4.476, p<.001, 95% CI = [-365;-.135], indicating a significant bias towards the distractor cup.

We also tested the difference between groups within belief condition. In the true belief condition, we found that there was no difference between the mentalising and non-mentalising participants in terms of mean DLTS, *t* = -.345, p = .732, 95% CI = [-.235;.166]. In the false belief condition, however, we found that DLTS of non-mentalising participants was significantly lower (i.e. more biased towards the distractor) than the mean DLTS of the mentalising participants, W = 324, p = .048, 95% CI = [.025;Inf].

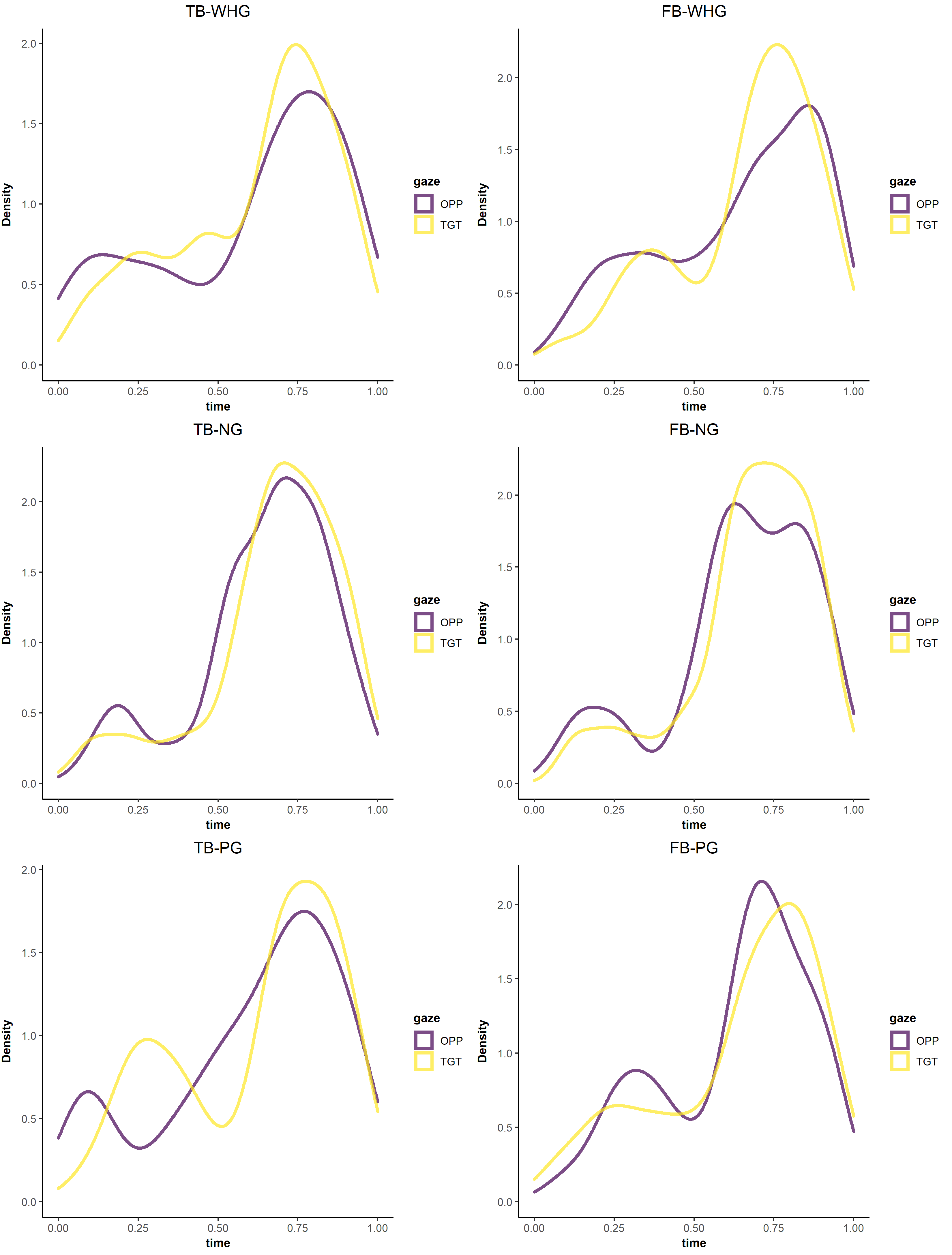
*2.7 Temporal visualisation of fixation density per method of prediction*

Finally, we wanted to have a look at the temporal variations in fixation density in function of method of prediction implement on the explicit response task, as self-reported by the participants. A similar methodology was implemented as in paragraph 2.4 (see above). An overview of these visualisations can be found below.

Option: removing the no-grip trials as they are irrelevant here and just drag means down to baseline

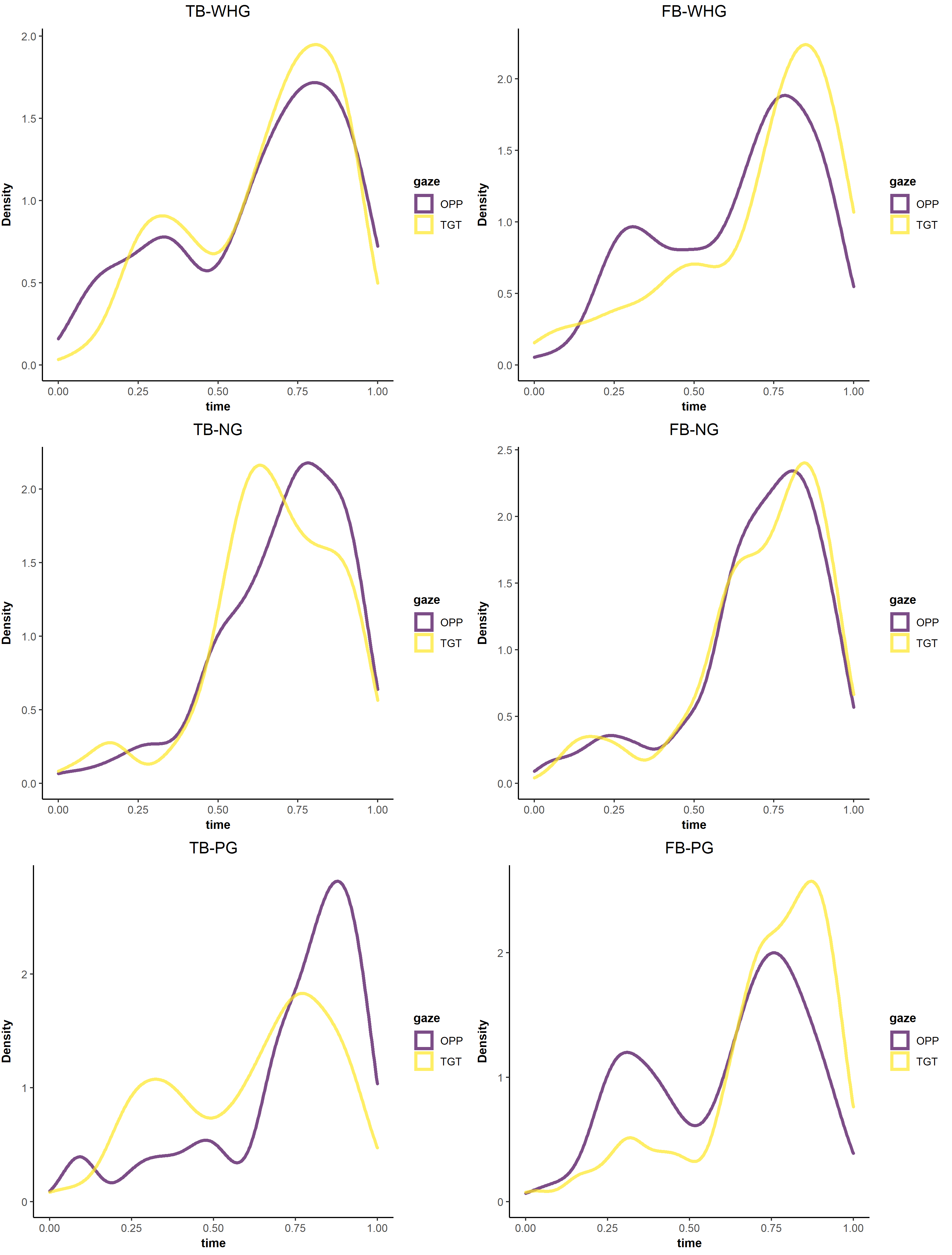
To do: include the estimates for every mean difference related to a CI?

*2.7.1 Mentalising group*

**

*Figure X. Density of fixations (weighted by their duration) over time in the mentalising group, per AOI, in function of belief and grip condition combinations.*

*2.7.2 Non-mentalising group*

**

*Figure X. Density of fixations (weighted by their duration) over time in the non-mentalising group, per AOI, in function of belief and grip condition combinations.*